This is a repository copy of *Phonetic imitation of the acoustic realization of stress in Spanish: production and perception*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/206210/

Version: Accepted Version

Article:
MacLeod, B. and Di Lonardo Burr, S.M. (2022) Phonetic imitation of the acoustic realization of stress in Spanish: production and perception. Journal of Phonetics, 92. 101139. ISSN 0095-4470

https://doi.org/10.1016/j.wocn.2022.101139

Article available under the terms of the CC-BY-NC-ND licence (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Phonetic imitation of the acoustic realization of stress in Spanish: production and perception

Bethany MacLeod[1]

Sabrina M. Di Lonardo Burr[2]

[1]School of Linguistics and Language Studies, Carleton University, Ottawa, ON, Canada

[2]Department of Cognitive Science, Carleton University, Ottawa, ON, Canada

Abstract

This study explores imitation of the acoustic realization of Spanish stress in disyllabic words produced in isolation, which is cued by three correlates: F0, duration, and intensity. Forty-eight native speakers of Mexican Spanish shadowed one of four model talkers of the same dialect. Differentials for each acoustic correlate of stress were generated by calculating the difference between the values of the first and second vowels for each of F0, duration, and intensity, for all recordings. Next, 87 Spanish speakers participated as listeners in a holistic perceptual assessment (4IAX task) of the shadowers' productions. Bayesian mixed-effects modelling was performed for both the acoustic and perceptual data. The results showed that the shadowers imitated the model talkers on all three differentials, but made the greatest shifts on the F0 differential, followed by duration, shifting the least on intensity. Analysis of the perceptual pattern showed that the listeners perceived imitation and that the shadowers' imitation on all three differentials contributed to the perceptual pattern. Lastly, the extent to which the listeners relied on imitation of the differentials roughly, but not exactly, aligned with how much the shadowers had converged on each differential, with listeners using imitation on duration the most, followed by F0, followed by intensity.

Keywords

phonetic imitation; phonetic convergence; stress; shadowing; acoustic analysis; perception; Spanish

This manuscript was accepted for publication in the *Journal of Phonetics* on February 21, 2022. This preprint is the peer-reviewed accepted version but has not yet been copyedited and may differ from the final version published in the journal.

Phonetic imitation of the acoustic realization of stress in Spanish: production and perception

1.0 Introduction & Background

Individuals who are exposed to the speech of another person often adjust their own speech to become more similar to the other person in a process known as linguistic accommodation. This process occurs at various levels of linguistic representation, including the pronunciation of sounds. When a speaker subconsciously imitates the acoustic-phonetic properties of another talker, this is known as phonetic imitation or phonetic convergence[1]. Talkers have been shown to imitate a wide variety of acoustic-phonetic properties of speech including vowel quality, vowel duration, word duration, speech rate, vowel nasalization, voice onset time, and fundamental frequency (F0), among others (Aubanel & Nguyen 2020; Babel 2010, 2012; Babel & Bulatov, 2012; Bonin et al. 2013; Brouwer, Mitterer & Huettig 2010; Clopper & Dossey 2020; Cohen Priva et al. 2017; Cohen Priva & Sanker 2018; Dufour & Nguyen 2013; Kim & Clayards 2019; Nielsen 2011; Pardo, Urmanche, Wilman & Wiener 2017; Phillips & Clopper 2011; Schweitzer & Walsh 2016; Shockley, Sabadini & Fowler 2004; Walker & Campbell-Kibler 2015; Zellou, Dahan & Embick 2017; Zellou, Scarborough & Nielsen 2016).

Much of the experimental research on phonetic imitation to date has focused on English and/or on monosyllabic words, looking at imitation of specific acoustic measurements of individual sounds such as vowel duration, F0, and vowel quality (Babel 2010, 2012; Babel & Bulatov, 2012; Babel et al. 2013; Clopper & Dossey 2020; Pardo et al. 2013; Pardo et al. 2017; Phillips & Cloppers 2011; Walker & Campbell-Kibler 2015). Other work has considered imitation at much higher levels, including at the level of the turn (Aubanel & Nguyen 2020; Gorisch, Wells & Brown 2012; Levitan & Hirschberg 2011; Schweitzer & Walsh 2016) and the entire conversation (Cohen Priva et al. 2017; Cohen Priva & Sanker 2018, 2020; Gregory & Hoyt 1982). However, we know less about how imitation might manifest at

[1] Different theoretical approaches or research questions often lead researchers to adopt a particular term to refer to this process including the above, but also alignment, accommodation, entrainment or synchrony (Babel, McAuliffe & Haber 2013; Babel, McGuire, Walters & Nicholls 2014; Cohen Priva, Edelist & Gleason 2017). In this study, we will use both imitation and convergence to refer to this process.

intermediate levels such as at the word or sentence level. In considering domains larger than the individual segment, but smaller than the entire conversation, we gain the opportunity to explore acoustic measures that change across the production of a word or that rely on relative measures. The current study contributes to filling in our understanding of how dynamic properties of the word might provide opportunities for imitation and how listeners might use variation in lexical properties when perceiving imitation. We focus specifically on the acoustic realization of stress in Spanish, where a stressed syllable tends to have greater duration, higher F0, and greater intensity than the unstressed syllables in the same word (Hualde 2012: 165; Llisterri, Machuco, de la Mota, Riera & Ríos 2003)[2]. That is, the cues to stress in Spanish are relative. Stress in Spanish is phonologically contrastive, creating many stress minimal pairs, which differ from each other only in the position of stress, such as verbs in the first person singular present (e.g. *salto* /ˈsal.to/ 'I jump') versus verbs in the third person singular past (e.g. *saltó* /sal.ˈto/ 'she/she/it jumped'). While English has stress minimal pairs too, the distinction between the words is complicated by the reduction of unstressed vowels, while in Spanish it is not (Ortega-Llebaria & Prieto 2011; Quilis 1981; Quilis & Esgueva 1983). As such, Spanish stress provides a nice test case for exploring how talkers might imitate properties of words that change as the word unfolds and that are inherently relative in nature.

1.1 Theoretical accounts of phonetic imitation

Research on phonetic imitation has shown that the changes speakers make are often subtle, yet statistically significant (e.g., Dufour & Nguyen, 2013; Nielsen, 2011) and perceptible to listeners (e.g., Pardo, Jordan, Mallari, Scanlon & Lewandowski, 2013). The findings that speakers make these changes suggest that they can perceive fine-grained variation, which influences their realization of the phonetic category. If such adjustments are made repeatedly, such as in a situation of dialect contact, the changes

[2] The relative strength of these cues depends on the context in which a word is embedded. See §1.3 for more details and Hualde (2002) for a review.

can become more permanently encoded in a speaker's phonetic repertoire, leading to second dialect acquisition. According to the Change-by-Accommodation model of sound change (Niedzielski & Giles, 1996), if enough individuals acquire these shifts, the result can be community-level sound change. There have been two main approaches to explaining why talkers might imitate in this way. In the first approach, talkers adjust their pronunciation during an interaction for social reasons. Under Communication Accommodation Theory (CAT: Giles, 1973; Giles, Coupland & Coupland, 1991), talkers converge towards an interlocutor (become more similar) to minimize social distance or diverge (become less similar) to increase social distance or show disdain for the interlocutor. Support for CAT is found in studies showing that the pattern of phonetic imitation can be influenced by social factors such as vocal attractiveness and prototypicality (e.g., Babel et al., 2014), gender (e.g., Namy, Nygaard & Sauerteig, 2002), age (e.g., Lin et al., 2021), and talker attitude towards a model talker (e.g., Yu, Abrego-Collier, & Sonderegger, 2013).

The second approach to explaining the mechanism behind phonetic imitation posits that imitation is the inevitable result of a direct connection between the perception and production systems and is therefore not socially motivated. For example, under an episodic memory approach (e.g., Goldinger, 1998; Pierrehumbert, 2001), exemplars are stored in rich acoustic detail. During articulatory planning, an exemplar is chosen as the production target, providing instructions for how the sound or word should be produced. An exemplar with high activation is more likely to be chosen as the target and a recently perceived exemplar will have relatively high activation. Episodic memory approaches have been used to explain patterns of phonetic imitation in previous work (e.g., Babel, McGuire, Walters & Nicholls, 2014; Goldinger, 1998; Tilsen, 2009). In Pickering and Garrod's (2013) interactive-alignment account, talkers generate a model of their interlocutor's speech through one of two processes: prediction-by-association, which relies on previously perceived speech, or prediction-by-simulation, which involves a process called covert imitation. Covert imitation necessarily involves the talker's own production system, creating a set of instructions for their next production. Since that production was based on covert imitation of an interlocutor, the prediction-by-simulation route in the interactive alignment account predicts that talkers'

pronunciation will shift to become more similar to the interlocutor; that is, they will show evidence of phonetic imitation. Such automatic accounts of imitation find support in studies showing the talkers tend to imitate model talkers even in non-social tasks, such as shadowing, in which there is no interaction between talkers and thus, no obvious social motivation to imitate. To better explain both the apparent social and non-social motivations for imitation, some studies have suggested the need for a hybrid account of imitation where the process is automatic, but one that can be modulated by social factors (e.g. Babel, 2012; Pardo et al., 2017). A hybrid account helps explain previous findings that linguistic factors, such as lexical frequency (e.g., Dias & Rosenblum, 2016; Goldinger, 1998; Goldinger & Azuma, 2004), phonological contrast (e.g., Nielsen, 2011), and perceptual salience (e.g., Babel, 2010; MacLeod, 2012b, 2014) also influence the pattern of imitation. As explained by Ross et al. (2021), existing work on imitation suggests that there is a tight link between perception and production, but that the extent to which this link will influence the pattern of phonetic imitation depends on various social and linguistic factors.

1.2 Imitation of relative measures and the effect of perceptual salience

The current study seeks to explore phonetic imitation in the acoustic realization of lexical stress in Spanish. This allows us to consider how the extent to which the three acoustic correlates are used to cue stress could influence the pattern of phonetic imitation. Based on work on the production and perception of Spanish stress, we expect F0 to be the most salient, relevant correlate cuing stress (Hualde, 2012: 165; Llisterri et al., 2003). Previous work considering how perceptual salience influences imitation have reached conflicting results. For example, Trudgill (1986) found that talkers imitate the most salient variables the most, a result that was partially supported by MacLeod (2012b, 2014), whereas Babel (2010) found that it was the least salient variable that talkers imitated the most, and Evans & Alshangiti (2011) found that the participants diverged on the most salient variables. Those studies considered imitation across dialects. This could mean that more salient differences between two dialects will encode social meaning, which talkers perceive. How talkers might respond to that social meaning (i.e. whether they

imitate it or not), would likely depend heavily on the associations of the variation, such as whether the dialectal difference is stigmatized or not (MacLeod, 2015). To our knowledge, no study has suggested that the acoustic realization of Spanish stress varies between dialects. As such, this allows an investigation into the effect of linguistic salience, without complicating the picture with social meaning.

As noted earlier, the realization of stress in Spanish involves considering the relative F0, duration, and intensity of vowels in the same word. Very few studies have considered the imitation of relative measures. An exception is Mantell & Pfordresher (2013), who explored relative F0 in their investigation of the imitation of melodies and spoken utterances. They found that among the spoken utterances, participants imitated the relative pitch contour more than absolute F0. According to Lin et al. (2021), this could mean that relative pitch is more central to speech processing than overall pitch. If so, this suggests that improving our understanding of the imitation of relative measures, such as relative F0, could have implications for improving our models of speech perception and phonetic imitation. Lin et al. (2021) also consider relative F0 in their exploration of the imitation of tone in Hong Kong Cantonese. They point out that lexical tone provides an example of using relative pitch to encode linguistic information. However, most work on imitation of F0 has considered absolute F0 of individual vowels in monosyllabic words, in non-tonal languages such as English (e.g. refs). While Spanish is also non-tonal, it does use F0 (along with duration and intensity) to encode information about stress. As such, studying the imitation of Spanish stress provides an opportunity to further our understanding of the imitation of relative measures, moving beyond only relative F0.

