Phonetic variability of the Spanish alveolar tap: spontaneous production and spoken word recognition

by

Scott James Perry

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics
University of Alberta

Abstract

The present dissertation investigated the spontaneous production of the Spanish alveolar tap and how the variability typical of spontaneous speech impacts the process of spoken word recognition. Our corpus analyses found that intervocalic taps vary in duration and intensity due to speech rate and phonetic environment. The intensity drop during taps is also associated with changes in lexical frequency, with higher-frequency words containing taps that are more likely to be reduced. These findings indicate that properties at the word level are related to systematic variation in Spanish tap production, aligning with similar findings in other languages. In the corpus analyses, automated methods did not reliably measure duration, while the same force-aligned boundaries were acceptable for measuring intensity differences. After qualitatively and quantitatively documenting the substantial variability in tap production, we designed an auditory lexical decision experiment to investigate how the reduction of the tap impacts how L1 and L2 Spanish listeners recognize words. In our initial planned analyses, we found that L1 listeners could exploit the advan-

tage of a canonical pronunciation to facilitate recognition accuracy but that no such effect was present for L2 listeners. When we changed our binary reduction variable to a tap-type coding based on previous literature, we saw an inhibitory effect associated with highly reduced perceptual taps. Our exploratory analyses outlined a potential confound between the expected production of the tap for specific words and our variable of interest - reduction. These findings indicate that L1 and L2 listeners recognize words faster when they contain taps that are more typical for that word and that after controlling for this experimentally or statistically, phonetic reduction is an inhibitory factor on spoken word recognition. Taken together, our production and perception data indicate that the variability of Spanish taps is due to lexical and phonetic variables and that word-specific patterns of variability in production shape the process of spoken word recognition regardless of language background.

Preface

The three body chapters of this dissertation are intended to be published as separate research articles. Each chapter has its own introduction and conclusion. As first author, I was responsible for the majority of the research. Benjamin V. Tucker supervised these studies and provided support and input throughout the process, from conceptualization to writing. Matthew Kelley provided computational support and contributed to the writing process for Chapters 2 and 3.

Chapter 2 has been published as: Perry, S. J., Kelley, M. C., & Tucker, B. V. (2023). Measuring and modelling the duration of intervocalic alveolar taps in Peninsular Spanish. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 699–703). Guarant International.

Chapter 3 has been published as: Perry, S. J., Kelley, M. C. & Tucker, B. V. (2024). Documenting and modeling the acoustic variability of intervocalic alveolar taps in conversational Peninsular Spanish. *Journal of the Acoustical Society of America, 155*(1), pp. 294–305.

Examining committee:

Benjamin V. Tucker, Supervisor

Anja Arnhold, Co-supervisor

Miquel Simonet, Supervisory Committee

Jennifer A. Foote, Examiner

Joseph V. Casillas, External Examiner

Dedication

For Adriana and Ollie

Acknowledgments

First and foremost, I must thank the members of my supervisory committee. To Ben Tucker, thank you mentoring me and collaborating with me for this thesis. I can't imagine a more supportive supervisor. From day one, you always treated me as a collaborator, and I'm looking forward to many interesting collaborations in the future. To Anja Arnhold, thank you for all of our great chats about phonetics and phonology, and for agreeing to co-supervise when the time came. Your support throughout the process, including when it came time to decide what to include, meant more that you may know. To Miquel Simonet, thank you for joining my committee, and for your input throughout the process. I look forward to seeing you in Barcelona.

Now that those who have been more directly involved with this thesis have been mentioned, I will attempt to do the rest of this more of less chronologically, in a vain attempt to not forget anyone. We then must start with my parents, who nurtured my childhood curiosity instead of squashing it, and also managed to instill a pattern of behaviour that lets me focus intensely and learn things quickly, so long as I happen to be interested enough to do so. This proved beneficial to my research. Additionally, without their love, support, and guidance, I never would have entered into...

...the business management degree at the University of Western Ontario. I'd like

to thank all of my professors from those classes for highlighting my vast dearth of interest in the topic. It was not where I was meant to be. Here I must also voice my appreciation of the university's breadth requirements, which led to me taking Spanish for beginners, and continuing on to intermediate courses the following year, where I had class with...

...Sílvia Perpiñán. It's possible that without her enthusiasm for linguistics and her care as an instructor, I would not have continued down the path to where I am now. Instead, I decided to major in Hispanic Linguistics, where I also took courses with Joyce Bruhn de Garavito and Yasaman Rafat. To Joyce, thank you for teaching me how to think about linguistic structure, as well as demonstrating to your students the mentality of a good researcher. To Yasaman, thank you for sharing with me your passion for L2 speech research, a field to which I hope to contribute for years to come. The overwhelmingly positive experience of my degree led to me continuing on to the graduate degree where I met...

...Adriana Soto Corominas. Adriana, without your love and support, there would have been no way I would have finished graduate school with my happiness and sanity intact. Your help and support with everything made all the difference. I also want to thank the other graduate students I was close with for making my first years of graduate school truly fantastic times full of learning, hard work, and tequila: David Brown, Itziri Moreno, Antonio Jiménez, Nandita Dutta, and Aya Ishai. To Ana García, thank you for being a fantastic boss who I am now happy to call a friend. My deepest thanks also go out to others in the department for making it easy for me to follow Adriana when she began her postdoc at...

...the University of Alberta, where I joined the Alberta Phonetics Laboratory. I must thank the members who overlapped with me many times over. To Matt Kelley, Filip Nenadic, Yoichi Mukai, Ryan Podlubny, and Annika Nijveld, thank you for not

only for making me the dumbest person in the room for the better part of the year, but for being so generous with your time and knowledge whenever I had a question. I sincerely doubt that I will ever learn as much so quickly as when I was working beside all of you. To Filip and Matt, thank you also for our Friday DnD sessions in the lab. Finding a better way to end the work week has proved somewhat difficult.

During the pandemic, the constant grind of working on my thesis while spending 23 hours a day in a small apartment would have been ten times as hard were it not for a large increase in the amount of virtual hangouts with old friends. Adam, Craig, Brendan, James, Simon, Dylan, Jared, thanks for all of the relaxing evenings playing together. A special thanks to Adam Jones for the short gaming sessions during the final writing stage years later that allowed me to disconnect and not go completely off my rocker.

I must now abandon my attempt to do things in order for the sake of including others who I feel I should mention. To David Heap, I gave us an idea and we gave me some of my first conference presentations, thank you for collaborating with us. Un agradecimiento especial a Carlos por todas las empanadas que me ayudaron a producir los modelos estadísticos del capítulo tres. Finally to SSHRC, thank you for funding my doctoral studies.

In the end, I did procrastinate, as I am wont to do on occasion, and that necessitated a final push on the writing side of things. This would have been impossible without help. Aquesta última empenta d'escriure no hauria estat possible sense el suport del Ramon, l'Antònia, i la Montse. Va ser fàcil concentrar-me sabent que l'Adriana i l'Ollie comptaven amb vosaltres tres. Us estimo molt.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Variation permeates every aspect of language. When it comes to speech, the variance in the speech signal has been a central problem for theoretical accounts of speech perception and production for decades (for a review, see Casserly & Pisoni, 2010). While this variability exists even in the carefully produced 'laboratory' speech that has been central to empirical speech science research, there is even more variation in the spontaneous, conversational speech that we encounter on a daily basis (Tucker & Mukai, 2023). Some of this variability in spontaneous speech is referred to as phonetic *reduction*, whereby a sound is altered or deleted compared to how it would be pronounced in more carefully articulated speech (Ernestus & Warner, 2011; Johnson, 2004; Warner, 2019). As this variation is ubiquitous in the speech we encounter daily, explaining why it happens and how we process it is a core question in phonetic research.

To illustrate the degree of variability possible in the speech signal, I provide two examples of massive reductions, one in English and the other in Spanish. The English example is taken from Warner (2019), which gives the example of the words *Friday*

night being produced as [fɹɛ̃ː]. A more careful pronunciation of these words by the same speaker could feasibly have been [ˈfɹaɪɾɛɪˈnaɪt̚]. The example from Spanish was taken from the Nijmegen Corpus of Casual Spanish (Torreira & Ernestus, 2010) by myself while inspecting the corpus for purposes related to the present dissertation. I found an instance of the words *hemos tenido* ('we had') being produced as [moseˈni], whereas a more careful production might of those words would be [ˈemosteˈniðo̪]. In both of these cases, listeners would be hard-pressed to correctly recognize these words out of context, and phoneticians would not recognize the words based on the phonetic transcription. And yet, given the context of the full sentence, many can recognize the intended message easily, and indeed, both of these examples were recorded during real communicative interactions.

