Taiwanese and Beijing Mandarin listeners' perception of English focus prosody

Jamie Adams, Sam Hellmuth

University of York

jamiecadams30@gmail.com, sam.hellmuth@york.ac.uk

Abstract

Post-focal compression (PFC) of F0 is a known cue to focus in English and Beijing Mandarin (BM), but PFC is neither present in Taiwan Mandarin (TM) production nor interpreted as a cue to focus in perception [1]. Studies of variation in L2 English production by BM and TM learners of English confirm transfer of some L1 patterns into their L2 English [2]. This paper explores for the first time how BM and TM listeners' interpret PFC in English. It also seeks to clarify L2 listeners' interpretation of PFC in contexts where discourse-new post-focal material carries a post-focal prominence in English [3]. Following [4] we presented L1 BM, TM and English listeners with a series of written discourse contexts and two prosodically congruous or incongruous audio responses in a between-participants design. One set of listeners in each language group judged SVO English stimuli produced in either all-new context (NN) or with initial narrow focus followed by discourse-given post-focal material (FG). Another set of listeners in each group judged the same all-new (NN) stimuli against recordings with initial narrow focus followed by discourse-new post-focal material (FN). Results indicate differential interpretation of on-focus and post-focal prosody matching a L1 perceptual transfer hypothesis.

Index Terms: focus, post-focal compression, L2 perception, Mandarin, English

1. Introduction

Languages differ in the prosodic exponents of semantic focus aimed at highlighting certain words to the listener, to signal a contrast or introduce new information. Compression of prosodic acoustic cues in the post-focal domain is a frequently observed cue to focus, alongside expansion of acoustic cues in the on-focus domain itself, but these post-focal cues are not present in all languages. Indeed, variation in the presence or absence of post-focal compression (PFC) is argued to be a parameter of variation in prosodic typology [5].

1.1. Focus prosody in Mandarin and English

Taiwan Mandarin (TM) and Beijing Mandarin (BM), although mutually-intelligible members of the same language family, have been shown to display features suggesting that they fall on different sides of this parameter of variation. BM speakers show 'on-focus' expansion of F0, intensity and duration in focused constituents, which is typically accompanied by compression of F0, duration and intensity in post-focal material. In contrast, Xu et al [1] showed that TM speakers do not use PFC to mark focus and instead rely more heavily on expansion of duration in the on-focus domain [1].

PFC is also observed as a focus cue in most inner circle varieties of English (see [1] for cross-linguistic PFC distribution), and PFC is conflated with de-accenting in the post-nuclear 'tail' in some accounts e.g. [5]. The presence of PFC in English does not always entail complete de-accenting of post-focal prominences, however. Katz and Selkirk [3] found that post-focal material new to the discourse carries prominence in English, albeit less prominent than the preceding nuclear accent. The effect of such post-focal prominences has received little attention in the debate about PFC. This paper thus investigates the effect of PFC on Mandarin listeners' perception of English focus prosody, in the context of both the presence and absence of post-focal accent on discourse-new material.

1.2. L2 acquisition of focus prosody

The extent to which typological differences in L1 focus prosody affect L2 acquisition has received relatively little attention. A notable exception is the work of Nava and Zubizarreta [6-8] who show that L1 Spanish/L2 English learners face a two-fold challenge in realising focus in English. Two aspects of focus realisation – both absent in L1 Spanish – must be acquired: first, the option to move the nuclear accent within utterances (following the Nuclear Stress Rule, NSR), and second, de-accenting of post-focal, discourse-given material. Some learners in their study showed mastery of NSR only (without post-focal deaccenting), but none showed deaccenting without ability to apply the NSR. Some intermediate and high-proficiency learners were able to successfully realign nuclear stress and also de-accent post-focally, however, suggesting that it is possible for learners to overcome transfer effects and produce L2 focus patterns similar to those of L1 speakers.