1.3 The acoustic realization of lexical stress in Spanish

In Spanish disyllabic words, one of the two syllables will bear primary stress and the other will be unstressed (Hualde 2005: 220). There are three main acoustic cues associated with the realization of stress in Spanish: F0, duration, and intensity (Hualde 2012: 165; Llisterri et al. 2003). Stressed syllables tend to have higher F0, greater duration, and higher intensity than unstressed syllables. However, previous work

has shown that the three cues are not weighted equally in the perception of stress and that which one is the strongest depends on the context in which the word is produced (Kim 2015, 2020). In accented contexts, stressed syllables serve as anchoring points for other prosodic elements, such as pitch accent, which, like pitch, is realized via F0 (Hualde 2005: 241; Hualde 2012: 164)[3]. For example, in declaratives, the most common prenuclear pitch accent is rising, causing F0 to begin to rise on the stressed syllable and frequently to peak on the post-tonic syllable. In this case, the stressed syllable in the word is not the one with the highest F0 (Hualde 2005: 243). Unaccented contexts work differently. In parentheticals, for example, F0 is mostly flat and the duration cue becomes the most robust (Ortega-Llebaria & Prieto, 2007). When produced in isolation (as in this study), each word belongs to an accented context, in which pitch accent is realized within the stressed syllable, making pitch accent and stress covary (Kim, 2015; Ortega-Llebaria, 2006). As such, F0 is expected to be the strongest cue to lexical stress when the words are produced in isolation. Across contexts, duration is the most robust cue since stressed syllables will be longer than unstressed in both accented and unaccented contexts (Kim, 2015; Ortega-Llebaria, 2006). Given this, we might expect duration to be the second strongest cue to lexical stress in isolation. Studies on the perception of lexical stress have typically found intensity to be the least robust cue, covarying with duration and F0 (Kim, 2020; Llisterri et al., 2003). Because of this, we might expect intensity to be the weakest cue in isolation. Taken together, although we expect that the three main cues will be present, with the stressed syllable being higher pitched, longer, and louder than the unstressed, we also predict that F0 will be the strongest cue, followed by duration, which is followed by intensity. Crucially, there is no specific value of these three measures that causes a syllable to be perceived as stressed. Instead, the values must be greater than those of another syllable. In this way, investigating the imitation of the acoustic realization of stress necessarily involves comparing the values of two (or more) syllables, rather than measuring only one. This will allow us to explore imitation patterns relating to measures that are

[3] Hualde (2012: 164) notes that there are deviations from this pattern in words with secondary stress; however, in this study, all of the stimuli are disyllabic and so they are not long enough to have a syllable bearing secondary stress.

otherwise problematic to explore in isolation. In particular, measuring imitation of the intensity of an individual vowel is likely meaningless because we cannot accurately determine how participants in a shadowing study or in a perception study perceive the intensity of a pre-recorded individual vowel. First, the intensity of individual segments depends on the specific configuration of the microphone (Gorisch et al., 2012: 67; Ortega-Llebaria & Prieto, 2011: 88), which is likely at least somewhat different between participants. Second, stimuli used in both types of studies are usually normalized in average or peak intensity (e.g., Pardo et al., 2017; Walker & Campbell-Kibler, 2015). Third, in some cases (such as in the current study) the listeners can adjust the volume to a comfortable level when performing a perception task. As such, exploring imitation of the intensity of individual vowels would be pointless and notably, no study of phonetic imitation has attempted to do so[4]. However, we can explore how talkers imitate the relative intensity of two different vowels within the same word. Relative intensity is also used in studies of consonant lenition to quantify the strength of a consonant relative to a neighbouring vowel (e.g., Cole, Hualde & Iskarous, 1999; Eddington, 2011; Ortega-Llebaria, 2004; Podesva, Eckert, Fine, Hilton & Jeong, 2015).

1.4 The current study

The purpose of this study is to investigate how talkers imitate the acoustic realization of lexical stress in Spanish and how listeners perceive this imitation. This study makes several contributions to the empirical literature on phonetic imitation by considering imitation of lexical stress in Spanish disyllabic words and the effect of perceptual salience on imitation. In addition, in §4.4 we explore connections between our findings and theoretical accounts of phonetic imitation.

We have five research questions:

[4] Levitan & Hirschberg (2011) look at convergence on intensity (along with pitch, voice quality, and speaking rate), but they do so at the level of the turn and the conversation, which likely correspond better with average speaking volume than would intensity measures of individual segments or words.

1. Do shadowers shift their production to become more aligned with the model talkers' acoustic realization of lexical stress?
2. If they do, do they imitate F0, duration, and intensity in relation to their strength as cues to stress (i.e. in relation to their perceptual salience as cues to stress)?
3. Can listeners perceive imitation in the shadowers' production?
4. If they can, does the shadowers' imitation of the three acoustic correlates of stress contribute to listeners' ability to perceive imitation?
5. If it does, does the extent to which the correlates contribute to the perception of stress reflect the extent to which the shadowers imitated those correlates?

To answer these questions, we conducted two experiments. The first is a shadowing experiment, described in §2.0, in which native Mexican Spanish speakers shadow pre-recorded model talkers of the same dialect. The baseline, shadowed, and model talker recordings are analysed acoustically to measure the differences made between the stressed and unstressed vowels in terms of F0, duration, and intensity. The analysis of the results determines whether the shadowers have imitated the three acoustic correlates of stress and whether they do so in relation to the strength of those correlates as cues to Spanish stress.

The second experiment, described in §0, is a perception task in which naïve judges listen to the baseline, shadowed, and model talker recordings from Experiment 1 and decide which of the baseline or shadowed production recordings sounds more like the model. The analysis of the results determines to what extent listeners can detect phonetic imitation in the recordings from Experiment 1 and to what extent the shadowers' imitation of the three acoustic correlates of stress contributes to the holistic perception of imitation. Our predictions for these two experiments are as follows:

Experiment 1:

1. Just as shadowers have been found to imitate the properties of individual vowels, we expect that they will show evidence of having imitated the model talkers' realization of stress in terms of the difference between the first and second vowel with respect to F0, duration, and intensity.

2. The shadowers will imitate the model talkers' realization of stress in terms of F0 differential the most, followed by duration, followed by intensity, reflecting the strength of these three cues to Spanish stress.

Experiment 2:

3. The listeners will be able to perceive imitation overall.
4. The shadowers' imitation of the three acoustic correlates of stress will contribute to the perception of imitation.
5. The extent to which imitation of the three acoustic correlates of stress contributes to the holistic perception of convergence will reflect the extent to which the shadowers have imitated those correlates.

2.0 Experiment 1: Shadowing

Shadowing has been used extensively to investigate phonetic imitation (Babel, 2010, 2012; Babel et al., 2014; Clopper & Dossey, 2020; Dufour & Nguyen, 2013; Goldinger, 1998; Kim & Clayards, 2019; Pardo al., 2013; Pardo et al., 2017; Phillips & Clopper, 2011; Walker & Campbell-Kibler, 2015). In the shadowing paradigm, talkers first read aloud a list of words to serve as a baseline of their pronunciation. Next, they hear another speaker, a pre-recorded model talker, reading the same words and the participants repeat after each word. The baseline, shadowing, and model recordings are then analysed.

2.1 Participants

In total, 57 talkers participated in a shadowing task. All were female[5] native speakers of Mexican Spanish. The majority were residents of Mexico, who were in the local Ottawa, ON, Canada area to

[5] Only female talkers were included in this study for two reasons. First, while the effect of gender of model talkers and shadowers has been considered before, the nature of the effect is far from clear (see Pardo et al., 2017 for a review). Second, since female talkers tend to have higher F0 than males, this would introduce a complicating level of variation in our investigation of F0 differential. It is important to note here that the findings of this study might not generalize directly to male modal+shadower pairs or to mixed-gender pairs.

complete a 3-week intensive language course in either English or French at a nearby campus of the Universidad Nacionál Autónoma de México. The first four of the participants (aged 21, 22, 26 and 29; all born and raised in Mexico City) provided model utterances, while the rest provided shadower utterances (mean age 27, median age 25). Five shadowers could not be included due to whispering or leaving long pauses after the word to be repeated. In the end, 48 participants remained, each randomly assigned to shadow one of the four model talkers, for a total of 12 shadowers per model. The majority of the shadowers were from inland South-Central Mexico including Mexico City (35) and Estado de Mexico (9), one was from Monterrey in the North-Central region, and one was from Guadalajara in the West. Only two hailed from coastal regions (one from Sinaloa and one from Veracruz). None of the participants reported any hearing issues and each was compensated $10 CAD.

2.2 Stimuli and recording

The stimuli consisted of 40 disyllabic Spanish words. In half of the words, the first syllable was stressed and in the other half, the second syllable was stressed. Across the two stress groups, the words were controlled for the quality of the first vowel, including the five monophthongs of Spanish /a, e, i, o, u/, and the consonants surrounding the vowel. The list of stimuli is provided in the appendix.

The model talker and shadower recordings were all made in a small sound-proof booth in front of a desktop monitor with the computer tower situated outside of the booth. The recordings were made using a Roland R-26 field recorder and an Audio-Technica AT831b lavalier microphone affixed to the participants' shirt to the left of centre near their collarbone. For all of the recordings (model recordings, participant baseline, and shadowing), the stimuli were presented in OpenSesame (Mathôt, Schreij & Theeuwes, 2012). To make the model talker recordings, the models read aloud the list of stimuli three times, each time in random order. The words appeared on the screen one at a time in a 48-point font every three seconds and advanced automatically. The recordings to be used in the shadowing and perceptual experiments were taken from the second reading to avoid any effects of hyperarticulation resulting from

lack of familiarity with the task or words in the first reading (Nielsen, 2011; Pardo et al., 2017). The third reading was used to replace any errors or anomalies, resulting in 16 (10%) of a total of 160 model talker words coming from the third reading. The sound files were spliced into individual words and were normalized in intensity to 70dB.

2.3 Procedure

The shadowing experiment was run in three phases: baseline reading, shadowing, and post-shadowing reading. There were 10 practice trials at the beginning of the first and second phases to familiarize the participants with the procedure. During the baseline reading, the participants read aloud the list of 40 words, three times in random order each time, following the same procedure as the model talkers. They were instructed to read aloud the words as they appeared on the screen and to speak as naturally as possible. During the shadowing phase, the shadowers were told that they would hear a voice saying words and that they should say the same word that they heard immediately after each word ended. Each shadower heard the same 40 words produced by one of the four model talkers[6] twice over high-quality speakers[7], each time in random order, and the participants repeated them. In the final phase, the shadowers read aloud the word list one more time, to serve as a postexposure phase to investigate any persistence in convergence. Only the baseline and shadowing phases are discussed in this paper.

2.4 Data analysis

[6] Other studies have had participants shadow multiple model talkers (e.g., Walker & Campbell-Kibler, 2015); however, this complicates the estimation of the baseline for model talkers after the first, since we expect that the participants' pronunciation might have shifted after exposure to the first model talker. To avoid this issue, in this study, the participants only shadowed one of the model talkers.

[7] Speakers were used instead of headphones for two reasons. First, it allows better monitoring of the participants' own voice (i.e., they do not hear their own voice through the physical barrier of the headphones). Second, it allows the recording to capture both the model voice and the shadowing, which means we can easily measure the duration of the interval between the offset of the model production and the onset of the shadowing, which could potentially influence how much the shadowers imitated or how much imitation the listeners perceive in Experiment 2. See footnote 13 for the results.

The shadowing experiment generated three sets of recordings: the model talkers' production and the shadowers' baseline and shadowed recordings. Since there were 48 shadowers producing 40 words, this resulted in 1,920 baseline and 1,920 shadowed words. To determine if a shadower imitated the model talker on a particular word, we need a model, baseline, and shadowed recording of that word. If any of these is missing, we cannot determine if the participant has imitated on that word. As noted above, each participant's baseline was estimated from the second reading of the word list in the baseline phase. Any words missing from this second reading (due to a mispronunciation or noise in the recording) were taken from the third reading (30 / 1,920 or 1.6%). The participants' shadowed productions were taken from the second shadowing[8], with any missing words sampled from the first shadowing (18 / 1,920 or 0.9%). Even with this procedure, there were 61 words in which one of the participants' baseline or shadowed words was not available, rendering them unusable in the experiment. A further 53 words were removed because they contained a vowel that had been completely devoiced, making measuring pitch and therefore calculating the difference in F0 between the two vowels impossible. In total, 1,806 complete sets remained.

2.4.1 Measuring phonetic imitation

There are two main approaches to assessing whether participants in a shadowing experiment have imitated the model talker or not. The first is to use acoustic analysis, where specific measurements are taken on the target sounds from the baseline, shadowing, and model recordings, such as formant frequencies in a study of vowel imitation (e.g., Babel 2012). This approach has the advantage of being able to pinpoint exactly what has changed from baseline to shadowing and to say whether it has shifted towards or away from the model talker and by how much. A disadvantage of acoustic analysis is that we cannot measure everything, so it could be that the participants imitated another variable that we did not

[8] The second repetition was chosen from the shadowing phase to maximize exposure to the model talkers and increase the likelihood of imitation (Goldinger, 1998).

consider. Another disadvantage is that acoustic analysis does not tell us whether imitation of the specific measures taken is relevant for how listeners would perceive phonetic convergence.

The second main approach is to use a perceptual assessment of imitation where third-party judges participate in a perception experiment in which they listen to the baseline and shadowed productions and decide which one sounds more like the model (e.g., Dias & Rosenblum, 2016; Kim et al., 2011; Namy, Nygaard & Sauerteig, 2002). The proportion of time that the listeners choose the shadowed token is taken to represent the strength of the evidence that they have perceived imitation. This approach essentially has the opposite pros and cons to those of acoustic analysis: while the perceptual method might align more closely with how listeners perceive imitation in the real world, it cannot tell us which acoustic properties have changed when the shadowers imitate or which changes the listeners use when making their judgements.

A related strand of work has aimed to alleviate the shortcomings of these two approaches by combining them. In what we will call the combined-analysis method, logistic mixed effects modelling is used to assess the extent to which the acoustic analysis can explain the results from the perception experiment. The combined-analysis method enjoys the benefits of both the acoustic and perceptual assessments, while also contributing to our understanding of how observed acoustic changes are related to the perception of imitation. This method was suggested by Pardo (2013) and has been applied in several recent studies (Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2013; Pardo et al., 2017; Walker & Campbell-Kibler, 2015). Some of these studies included more than one acoustic measure as predictors in the statistical analysis, allowing a comparison of how much the listeners relied on the different acoustic measures in their perception. The combined-analysis method provides a way of generating evidence about which measures influence listeners' holistic perception of imitation.

In this study, we used both the acoustic analysis (in §2.6) and perceptual assessments of imitation (in §3.4) and then used the combined analysis method to explore whether listeners in the perception task use imitation of the correlates of Spanish stress when perceiving imitation overall (in §3.5).

2.4.2 Acoustic analysis

For each of the three groupings of recordings (model, baseline and shadowed), a Textgrid in Praat (Boersma & Weenink, 2021) was created in which the first and second vowels were manually marked, using cues from both the waveform and the spectrogram, as the interval between the first regular vocal pulse and the offset of F2 (Chitoran, 2002; MacLeod, 2012a). In the case of the shadowed words, the interval between the end of the model's production of the word and the onset of the shadowed word was also marked. The duration, mean intensity, and mean F0 were measured for both the first and second vowels using a Praat script. Mean F0 was measured using the autocorrelation method by averaging the F0 measurements from 10ms time steps across the duration of the vowel. Null values were not included in the calculation of the mean. F0 measurements were transformed to the ERB scale, following Babel & Bulatov (2012).