Examples such as those presented in the previous paragraph are effective illustrations, but variability need not be so extreme to be considered reduction. The shortening of words (e.g., Bell et al., 2009; Gahl et al., 2012; Seyfarth, 2014), syllables (e.g., Aylett & Turk, 2004; Hilton et al., 2011), and segments (e.g., Cohen Priva, 2015; Cohen Priva & Gleason, 2020; DiCanio et al., 2022; Warner & Tucker, 2011) is the most-studied form of reduction, likely due to the ease and universality of measurement. Reduction is not just shorter speech, however, and other studied patterns include vowel centralization (e.g., Aylett & Turk, 2006; Hernandez et al., 2023; Munson & Solomon, 2004; Wright, 2004), consonant lenition (Broś et al., 2021; DiCanio et al., 2022; Katz, 2021; Warner & Tucker, 2011), and syllable or segment deletion (Hilton et al., 2011; Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, Bell, Gregory, & Raymond, 2001). While reduction occurs in all languages, we have evidence that patterns of reduction may vary cross-linguistically (Torreira & Ernestus, 2011). This means that gathering information about phonetic variability in as many languages as possible is of crucial importance if we are to tease apart language-

specific trends from the general cognitive mechanisms of language that lead to the observed phonetic variability.

1.1 Theoretical background

Theories and hypotheses that attempt to explain and account for the patterns of phonetic variability we observe, that is to say, why speech is sometimes more carefully enunciated while other times it is more reduced, are often grouped into being based on the talker or based on the listener (Gahl et al., 2012). Talker-oriented accounts ascribe phonetic variation to cognitive factors such as the speed of lexical access during speech production (e.g., Bell et al., 2009; Gahl, 2008), while the listener-oriented accounts argue that talkers are prioritizing intelligibility or attempting to maintain a constant transmission of information in order to facilitate speech perception (e.g., Aylett & Turk, 2004, 2006; Jaeger, 2010). Gahl et al. (2012) points out that these accounts often have similar predictions, namely that high-frequency and predictable parts of speech will be reduced. It is also worth pointing out that even listener-oriented accounts must be based on the talker's perception of listener needs, and it is not clear to what extent this information is taken into account (Jaeger, 2010).

Instead of situating the present dissertation within one of the proposals described above, this work is guided by the more general outline of H&H Theory (Lindblom, 1990). H&H Theory provides a general framework for viewing phonetic variability and reduction that can account for why speech can vary from carefully articulated *hyperspeech* to more reduced *hypospeech*, even within a single conversation. While H&H Theory does provide less specific claims than more recent hypotheses, it is compatible with many of them and has decades of empirical support. Below, I provide a general overview of the theoretical assumptions of the theory, followed by

a brief discussion of related hypotheses and empirical evidence.

In comparison with the dichotomy between production- and listener-driven variation described above, H&H Theory assumes that there are both production and reception constraints active during speech production. On the production side, the theory discusses both physiological and cognitive factors in production, with cognitive factors not being discussed in depth. In terms of physiological constraints, H&H Theory assumes that speech, like other forms of physical movement, obeys the principles of economy of effort while also being inherently flexible or 'plastic'. Reception constraints are divided into social and communicative. The social aspect allows for well-attested variation according to social and cultural aspects of communication (e.g., Ernestus et al., 2015; Labov, 1972; Sanchez et al., 2015; Wagner et al., 2015). The communicative constraints are simple in that it is assumed that we give interlocutors enough information in the acoustic signal so that our speech is sufficiently able to be *discriminated* in a given context. A key component of this perceptual discrimination is that we do not assume all of the information required for successful perception is located within the acoustic signal. The potentially competing constraints discussed above can lead to a continuum between hypo- and hyperspeech as we produce speech as efficiently as possible while still being understood.

Since H&H Theory was proposed, the use of large data sets of spontaneous speech has become commonplace, and findings from these studies have found several factors associated with reduction. The findings of more reduction in more predictable or frequent words (e.g., Bell et al., 2009; Jurafsky, Bell, Gregory, & Raymond, 2001) as well as in words with few phonological neighbours (e.g., Gahl, 2008; Wright, 2004) is compatible both with cognitive constraints during production as well as communicative constraints. While Lindblom (1990) mostly discusses physiological effects in terms of economy of effort, there is an emphasis on the fact that speakers

have a *choice*, with the example being given that decreased duration need not lead to articulatory undershoot (Lindblom, 1963) if the speed of articulation (and therefore effort) is increased (see Ennever et al., 2017, for a related proposal quantifying this). Although the general finding associated with higher frequency is shorter duration, some research has found more peripheral vowels in high-frequency words (Tomaschek et al., 2018) and bigrams (D. Kim & Smith, 2019), with one explanation being that the practice involved with saying high-frequency words leads to peripheral vowel articulations being easier to complete (Tomaschek et al., 2018). D. Kim and Smith (2019) discuss how even though they find temporal reduction at the word level, this may not lead to uniform reduction within the word. This is consistent with findings that not all parts of a word are equally important for successful recognition (van de Ven & Ernestus, 2018).

Now that I have laid out a general theoretical framework through which the present dissertation will view and discuss reduction phenomena, I turn to previous work that is directly relevant to the present dissertation, which investigates the phonetic variability of the Spanish alveolar tap from the point of view of both production and perception. In the following sections of this chapter, I will focus on reviewing previous empirical work that has investigated stop-consonant variability, reduction and spoken word recognition, as well as research looking at how second-language learners produce and perceive reduction. This will provide the empirical foundation that supports the planned dissertation, which is outlined in the last section of this chapter.

1.1.1 Variability in stop production

Many sounds that are described as stop-consonants in phonetic and phonological descriptions are often realized quite differently in spontaneous speech (e.g., Barry & Andreeva, 2001; Broś et al., 2021; Mukai, 2020; Torreira & Ernestus, 2011; Warner & Tucker, 2011). Reduced forms, sometimes referred to in the literature as 'lenited', are realizations that include fricatives, approximants, and deletions. This has been documented across several languages and should be considered the 'norm' in spontaneous speech, not an exception. For example, Mukai (2020) found that 59% of word-medial voiced stops in Japanese lacked a complete stop closure, with variation in this percentage based on the place of articulation (35% for /d/, 83% for /g/). In American English, Warner and Tucker (2011) reported that less than 40% of voiced stops had a formant break in F2-F3, and that only a quarter of these stops had a burst release. Broś et al. (2021), who collected data from Spanish spoken in Gran Canaria, found that even *voiceless* stops, which unlike voiced stops in Spanish are not reported to lenite phonologically (Hualde, 2005), were produced with voicing or as approximants almost half of the time.

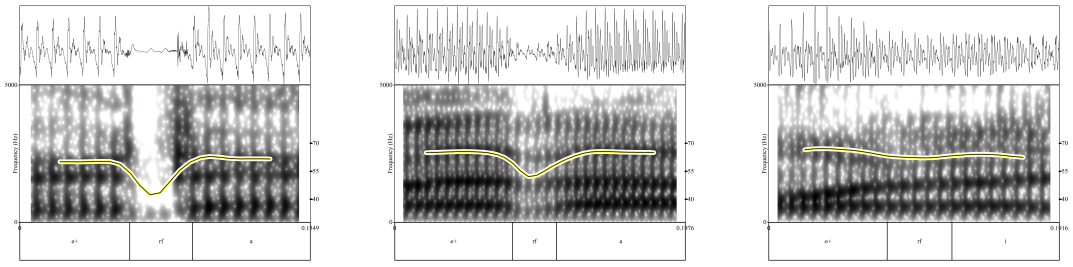
The general rates of stop-reduction reported for different languages are convincing evidence that reduced stops are everywhere (see also Barry & Andreeva, 2001). Many factors have been associated with variation in stop reduction, but some of the more consistent differences found in the literature are those whose effects stem from articulation: speech rate and place of articulation. Increased speech rate is one of the most robust predictors of reduction in general (Ernestus, 2014), and studies have found increased reduction of stops as speech rate increases (e.g., Cohen Priva & Gleason, 2020; Narayan, 2023), which is consistent with proposals that reduction is due to articulatory undershoot (Lindblom, 1963). Similarly, studies which examine

the reduction of multiple segments find that there are differences by place of articulation (Cohen Priva & Gleason, 2020; DiCanio et al., 2022; Mukai, 2020; Warner & Tucker, 2011). While differences by place of articulation may be due to undershoot not affecting different points of the vocal tract equally, part of the difference may also be how informative the different sounds are in the language. Cohen Priva (2015) argued that a segment's average informativity impacts how it is produced, with segments that matter less for the discrimination of the intended meaning seeing increased reduction.

Consistent with the theoretical assumption that information not present in the acoustic signal is related to speech perception, several studies have investigated the effects of word frequency and contextual probability on stop reduction. These effects appear to be inconsistent, with some studies finding effects of word frequency but not bigrams (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001), while others find an effect of bigrams but not word frequency (e.g., Warner & Tucker, 2011). More research is needed if we are to discern when these variables are reliably related to reduction and when they are irrelevant.