In a production study of TM-speaking and BM-speaking learners of English, Visceglia et al [2] found mixed patterns of L1 transfer. In their English productions TM speakers showed reduced duration on post-focal materials, and both TM and BM speakers exhibited compression of post-focal intensity. However, neither TM nor BM speakers produced post-focal compression of F0, which was unexpected. While L1 transfer might explain the absence of PFC of F0 in the TM group, it does not explain why L1 BM speakers would not transfer the pattern of post-focal compression of pitch range from their L1 to L2.

1.3. The present study

This paper explores TM and BM listeners' perception of English focus prosody, for the first time. We address two research questions: 1) Are the differences observed by [1], in L1 TM and BM listeners' L1 perception of utterances realized with PFC in Mandarin, reflected in L1 TM and BM listeners' L2 perception of utterances realized with PFC in English? 2) Is listeners' ability to identify the intended focus of an utterance reduced by the presence of post-focal accents on non-discourse-given material? We predict that 1) TM listeners will be less able to identify the intended focus of an utterance realized with PFC than BM listeners, due to the lack of PFC in their L1 TM, and 2) that both BM and TM listeners may show reduced accuracy in interpreting utterances with post-focal prominences.

2. Methods

2.1. Materials

Stimuli for the perception task were recorded on a Marantz professional solid-state recorder PMD661 MKII at 44.1kHz 16 bit. The phonetically-trained, female native speaker of British English wore a Shure SM10 professional unidirectional head-worn dynamic microphone. Context utterances were read first, followed by the target utterance, to ensure production was as natural as possible. Each target utterance contained a disyllabic subject noun, past tense verb, and object noun phrase. Lexical items containing sonorants were selected to facilitate phonetic analysis. The stimuli were elicited in 3 focus conditions (see Table 1) and comprised 8 lexical sets (see Table 2).

Table 1: *Sample set of contexts; target phrase in bold.*

Condition	Text
All New (NN)	*Why did you do that?* It turns out that **Gary needed the money** after all.
Initial Narrow Focus then Given (FG)	*What did the man at the bank say about the money?* He said that only **Gary needed the money** in the end.
Initial Narrow Focus then New (FN)	*What did your parents say yesterday?* They said that only **Gary needed the money** this time.

Table 2: *Target sentences.*

Lexical set	Target sentence
Abby	Abby mended her mobile.
Anna	Anna opened the window.
David	David managed the money.
Gary	Gary needed the money.
Jenny	Jenny ended the marriage.
Lily	Lily boarded the ferry
Manny	Manny loaded the weapon.
Nora	Nora ordered the dinner.

Figure 1 shows the time-normalized F0 contour over the three constituents in the sentence (S-V-O) for all individual stimuli, plus a smoothed curve across all stimuli, by focus condition. The NN stimuli (n=8) show steady declination through the utterance. The FG stimuli (n=8) show a steep fall in pitch after the subject, with the verb and object realized in a compressed pitch range and de-accented. Stimuli were recorded in their embedded contexts, as in Table 1, and resulted in two types of FN realization: FN1 is similar to FG in having a steep fall after the subject followed by post-focal compression of pitch range on verb and object, but declination is then suspended; in FN2 realizations there is a steep fall in pitch after the subject followed by a full pitch accent on the object.

These two different realizations of post-focal discourse new material were produced somewhat evenly across lexical sets by our speaker, in the same embedded contexts. This supports the claim in [3] that post-focal new information may elicit a post-focal accent, and we took the decision to include both types of FN realization in the experiment to find out how these variant realizations are interpreted by listeners. Listeners in the NN-FG experimental block thus heard four FN target sentences with a FN1 contour shape and four with a FN2 contour shape.