Next, three "differential" measures were calculated, one for each of F0, duration, and intensity, by taking the difference between the value of the first vowel and the value of the second vowel in each word (Kim, 2015; Ortega-Llebaria & Prieto, 2011; Torreira, Simonet & Hualde, 2014). The formulae for the differentials are given in (1) – (3).

(1) $\text{F0 differential} = \text{F0}_{\text{V1}} - \text{F0}_{\text{V2}}$

(2) $\text{duration differential} = \text{duration}_{\text{V1}} - \text{duration}_{\text{V2}}$

(3) $\text{intensity differential} = \text{intensity}_{\text{V1}} - \text{intensity}_{\text{V2}}$

The differentials are negative when the second vowel has a higher value (higher F0, longer duration or greater intensity) than the first and positive when the first vowel has a higher value than the second. Our general expectation is that the model talkers and shadowers will produce words with positive differentials when the first syllable is stressed (paroxytonic words) and negative differentials when the second syllable is stressed (oxytonic words). However, there might be variation, with many tokens having positive differences even when the second syllable is stressed or negative differences when the first syllable is stressed (Kim, 2020; Torreira et al., 2014). We also expect inter-speaker variation in the magnitude of the

difference that the participants make between the syllables (for both stress patterns). It is this variation that we expect the shadowers might target for imitation when shadowing the model talkers. For simplicity, we will refer to the paroxytonic words as SU words, meaning that the first syllable is stressed and second is unstressed, and to oxytonic words as US words.

With respect to outliers in the model productions, we took a conservative approach, only removing those that were clearly outliers as demonstrated through plotting in order to avoid the shadowers having to approximate extreme targets. We found no clear outliers for the models' production of the duration and intensity differentials, but for F0 differential there were 7 outliers on the high end. Four of these were caused by the second vowel being produced with creaky voice[9], which causes F0 to drop (Johnson, 2003: 138), while the others were simply outliers. We removed all shadower production associated with those outliers, resulting in 71 datapoints being removed from the dataset (4%). For the shadowers, we only removed tokens where either the first or second vowel was produced with creaky voice in either the baseline or shadowed productions. This resulted in 52 words[10] being removed from the shadowers' baseline data, leaving a total of 1,682 in the dataset.

2.4.3 Statistical analyses

Most existing work on phonetic imitation has used frequentist mixed effects modelling during statistical analysis of both acoustic and perceptual data, typically fit using the *lme4* package (Bates, Mächler, Bolker & Walker, 2015) in the software R (R Development Core Team, 2016). In the current study, we use Bayesian inference instead of the frequentist approach for several reasons. First, Bayesian statistics offer advantages over frequentist models, allowing researchers to make claims about the likelihood of both the null and alternative hypotheses, given the evidence (Jarosz & Wiley, 2014), unlike frequentist models in

[9] Where creaky voice was defined as having F0 below 110Hz (Wagner et al., 2021).

[10] Nine were removed for the first vowel being creaky in the baseline, 23 for the second vowel being creaky in the baseline, 1 for the first vowel being creaky in the shadowing, and 19 for the second vowel being creaky in the shadowing.

which indirect evidence is collected against only the null hypothesis. Second, Bayesian models do not rely on hard cut-offs, such as $p < 0.05$, which allows for interpretation of the evidence for or against a hypothesis in a continuous manner. Third, frequentist models can suffer from problems with convergence (Nicenboim & Vasishth, 2016; Vasishth, Nicenboim, Beckman, Li & Kong, 2018), especially when a maximal random-effects structure is specified (Barr, Levy, Scheepers & Tily, 2013). In the current study, frequentist linear mixed effects models for measures of imitation of the differentials fit with *lme4* failed to converge. In contrast, Bayesian models always converge once regularizing priors are used (Vasishth et al., 2018). Finally, with Bayesian models, weakly informative priors can be set, which allow for a more conservative estimate of effects and a more natural interpretation of the findings through interpretable answers, such as the probability of a parameter falling within a credible interval.

All Bayesian multilevel regression models in this study were fitted using the Stan modelling language (Carpenter et al., 2016) and the package *brms* (Bürkner, 2016) in R (R Development Core Team, 2016). All plots were created using the package *ggplot2* (Wickham, 2016).

2.5 Model talker and shadower production of differentials

2.5.1 Model talker production of differentials

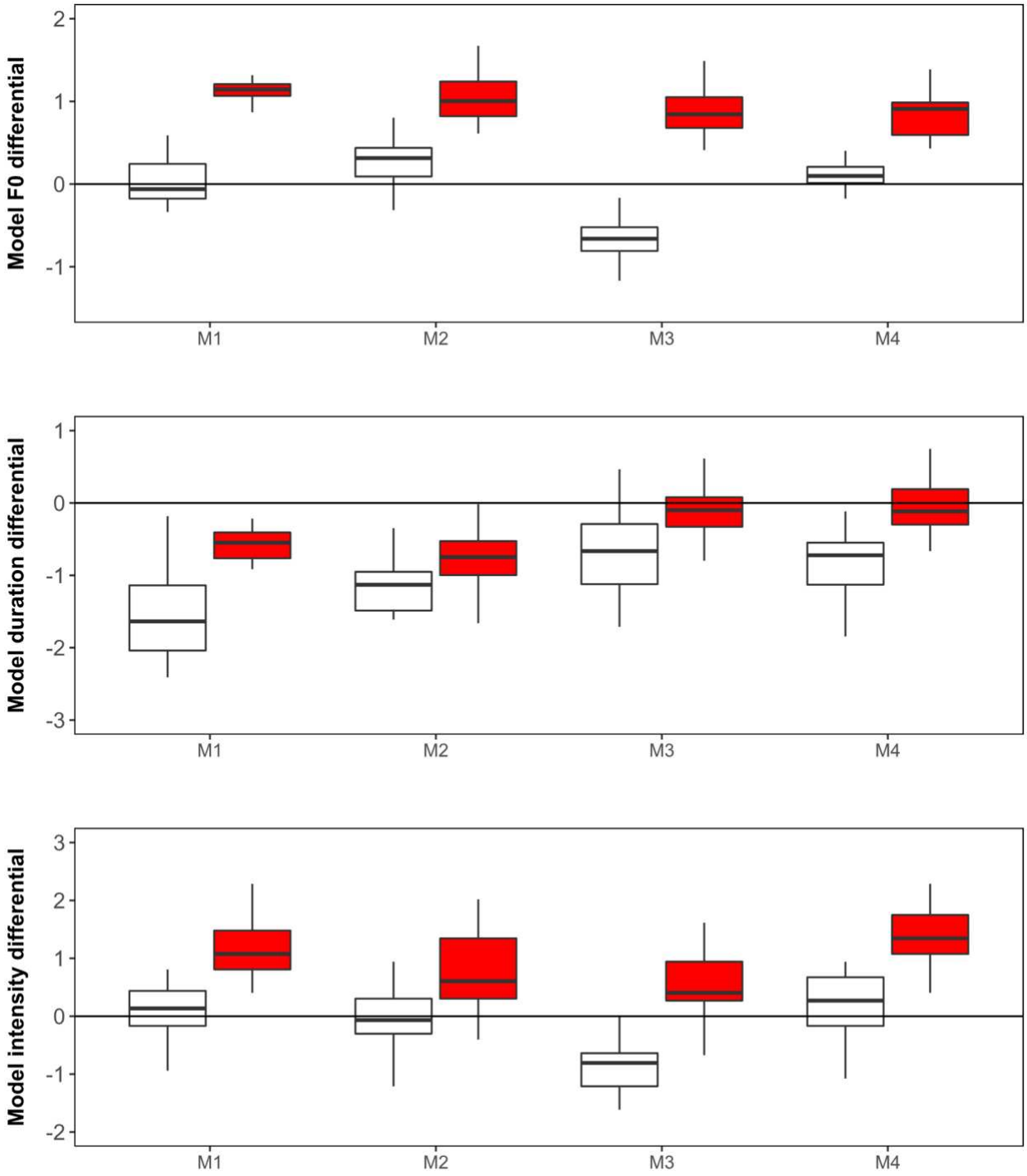
In this section, we present the model talkers' realization of stress in terms of how they produce the three differentials. The three differentials are scaled to be comparable, but are not centred, such that the sign of the differential corresponds between unscaled and scaled. Figure 1 shows the distribution of each of the three differentials, split by model talker. The black horizontal line shows where the differential is 0. For the F0 differential, three of the four models tend towards having positive differentials for both stress patterns, while Model 3 has the expected pattern with positive F0 differential for SU words and negative for US words.

For the duration differential, the models differ somewhat in how they distinguish between the two stress patterns, but all four tend to have a negative duration differential in all words, indicating that the

second vowel is longer than the first. This is likely the result of final lengthening, where the vowel in a phrase-final syllable is lengthened, causing the duration differential to be negative for both stress patterns (Hualde 2012: 165; Kim 2020; Nadeu 2013: 13). Model 4's pattern is closer to the expected one, with a greater proportion of the SU words being realized with a positive duration differential.

For intensity differential, the patterns are similar except that they tend towards positive differentials across the board. This suggests that most of the model talkers had falling intensity as the word unfolded. Note that while the differentials do not always follow the expected pattern in terms of positive for SU words and negative for US words, each model makes a distinction between the SU and US words, where the SU words tend to have a higher differential than the US words.

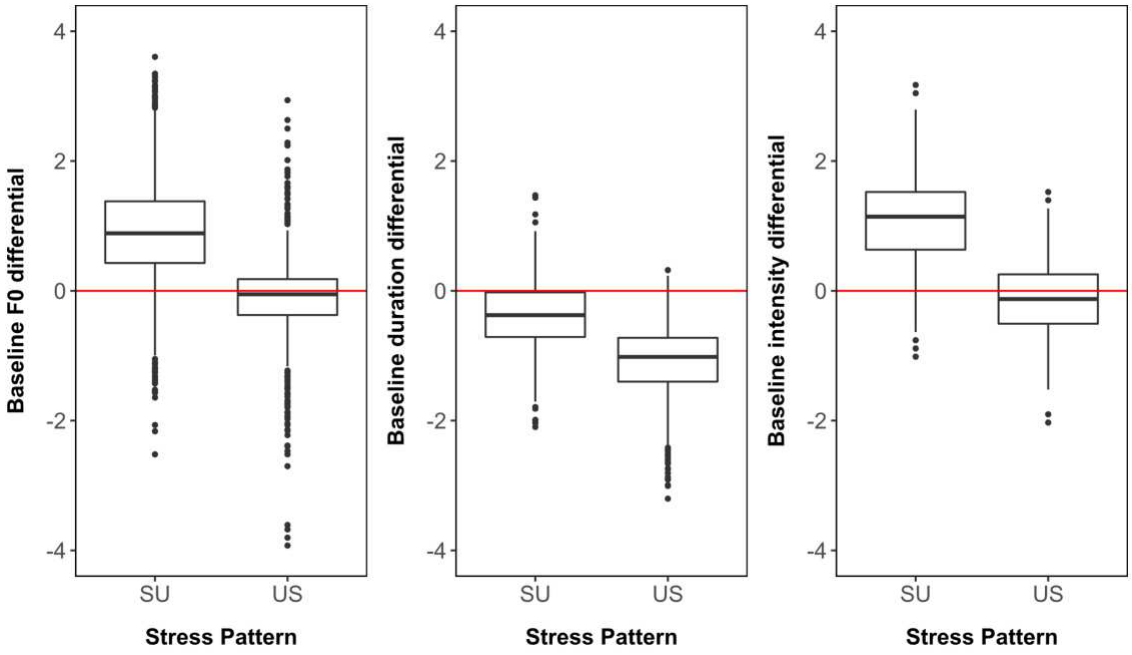
Figure 1 Boxplots of the three differentials by stress pattern for each model talker. White boxes indicate US words and red boxes indicate SU words.