As with other stop-consonants, a wide degree of variation has been documented in the Spanish tap for several dialects and speaker populations (Amengual, 2016; Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008). Most studies have employed a qualitative coding scheme based on the visual interpretations of the tap's presence in a spectrogram. The most common categories included are true taps, approximant taps, and perceptual taps. Examples of these different tap realizations are visualized in Figure 1.1. True taps are stop-like taps that are produced with a clear stop closure and/or burst release (Figure 1.1 A), consistent with textbook descriptions of the pronunciation of this sound (e.g., Hualde, 2005). Approximant taps have a visible presence in

a spectrogram, as the intensity drop is apparent in the colour gradient, but have an unbroken formant structure (Figure 1.1 B). Perceptual taps are not visible on a spectrogram, and instead, their presence is inferred based on a mild intensity drop that is only visible in the waveform and/or the overlaid intensity curve (Figure 1.1 C).



(A) (B) (C)

Figure 1.1: Spectrograms depicting the three Spanish tap types most often coded in previous literature. Panel A depicts a true tap, which contains a stop closure and a burst release. Panel B depicts an approximant tap, with a visible presence in the spectrogram but a continuant formant structure. Panel C depicts a perceptual tap, without a clear visible presence in the spectrogram but an intensity dip present. Overlaid on each spectrogram is the intensity curve, which is the largest for the true tap and the smallest for the perceptual tap.

Experiments which have elicited speech through reading target words embedded in carrier sentences find that true taps make up the majority of productions (Amengual, 2016) while tasks such as picture naming see a more even distribution of the variants (J. Y. Kim & Repiso-Puigdelliura, 2020). Elicited narration tasks in two varieties of Spanish have found that highly-reduced perceptual taps make up approximately half of all tokens (Bradley & Willis, 2012; Willis & Bradley, 2008). Comparisons across studies do need to be taken with a grain of salt, as much of the work on this sound has investigated heritage speakers, whose experience with

the language is known to be somewhat heterogeneous (Chang & Yao, 2016), and factors related to differences in language experience have been argued to impact tap production in heritage Spanish (J. Y. Kim & Repiso-Puigdelliura, 2020).

The evidence we have indicates that tap reduction is likely an extremely common phenomenon in Spanish that changes according to speech style. However, we are unsure how this variability is shaped by cognitive and communicative factors, as the focus of most studies looking at Spanish tap variability has been inter-speaker variation, such as comparing groups with different language backgrounds in the heritage or L2 contexts (Amengual, 2016; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020). The only findings related to intra-speaker tap variation we are aware of are reported by J. Y. Kim and Repiso-Puigdelliura (2020), who found significant effects of the phonetic environment on the intensity difference of taps in heritage speakers of Spanish living in Southern California.

1.1.2 Phonetic variability and spoken word recognition

The process of how humans recognize the acoustic forms of words, referred to as spoken word recognition, is far from a solved problem, as evidenced by the various models that differ in their theoretical assumptions (e.g., see Weber & Scharenborg, 2012, and the references therein). There are, however, certain robust experimental findings, as well as commonalities across models, that we must consider during experimentation and when interpreting our results. The first 'settled' matter in spoken word recognition is that we do not wait for a word to end to recognize it; the process begins immediately and automatically, and several potential candidate words can be 'activated' in parallel (e.g., Luce & Cluff, 1998; Vroomen & De Gelder, 1997). As the time-course of the speech moves forward, additional acoustic information is

processed which matches with some words and not others, until one word is recognized over all other candidates. Several word-level properties not present in the acoustic signal have been shown to impact spoken word recognition, consistent with the assumptions of H&H Theory. Undoubtedly, the most well-known of these is lexical frequency, and results from many languages indicate that more frequent words are recognized faster and more accurately (e.g., Brysbaert et al., 2011; Ernestus & Cutler, 2015; González-Alvarez & Palomar-García, 2016; Lõo et al., 2018; Tucker et al., 2019). Consistent with the competition process that has been described above, several studies have also found that words that sound like many other words are recognized more slowly (Dufour & Frauenfelder, 2010; Gahl & Strand, 2016; Kelley & Tucker, 2022; Luce & Pisoni, 1998), although data from Spanish indicates that all languages may not behave the same way in this regard (Vitevitch & Rodríguez, 2005).

Studying spoken word recognition in an experimental setting differs in many respects from how the process typically unfolds during one's daily routine. The obvious differences include the fact that participants listen to recordings of an unfamiliar voice while situated in an unfamiliar setting. Beyond this, several other types of signal-independent information are conspicuously absent, as most experiments test reactions to individual words, while most language recognition has the advantage of additional context, which can facilitate predictive processing (e.g., Boudewyn et al., 2015; Grüter et al., 2020; Kuperberg & Jaeger, 2016). Pertinent to viewing spoken word recognition through the lens of H&H Theory, the talker in any given experiment recorded the stimuli while being removed in time (and possibly space) from the experimental participants. H&H Theory proposes that there is a feedback loop that allows the talker to modulate the speech signal based on the perceived success of communication, whereas this is not possible with recorded stimuli. This leaves only

word-level properties – and general tendencies of adaptations from repeated motor routines – present in the stimuli to impact the process of lexical retrieval. Experimental investigation into how reduction impacts spoken word recognition involves the manipulation of the signal-dependent information (i.e. the acoustics) while experimentally controlling for signal-independent information (lexical properties).

Phonetic reduction in an experimental setting has most often been related to less successful spoken word recognition (Ernestus et al., 2002; Ernestus & Baayen, 2007; Mukai, 2020; Pitt, 2009; Ranbom & Connine, 2007; Tucker, 2011; Wanrooij & Raijmakers, 2020). This is consistent with the view of reduction containing a less robust acoustic signal and, therefore, transmitting less signal-driven information to the listener. In contrast to these findings above, van de Ven and Ernestus (2018) found that unstressed vowel deletion may facilitate recognition if additional context is present.

The research on reduction normally compares a 'reduced' variant of the word to what is sometimes referred to as a 'canonical' variant, i.e. the variant that would typically be produced in careful speech and may be included in a dictionary transcription. While some studies have found an advantage of the canonical over a reduced variant (Pitt, 2009; Tucker, 2011), others have *also* found effects of variant frequency (Pitt, 2009) or failed to find an advantage of the canonical variant once controlling for frequency (Bürki et al., 2018) or experimental factors (Sumner, 2013). Variant frequency is something that will likely vary by language and dialect. It is, therefore, important to incorporate it into experimental studies of spoken word recognition in order to see if a more general explanation can be found for cross-linguistic differences. It is easy to see how both reduction and variant frequency could be involved in spoken-word recognition and how a 'tug-of-war' could arise: if the reduced variant is also more frequent, the direct effect of the change in the acoustic signal may be

confounded with variant frequency, potentially canceling each other out depending on how the effects are examined.

1.1.3 Reduction and second-language learners

We purposely limited the scope of the previous sections to reporting research into phonetic variability and reduction that has focused on adult, first-language speakers. These speakers have substantial amounts of experience with the language, and their acquisition process begins during infancy. In contrast, second-language speakers of a language have an experience that is both qualitatively and quantitatively different, with their acquisition starting after the use of another language was firmly established. In this section, we will discuss the small but growing body of research that investigates the production and perception of reduced forms by second-language learners.

Studies have shown that L2 listeners have a hard time recognizing reduced speech (Ernestus, Dikmans, & Giezenaar, 2017; Ernestus, Kouwenhoven, & van Mulken, 2017; Wanrooij & Raijmakers, 2021). These studies indicate that the difficulty associated with reduction is larger for L2 listeners than for L1 listeners (Ernestus, Dikmans, & Giezenaar, 2017; Ernestus, Kouwenhoven, & van Mulken, 2017; Wanrooij & Raijmakers, 2021), and that this difficulty persists even for those with advanced proficiency in the L2 (Ernestus, Dikmans, & Giezenaar, 2017). This is not surprising, as comparisons between adolescent and adult L1 speakers of German indicate that teenagers are still improving with respect to recognizing reduced speech, even after more than a decade of consistent language use and exposure (Wanrooij & Raijmakers, 2020). Explanations for why L2 listeners differ include reduced language experience and input (Wanrooij & Raijmakers, 2021) with recent work by Morano et al. (2023)

indicating that exposure to reduced word forms is critical for successful recognition. We also have evidence that the sound system of the first language impacts their processing of reduced speech in the second language (Ernestus, Kouwenhoven, & van Mulken, 2017). Thus, the general picture painted for L2 speakers and reduction is the following: they differ from L1 listeners in terms of their perceptual abilities, properties of the L1 sound system can affect their performance in the L2, and increased exposure improves their abilities in the L2. This is similar to theoretical assumptions of segmental learning (Best & Tyler, 2007; Flege, 1995; Flege & Bohn, 2021; Van Leussen & Escudero, 2015), with the possible caveat that dealing with the variability in spontaneous speech is a more difficult task.