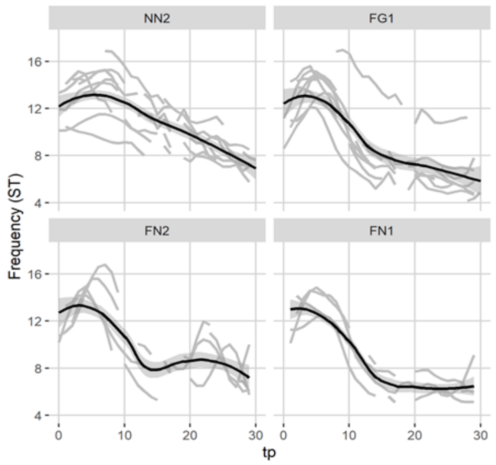

Figure 1: *Time-normalised smoothed F0 contour of individual stimuli (in grey) and GAM (REML) smoothed f0 (in black), by condition, showing split by realization of FN condition with (FN2) or without (FN1) an accent on the last lexical item.*

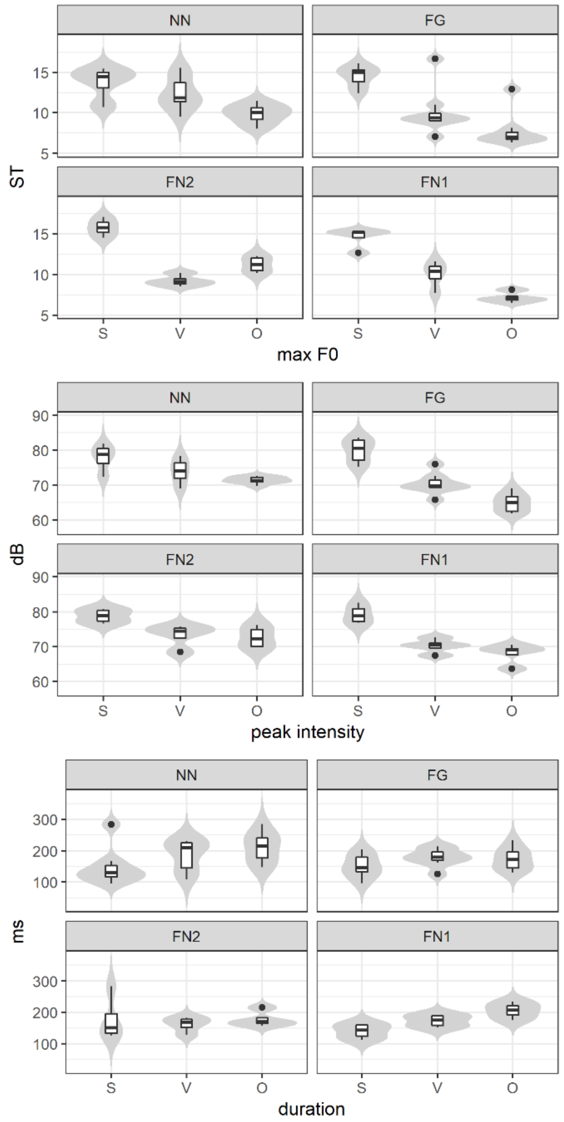

Figure 2: *Duration, peak intensity and maximum F0 in the stressed syllable of each word in stimuli, by condition.*

Figure 2 shows duration, peak intensity and maximum F0 in the stressed syllable of the subject, verb and object in all stimuli. PFC of duration and intensity are visible in all FG and FN stimuli, including when a post-focal prominence is present (in FN2). PFC of F0 is also visible in the FG and FN1 cases, but is reduced in FN2 stimuli containing a post-focal prominence.

2.2. Procedure

Following [4] Experiment 1 we used a listening paradigm in which, for each trial, participants see a written context (the part in *italics* in Table 1) and a two alternative forced choice between two different audio recordings of the target sentence associated with that context (the part in **bold** in Table 1). This 'one context two prosodic realizations' (1C2P) presentation yielded higher accuracy rates by American listeners in match of a recording to its intended context than a paradigm in which listeners try to match a single prosodic realization to one of a choice of possible contexts [4]. We adopted the 1C2P paradigm, therefore, to make the listening task as manageable as possible for speakers of English as an additional language.