2.5.2 Shadower baseline production of differentials

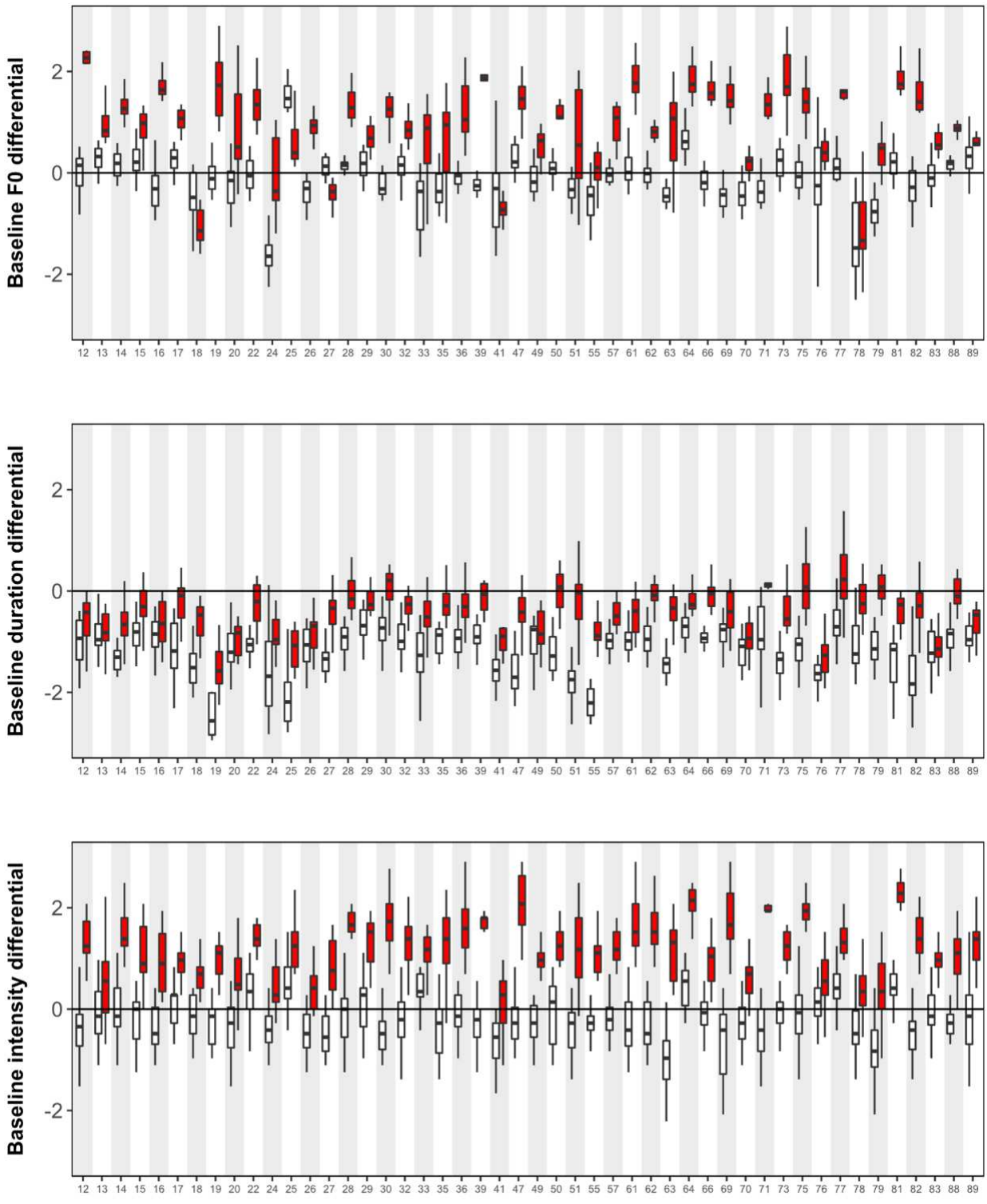
Next, we look at how the shadowers realize stress in terms of their production of the three differentials at baseline. Figure 2 shows the distribution of the three differentials for the shadowers' baseline production taking all shadowers together. The overall patterns are very similar to the model talkers' in that again there is a distinction between the two stress patterns for each differential with the SU words having the higher mean value than the US words. Furthermore, as with the model talkers, most of the tokens for both stress patterns have negative duration differentials, indicating that the second vowel is longer than the first, again likely the result of final lengthening (Hualde, 2012: 165; Kim, 2020; Nadeu, 2013: 13).

Figure 2 Boxplots of the three differentials by stress pattern as produced by the shadowers in the baseline phase.



While the group-level pattern in Figure 2 seems generally similar to the four model talkers, it might obscure significant individual variation. Figure 3 splits out the pattern by individual shadower to give a sense of the range of variation in how the shadowers realize the differentials in SU and US words. The shadower numbers are kept small to maintain readability of the plot and the vertical alternating bands of grey and white attempt to help separate the pairs of boxes by shadower.

Figure 3 Boxplots of differentials by individual shadower



The first panel of Figure 3 shows that there is a lot of variation in how the shadowers realized the distinction between SU and US words in terms of the F0 differential. Some shadowers, such as the first few on the left-hand side, produced a positive differential for almost all the words, while others, such as 18, 41, and 78, produced a negative differential for almost all the words. Others, such as 63, 69, and 82, followed essentially the expected pattern, with positive F0 differential for SU words and negative differential for US words. This individual variation might have resulted from differences in how the participants interpreted the words as belonging to a higher-order prosodic domain. F0 is implicated in prosodic patterns other than lexical stress, including sentence-level intonation and accent. When the shadowers made their baseline recordings, the words were presented one at a time, with a three second interval in between each word. As such, they were in isolation in that they were not part of a carrier phrase, but some shadowers might have interpreted the words as belonging to a list, and therefore produced list intonation. According to Morrill, Dilley & McCauley (2014), both high falling (HL: high-low) and low rising (LH: low-high) repeating patterns are common intonational contours associated with list constructions, at least for English. Other shadowers might not have applied a higher-level intonation pattern, treating each word as in isolation, producing roughly the expected 50/50 pattern between the stress patterns (i.e., HL for SU and LH for US). As noted above, we can see examples of all three of these patterns in the first panel of Figure 3.

For the duration differential, Figure 3 shows that most of the shadowers produce the majority of the words with a negative differential, regardless of stress pattern. This suggests that most of the shadowers realized the second syllable with final lengthening, as did the model talkers. There are a few exceptions, such 30, 50, and 71, whose SU words are primarily produced with a positive duration differential. Despite the tendency to negative duration differentials, there is still variation among the shadowers in terms of the magnitude of the distinction they make between SU and US words and in terms of how variable their differentials are within stress patterns.

In the intensity differential, most of the shadowers follow the expected pattern, with negative intensity differential for the US words and positive for SU. There are a few exceptions such as 22, 25, and 33, who have positive differentials for all the words, but notably no shadowers who produce all the words with a negative intensity differential. We see variation among the participants again in terms of the magnitude of the distinction between the stress patterns and the variability of the differentials within stress pattern.

Taken together, the panels in Figure 3 indicate that the shadowers vary in how they realize stress via the three differentials. Some of them would be similar to the model talker they shadow, but others will be quite different. Despite the shadowers and model talkers all being female speakers of the same dialect, Figure 3 shows that there is sufficient variability between them to create an opportunity for shadowers to shift their pronunciation to become more similar to the model talker.

2.6 Acoustic analysis assessment

There are two main approaches to using acoustic analysis to determine whether participants have imitated, both of which have pros and cons. The first is to calculate and summarize the Difference-in-Distance (DID) measure, which is intended to capture the amount of change that participants make from baseline to shadowed, relative to the model talker, using the formula |(baseline - model)| - |(shadowed - model)|. If the distance between the model and the participant decreases from the baseline to shadowing, then DID is positive, and we take this to mean convergence. If the distance increases from baseline to shadowing then DID is negative and we take this to mean divergence. DID has been commonly used and provides a word-by-word measure of phonetic imitation (Babel 2012; Clopper & Dossey 2020; Pardo et al. 2013; Pardo et al. 2017; Walker & Campbell-Kibler 2015).

However, there are problems with DID. First, DID is biased in two ways: (a) that talkers whose baseline is more distinct from an interlocutor or model talker will be found to converge more (i.e. have higher DIDs) than those who are more similar at baseline (Cohen Priva & Sanker, 2019; MacLeod, 2021); and (b) that the more extreme a baseline production is, the more likely it is to be counted as a

convergence when using DID (Cohen Priva & Sanker, 2019). Second, when considered at the group level, positive and negative DIDs can effectively cancel each other out (MacLeod, 2021). That is, the group-level mean DID can be found not to be significantly different from 0, even if there are many negative and positive shifts at the individual level, which could represent genuine convergences and divergences.

The second approach to measuring phonetic imitation using acoustic analysis is linear combination (Cohen Priva et al., 2017; Cohen Priva & Sanker, 2018, 2019; MacLeod, 2021). Linear combination uses linear mixed effects modelling to explore the extent to which the shadowed productions can be predicted by a combination of the participants' own baseline and the model talkers' productions. Linear combination avoids the problems associated with DID described above, making it a better choice for determining if shadowers have imitated at the group level, but it does not provide a word-by-word measure of phonetic imitation. Since one of the goals of this study is to investigate the extent to which listeners use acoustic changes in their perception of imitation, we need the word-specific estimation of imitation that DID provides. We do not know of any available alternative to DID for this purpose. As a result, in this study we use linear combination to assess imitation at the group level and we use DID to provide an estimate of imitation at the word level that is used in §3.5 when we explore the relationship between perception and production of imitation.

In phonetic imitation studies, DIDs are typically calculated for individual acoustic measurements, such as the pitch of a single vowel. In this study, the DIDs were calculated for the differentials, meaning that we are determining the extent to which the shadowers imitated the model talkers on the magnitude of the difference between the first and second vowels for pitch, duration, and intensity. DIDs more than three standard deviations away from the mean were removed from the dataset, separately by shadower and stress pattern. This resulted in 58 DIDs or 3% of the data being removed, leaving 1,625 datapoints. The means and standard deviations for each differential DID are given below in Table 1.

Table 1 Mean and standard deviation of the DIDs for each differential.

Variable	Mean (SD)

F0 differential DID[11]	0.068 ERB (0.47)
duration differential DID	6.09 ms (30.05)
intensity differential DID	0.43 dB (3.54)

2.6.1 Linear combination assessment of imitation

As noted above, in the linear combination approach, the shadowed productions are predicted using the shadowers' baseline productions and the model talkers' productions. If the model talker production is a significant predictor of the shadowed production, this indicates that the model talker has influenced the way the participants produced the words in the shadowing phase. If the coefficient for model talker is positive, then the participants have converged, on average as a group. If the coefficient is negative, then the participants have diverged. However, we would expect a relationship between the way the shadowers produced the words in the shadowing phase and how the model talkers produced them simply due to the same list of words being produced in both cases. To control for this, the shadowers' own baseline productions are also included as a predictor[12]. This predictor captures the extent to which the participants are consistent in their productions of the words from baseline to shadowing, while the model talker predictor captures convergence, or the extent to which the model talkers influence shadowing.

In this analysis, we use Bayesian linear mixed-effect modelling. We constructed three linear mixed effects models, one for the shadowed production of each differential. For each model, there were three fixed effects and two interactions. The fixed effects were the shadower's own BASELINE, the MODEL TALKER production, and STRESS PATTERN. STRESS PATTERN has two levels, SU and US, corresponding to

[11] The mean F0 differential in Hz is 3.66 Hz with a standard deviation of 24.27.

[12] However, we would also expect the model talkers' production and the shadowers' baseline production to be related, again simply due to it being the same list of words being read in both cases. As such, it is possible that these two predictors are collinear, which can cause problems for the interpretation of model estimates. We tested for multicollinearity using the collin.fnc function in the *languageR* package (Baayen & Shafaei-Bajestan, 2019), which provides a measure of collinearity, known as the condition number κ. According to Tomaschek, Hendrix & Baayen (2018), κ values higher than 15 indicate harmful levels of collinearity, while values higher than 30 indicate severe collinearity, for which corrective action is needed. For our study, κ values were well below 15 (F0: 2.36; duration: 3.58; intensity: 2.40), indicating that the baseline and model talker productions are not strongly collinear in any of the differentials.

the position of stress in either the first or second syllable, with US set as the reference level. STRESS PATTERN was sum-coded and centred at 0 to ensure that the coefficient for SU was compared to the mean of the reference level (US), but that the overall intercept of the model was calculated as the grand mean of both levels. Two interactions were also included to explore two influences on the amount of imitation. The first was included to determine whether shadowers imitated to different extents depending on whether the first or second syllable was stressed. This possibility was tested by including an interaction between MODEL TALKER and STRESS PATTERN. A second interaction was included to explore whether shadowers who started out as more distinct from the model talker imitated more. That is, we aimed to test whether greater starting distance gives rise to greater imitation, as has been suggested in previous work. For example, in Babel's (2010) study, New Zealand English speakers converged towards an Australian English model talker the most on the DRESS vowel, which is the most distinct between the two varieties of English. Other work has come to similar conclusions that greater baseline distance gives rise to greater imitation (Babel, 2010; Babel, 2012; Walker & Campbell-Kibler, 2015). However, as noted in the previous section, recent work has shown that some of these findings might be due to the Starting Distance Bias (Cohen Priva & Sanker, 2019; MacLeod, 2021), due to the way that DID restricts the range of possible convergences as a function of starting distance (MacLeod, 2021). Given this, it is not clear that greater starting distance should give rise to greater imitation. Indeed, Cohen Priva & Sanker (2019) and MacLeod (2021) found no evidence of such a relationship when using an alternative method of measuring imitation (linear combination), which is not subject to the Starting Distance Bias. However, the idea that a shadower who starts out as more distinct from a model talker might imitate more is intuitive: if there is more room to move, more imitation can take place. We test the idea here by including an interaction between MODEL TALKER and the difference between the model talker and shadower at baseline. Following Cohen Priva & Sanker (2019) and MacLeod (2021), the interaction is MODEL TALKER : ABS (BASELINE – MODEL TALKER). This interaction tests whether the influence of the model talker production on the shadowed production depends on how different the shadower and model talker are at baseline.

The models also included random effects of *shadowerID*, and *word*, with by-*shadowerID* slopes for the effects of BASELINE and MODEL TALKER. For each differential, we tested a model with a by-*modelID* slope for MODEL TALKER[13]. That is, we included a random slope that would capture variation in how much the model talker's production influenced the shadowed production among the four model talkers. We then compared model fit with that of a simpler model without a random slope for *modelID* using the Watanabe Akaike Information Criterion (WAIC: Watanabe, 2010). If this random effect improved model fit, this would indicate that the shadowers imitated to different extents depending on which model they shadowed. If it did not improve model fit, we took this to mean that there was no significant variation in how much imitation each model talker elicited. The dependent variable and continuous fixed effects were scaled and centred. Priors were set using a Cauchy distribution, with a mean of 0 and standard deviation of 2.5. The Cauchy distribution is a special case of the Student-*t* distribution with 1 degree of freedom. This distribution is recommended as a prior for normal mean parameters (Jeffreys, 1961). Two sampling chains ran for 3,000 iterations with a warm-up period of 1,000 iterations for each model, yielding 4,000 samples for each parameter tuple. In Table 2, we report the estimates and their 89%[14] credible intervals (CIs). The estimate is the mean of the posterior distribution. If a factor has an estimate of 0, this suggests that that factor did not influence the dependent variable. More meaningful, however, is the interpretation of the CI. A CI that does not contain zero suggests there is evidence that a given effect is non-zero. From Table 3, we can see that for each differential, the estimates for BASELINE are positive, with corresponding CIs that do not contain 0 in all cases, indicating that the participants show consistency from baseline to shadowing, as expected. The estimates for MODEL TALKER are also positive with CIs not containing 0 for all three differentials, meaning that as a group the participants

[13] Note that including a random intercept for *modelID* is not meaningful here. This is no reason to expect variation in how the shadowed productions are realized between model talkers, rather it is the *influence* of the model talker on the shadowed production where we might find variation. To this end, the slopes were specified as (0 + MODEL TALKER | *modelID*) for each differential.