While the patterns of results concerning L2 abilities and reduction look similar to general theoretical assumptions, most popular models of L2 speech are concerned with specific sounds (Flege, 1995; Flege & Bohn, 2021; Van Leussen & Escudero, 2015) or sound contrasts (Best & Tyler, 2007), and make no predictions regarding phenomena such as gradient lexical effects on second-language segment production. Speech research has historically been focused on the more carefully articulated 'laboratory' speech rather than more natural spontaneous speech (Tucker & Mukai, 2023), and this is especially true of L2 speech learning research (Zampini, 2008). However, the Speech Learning Model and its revised version (Flege, 1995; Flege & Bohn, 2021) assume that whatever processes or mechanisms lead to first-language speech perception and production are available for the second-language learner, so in principle, L2 learners should be able to follow similar patterns in production and perception to the input they are exposed to, given enough time and exposure. Whether or not the effects of frequency and predictability influence second-language speakers in a way similar to first-language speakers is an unanswered empirical question.

Compared to the body of research looking at the L2 perception of reduction, the

body of research investigating the production of reduced forms is smaller and largely focused on suprasegmental aspects (van Dommelen, 2018). For example, it is well-documented that L2 speakers generally have a slower rate of speech than L1 speakers (e.g., Bradlow et al., 2011; MacKay & Flege, 2004). Work at the syllable level has found that while both L1 and L2 speakers reduce, L2 speakers reduce less (Bradlow, 2022; Bradlow et al., 2011). Consistent with current theoretical proposals, work at the segment level has found effects of L1 background on L2 segmental reduction (Spilková, 2014). Of particular relevance to the present work, van Dommelen (2018) looked at stop reduction in L1 and L2 speakers of English, finding reduction to be frequent in both groups but with significant between-group differences in terms of average acoustic realizations. Overall, it seems that L2 speakers do reduce, but may do so less, or do so differently, than L1 speakers.

1.2 The present dissertation

The primary goal of this dissertation is to contribute observational and experimental data to the body of research that investigates the patterns of phonetic variability present in the world's languages. This contribution is meant to document how the Spanish tap varies in spontaneous speech, providing data from a variety of Spanish where this has not yet been documented. We also investigate patterns of *intra*-speaker variation of the Spanish tap, something which previous studies of the segment have not focused on. We complement this analysis of spontaneous production data with experimental data that investigates how the variation we find in spontaneous speech impacts how both L1 and L2 Spanish speakers recognize words.

A secondary goal of this dissertation was to set the stage for future work on the spontaneous production of the Spanish tap by L2 learners. Previous studies of the

acquisition of this sound often analyze L2 productions evaluated against an idealized 'canonical' variant (e.g., Face, 2006; Herd, 2011; Patience, 2018) even though there is evidence that other variants may be as frequent or more frequent than the more careful pronunciation of the sound (Bradley & Willis, 2012; Willis & Bradley, 2008). As L2 learners in situations of naturalistic acquisition will be exposed to this type of variability in the input, a deeper understanding of the patterns present in the spontaneous speech of L1 speakers, as well as how L2 learners are impacted by this variation in spoken word recognition, would help better to contextualize future results from analyses of spontaneous L2 productions.

The second chapter of this dissertation (published as Perry et al., 2023) investigates variability in the *duration* of the Spanish alveolar tap in spontaneous speech by analyzing a publicly available corpus. This chapter aims to investigate how lexical and phonetic factors impact variability in the duration for the subset of tokens where the duration can be consistently measured. As duration could only be hand-measured for approximately half of our tokens, this chapter's secondary goal is methodological. There are alternatives to manual measurements of duration, so we evaluate three more automated methods to see whether or not they are reliable approximations of hand-placed boundaries. This is important, as due to the large amount of censored/missing data, the generalizability of the duration results to the segment overall is compromised (the missing data mechanism is unclear, but likely not Missing Completely at Random). Unlike manual markup of phone boundaries, these other automatic measurement methods can use the acoustic signal to estimate the onset and offset of all tokens in principle. This means that if the methods are reliable, they would play a vital role in investigating the acoustic duration of the Spanish tap.

Chapter 3 (published as Perry et al., 2024) complements our analysis of duration

by documenting other types of acoustic variation in the Spanish tap in the same corpus analyzed in Chapter 2. The first goal of this chapter is to document the phonetic variability of a random subset of the corpus. The qualitative coding we employed is based on visual spectrographic interpretation and was discussed earlier in the present chapter (and shown in Figure 1.1). This coding facilitates comparison with previous studies on the production of Spanish taps. These are accompanied by descriptive statistics regarding duration and intensity differences based on these coding categories. Our hope is that this reporting, together with the provided raw data, will allow for future studies of Spanish taps to leverage this information in pilot analyses or in defining informative prior models for analyses of future data.

The second goal of the study in Chapter 3 was to model the intensity difference between the tap and surrounding vowels in order to understand potential patterns of variability beyond durational differences. The modelling approach in this paper attempts to reasonably approximate the data-generating process, which led to the adoption of non-standard statistical models. Our predictors are the same ones explored in Chapter 2: a combination of lexical and phonetic variables that have been associated with phonetic variability in other languages.

Chapter 4 complements the production data for our corpus analyses with perceptual data of an experimental nature. Given that our production data, which contains substantial amounts of reduction, came from real conversations where people understood each other, we know that the variability of the tap doesn't cause communication breakdown in running speech. However, no previous studies have examined how Spanish tap reduction impacts the recognition of isolated words, despite variation in production being well-documented for some time. We investigate how the phonetic variability of the tap impacts spoken word recognition using an auditory lexical decision task. Two groups of listeners participated in this study: L1

Spanish speakers and L2 Spanish speakers whose L1 was English. The goal is to investigate the effect of reduction separately for both groups and compare the effect of reduction across two populations whose experience with the language undoubtedly differs along a number of dimensions.

Chapter 2

Measuring and modelling the duration of intervocalic alveolar taps in Peninsular Spanish

This chapter has been published as:

Perry, S. J., Kelley, M. C., & Tucker, B. V. (2023). Measuring and modelling the duration of intervocalic alveolar taps in Peninsular Spanish. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 699-703). Guarant International.

Abstract

Factors predicting acoustic variation in the production of Spanish taps have yet to be investigated outside of their relationship to the tap-trill contrast. The present study models the duration of alveolar taps with occlusions visible on a spectrogram in spontaneous Spanish and compares durations measured by automated methods to hand-placed boundaries. We model tap duration from the Nijmegen Corpus of Casual Spanish (Torreira & Ernestus, 2010) with a combination of lexical, phonetic, and phonological predictors. Results indicate a high degree of uncertainty regarding the relationship between most of our predictors and tap duration. However, we are confident that faster speech rates are associated with decreased duration. Our automated measurements show deviations from hand-measured duration, indicating a need to evaluate the performance of the automated methods in future research.

2.1 Introduction

Studies analyzing the duration of Spanish alveolar taps (hereafter taps) have generally focused on comparing productions across speaker groups (Henriksen, 2015), investigating acoustic correlates of the tap-trill contrast (Bradley & Willis, 2012; Willis & Bradley, 2008) or both (Amengual, 2016). The present study's primary goal is to augment our understanding of variation in tap production. We model tap duration with phonetic, phonological, lexical, and predictability-related factors. Our secondary goal is to compare tap durations based on experimenter markup to three automated methods that measure duration with minimal researcher markup and evaluate the methods' impact on model estimates.

In research on other languages, various predictors have been associated with changes in duration at the segmental, syllabic, and word levels. Increased frequency and predictability have been associated with decreased duration (Aylett & Turk, 2004; Bell et al., 2009). Duration differences by phonetic factors such as phonetic environment and speech rate have been attested (Cohen Priva, 2015; Warner & Tucker, 2011). Lexical stress and pitch accents have also been associated with changes in stop closure duration (Cho & McQueen, 2005).

When measuring the duration of segments without an apparent onset or offset, which includes many Spanish taps (Bradley & Willis, 2012; Willis & Bradley, 2008), it is desirable to have a measurement method that applies to most realizations. One alternative to human markup includes using boundaries placed by an acoustic model through forced alignment, although this method has some known issues (Tucker & Mukai, 2023). Another option is to use intensity to measure duration, which can be done in various ways (Katz & Pitzanti, 2019; Kingston, 2008; Warner & Tucker, 2011). Before applying these methods to Spanish taps, we believe it is important to compare them to hand measurement for tokens with visual cues to onset and offset, where experimenter markup can be consistent.