There was one practice trial, involving a context and related audio recordings which did not appear in test trials. The task instructions in the practice trial were: "For each question, read the short text then listen to two different recordings of part of the text in bold: which recording matches how the bold text should sound?". Participants received feedback on their response to the practice trial, but not to test trial responses.

For test trials, the experiment was divided into two blocks, so that each participant was exposed to one pair of focus conditions only: either all-new *versus* narrow-focus-then-given (NN-FG) or all-new *versus* narrow-focus-then-new (NN-FN). Each block comprised 16 trials; the relevant pair of audio recordings for each lexical set was presented twice, once with the congruent written context for the all-new recording (NN) and once with the congruent written context for the focus recording (FG or FN, depending on the block). Following [4], again, the order of presentation of trials was fixed for all participants in each block and was pseudo-randomized to ensure that no trials from the same lexical set were adjacent.

The experiment was implemented online using [9]. All instructions were in English. Participants read an information sheet and provided informed consent on the experiment landing page, then completed a demographic and language background questionnaire asking them to self-report their age, gender, level of general education, country and region of origin, first language, number of years learning English and % level of English across four skills (speaking, listening, reading and writing). Participants were asked to report whether they were using earphones or their device speakers, with an invitation to use earphones if available. The mode test duration was less than 10 minutes for all listener groups.

2.3. Participants

We recruited 342 participants who were speakers of Beijing Mandarin (BM), Taiwanese Mandarin (TM) or English (EN) as a first language. Recruitment (by email and social media) generated uneven sample sizes across listener groups as shown in Table 4. Participants were randomly assigned to one of the two experimental blocks in a between-listeners design. Due to technical difficulties randomization was uneven for the TM listener group. Table 4 sets out the number of participants by group and condition.

Table 4: *Participant counts by experimental condition*

Group	NN-FG	NN-FN	Total
BM	117	111	228
TM	51	25	76
EN	19	19	38

The BM group had a mode age range of 18-24, whereas mode age range in the TM and EN groups was 25-34. The majority of BM and TM participants reported more than ten years' experience learning English, consistent with learning English in formal educational settings. However, proficiency scores varied more widely, from 20-100% in both learner groups, and were normally distributed, though with slightly higher proficiency scores on average in the TM group (TM: mean 61.4 sd 16.9; BM: mean 51.7 sd 16.4). Self-reported proficiency is thus likely to account for any observed variation in performance across L2 participants better than years of English learning experience. Two-thirds of BM listeners performed the task using earphones or headphones, whereas all TM listeners reported listening via the speaker of their device.

2.4. Analysis

Participant responses in test trials were binary coded for *accuracy* with levels 'correct' where the audio recording congruent to the written context was selected and 'incorrect' where the competitor recording was selected. Results for the two learner groups were explored in a generalized logistic mixed effects model (glmer) using lme4 [10], with *group* and *condition* plus the interaction between them as fixed factors, with further fixed factors for *age*, *gender* and English *proficiency* level. Mean values of self-reported speaking/reading/writing/listening level by participant were binned equally across all participants into four *proficiency* levels (<32.5%; 32.5-55; 55-77.5; 77.5-100). We included random intercepts for *participant* and *item* (defined as the lexical sets in Table 2) and random slopes by-*item* for *condition*. Model pairwise predictions are estimated using emmeans [11].

3. Results

3.1. English listeners

Figure 3 shows accuracy across English listeners in response to trials, by condition. The results confirm that English listeners perform well above chance in both conditions, but that accuracy is overall lower in NN-FN condition. We know from [4] that English listeners can match prosodic realization to context in a 1C2P paradigm with stimuli that encode a NN-FG contrast. The results in NN-FN condition confirm that English listeners are also able match prosodic form to context in response to the NN-FG stimuli used in the present experiment. Due to low EN sample size we do not explore the predictions of these results further, but note the reduced accuracy in NN-FN condition.