[14] Eighty-nine percent CI were chosen since 95% CI may not be appropriate for Bayesian posterior distributions due to a potential lack of stability if too few posterior samples are drawn (Kruschke, 2015; McElreath, 2020).

converge towards the model talker on these measures. The magnitude of the estimate for MODEL TALKER is greatest in the F0 differential model (0.33) followed by the duration differential model (0.26) and is the least in the model for intensity differential (0.12). This suggests that while the shadowers converged towards the models on all three differentials, they did so to the greatest extent on F0 differential, followed by duration differential, and the least on intensity differential. This order aligns with our prediction, where we expected that shadowers would imitate the differentials in relation to their strength as cues to stress in Spanish. Furthermore, note that for the duration and intensity differentials, the magnitude of the estimate is higher for BASELINE than for MODEL TALKER, suggesting that the participants' shadowed productions are more similar to their own baseline productions than they are to the model for those differentials. In the case of F0 differential, however, the influence of the MODEL TALKER is much greater than that of the BASELINE. This indicates that for this differential, the shadowers were less consistent between the phases and converged more.

Table 2 ***Summary of output of Bayesian linear combination models to assess imitation in the three differentials***

Differential	Predictor	Estimate	Std. Error	*Credible Interval*
F0 differential	intercept	0.34	0.07	[0.23, 0.46]
	baseline	0.19	0.04	[0.14, 0.25]
	model talker	0.32	0.08	[0.21, 0.46]
	stress_SU	0.47	0.12	[0.28, 0.68]
	model talker : stress_SU	-0.04	0.12	[-0.21, 0.14]
	model talker : baseline distance	-0.03	0.04	[-0.09, 0.03]
duration differential	intercept	-0.37	0.05	[-0.44, -0.30]
	baseline	0.37	0.03	[0.32, 0.42]
	model talker	0.27	0.04	[0.21, 0.32]
	stress_SU	0.22	0.05	[0.14, 0.31]
	model talker : stress_SU	0.06	0.05	[-0.02, 0.14]
	model talker : baseline distance	-0.02	0.03	[-0.06, 0.03]
intensity differential	intercept	0.31	0.06	[0.22, 0.40]
	baseline	0.31	0.03	[0.26, 0.37]
	model talker	0.14	0.09	[0.03, 0.27]
	stress_SU	0.66	0.09	[0.52, 0.80]
	model talker : stress_SU	0.09	0.05	[0.02, 0.17]
	model talker : baseline distance	-0.03	0.03	[-0.08, 0.01]

In each model, STRESS PATTERN has the expected effect. When STRESS PATTERN is SU, the differentials are much higher than when the STRESS PATTERN is US. For all three differentials, STRESS PATTERN does

not seem to influence how much the shadowers will imitate, since the CIs for the interaction between MODEL TALKER and STRESS PATTERN contain 0. Lastly, the interaction between MODEL TALKER and BASELINE DISTANCE has a small estimate and a narrow CI containing 0 for each differential. This suggests that the shadowers who started out as more distinct from the model talkers on the differentials did not imitate more than those who started out as more similar to the model talker.

Including the by-*modelID* slope for MODEL TALKER did not improve model fit in the cases of F0 and duration differentials, suggesting that the four model talkers did not elicit different amounts of imitation on these measures. However, model fit was improved by including the by-*modelID* slope in the model for intensity differential. Inspection of the coefficients for the slope showed that they ranged from 0.08 (for Model 2) to 0.18 (for Model 3), indicating that while there was variation in how much the models were imitated, none showed divergence overall and they were imitated within a relatively small range.

3.0 Experiment 2: Perception experiment

The second method of assessing imitation employed in this study is a perceptual assessment. This section provides details of this experiment including the materials used in the perception task (§3.1), the listeners who participated in the study (§3.2), and the procedure (§3.3). The results are detailed in two subsections, first focusing on the overall perceptual pattern in §3.4 and next the relation between the acoustic measures and the perception pattern using the combined-analysis method in §3.5.

3.1 Materials

As explained in §2.4, the model, baseline, and shadowed recordings from Experiment 1 comprise the materials for the perceptual experiment. These recordings were split into individual sound files and normalized in intensity to an average of 70dB.

3.2 Participants

Eighty-seven (35 females, 52 males) native Spanish speakers (a different group from the production experiment) participated as listeners in the perception task. They ranged in age from 28 to 60 years (average 28, median 24) and were mostly from Mexico (74)[15], with the remainder from other countries (Colombia: 3, Peru: 3, Spain: 3, Canada: 2, Ecuador: 1, Venezuela: 1). To our knowledge, there is no research that would suggest that speakers' use of the acoustic correlates to perceive and produce stress would depend on the variety of Spanish that they speak. None reported any hearing problems, and all were compensated with $10 for their participation.

3.3 Procedure

Most studies using a perceptual assessment of phonetic imitation use a variant of the AXB task (e.g., Babel et al., 2014; Namy et al., 2002). In the current study, we chose to use a four-interval forced choice (4IAX; Pisoni & House Lazarus, 1974; Tuninetti, Whang & Escudero, 2019) task for two reasons. First, preliminary testing with an AXB task suggested that participants were sometimes attempting to compare the A and B tokens, which is not the goal of the task. The 4IAX task avoids this difficulty by pairing the X token with A and B separately. Second, the 4IAX task might allow listeners to access auditory information that is less available in an AXB task (Pisoni & House Lazarus, 1974). Although the AXB is the more standardly used task, we expect the 4IAX task to be just as reliable.

On each trial, the listeners heard a word repeated four times over high-quality headphones, resulting in two pairs: XA XB. X was always the model talker and A and B were baseline and shadowed words (counterbalanced) from one of the shadowers who shadowed that particular model talker. The listeners' task was to decide which of X and A (in the first pair) or X and B (in the second pair) were more similar to each other. If they thought that X and A were more similar, they pressed 'A' on the keyboard and if they thought X and B were more similar, they pressed 'L'. We took the proportion of

[15] Like the majority of the shadowers, the Mexican Spanish speakers from the perception task were residents of Mexico and were in the local Ottawa, ON, Canada area to complete a 3-week intensive language course in either English or French at the Universidad Nacionál Autónoma de México.

trials in which the listeners chose the pair involving the shadowed token to reflect the proportion of words in which the shadowers imitated the model talker. The stimuli in each pair were separated by a 100ms interval and each pair was separated by a 500ms interval. Each listener completed two blocks in the experiment, where each block included trials from one shadower + model pairing from Experiment 1. The 40 words were included twice, once with the shadowed token in the XA pair and once with it in the XB pair, for a total of 80 trials per block and 160 per listener. In total there were 24 versions of the experiment, with two shadower + model pairs per version. Each version was evaluated by between three and six listeners[16]. The experiment took approximately 20 minutes to complete and was run in OpenSesame (Mathôt, Schreij & Theeuwes, 2012).

The 87 listeners evaluated roughly 160 trials each[17], for a total of 13,264 trials. We removed DIDs more than three standard deviations away from the mean, separately by shadower and stress pattern. This resulted in 58 words being removed, corresponding to 503 trials in the perception data. We also removed trials with response times below 150ms or above 4000ms (572 rows), those corresponding to the 53 words that contained a devoiced vowel (416), and those corresponding to the 71 model, baseline or shadowed words produced with creaky voice (814). In total, 2,305 trials (17%) were removed from the perception data, leaving 10,959 trials.

3.4 Results of perceptual assessment

Altogether, the listeners chose the pair involving the shadowed token in 53.7% of trials. Perceptual analyses of phonetic imitation are typically subtle and highly variable (Pardo et al., 2018; Wagner et al., 2021). According to Pardo et al.'s (2018) survey of existing findings, the average proportion of shadowed tokens chosen as being more similar to a model talker or interlocutor is 56%. As such, the proportion of

[16] 18 versions were evaluated by 3 listeners, 4 versions were evaluated by 4 listeners, 1 version was evaluated by 5 listeners, and 2 versions were evaluated by 6 listeners.

[17] Depending on which shadower + model pair they listened to since some complete sets of data were not available; see §2.4.

53.7% found here falls into line with previous studies using the AXB task, in which listeners typically chose the shadowed token in 51% to 58% of trials (Kim, 2012; Pardo et al., 2017; Shockley et al., 2004; Wagner et al., 2021; Walker & Campbell-Kibler, 2015). Furthermore, since the current study considers imitation within a dialect and in same-sex pairs, we might expect the magnitude of imitation to be smaller than in studies of cross-dialectal imitation or among mixed-sex pairs, possibly resulting in that imitation being less easily perceived by the listeners in the perception task. Note, however, that some previous work has considered cross-dialect imitation and still had proportions of imitation detected that are not much higher than 50%. For example, listeners in Walker and Campbell-Kibler's (2015) study of imitation of New Zealand, Australian, and two American English varieties only identified imitation in 52.7% of trials.

To determine whether the proportion of 53.7% reflects evidence that the listeners perceived imitation, an intercept-only Bayesian multilevel logistic regression model was fitted to the responses from the perception task in *brms* (Bürkner, 2016). Two sampling chains ran for 2,000 iterations with a warm-up period of 500 iterations for each model, yielding 3,000 samples for each parameter tuple. The participants' responses were coded as a binary dependent variable: 0.5 on trials where the listener selected the pair that involved the shadowed token and -0.5 where the listener selected the pair that involved the baseline token. The intercept had an estimate of 0.14 with a credible interval of [0.11, 0.17]. This finding indicates that, at the group level, the listeners were more likely to choose the pair involving the shadowed token than they were to choose the pair involving the baseline token. Overall, the listeners did perceive imitation in the shadowers' recordings.

Next, four random effects were tested one by one: *listenerID, shadowerID, modelID*, and *word*. We were interested if adding the effect significantly improved model fit. After testing each of the four random effects by comparing the Watanabe Akaike Information Criterion (WAIC: Watanabe, 2010) it was determined that only *shadowerID* and *listenerID* significantly improved model fit. Because successive additions made to an original model may lead to overfitting, the simplest best-fitting model was selected. This model did not include *modelID* or *word* as random effects. Since including *word* as a random

intercept did not improve model fit, this suggests that the listeners did not perceive more imitation in certain words. Similarly, that a random intercept for *modelID* did not improve model fit suggests that the listeners did not perceive different amounts of imitation depending on which model talker had been shadowed. Since the random intercept of *shadowerID* did improve model fit, this suggests that there is a significant amount of variation in how much imitation the listeners perceived that depends on which shadower they were listening to. This finding aligns with previous work that has shown individual variation in the extent to which shadowers imitate (Babel et al., 2013; Lewandowski & Nygaard, 2018; Pardo et al., 2013; Wagner et al., 2021). Similarly, that the random intercept of *listenerID* improved model fit indicates that the listeners differ in how much imitation they perceive, suggesting that some listeners are better able to perceive the subtle changes than others. Variation in listeners' ability to perceive imitation has only been touched upon briefly in existing work, such as Babel & Bulatov (2012) and Pardo (2013), and could depend on factors such as language learning experience (e.g. Tremblay & Sabourin, 2012).

3.5 Combined analysis

The next step was to determine to what extent the listeners' performance in perceiving imitation was related to the shadowers' imitation of the three differentials. In other words, which of the differentials were the listeners using to perceive imitation? A Bayesian logistic multilevel regression model was tested that included the three differential DIDs, centred, as fixed effects, along with PAIR ORDER and STRESS PATTERN. PAIR ORDER refers to the position of the shadowed token relative to the baseline token in the XA XB pairs in the perception task. PAIR ORDER is coded as SB when A is the shadowed token and B is the baseline token and as BS when A is the baseline token and B is the shadowed token. STRESS PATTERN has two levels, SU and US, corresponding to the position of stress in either the first or second syllable. Both PAIR ORDER and STRESS PATTERN were sum-coded and centred at 0 to ensure that the coefficient for each was compared to the mean of the reference level, but that the overall intercept of the model was

calculated as the grand mean. For PAIR ORDER the reference level was BS and for STRESS PATTERN it was US. Weakly informative priors were used for all fixed effects and consisted of normal distributions with mean 0 and standard deviation 0.15. Weakly informative priors serve as a method of statistical regularization, shrinking the parameter estimates towards zero unless there is sufficient evidence for a large effect (McElreath, 2020). The model also included random intercepts for *shadowerID* and *listenerID* with by-*listenerID* slopes for the effects of PAIR ORDER and STRESS PATTERN[18]. In Table 3, we report the median effects under the posterior distribution and their 89% CI.