2.2 Method

The data and the script documenting the analysis are available through the University of Alberta Education and Research Archive here: https://doi.org/10.7939/r3-5k3f-4t63

2.2.1 Data coding and measurement

We analyzed a random 10% sample of intervocalic taps from the Nijmegen Corpus of Casual Spanish (Torreira & Ernestus, 2010), containing 20 conversations between groups of three university students from Madrid, Spain. Of the 2,606 hand-coded taps, 1,312 had occlusions visible in a spectrogram ('True' or 'Approximant' taps following previous studies (Amengual, 2016; Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008)).

Each tap's duration was measured in four ways using a script in Praat (v 6.1.47; (Boersma & Weenink, 2022)). The first method was the manual placement of boundaries, carried out by the first author. For stop-like taps, the onset was placed at the beginning of the stop closure, and the offset was placed at the onset of periodic energy after the burst release. For approximant taps, onset and offset were placed where the spectrogram abruptly changed intensity. The second method used force-aligned boundaries from the Montreal Forced Aligner (McAuliffe et al., 2017) trained on the corpus under analysis. The third method took the tap onset and offset as the midway points between the maximum intensities of the surrounding vowels and the minimum intensity during the tap (Warner & Tucker, 2011). The final method placed the onset and offset of the tap at the largest absolute values of the spline-smoothed intensity slope in and out of the tap, using methods from Katz and Pitzanti (2019) and Kingston (2008).

When using automated measurements, outliers that are likely measurement errors are removed based on domain knowledge. To fairly compare methods, we removed observations that we judged to have impossible values in any of the measurement methods. Spanish taps have average durations below 50 ms (Bradley & Willis, 2012; Willis & Bradley, 2008), and virtually all taps reported in Amengual (2016) were

under 150 ms. Therefore, we removed nine observations with values of 200 ms or more and seven with negative values. This left a total of 1,296 tokens for statistical analysis.

2.2.2 Statistical analysis

To analyze tap duration, we fit four hierarchical Bayesian models (one for each measurement method) with lognormal likelihoods using `brms` (v 2.18.0; (Bürkner, 2017)) in `R` (v 4.1.1; (R Core Team, 2021)). Priors were weakly-informative in the context of Spanish taps, and their assumptions were assessed through prior predictive simulation. Following best practices (Kruschke, 2021), we checked our priors' influence on the posterior by using wider priors, which resulted in identical posteriors upon visual inspection.

The population-level predictors for our models appear below. Group-level effects were varying intercepts by speaker and correlated varying slopes for all population-level effects by speaker. Unigram and bigram frequencies from the corpus under analysis were added to counts from the Spanish OpenSubtitles corpus (Lison & Tiedemann, 2016), and conditional probabilities derived therefrom. We extracted speech rate (syllables/second), surrounding vowels, and lexical stress from force-aligned text grids. Word length and content/function status were derived from the corpus data file.

The following predictors were used in the models:

- **Unigram freq.** Log unigram frequency for word containing tap
- **Bigram freq. prev.** Log bigram frequency for word and previous word
- **Bigram freq. fol.** Log bigram frequency for word and following word

- **Cond. prob. prev.** Conditional probability of the word based on previous word
- **Cond. prob. fol.** Conditional probability of the word based on following word
- **Number of syllables** Word length in number of syllables (log-transformed)
- **Function word** Sum-coded, difference between function (0.5) vs. content words (-0.5)
- **Local speech rate** Log average speech rate of the utterance containing tap (syllables/s)
- **Prev. vowel** Preceding vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Fol. vowel** Following vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Prev. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) previous vowel
- **Fol. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) following vowel

To measure the similarity between posterior distributions from the manual model to those from the automated methods, we calculated overlap in the population-level posteriors using the `overlap()` function from package `bayestestR` (v 0.11.0; (Makowski et al., 2019)). We entered the overlap values as the dependent variable in a hierarchical Beta regression predicting posterior overlap with the manual model by method. We included varying intercepts for predictors with varying slopes by method.

2.3 Results

For complete model summaries, the reader is directed to the materials hosted in the repository, which include the saved models for convenience.

2.3.1 Descriptive comparison of methods

Figure 2.1 contains visualizations that compare manual duration to the three automated methods. In the top row (A, B, & C) are scatter plots between manual measurements on the y-axis and each automated method on the x-axis. Marginal density distributions are placed along the edge of the plots. All automated methods were weakly correlated with manual measurements. In Figure 2.1 D, we plot the density distributions of the difference between the manual measurement and the automated methods, subtracting the automated duration from the manual measurement for each token. The red dotted line at zero is where the methods had the same value.

2.3.2 Factors affecting duration

For the model of manual durations, we report in Table 2.1 the percentages of each posterior that fell below, within, and above a Region of Practical Equivalence (ROPE). A ROPE establishes a minimum effect size the researchers consider practically different from zero. We chose a ROPE of -0.05 to 0.05 for log-scale duration, which for our model states that if the total effect of a predictor is less than $\approx$ 1.2ms, then we consider the effect to be negligible. This approach allows us to consider the evidence from our model in terms of the probability of both the existence and direction of an effect. An effect below the ROPE is associated with shorter taps,

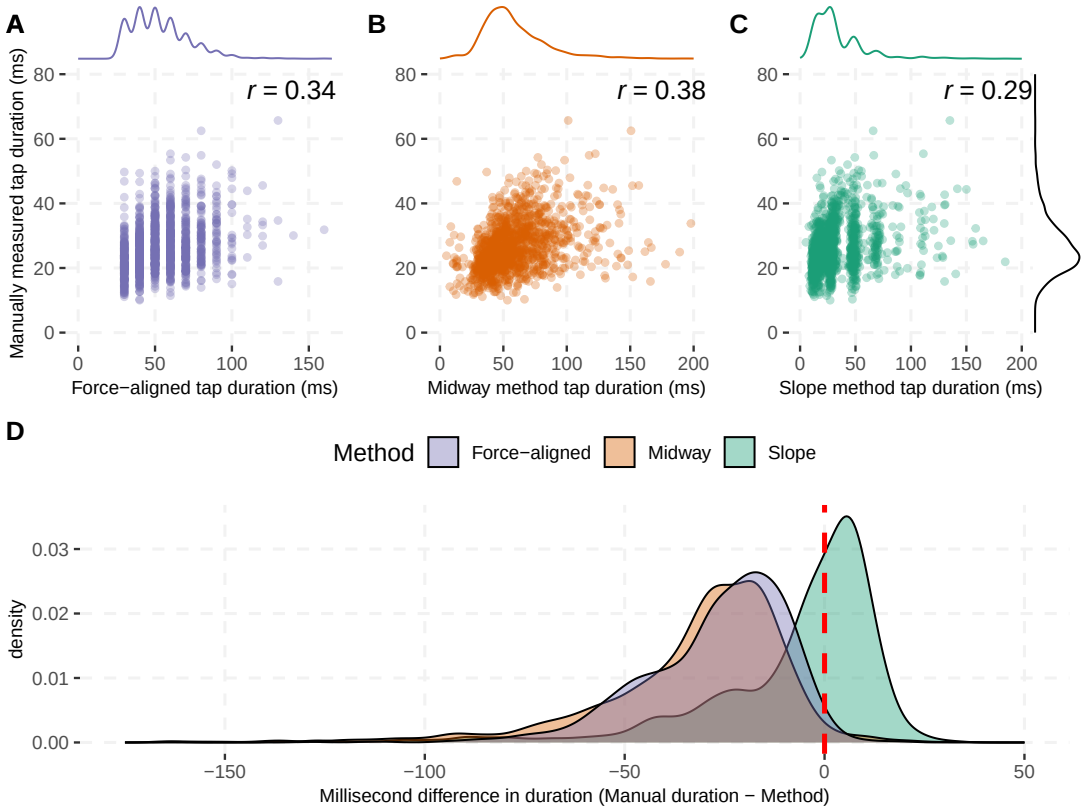


Figure 2.1: Scatter plots for comparisons between the manually measured durations and the durations taken from forced alignment (A), the midway method (B), and the slope method (C). The density distribution of manual measurements are along the right-hand y-axis of C. Pearson's correlation coefficients are printed on the scatterplot for each comparison. D displays density distributions of by-token differences.

and an effect above the ROPE signals an association with longer taps. For example, our models suggest we cannot be sure if function words contain shorter taps than content words, as roughly 68% of the posterior is within the ROPE, and 32% is below it. We interpret this as a 2/3 chance that there is no difference in tap duration for function words and a 1/3 chance that function words contain shorter taps. The posterior distribution did not extend above the ROPE, meaning we are confident function words don't contain longer taps.

Table 2.1: Percentages of posteriors from the model of manually-measured duration that fall below, within, and above the established Region of Practical Equivalence. The ROPE range of (-0.05, 0.05) was divided by the range of continuous predictors to evaluate the total effect size.