3.2. BM and TM listeners

Figure 4 shows accuracy across BM and TM participants by group and by condition. The raw results suggest little difference between the two groups in performance accuracy, in either experimental condition. Accuracy is lower in both groups in NN-FN condition, however, mirroring the pattern seen for English listeners. There is wide variation in accuracy between participants within both listener groups. Nevertheless, although the overall distribution of accuracy scores for BM and TM listeners (in Figure 4) is lower than those of the small set of English listeners (in Figure 3), the majority of BM and TM listeners are performing above chance level.

These patterns were explored in a generalized logistic mixed effects model run on the data from BM and TM listeners only, with the following structure: accuracy ~ *group***condition* + *proficiencylevel* + (1 + *condition* | *lexset*) + (1 | *participant*). Figure 5 visualizes the model predictions and coefficients are listed in Table 5. The intercept is positive and significant (β= 0.39786; SE= 0.13163; z= 3.022; p= .0025) confirming the descriptive impression that learners' performance tended to be accurate. There is a small but significant main effect of condition (β= -0.24822; SE= 0.12328; z= -2.013;p= .044), with lower accuracy predicted in NN-FN. There is no main effect of listener group nor any significant interaction between group and condition. There is a large main effect of proficiency level (β= 0.74992; SE= 0.20188; z= 3.715; p=.000204): those who self-report very high overall proficiency (>77.5%) in English are significantly more likely to perform accurately in the task.

Table 5: *Coefficient Estimates of Model Parameters*

Fixed effects:	Estimate	Std. Error	z
Intercept	0.39786	0.13163	3.022
groupTM	-0.0821	0.11203	-0.733
condition NN-FN	-0.24822	0.12328	-2.013
proficiency mid	-0.04045	0.12059	-0.335
proficiency high	0.109	0.12064	0.904
proficiency veryhigh	0.74992	0.20188	3.715
group:condition TM:NN-FN	-0.1595	0.18114	-0.881

4. Discussion

Both BM and TM listeners tended towards accurate perception of intended meaning of focus prosody in English. Proficiency in English (inasmuch as participants' self-reports are realistic) is the strongest predictor of variation in task accuracy, matching the findings of Nava and Zubizaretta: English focus prosody can be successfully acquired. We did not find a significant difference between TM and BM listeners, going against hypothesis for our first research question; this may be explained by the consistent availability of durational cues in the stimuli, as in L1 TM. The TM listeners' performance was consistently lower than BM listeners, however, across proficiency levels, albeit not to a significant extent. A further study in which the sample size and proficiency of the two listener groups is more evenly matched would fully exclude this hypothesis.

Our prediction that post-focal accents in FN condition would result in reduced accuracy was borne out, though this tendency was seen also among English listeners. Figure 6 shows mean proportion of accurate responses to congruent for all listeners in NN-FN condition, split by focus condition of the congruent prosodic realization presented. The sample sizes are small but hint at a difference between BM and TM listeners in which of the two FN realizations is harder to rule out as being NN congruent. Overall, this result calls for awareness of potential interaction of post-focal prominence with PFC, and lends tentative support to Nava and Zubizaretta's claim that on- and post-focal prosody are independently acquired.

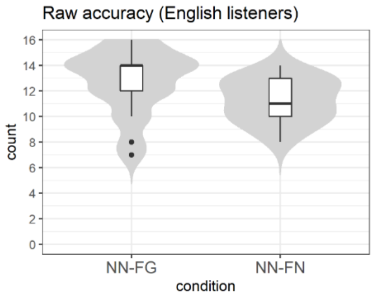

Figure 3: *Frequency plot of mean accuracy score across English participants, by condition.*