Table 3 Summary of logistic Bayesian multilevel regression model: combined-analysis method

Predictor	Median Estimate	Std. Error	Credible Interval
intercept	0.11	0.07	[0.00, 0.21]
stress	-0.06	0.05	[-0.15, 0.02]
pair order	0.17	0.08	[0.04, 0.29]
duration differential DID	0.08	0.02	[0.05, 0.12]
F0 differential DID	0.07	0.02	[0.04, 0.11]
intensity differential DID	0.05	0.02	[0.01, 0.08]

First, Table 3 shows that the median estimate for the intercept was 0.11 with a credible interval of [0.00, 0.21]. Notably, the lower bound of the CI rounds to 0. However, the probability of direction of the intercept is 94.4%. This means that 94.4% of the posterior distribution has the same sign as the estimate (i.e. positive). So, while the extent of the uncertainty around the intercept does encompass 0, most of the posterior is, in fact, positive for the intercept. This could indicate that, overall, the listeners were more likely to choose the pair involving the shadowed token than the pair involving the baseline token, even when taking the rest of the effects into account. It is likely the case that there are other influences beyond those investigated here that contribute to the perceptual pattern, as has been found in other investigations using the combined analysis method (refs). All three differential DIDs had CIs that only contained positive values. This means that the more closely the shadowers imitated the models on each of the three

[18] The effects of lexical frequency and the duration of the interval between the offset of the model talker production and the onset of the shadower repetition were also tested as fixed effects, but both were found to have 89% credible intervals that contained 0 (frequency: [-0.04, 0.02]; lag duration: [-0.01, 0.10]), suggesting that they have no significant influence on the perceptual pattern. Pardo et al. (2017) explains that previous results about the effect of lexical frequency on imitation have been mixed and their study also found no effect of frequency.

differentials, the more likely the listeners were to choose the shadowed token. However, the differentials varied in the extent to which they were related to the perceptual pattern. The duration differential DID had the greatest effect on perception with a median estimate of 0.08, followed by the duration differential DID at 0.07, with intensity differential DID having the least effect with an estimate of 0.05. These effect sizes fall in line with those found in previous studies that have investigated the relationship between acoustic and perceptual measures of phonetic imitation, as we do here. For example, Pardo et al. (2013) explored the influence of lexical factors on phonetic imitation. They tested the extent to which three DIDs (vowel duration, F0, and Euclidean distance) predicted the pattern of perception of imitation from an AXB task. Their analysis showed that all three DIDs were significantly related to the perceptual pattern, with the following coefficients: 0.065 for duration DID, 0.065 for F0 DID, and 0.057 for Euclidean distance[19]. In a later study, Pardo et al. (2017) considered the extent to which the same three DIDs influenced the listeners' perception of imitation in their study on the impacts of lexical frequency, talker sex, and model talker on phonetic imitation. The coefficients for the three DIDs were 0.08 for duration, 0.073 for F0, and 0.057 for vowel formants. Walker & Campbell-Kibler (2015) also used the combined analysis approach in their cross-dialectal study of phonetic imitation to determine to what extent Euclidean distance DID predicted the AXB results. The analysis showed that Euclidean distance DID only significantly predicted the AXB pattern in the New Zealand monophthong model with a coefficient of 0.08422. Taken together, the results for the differential DIDs in the current study suggest that the listeners make use of shifts on all three acoustic correlates of stress when perceiving imitation in the shadowers' productions, but not all to the same degree, using duration the most, followed by F0, and intensity the least.

The second line in Table 3 shows that there is a 0.89 probability that the value of STRESS PATTERN lies in the interval [-0.15, 0.02] and that this interval contains zero, which suggests that stress pattern did not influence the likelihood that the listeners would perceive imitation in the shadowers' productions. Our analysis also found that the order in which the pairs were presented affected the likelihood that the

[19] In that study, an interaction between duration DID and F0 DID was also significant with a coefficient of 0.076.

listeners would choose the pair involving the shadowed token over the pair involving the baseline token. When the shadowed token was presented in the first pair (PAIR ORDER = SB), the listeners selected the first pair as being more similar than the second pair in 57% of trials. In contrast, when the shadowed token was in the second pair (PAIR ORDER = BS), the participants were equally likely to select the two pairs. This finding parallels that of Pardo et al. (2013) in which the listeners chose the shadowed token more often when it was presented as the first item in the triad (A) than when it was the last (B). However, it is the opposite of Walker & Campbell-Kibler (2015) who found that their AXB listeners were more likely to choose the third token (B) than the first (A). However, in that study, the participants seemed to have a general tendency to choose the third token, since they were more likely to choose it regardless of the position of the shadowed token. In the current study, as in Pardo et al. (2013), the listeners were only more likely to choose the first pair when it contained the shadowed token (i.e., when pair order was SB). This suggests that rather than simply a response bias in favour of the first pair, the participants were better able to perceive imitation when the pair involving the shadowed token was presented first. It is beyond the scope of this paper to establish a definitive explanation of this effect, but one possibility is that the items in the first pair benefit from a primacy[20] effect and thus, the listeners are better able to identify imitation in this pair.

4.0 Discussion

4.1 Imitation of the acoustic correlates of stress

[20] A primacy effect reflects a robust finding in studies of memory for lists, where participants typically are better able to recall items presented earliest in the list as compared to those in the middle of the list (e.g. Postman & Phillips, 1965; Glanzer & Cunitz, 1966). While the current study is not a recall study, it could be that the primacy effect is caused by greater focus on earlier items (e.g., Morrison, Conway & Chein, 2014), which might help explain our participants being better able to perceive imitation in the first pair. Of course, memory studies also find recency effects, where items presented latest in a list are also more accurately recalled than those in the middle, which might have suggested that participants in the current study would be better able to perceive imitation when the shadowed token appeared in the second pair. Although we cannot explain the tendency for our participants to be better able to perceive imitation in the first pair, this variation is captured in our model by the predictor PAIR ORDER.

The results of the acoustic analysis of the shadowing data allow us to answer our first and second research questions. The first concerned whether shadowers would shift their production to become more aligned with the model talkers' acoustic realization of lexical stress (we predicted they would) and the second concerned whether the amount of imitation of the acoustic correlates would reflect their strength as cues to stress. In general, we might expect that shadowers would imitate the most linguistically salient cues more than less salient cues. Previous work on the production and perception of Spanish stress has suggested that, in accented contexts, including in isolation, F0 is the most relevant cue in perceiving which syllable is stressed, followed by duration, and then intensity (Hualde, 2005: 245, 2012: 165, Kim, 2015, 2020; Llisterri et al., 2003; Prieto & Torreira, 2007; Quilis, 1981). Based on this work, if shadowers do imitate how stress is realized, we might expect that they would imitate the F0 differential most and intensity differential the least, with duration differential falling in between. As discussed in §2.6.1, using Bayesian linear combination, both predictions were confirmed. The shadowers converged towards the model talker on all three differentials, imitating F0 differential the most, followed by duration differential, and then intensity differential.

The results in §2.6.1 showed that the position of stress in the word had a strong influence on the magnitude of the differentials, in the direction that we expected: SU words had higher differentials than US words. However, it does not seem that shadowers imitate the differentials to different extents depending on the stress pattern. As noted in §2.6.1, for the intensity and duration differentials, the shadowers' own baseline exerted more of an influence on the shadowed production than the model talker did. This means that while the shadowers did imitate the model talkers on those differentials, the shadowed production was still highly related to their baseline production. That is, the shadowers were still more consistent than convergent. However, the same was not true for the F0 differential. There, the influence of model talker was much higher (0.32) than the influence of the shadowers' own baseline (0.19). Why does F0 differential pattern differently from the other differentials in this respect? One possibility stems from the variation in the shadowers' baseline production of F0 differential that we noted in the first panel of Figure 3. Some shadowers produced all the words with a falling pitch pattern, where

the pitch of the first vowel was higher than the second. Others did the opposite, using a rising pitch pattern in all words. Some shadowers produced the SU words with falling pitch and the SU words with rising pitch. As we explained in §2.5.2, this variation could have been caused to some extent by differences in the higher order domain that the shadowers determined the words to belong to. Some produced a list intonation, in which pitch rises across each word, until falling on the final word[21]. However, none of the four model talkers used this same rising pitch pattern on SU words. As such, whenever a shadower used rising pitch on an SU word in the baseline, the model talker's production of the same word would not "match". That is, where the shadower would have a rising pitch, the model talker they shadowed would have a falling pitch. This mismatch creates an opportunity for a very large imitation, in which not only does the F0 differential come to be more similar to the model talker's, but it reverses its sign, from a negative to a positive. The opposite also occurred where the model talker produced a US word with rising pitch and the shadower produced it at baseline with a falling pitch. In those cases, imitation could cause another large shift, but from positive to negative. These opportunities for large shifts on F0 differential are made possible by the variety of ways that the shadowers produced their baseline word list, using list intonation with rising or falling intonation or by producing each word purely in isolation without imposing an intonation contour across them. In contrast, duration and intensity differentials are not afforded so many opportunities for reversing the sign of the differential given that intonation is much more strongly cued by F0 than by duration or intensity. Taken together, we argue that the model talker is able to exert a much bigger influence on the shadower on F0 differential than for the other differentials due to F0's tight connection with the realization of intonation. This position is supported by the relatively wide credible interval for the influence of the model talker on the shadowed F0 differential ([0.21, 0.46]) as compared to those for the duration ([0.21, 0.32]) and intensity differentials ([0.03, 0.27]). The wider CI for the F0 differential indicates that there is more uncertainty in

[21] In fact, in this study, participants were not aware of which word was the final one while producing it and so we do not expect to see falling intonation on the final word in the list.

the model about the extent to which the model talker influences the shadowed production; that is, there is more uncertainty around how much the shadowers imitate. This could be due to some shadowers imitating the F0 differential more than others. In contrast, in duration and intensity differentials, there is less uncertainty surrounding how much the shadowers imitate, suggesting that they were less variable in this behaviour.

Our acoustic analysis assessment of imitation also explored the role of baseline distance in explaining variation in how the shadowers would imitate. Some studies have found a relationship between baseline distance and imitation, where shadowers who are more distinct from the model talker at baseline are found to converge towards the model talker more than those who start out as more similar (e.g. Babel, 2010, 2012; Clopper & Dossey, 2020; Kim & Clayards, 2019; Walker & Campbell-Kibler, 2015). However, much of this work has used the DID metric, as discussed earlier, which has been shown to be biased to find that greater starting distance gives rise to greater convergence (Cohen Priva & Sanker, 2019; MacLeod, 2021). In this study, we did not use DID to assess imitation, but instead used linear combination, which has shown not to have this bias (Cohen Priva & Sanker, 2019; MacLeod, 2021). We tested whether starting distance influenced the amount of imitation by including interaction terms in our three differential models. As explained in §2.6.1, for all three differentials, the effect of the interaction was small, and each had a wide CI that encompassed 0. These results provide very little evidence that the size of the starting distance affects how much the shadowers imitate any of the differentials, paralleling the findings of Cohen Priva & Sanker (2019) and MacLeod (2021).

4.2 Relationship between perceptual and acoustic measures of imitation

Our third research question concerned whether listeners could perceive imitation in the shadowers' production. The findings of the analysis in §3.4 determined that, overall, listeners were more likely to choose the pair involving the shadowed token than the pair involving the baseline token. This finding suggests that listeners are sensitive to acoustic variation caused by imitation, confirming our prediction

and supporting previous work that reached the same conclusion (Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2013; Pardo et al., 2017; Walker & Campbell-Kibler, 2015). Our fourth research question asked whether listeners would use imitation of the three acoustic correlates when making their judgements[22] and the fifth asked whether they would do so in relation to how much the shadowers had imitated the correlates. The combined analysis in §3.5 determined that listeners did use imitation on all three acoustic correlates of stress in the perception task, as predicted, supporting previous work that has also found F0 and duration measures to be related to listeners' perception of imitation (Pardo et al., 2013; Pardo et al., 2017; Wagner et al., 2021). Furthermore, the analysis showed that the listeners made use of imitation on the duration differential the most in making their judgements, followed closely by the F0 differential, with intensity differential used the least. We predicted that the listeners would rely on the differentials in their perception of imitation in relation to the extent to which the shadowers had imitated them. As we have seen, the order is roughly aligned, but not exactly the same. Whereas the shadowers imitated F0 differential the most, followed by duration differential, the listeners used these two differentials to the same extent in the perception task. Intensity differential was the least imitated and least used in the perception task.

Why might the listeners not have used F0 differential the most, given that the shadowers imitated that differential more than the others? One possibility could be that some of the changes that shadowers made are more salient than others, leading to variation in how much the listeners can rely on them. Imitation of F0 differential would likely be most salient when it creates a change in the direction of pitch contour from baseline to shadowing that causes the shadowing to align with that of the model talker (Wagner et al., 2021). This would happen when shadowers changed from rising to falling pitch or from falling to rising to align with the pattern used by the model talker. However, such direction-changing

[22] However, note that studies that use the combined-analysis method as in the current study do not directly test which acoustic cues the listeners are using, such as by generating highly controlled stimuli for the perception task that systematically adjust the various acoustic parameters, such as in Kim & Clayards (2019). The Bayesian logistic mixed effects regression model shows that there is a relationship between the three acoustic DIDs and the perception pattern, which suggests, but cannot conclusively state, that the listeners are using those measures in their perception of imitation.

shifts were not available on all trials since many of the baseline productions already aligned in the general direction of the pitch pattern in the model talker. As a result, the shadowers would be variable in how much they imitated F0 differential, this could lead to less consistently available shifts on F0 differential for the listeners in the perception task to make use of. This would cause the overall effect of F0 differential DID to be lower in the combined analysis model.

Another possible reason why F0 differential DID was not the strongest influence on perception could result from differences among the participants in terms of auditory perception bias. Postma-Nilsenová & Postma (2013) explored the F0 imitation performance of shadowers who differed in their perceptual processing of F0. One group was composed of so-called "spectral listeners", meaning they decomposed the sound signal into individual harmonics, while another group, "fundamental listeners", perceived the harmonics as a whole. Postma-Nilsenová and Postma found that fundamental listeners were better able to imitate F0 of a model talker than were spectral listeners. They point out that their findings could partly explain observed individual variation in imitation of F0, such as in Babel & Bulatov (2012). In addition to differences between spectral and fundamental listeners in terms of how much they imitate, we might also expect differences between the two listener types in their performance in perceiving imitation. As such, variation between listeners in the current study in terms of the extent to which they used imitation of F0 differential to perceive imitation might be partly explained by auditory processing bias. If enough of the participants in the perception task were spectral listeners, this might mean that the reliance of the group on F0 differential to perceive imitation could be lowered. However, this cannot be investigated further since we cannot determine which type of listeners the participants in this study are.