Predictor	% Below/Within/Above ROPE
Unigram freq.	2.8 / 26.1 / 71.0
Bigram freq. prev.	56.6 / 42.5 / 0.9
Bigram freq. fol.	17.3 / 76.2 / 6.5
Cond. prob. prev.	70.1 / 7.8 / 22.2
Cond. prob. fol.	91.8 / 7.2 / 1.0
Number of syllables	56.3 / 39.1 / 4.6
Function word	32.4 / 67.6 / 0.0
Local speech rate	98.9 / 1.1 / 0.0
Prev. vowel /i/	0.0 / 10.0 / 90.0
Prev. vowel /a/	4.7 / 95.3 / 0.0
Prev. vowel /o/	23.8 / 76.1 / 0.1
Prev. vowel /u/	22.4 / 73.3 / 4.3
Fol. vowel /i/	0.0 / 0.0 / 100
Fol. vowel /a/	0.7 / 93.0 / 6.3
Fol. vowel /o/	0.0 / 40.2 / 59.7
Fol. vowel /u/	83.1 / 15.7 / 1.3
Prev. stress	14.4 / 83.9 / 1.7
Fol. stress	5.3 / 93.4 / 1.3

2.3.3 Modelling differences & posterior overlap

In Figure 2.2, we plot the population-level posteriors from the models fit to durations from the four methods. The mean of the posterior is plotted by shape, and the two-stage lines visualize the most probable 80% and 95% of each posterior. The posteriors for the same predictor from the four models show variable levels of overlap. For some predictors (e.g., previous bigram frequency), all methods have similar posterior distributions. For others, there are larger differences between the methods, sometimes in a way that could change model interpretation as compared to hand-measured duration or compared to other automated methods.

The posteriors from the Beta regression estimating overlap as well as group-level standard deviations among the predictors are plotted by method in Figure 2.3. We cannot be confident that any automated method overlaps more or less with our manual model, although the Slope method had the highest estimated overlap (Figure 2.3 A). We are confident the Slope method showed less variability by predictor than the other methods (Figure 2.3 B).

2.4 Discussion & Conclusions

The present study measured tap durations in conversational Spanish and modelled this duration with several predictors. As many taps lack visible occlusions, we wanted to evaluate alternative methods of measuring duration that could be applied to more variable realizations. In predicting manually-measured taps, most predictors are highly uncertain regarding the presence or direction of an effect, although we can rule certain patterns out. Measurements from all automated methods correlated similarly with manual measurements, and the Slope method had the lowest absolute

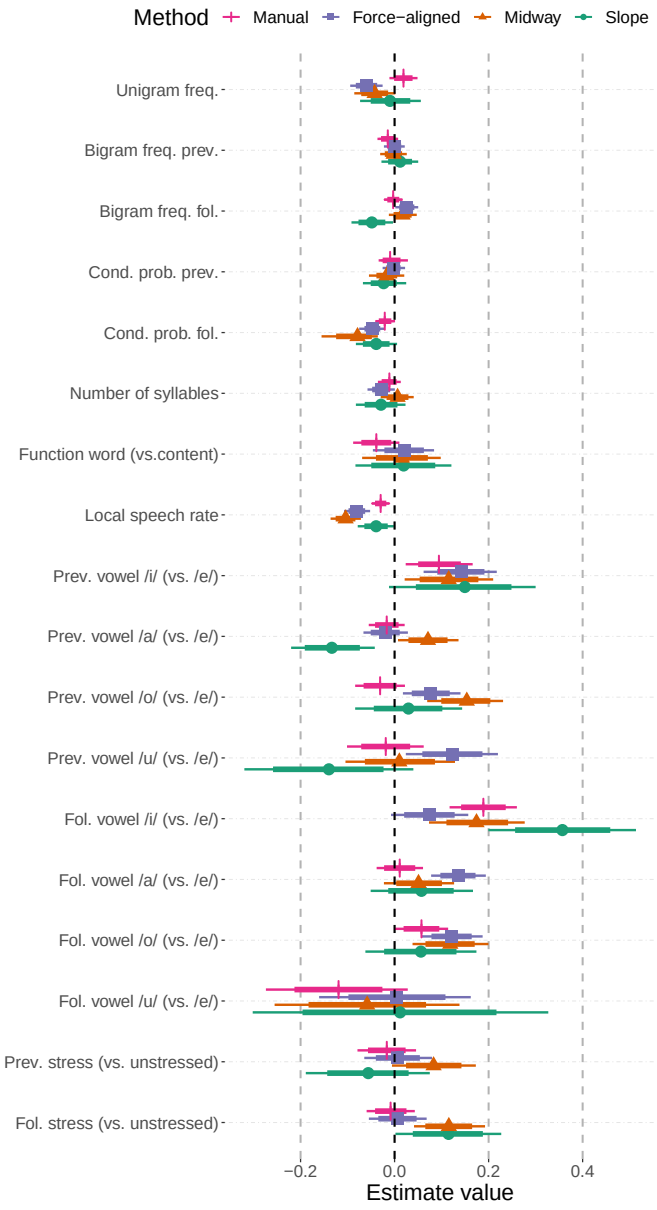


Figure 2.2: Posteriors for population-level effects models of tap duration calculated by four methods: manually-placed boundaries, force-aligned boundaries, using the midway points of the intensity curve, and using the intensity slope. The thick line corresponds to 80% of the posterior, and the extended, thin line corresponds to 95%.

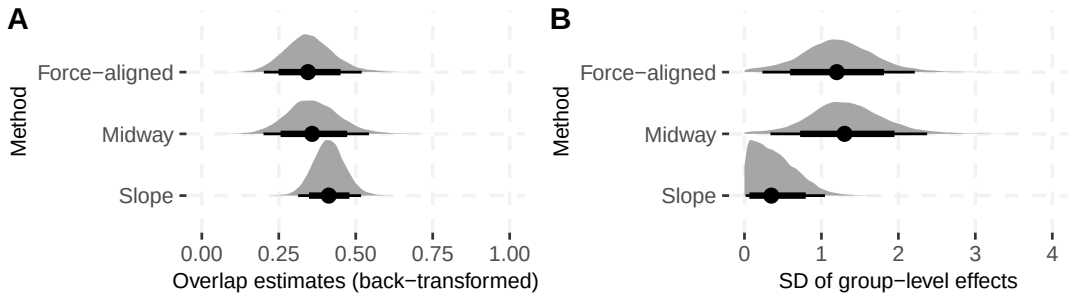


Figure 2.3: Posterior distributions of estimated overlap with model estimates from manual durations. The thick lines correspond to 80% of the posterior and the extended thin line 95%. 'A' plots predicted overlap, zero meaning posteriors do not overlap and one meaning posteriors are identical. 'B' plots the standard deviations of group-level effects (variation across predictors).

error. The force-aligned durations were multiples of 10ms with a floor at 30ms. The midway method had a similarly-shaped distribution to manual measurements but with longer values, while the Slope method had clusters at multiple modes.

In our model of manually-measured tap duration, we can only generalize results to taps that have a visible presence in a spectrogram. We only claim with confidence that increased speech rate is associated with shorter taps, and that taps are longer before /i/ than before /e/. Other specific effects show uncertainty regarding the presence of a meaningful association, but effects in specific directions are incompatible with our model. For some effects, large percentages of the posterior fall within the ROPE, indicating the most likely interpretation should be that these predictors are not associated with changes in tap duration.

When comparing estimates from our four models, the patterns of similarity and difference varied depending on the predictor. For some, like speech rate, the effects from all models are reliably negative, but some methods overestimate the effect's size, which is likely due to overestimating tap duration overall. If we consider our

manual model the gold standard, other methods make both Type S errors (getting the direction wrong) and Type M errors (wrong effect magnitude). This assumption is reasonable, as retrodictive checks showed the manual model fit the data well, while other models did not. Several predictors with posteriors centered around zero in the manual model show reliably negative or positive effects in one or more automated methods (e.g., Bigram freq. fol, Fol. stress). A potential explanation is that some variables are related to intensity changes independent of duration.

From our model of posterior overlap, all automated methods had less than 60% overlap with the manual model for an average predictor. This result is not reassuring, although the slope method, which may have slightly more overlap, also showed less variation across our predictors, possibly due to having wider posteriors than the other methods. Based on these results, we cannot recommend these automated methods be used to measure Spanish taps. We also must question the reliability of measuring segment duration using intensity more generally, and recommend researchers evaluate their measurement methods as standard practice.

Hand-correcting data is costly, but building knowledge on results skewed by measurement error will be more so. Sharing hand-corrected data publicly will allow for data to be used by the wider research community to answer research questions and generate informative priors that allow them to use their data more efficiently.