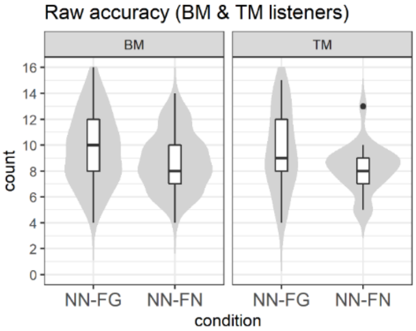

Figure 4: *Frequency plot of mean accuracy score across BM and TM participants, by group and condition.*

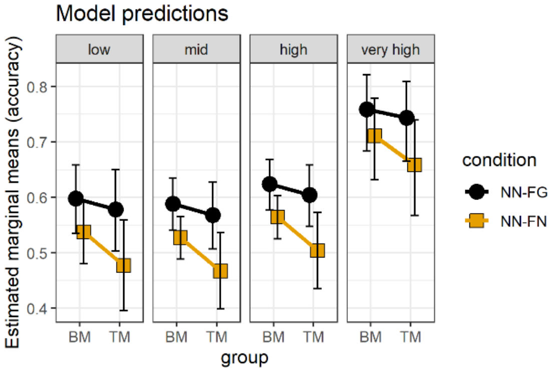

Figure 5: *Estimated marginal means in GLMM for accuracy by BM/TM listeners, by group, condition and proficiency level.*

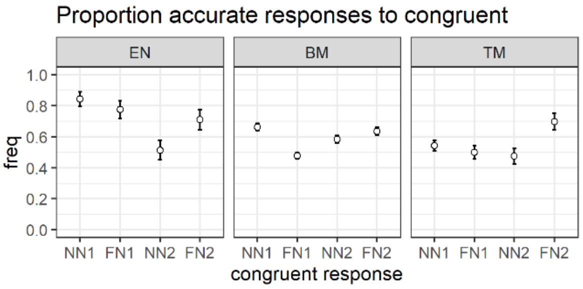

Figure 6: *Mean and 95% CI of proportion of accurate responses to congruent, in NN-FN condition, by group.*

5. References

[1] Xu, Y., S.-W. Chen, and B. Wang. 2012. Prosodic focus with and without post-focus compression: A typological divide within the same language family? *The Linguistic Review*, **29**(1): p. 131-147.

[2] Visceglia, T., C.-y. Su, and C.-y. Tseng. *Comparison of English narrow focus production by L1 English, Beijing and Taiwan Mandarin speakers*. in *2012 International Conference on Speech Database and Assessments*. 2012. IEEE.

[3] Katz, J. and E.O. Selkirk. 2011. Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, **87**(4): p. 771-816.

[4] Roettger, T.B., T. Mahrt, and J.J. Cole. 2019. Mapping prosody onto meaning–the case of information structure in American English. *Language, Cognition and Neuroscience*, **34**(7): p. 841-860.

[5] Xu, Y. *Post-focus Compression: Cross-linguistic Distribution and Historical Origin*. in *ICPhS*. 2011.

[6] Nava, E. and M.L. Zubizaretta. 2008. Prosodic Transfer in L2 Speech: Evidence from Phrasal Prominence and Rhythm. *Proceedings of Speech Prosody 2008*: p. 335-338.

[7] Nava, E. and M.L. Zubizarreta. *Order of L2 acquisition of prosodic prominence patterns: Evidence from L1 Spanish/L2 English speech*. in *Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition North America (GALANA 2008)*. 2009. Citeseer.

[8] Zubizarreta, M.L. and E. Nava. 2011. Encoding discourse-based meaning: Prosody vs. syntax. Implications for second language acquisition. *Lingua: International Review of General Linguistics*, **121**(4): p. 652-669.

[9] 2021. Qualtrics©. https://www.qualtrics.com: Provo, Utah, USA.

[10] Bates, D., et al. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1): p. 1-48.

[11] Lenth, R.V. 2021. Emmeans: estimated marginal means, aka least-squares means. R package version 1.6. 1.