Another contributing factor could be that, while F0 is expected to be the most salient cue to Spanish stress, imitation of F0 is perhaps more perceptible to listeners when it conveys social information. For example, Gregory et al. (1997) suggested that F0 is commonly imitated when it conveys emotion and attitude. In the current study, F0 differential was imitated more than the other differentials, but it was not used the most by listeners in their holistic perception of imitation. Perhaps the reason for this is because F0 was conveying a grammatical concept (i.e. stress) instead of social information.

While the idea that listeners would use imitation of different measures to the extent that shadowers imitated them is intuitive, empirical evidence about whether this pattern emerges is unclear. On the one hand, some have found a general alignment between amount of imitation on certain acoustic measures and use of that imitation in perception. For example, Pardo et al. (2017) investigated the effect of lexical frequency, talker sex, and model talker on the imitation of F0 and vowel duration (and F1 and F2) of English vowels. The acoustic analysis found that the shadowers had converged towards the model talkers at the group level for vowel duration, but not for F0. The results of the combined analysis indicated that the listeners used imitation on vowel duration to a greater extent than imitation on F0 when making their judgements, showing a rough alignment between amount of imitation and use of that imitation in perception. Clopper & Dossey (2020) explored convergence on word duration, /ai/ monophthongization, and vowel fronting. The acoustic analysis showed that the shadowers imitated the most on word duration, followed by vowel fronting, with the least imitation on /ai/ monophthongization. The combined perceptual analysis suggested that the listeners used the acoustic cues available to them in relation to the extent to which the shadowers imitated those cues. However, in Clopper & Dossey (2020), the amount of imitation was defined in terms of how consistently the shadowers imitated on each variable, rather than by magnitude of imitation. For example, imitation on word duration was the most consistent in that it was detected in words containing three of the four vowels included in the study, whereas imitation of vowel fronting was less consistent because it was detected in two of four vowels. This approach to defining amount of imitation might generate different predictions regarding the relationship between the production and perception of imitation than if the magnitude of the shifts were used. On the other hand, other work has found no relationship between amount of imitation and use of that imitation in perception. Babel & Bulatov (2012) assessed imitation of F0 using both acoustic analysis and an AXB task. Although imitation was found in both assessments, the amount of imitation was not correlated between the two assessments. They note that this might mean that AXB tasks do not "reflect the robustness with which participants accommodate to particular aspects of the acoustic signal" (Babel & Bulatov, 2012: 16), an assertion that would certainly apply to the 4IAX task as well. Pardo, Jordan et al.

(2013) found that while vowel duration and F0 were not imitated when taking all the shadowers together, in the combined analysis, imitation of these two measures was equally used (both with estimates of 0.065).

The results of the current study fall somewhere in between previous findings: the alignment between amount of imitation and use of that imitation is rough, but not exact.

4.3 Other contributors to the perception of imitation

As noted in §3.5, the intercept in the combined-analysis model was positive, with a probability of direction of 94.4%. We took this to mean that there was some evidence that listeners were more likely to choose the pair involving the shadowed token, even with the differential DIDs and the effect of pair order included in the model. This could mean that the listeners were making use of other factors beyond the amount of imitation of the differentials in making their judgements. Such factors could include imitation of other aspects of the vowels, including the first and second formants. These measures were not included in the current study since the goal was to explore imitation of the acoustic realization of stress, and, as noted earlier, Spanish vowels are typically not reduced to nearly as great an extent in unstressed syllables as in English (Ortega-Llebaria & Prieto, 2011; Quilis, 1981; Quilis & Esgueva, 1983). Other factors could include imitation of segments not considered here, including the consonants. For example, in Spanish, the voiced stops /b, d, g/ in onset position are realized as full stops [b, d, g] when they are utterance-initial or after nasals (or after laterals in the case of /d/), but lenite to the approximants [β, ð, ɣ] in other contexts (Harris, 1969; Mascaró, 1984; 1991; Romero, 1995). Furthermore, the extent to which they lenite is variable and depends on a range of factors, including stress. In contexts where lenition is expected, voiced stops will weaken to a greater extent when they are the onset of an unstressed syllable than in a stressed syllable (Carrasco, Hualde & Simonet, 2012; Colantoni & Marinescu, 2010; Cole et al., 1999; Eddington, 2011; Ortega-Llebaria, 2004; Simonet, Hualde & Nadeu, 2012). Given this, it is possible that shadowers could imitate variation in the realization of the consonants, perhaps especially in the voiced stops, which

are found in many of our stimuli, and listeners could be responding to such imitation in making their judgements. In cross-dialectal imitation studies, other factors relating to social salience are likely also at play (MacLeod, 2012b; Clopper & Dossey, 2020), but we expect these types of factors to be less relevant here since this is a within-dialect study.

Although we have argued here that our results show evidence of the shadowers having imitated the acoustic realization of stress, it is possible that they are instead imitating the raw values of duration, F0, and intensity of the first and second vowels. If they did this, the differentials would also appear to be imitated, but the target of imitation would in fact be the individual vowels. Previous work has shown that shadowers do imitate characteristics of individual vowels, such as F0, duration, and formant frequencies (e.g., Pardo et al., 2017; Walker & Campbell-Kibler, 2015). However, most of that work has focused on monosyllabic words, where it would not be possible to investigate relative measures as we did in this study. Here, we cannot conclusively rule out the possibility that shadowers are not targeting the realization of stress, but instead the individual realizations of the vowels. One of the main reasons for this is that it is essentially impossible to explore the imitation of intensity without the measure being relative to another segment in the word. That is, we cannot investigate imitation of the intensity of individual vowels. This is because, as noted in §1.3, measurements of intensity will be influenced by a variety of factors including the configuration of the microphone, any normalization procedure of recordings, and participant ability to adjust the volume. For the shadowers, the specific intensity of the individual vowels as produced by the model talkers is effectively lost by the time the recordings are presented in the shadowing task. The same thing applies to both the model talker and shadower productions when they are presented to the listeners in the 4IAX task. As such, intensity of individual vowels can be measured in Praat, but we cannot determine how a participant hearing the recording would perceive the intensity. Given this constraint, we can explore the imitation of relative measures involved in cuing stress, but not the measures on individual vowels.

4.4 Theoretical implications

The results of this study provide insight into three of the main theoretical accounts of phonetic imitation. First, under Communication Accommodation Theory, the reason that talkers imitate is social: they shift their pronunciation (or other aspects of speech) to become more similar to an interlocutor in order to decrease social distance. If phonetic imitation is indeed socially motivated, we would not expect it to occur in situations where social interaction is minimized or eliminated, such as in a shadowing task. Although shadowers are hearing and repeating another talker's voice, there is no interaction between them. That our study found evidence of imitation even in the absence of social interaction suggests that the mechanism behind imitation is not purely socially motivated, providing support for previous work that indicated the same (e.g. refs).

Other approaches to accounting for phonetic imitation posit that there is a direct connection between the perception and production systems, with the general idea that perception of a sound or word provides a set of instructions that are used by the production system, making phonetic imitation inevitable. As explained in §1.1, in Pickering and Garrod's (2013) interactive-alignment account, when talkers use covert imitation to generate a model of their interlocutor's speech, this creates a set of instructions for the talker's next production, leading to phonetic imitation. Pickering and Garrod's (2013) account also predicts that talkers will covertly imitate others who they perceive to be more similar to themselves in terms of linguistic or social characteristics. Applying this prediction to the current study, we might expect that shadowers who are more similar to the model talker at baseline would imitate more than shadowers who are more distinct from the model talker. However, it is not clear whether the similarity between talkers should be evaluated on the basis of individual acoustic measures, or whether it was intended to apply at a more holistic level. Furthermore, the prediction that talkers who are more similar to a model talker will converge more is the opposite of several previous shadowing studies in which talkers who were more distinct from the model talker at baseline were found to converge more (e.g. refs, although see refs). However, as discussed in MacLeod (2021), some of those findings might be due to the Starting Distance Bias introduced by using the DID measure. As such, the analysis of the extent of imitation in §2.6.1, which did not use DID, provides an opportunity to test this prediction of the automatic

alignment account. As explained in §2.6.1, each statistical model included a predictor designed to capture any effect of starting distance on the extent of imitation. If the prediction of the interactive alignment account holds, then we would expect the estimate for these predictors to be negative. This would mean that as starting distance increases, the influence of the model talker on the shadowed production decreases. For all three differentials, the results showed estimates that were very small and credible intervals that contained 0. As such, our findings do not provide evidence for greater imitation when starting distance is small. Note, as well, that our findings do not support the idea that greater imitation will occur when starting distance is large, either. In our results, starting distance did not influence imitation at all.

As noted in §1.2, Mantell & Pfordresher (2013) found that participants imitated relative pitch more than absolute pitch in their study of imitation of spoken utterances. Lin et al. (2021) note that this could mean that relative pitch is more central to speech processing than overall pitch. How might such a finding be integrated into theories of imitation, such as episodic memory models? In those approaches, during perception of a word or sound, exemplars are activated to different extents depending on their similarity to the incoming stimulus. A production target is then chosen based on the activation of traces in memory. Perhaps talkers not only compare similarity of acoustic properties of incoming stimuli to stored exemplars, but also compare the relation between those properties of syllables within the same word. If they do, this could help account for how a relative measure becomes a target of imitation, such as was proposed by Mantell & Pfordresher (2013) for relative pitch. The findings of the current study also support that relative measures can be the target of imitation; although, as noted in §4.3, we cannot determine conclusively that the participants did not target each vowel's acoustic properties individually.

In addition to being implicated in theories of speech perception and production, phonetic imitation is also centrally involved in the Change-by-Accommodation model of sound change (Niedzielski & Giles, 1999). In this account, short-term phonetic convergence is predicted to be a precursor stage to second dialect acquisition and community-level sound change. Several studies of phonetic imitation have considered how phonetic imitation might be related to sound change including Babel, McAuliffe & Haber

(2013) and Lin et al. (2021). However, sound change is not always occurring on all possible sounds and even if "given enough time, language change is inevitable…" (Blust, 2007:40), at any given moment, many sounds will not be undergoing change. To further our understanding of sound change and imitation's role in it, we need to also understand patterns of phonetic imitation in situations where change is not underway. Backus (2004) notes that demonstrating what does not change is an important aspect for theories of sound change that has been overlooked in the literature. It could be that in such situations, phonetic imitation works as an inhibiting influence on change, where over time individuals make relatively small shifts towards each other with the effect of remaining similar to each other. On the other hand, sound change does sometimes occur even without obvious influences of contact from other varieties or languages. As Lin et al. (2021) point out, studying phonetic imitation between speakers of the same dialect allows us to explore changes with internal origins. Accordingly, we argue that studying within-dialect imitation provides an opportunity to explore the malleability of pronunciation in precisely the context in which many speakers find themselves on a regular basis – that is, interacting with speakers of the same dialect. Understanding more about how phonetic imitation manifests within this context can help us to continue developing our theories of sound change.

4.5 Application to L2 acquisition

The current study contributes to the growing body of knowledge about phonetic imitation by showing that shadowers perceive fine-grained variation in the acoustic realization of stress, a suprasegmental property of words. The realization of stress provides another dimension that talkers can imitate and, as we saw in the combined analysis in §3.5, imitation of the acoustic correlates contributes to listeners' perception of convergence. Our findings also make some predictions about how English-speaking second language (L2) learners of Spanish might imitate the acoustic correlates of stress. For example, Kim (2020) found that whereas monolingual Spanish speakers realized stress in accented contexts using a combination of the three acoustic correlates, English-speaking L2 learners of Spanish relied primarily on duration to

distinguish between stressed and unstressed syllables and still were found to have a lot of overlap between oxytones and paroxytones in terms of duration differential. With respect to perception, while L2 learners were fairly accurate in identifying the position of lexical stress with paroxytones, they were significantly less accurate with oxytones. These findings from Kim (2020) might suggest that if L2 Spanish learners participated in a shadowing task such as the one in our study, they would be likely to imitate only the duration differential and not the other correlates of stress. However, this prediction could be different for learners with a high level of phonetic talent. Lewandowski and Jilka (2019) explored the influence of phonetic talent of German-speaking L2 learners of English on convergence with native speakers of English, where phonetic talent was assessed using a combination of speech perception, production tasks, and imitation tasks. They found that the phonetically talented learners converged more towards the native English speakers in a conversational accommodation task than the less talented learners. The authors conclude that phonetic talent is a requirement for convergence. Taken together, the results of Lewandowski & Jilka (2019), Kim (2020) and the current study suggest that English-speaking L2 learners of Spanish with greater phonetic talent would be more inclined to imitate all three phonetic correlates of stress in Spanish, whereas those with less phonetic talent might only imitate the duration differential. This connection could help explain variation between learners in their trajectory and ultimate attainment of the acoustic realization of lexical stress; individuals who demonstrate higher than average ability to imitate the acoustic correlates might be expected to acquire the realization of stress faster or more accurately.

5.0 Conclusions

The current study adds to the growing body of work showing that talkers imitate acoustic properties of model talker speech and that listeners can perceive that imitation. Our findings show that Spanish speakers imitate the acoustic realization of stress in Spanish and that the extent to which they do so mirrors the salience of each correlate as a cue to stress. We also found that listeners can perceive imitation and that they use imitation of the acoustic correlates of stress roughly in relation to how much they were imitated.

The current study makes several contributions to the phonetic imitation literature. First, it concentrates on the intermediate level of the word, helping to develop our knowledge of how dynamic properties of the word might provide opportunities for imitation and how listeners might use variation in these properties when perceiving imitation. Second, our study focuses on the imitation of Spanish stress, allowing an investigation into how the perceptual salience of different acoustic cues might influence the pattern of imitation. Third, we included Spanish disyllabic words, expanding our understanding of patterns of phonetic imitation outside of English monosyllables. Lastly, it provides a novel application of the 4IAX task to the perceptual assessment of phonetic imitation.