Chapter 3

Documenting and modeling the acoustic variability of intervocalic alveolar taps in conversational Peninsular Spanish

This chapter has been published as:

Perry, S. J., Kelley, M. C., & Tucker, B. V. (2024). Documenting and modeling the acoustic variability of intervocalic alveolar taps in conversational Peninsular Spanish. *The Journal of the Acoustical Society of America, 155*(1), 294-305

Abstract

This study constitutes an investigation into the acoustic variability of intervocalic alveolar taps in a corpus of spontaneous speech from Madrid, Spain. Substantial variability was documented in this segment, with highly reduced variants constituting roughly half of all tokens during spectrographic inspection. In addition to qualitative documentation, the intensity difference between the tap and surrounding vowels was measured. Changes in this intensity difference were statistically modeled using Bayesian finite mixture models containing lexical and phonetic predictors. Model comparisons indicate predictive performance is improved when we assume two latent categories, interpreted as two pronunciation variants for the Spanish tap. In interpreting the model, predictors were more often related to categorical changes in which pronunciation variant was produced than to gradient intensity changes within each tap type. Variability in tap production was found according to lexical frequency, speech rate, and phonetic environment. These results underscore the importance of evaluating model fit to the data as well as what researchers modeling phonetic variability can gain in moving past linear models when they do not adequately fit the observed data.

3.1 Introduction

The production of stop-consonants has been shown to be highly variable in a number of languages, with segments that we would expect to have closures and burst-releases being realized as fricatives, realized as approximants, or being deleted entirely (e.g., Broś et al., 2021; Davidson, 2011; DiCanio et al., 2022; Katz & Pitzanti, 2019; Mukai,

2020; Torreira & Ernestus, 2011; Warner & Tucker, 2011). This variation is often defined as *phonetic reduction*, a term that can refer to phonetic variation that is not predictable through phonological processes. Research on phonetic reduction involves both documenting the variation present in the sounds of the world's languages and attempting to develop insight into what factors may predict this variation. Other reduction phenomena include vowel centralization (e.g., Aylett & Turk, 2006), segment deletion (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001), and durational shortening (e.g., Bell et al., 2009; Pluymaekers et al., 2005; Seyfarth, 2014). As pointed out by Tang and Bennett (2018), much of the research on this topic has been conducted on English and Dutch. More research is needed to determine cross-linguistic similarities and differences in patterns of reduction and the factors that may drive them. The two main goals of the present paper are to document the variability of intervocalic alveolar taps (hereafter taps) in a conversational corpus of Madrilenian Spanish (Torreira & Ernestus, 2010) and to statistically model an acoustic correlate of stop reduction in order to gain additional insight into what may predict variation in Spanish tap production.

Broadly speaking, the theoretical motivation for the present study falls under Hyper and Hypo-articulation theory (H&H Theory; Lindblom, 1990). H&H Theory claims that the phonetic variability we observe stems at least partially from the fact that we obey economy of effort in granting lexical access to interlocutors. A related hypothesis that follows logically from H&H Theory is the Probabilistic Reduction Hypothesis, which posits that more predictable parts of speech are more likely to be reduced (Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, Bell, Gregory, & Raymond, 2001). Another related hypothesis is the Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004, 2006), which claims that phonetic variation stems from attempts to maintain a constant flow of information while communicating. This

leads to less informative syllables having shorter durations (Aylett & Turk, 2004) and more centralized vowels (Aylett & Turk, 2006). We believe that currently, calculating predictability in a way that integrates all sources of information available to humans communicating is an intractable problem, as it would necessarily involve information at all levels of linguistic analysis together with contextual information about the environment and interlocutor(s). Given this problem, it is perhaps unsurprising that several variables have been operationalized to study predictability, such as lexical frequencies, conditional probabilities, mutual information, relative entropy, and language model probabilities (e.g., Aylett & Turk, 2004; Balling & Baayen, 2012; Bell et al., 2009; Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, Bell, Gregory, & Raymond, 2001).

Increased predictability, as operationalized by these measures, has been generally associated with more reduced speech. This comes in the form of decreased duration in syllables (e.g., Aylett & Turk, 2004; Tang & Bennett, 2018), words (e.g., Bell et al., 2009; Seyfarth, 2014) and multi-word utterances (Tremblay & Tucker, 2011). Beyond duration, increased predictability has been associated with higher rates of segment deletion (Jurafsky, Bell, Gregory, & Raymond, 2001), more centralized vowels (Aylett & Turk, 2006; Munson & Solomon, 2004; Wright, 2004), and less stop-like stops (e.g., Cohen Priva & Gleason, 2020; Warner & Tucker, 2011). However, some research has reported the opposite, finding less reduced productions for more frequent or predictable words (e.g., D. Kim & Smith, 2019; Tily & Kuperman, 2012).

3.1.1 Variation in Spanish tap production

Spanish taps have been documented to have variable productions across different dialects and speaker populations (e.g., Bradley & Willis, 2012; Henriksen, 2015;

J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008). Research that has elicited speech using narration and picture naming (e.g., J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008) reports higher rates of reduced taps than studies eliciting productions with carrier sentences (e.g., Amengual, 2016), which is consistent with the variation we would expect according to speech style (Tucker & Mukai, 2023). Previous investigations into variation in Spanish tap production have employed a qualitative coding scheme based on the visual interpretation of spectrograms. While discretely categorizing speech productions can be problematic, using a qualitative evaluation with clear guidelines can improve communication about phonetic variability in written research (e.g., Davidson, 2016), and the coding scheme used for Spanish taps has facilitated comparisons across several published studies (Amengual, 2016; Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008). First, if taps have a closure and/or burst release, they are called 'true taps', visualized in Figure 3.1 A. Approximant taps are denoted by a visible intensity change on a spectrogram, but with a continuous formant structure, as in Figure 3.1 B. Perceptual taps have no visible intensity change on a spectrogram but show a dip in amplitude in the intensity curve and waveform, as in Figure 3.1 C. Deletions, a category not always included, show no evidence of a tap, as shown in Figure 3.1 D. In addition to the categories above, some studies code 'non-tap productions,' which are most often fricatives (J. Y. Kim & Repiso-Puigdelliura, 2020).

Previous acoustic measurements of taps have largely come in the form of duration, although as pointed out by previous studies (Amengual, 2016; Bradley & Willis, 2012), duration may be meaningful only for taps with a visible occlusion, as the onset and offset of perceptual taps have no clear markers. Consistent with its categorization as a tap, rather than a plosive, tap durations are relatively short. Average durations

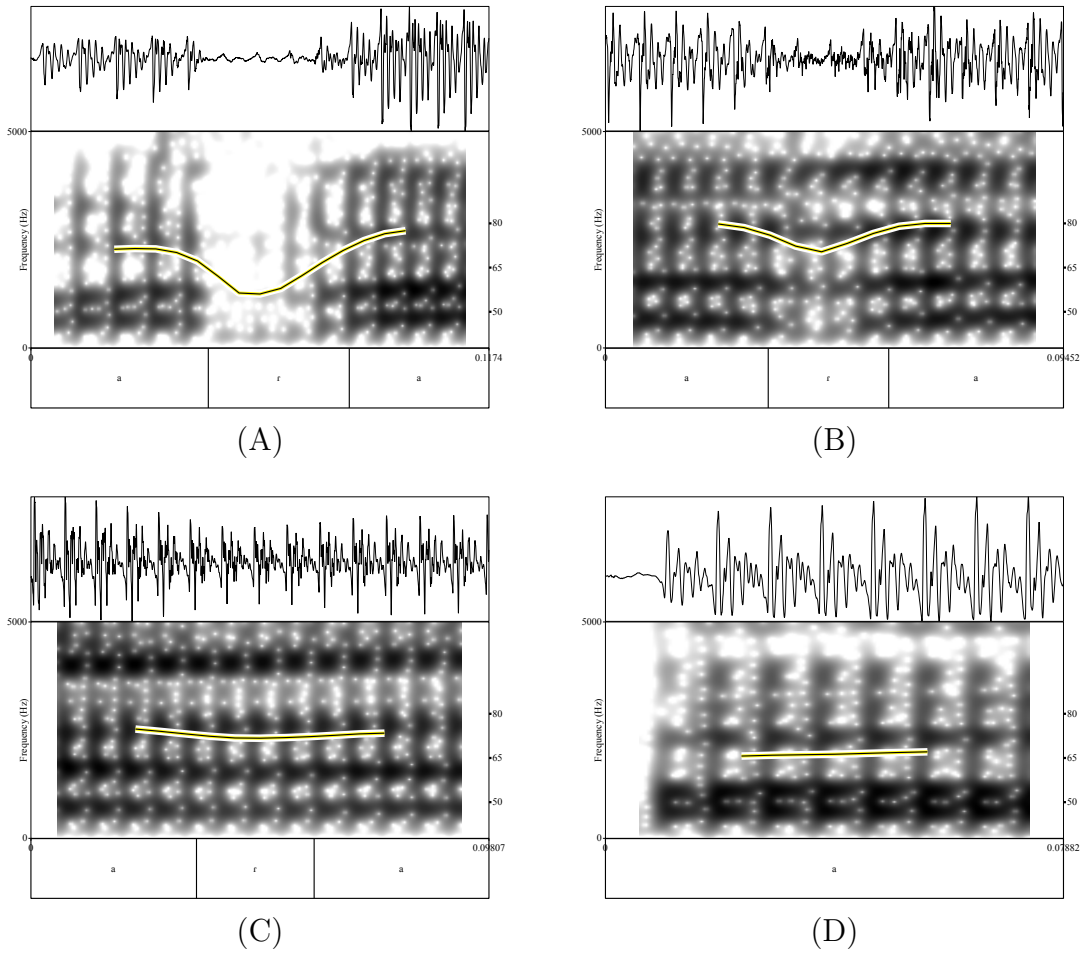


Figure 3.1: Spectrograms of the four main tap types in Spanish with intensity curves overlaid. Right axis indicates intensity in dB. Right axes are constant across the spectrograms to aid visual comparison. A shows a true tap. B shows an approximant tap. C is a perceptual tap. D is a deleted tap, with no visible presence in the spectrogram. All examples come from one male speaker. The stress of the surrounding vowels was not controlled, leading to intensity differences.