6.0 Acknowledgements

This work was supported by an Early Career Research Award from the Faculty of Arts and Social Sciences at Carleton University. We would like to thank all the participants who took part as model talkers, shadowers and listeners, Tania Freudenthaler for running parts of the experiment, Richard Edwards-Jones for help preparing the perception materials, and Kaya Gouin for careful work with the acoustic analysis.

7.0 Appendix

	Stressed vowels			Unstressed vowels		
	word	**IPA**	**meaning**	**word**	**IPA**	**meaning**
/i/	quito	[ˈki.to]	‘take away.1sg.pres.ind.’	quitó	[ki.ˈto]	‘take away.3sg.past.ind’
	vino	[ˈbi.no]	‘wine’	vital	[bi.ˈtal]	‘vital’
	dice	[ˈdi.se]	‘say.3sg.pres.ind’	diré	[di.ˈɾe]	‘say.1s.fut’
	cita	[ˈsi.ta]	‘appointment’	citar	[si.ˈtaɾ]	‘to make an appointment’
/e/	quedo	[ˈke.ðo]	‘remain.1sg.pres.ind’	quedó	[ke.ˈðo]	‘remain.3sg.past.ind’
	bebo	[ˈbe.βe]	‘drink.3sg.pres.ind’	bebé	[be.ˈβe]	‘baby’
	deja	[ˈde.xa]	‘leave.3sg.pres.ind’	dejar	[de.ˈxaɾ]	‘to leave’
	cero	[ˈse.ɾo]	‘zero’	cerró	[se.ˈro]	‘close.3sg.past.ind’
/a/	casa	[ˈka.sa]	‘house’	casar	[ka.ˈsaɾ]	‘to marry’
	base	[ˈba.se]	‘base’	balón	[ba.ˈlon]	‘ball’
	datos	[ˈda.tos]	‘data’	dará	[da.ˈɾa]	‘give.3sg.fut’
	saco	[ˈsa.ko]	‘throw.1sg.pres.ind’	sacó	[sa.ˈko]	‘throw out.3sg.past.ind’
/o/	cose	[ˈko.se]	‘sew.3sg.pres.ind’	coser	[ko.ˈseɾ]	‘to sew’
	vota	[ˈvo.ta]	‘vote.3sg.pres.ind’	votar	[bo.ˈtaɾ]	‘to vote’
	doce	[ˈdo.se]	‘twelve’	dolor	[do.ˈloɾ]	‘pain’
	sopa	[ˈso.pa]	‘soup’	soplar	[so.ˈplaɾ]	‘to blow’
/u/	cura	[ˈku.ɾa]	‘cure’	curar	[ku.ˈɾaɾ]	‘to heal’
	buzo	[ˈbu.so]	‘scuba diver’	buzón	[bu.ˈson]	‘mailbox’
	dura	[ˈdu.ɾa]	‘last.3sg.pres.ind’	durar	[du.ˈɾaɾ]	‘to last’
	suma	[ˈsu.ma]	‘addition’	sumar	[su.ˈmaɾ]	‘to equal’

8.0 Reference List

Ashby, M. G. (1978). A study of two English nuclear tones. *Language and Speech*, 21(4), 326–336.

Aubanel, V., & Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, 52(6), 577–586.

Aubanel, V., & Nguyen, N. (2020). Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task. *PLoS ONE*, 15(5), 1–16.

Baayen, R.H, & Shafaei-Bajestan, E. (2019). languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics. R package version 1.5.0. https://CRAN.R-project.org/package=languageR

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, 39, 437–456.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1), 178–189.

Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, 55(2), 231–248.

Babel, M., McAuliffe, M., & Haber, G. (2013). Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in Psychology*, 4(SEP).

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, 5(1), 123–150.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer Version 6.1.08, retrieved 9 December 2019 from http://www.praat.org/ (D. Weenink, Ed.).

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., … Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 539–543.

Brouwer, S., Mitterer, H., & Huettig, F. (2010). Shadowing reduced speech and alignment. *The Journal of the Acoustical Society of America,* 128(1), EL32–EL37.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–37.

Carrasco, P., Hualde, J. I., & Simonet, M. (2012). Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica*, 69(3), 149–179.

Chitoran, I. (2002). A perception-production study of Romanian diphthongs and glide-vowel sequences. *Journal of the International Phonetic Association*, 32(2), 203–222.

Clopper, C. G., & Dossey, E. (2020). Phonetic convergence to Southern American English: Acoustics and perception. *The Journal of the Acoustical Society of America*, 147(1), 671–683.

Cohen Priva, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, 141(5), 2989–2996.

Cohen Priva, U., & Sanker, C. (2018). Distinct behaviors in convergence across measures. Proceedings of the Annual Conference of the Cognitive Science Society, (July), 1518–1523.

Cohen Priva, U., & Sanker, C. (2019). Limitations of difference-in-difference for measuring convergence. Laboratory Phonology: *Journal of the Association for Laboratory Phonology*, 10(1), 15.

Cohen Priva, U., & Sanker, C. (2020). Natural leaders: some interlocutors elicit greater convergence across conversations and across characteristics. *Cognitive Science*, 44(10).

Colantoni, L., & Marinescu, I. (2010). The scope of stop weakening in Argentine Spanish. Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology, 100–114.

Cole, J., Hualde, J. I., & Iskarous, K. (1999). Effects of prosodic and segmental context on /g/-lenition in Spanish (O. Fujimura, B. D. Joseph, & B. Palek, Eds.). 4th International Linguistics and Phonetics Conference, pp. 575–589.
Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, and Psychophysics*, 78(1).
Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology,* 4(Article 346).
Eddington, D. (2011). What are the contextual phonetic variants of /β, ð, ɣ/ in colloquial Spanish? *Probus*, 23, 1–19.
Giles, H. (1973). Accent mobility: a model and some data. *Anthropological Linguistics*, 15(2), 87–105.
Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, N. Coupland, & J. Coupland (Eds.), Contexts of Accommodation (pp. 1–68). Cambridge: Cambridge University Press.
Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 5(4), 351–360.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.
Gorisch, J., Wells, B., & Brown, G. J. (2012). Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation. *Language and Speech*, 55(1), 57–76.
Gregory, S. W., & Hoyt, B. R. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psycholinguistic Research*, 11(1), 35–46.
Harris, J. (1969). Spanish phonology. Cambridge: MIT Press.
Hualde, J. I. (2002). Intonation in Spanish and the other Ibero-Romance languages: Overview and status quaestionis. In C. W. and J. Camps (Ed.), Romance phonology and variation. Selected papers from the 30th Linguistic Symposium in Romance Languages, Gainsville, Florida, February 2000 (pp. 101–116). Amsterdam: Benjamins.
Hualde, J. I. (2005). The Sounds of Spanish. New York: Cambridge University Press.
Hualde, J. I. (2012). Stress and Rhythm. In The Handbook of Hispanic Linguistics (pp. 153–171).
Hualde, J. I., Simonet, M., & Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Laboratory Phonology*, 2(2), 239–301.
Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting BF10s. *The Journal of Problem Solving*, 7(1).
Jeffreys, H. (1961). Theory of probability (3rd edition). Oxford: Clarendon Press.
Kim, D., & Clayards, M. (2019). Individual differences in the relation between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, 34(6), 769–786.
Kim, J. (2015). Perception and production of Spanish lexical stress by Spanish heritage speakers and English L2 learners of Spanish. Selected Proceedings of the 6th Conference on Laboratory Approaches to Romance Phonology, (January 2015), 106–128.
Kim, J. (2020). Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of Spanish lexical stress. *Bilingualism: Language & Cognition*, 23(2), 233–250.
Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, 2, 125–156.
Kruschke, J. K. (2015). Doing Bayesian Data Analysis: A Tutorial with R and BUGS (2nd ed.). Academic Press, Inc.
Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH, 3081–3084.
Lewandowski, E. M., & Nygaard, L. C. (2018). Vocal alignment to native and non-native speakers of English. *The Journal of the Acoustical Society of America*, 144(2), 620–633.
Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, 4(18), 1–19.

Llisterri, J., Machuca, M., de la Mota, C., Riera, M., & Ríos, A. (2003). The perception of lexical stress in Spanish. In M. J. Solé, D. Recasens, & J. Romero (Eds.), Proceedings of the 15th International Congress of Phonetic Sciences (pp. 2023–2026). Barcelona: Causal Productions.
MacLeod, B. (2012a). Investigating L2 acquisition of Spanish vocalic sequences. *Studies in Hispanic and Lusophone Linguistics*, 5(1), 103–148.
MacLeod, B. (2012b). The effect of perceptual salience on cross-dialectal phonetic convergence in Spanish. PhD dissertation. University of Toronto.
MacLeod, B. (2021). Problems in the Difference-in-Distance measure of phonetic imitation. *Journal of Phonetics*, 87, 101058.
Mascaró, J. (1984). Continuant spreading in Basque, Catalan, and Spanish. In M. A. and R. T. Oehrle (Ed.), Language Sound Structure (pp. 287–298). Cambridge, MA: MIT Press.
Mascaró, J. (1991). Iberian spirantization and continuant spreading. Catalan Working Papers in Linguistics, 1, 167–179.
Mathôt, S., Schreij, D., & Theeuwes, J. (2012, June). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, Vol. 44, pp. 314–324.
McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN. Milton: CRC Press LLC.
Morrison, A. B., Conway, A. R. A., & Chein, J. M. (2014). Primacy and recency effects as indices of the focus of attention. *Frontiers in Human Neuroscience*, 8(JAN), 1–14.
Morrill, T. H., Dilley, L. C., & McAuley, J. D. (2014). Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics*, 46(1), 68–85.
Nadeu, M. (2013). The effects of lexical stress, intonational pitch accent, and speech rate on vowel quality in Catalan and Spanish. PhD dissertation. University of Illinois at Urbana-Champaign.
Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: the role of perception. *Journal of Language and Social Psychology*, 21(4), 422–432.
Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11), 591–613.
Niedzielski, N., & Giles, H. (1996). Linguistic accommodation. In H. Goebl, P. Nelde, Z. Starý, & W. Wölck (Eds.), Kontaktlinguistik – Ein internationales Handbuch zeitgen¨ossischer Forschung (pp. 332–342). Berlin/New York: Mouton de Gruyter.
Ortega-Llebaria, M. (2004). Interplay between phonetic and inventory constraints in the degree of spirantization of voiced stops: Comparing intervocalic /b/ and intervocalic /g/ in Spanish and English. In T. L. Face (Ed.), Laboratory Approaches to Spanish Phonology (pp. 239–253). Berlin: Mouton de Gruyter.
Ortega-Llebaria, M. (2006). Phonetic Cues to Stress and Accent in Spanish. *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology*, 104–118.
Ortega-Llebaria, M., & Prieto, P. (2011). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech*, 54(1), 73–97.
Ortega-Llebaria, M., & Prieto, P. (2007). Disentangling stress from accent in Spanish. Production patterns of the stress contrast in deaccented syllables. In P. Prieto, J. Mascaró, & M. J. Solé (Eds.), Segmental and prosodic issues in Romance phonology (pp. 155–176).
Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4(AUG), 1–5.
Pardo, J. S., Cajori Jay, I., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72(8), 2254–2264.
Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, 69(3), 183–195.
Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, and Psychophysics*, 79(2), 637–659.

Phillips, S., & Clopper, C. G. (2011). Perceived imitation of regional dialects. Proceedings of Meetings on Acoustics, 12.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences,* 27(2), 169–226.
Pisoni, D. B., & House Lazarus, J. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 55(2), 328–333.
Podesva, R. J., Eckert, P., Fine, J., Hilton, K., Jeong, S., King, S., & Pratt, T. (2015). Social influences on the degree of stop voicing in Inland California. University of Pennsylvania Working Papers in Linguistics, 21(2), 1–10.
Postma-Nilsenovà, M., & Postma, E. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology*, 4(NOV), 826.
Postman, L., & Phillips, L. W. (1965). Short-term Temporal Changes in Free Recall. *Quarterly Journal of Experimental Psychology,* 17(2), 132–138.
Prieto, P., & Torreira, F. (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics*, 35(4), 473–500.
Quilis, A. (1981). Fonética acústica de la lengua española. Madrid: Gredos.
Quilis, A., & Esgueva, M. (1983). Realización de los fonemas vocálicos españoles en posición normal. In M. (Margarita) Cantarero & M. (Manuel) Esgueva (Eds.), Estudios de fonética (pp. 137–252). Consejo Superior de Investigaciones Científicas, Instituto "Miguel de Cervantes."
R Core Development Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
Romero, J. (1995). Gestural organization in Spanish: an experimental study of spirantization and aspiration. PhD dissertation. The University of Connecticut.
Schweitzer, A., & Walsh, M. (2016). Exemplar dynamics in phonetic convergence of speech rate. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 08-12-Sept, 2100–2104.
Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics,* 66(3), 422–429.
Simonet, M., Hualde, J. I., & Nadeu, M. (2012). Lenition of /d/ in spontaneous Spanish and Catalan. Proceedings of InterSpeech, (October), 1416–1419. Portland, OR.
Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71, 249–267.
Torreira, F., Simonet, M., & Hualde, J. I. (2014). Quasi-neutralization of stress contrasts in Spanish. *Proceedings of the International Conference on Speech Prosody*, July, 197–201.
Tremblay, M.-C., & Sabourin, L. (2012). Comparing behavioral discrimination and learning abilities in monolinguals, bilinguals and multilinguals. *The Journal of the Acoustical Society of America*, 132(5), 3465–3474.
Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71, 147–161.
Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, 6(Article 546), 1–18.
Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11, 3571–3594.
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag.
Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience*, 32(6), 776–791.
Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *The Journal of the Acoustical Society of America*, 140(5), 3560–3575.