are reported between 30-50ms (Amengual, 2016; Bradley & Willis, 2012; Willis & Bradley, 2008) with the vast majority of productions being under 100ms (Amengual, 2016). Another acoustic correlate of stop variation is the intensity difference (IntDiff) between the surrounding vowels and the minimum during the consonant. IntDiff has

the advantage of being meaningful and interpretable in the presence of variation that spans from clearly-articulated stop-like phones all the way to deletions (Warner & Tucker, 2011). Similar measurements have been used in studies of other Spanish sounds (e.g., Hualde et al., 2011), but to our knowledge J. Y. Kim and Repiso-Puigdelliura (2020) is the only study that has used IntDiff to investigate variability in Spanish taps. J. Y. Kim and Repiso-Puigdelliura (2020) found that IntDiff ranges from 0-20 dB, with the majority of values under 15 dB. In their study, true taps had larger average intensity drops than approximant taps, which in turn had a larger drop than perceptual taps, with overlap between the categories.

While previous studies have given the field knowledge of the acoustic variation of Spanish taps, previous work has focused on examining this variation in the context of the tap-trill contrast (Bradley & Willis, 2012; Willis & Bradley, 2008), or looking at between-group differences in terms of language experience in the heritage and L2 contexts (Amengual, 2016; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020). The present study approaches this variation from a phonetic and lexical point of view, investigating the sort of variation that can occur within the same speaker in a single conversation.

3.1.2 Study aims

As mentioned briefly above, the first goal of this paper is to document the variation of Spanish taps in conversational speech. To maximize comparisons with previous studies on Spanish taps, we use the categorical coding scheme from spectrographic studies of Spanish taps conducted on speech samples from other dialects and populations. For each qualitative category, we provide visualizations and descriptive statistics of IntDiff, and, when appropriate, duration. Additionally, we provide raw

and processed versions of our data in an online repository.

The second goal of the paper is to model our chosen correlate of stop reduction: IntDiff. While modeling IntDiff using all predictors previously used by research on reduction is not feasible in a single paper, we include a set of phonetic and lexical predictors that have been shown to influence phonetic variation and reduction patterns in other languages. We believe this constitutes a strong initial pass at modeling this variability. While we did not initially plan on employing such models, the present paper also makes a case for using Bayesian finite mixture models as a flexible option for analyzing acoustic data in situations where linear models are not appropriate.

3.2 Methods

The online repository includes the raw and processed versions of the data, the code used to process and analyze the data, and supplementary documents discussing certain aspects of the data and analysis in further detail. See Section 3.6 for details.

3.2.1 The Nijmegen Corpus of Casual Spanish

The speech recordings analyzed for the present study are from the Nijmegen Corpus of Casual Spanish (`NCCSp`; Torreira & Ernestus, 2010). This corpus contains recordings of monolingual speakers of Spanish from Madrid, Spain (N=52, 27 females & 25 males). The speakers, all university students between the ages of 19-25 years, were recorded using head-mounted microphones in a sound-attenuated booth while holding conversations in groups of three people who knew each other well. Eight speakers were recorded on two occasions. Twenty conversations are included in the

corpus for a total of 60 recordings of approximately 90 minutes each. A professional company (Verbio Speech Technologies S.L.) orthographically transcribed the corpus, and these transcriptions are included with the corpus as `Praat` text grids segmented at the utterance level.

3.2.2 Automatic alignment

Segmentation of the audio files at the word and phone levels was done automatically using the Montreal Forced Aligner (`MFA`; v1.0.1; McAuliffe et al., 2017). First, a pronunciation dictionary was generated using the Spanish grapheme-to-phoneme (G2P) model available with the `MFA`. This output a broad phonetic transcription of every word in the corpus data file. The transcriptions output by the G2P model were checked to ensure they were consistent with the Madrilenian dialect, and edited if needed. This included, for example, changing many instances of the voiceless alveolar fricative /s/ to the voiceless interdental fricative /θ/, as the G2P model was inconsistent with respect to the distinction between /s/ and /θ/. As stress was marked in the phonetic transcription for words with orthographically marked stress but not for default stress patterns that did not contain orthographic accents, this information was added to the transcriptions using a custom programming script to ensure that the transcriptions of all words had lexical stress marked. We then aligned the audio files using the `train_and_align` function from the `MFA`, which aligned the audio files using the data itself, not employing a pre-trained acoustic model. This was chosen as a better alternative than the available pre-trained acoustic model as it was not documented which dialect(s) of Spanish was used to train the model, and the `NCCSp` contains considerably more than the one hour of speech required to produce desirable alignment.

3.2.3 Acoustic analysis

Acoustic measurements were taken automatically from each tap in the corpus with a custom script in `Praat` (Boersma & Weenink, 2022) using the force-aligned boundaries. The information extracted included the duration of the segment interval from the `MFA` phone tier, the word identity, the previous and following segments, and previous and following words, in addition to the following intensity measurements. The script measured the maximum intensity of the previous and following segments, the minimum intensity during the tap, and the corresponding timestamps. This information was used to calculate the intensity difference between the tap minimum and the average of the surrounding vowels.

In addition to taking acoustic measurements, the script was written to save a subset of tokens for the manual adjustment of force-aligned boundaries. Each token had a 10% chance of being included in this subset. The full data-set of intervocalic taps included 28,193 taps that were produced between two monophthong vowels, with the script saving 2,786 tokens (9.88% of the data). Hand-correcting this subset allowed for direct comparisons between the same acoustic measurements taken from the force-aligned vs. hand-corrected boundaries, which also permits us to measure the noise being introduced by using force-aligned boundaries and decide whether the automatic measurements were reliable. This was also a manageable random sample for hand-coding tap type following previous studies.

3.2.4 Hand-correction and coding

The hand-coding and boundary correction were performed by the first author. The coding scheme closely followed previous work on Spanish taps (Amengual, 2016; Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020;

Willis & Bradley, 2008). Taps containing a visible stop closure and/or burst release were coded as true taps. Taps containing a visible drop in intensity on the spectrogram, but without a break in the formants, were classified as approximant taps. Taps whose only evidence of existence was a drop in intensity visible in the `Praat` intensity contour and the waveform were coded as perceptual taps. We also coded as perceptual taps tokens where there was movement in F3/F4, even if there was no visible intensity drop, as previous work indicates that this may be a relevant perceptual cue for English flaps (Warner et al., 2009). Taps that were produced as fricatives, as well as taps that did not fall into the other categories, were coded as non-tap productions. If there was no intensity drop, no formant movement, or no other visual or auditory evidence to indicate that any segment was produced, it was coded as deleted. Regarding deletions, we do acknowledge that an apparent lack of acoustic evidence does not mean that no articulatory movement took place. As such, these segments may not have truly been deleted.

Of the 2,786 taps that were hand-coded, 1.5% were instances of the corpus' orthographic transcription being incorrect. Due to some other reduction, 1.9% of taps were not produced within an intervocalic environment. Noise interference, such as loud laughing from another participant, was present in 2.9% of taps, and 0.1% of recordings extracted were too short to extract acoustic information or code the tap type. These observations were removed before calculating the percentages in the results section. This left us with 2,606 tokens in the hand-corrected data, which was 93.5% of the original data.

3.2.5 Calculation of other variables

We extracted unigram and bigram frequencies from the Spanish data included in the `OpenSubtitles` corpus (Lison & Tiedemann, 2016) using a custom script. This corpus is comprised of Spanish movie and television subtitles and contained approximately 1.1 billion words at the time of download. To ensure that all words under analysis had frequency counts, the unigram and bigram counts from the data we extracted from the `NCCSp` were added to those extracted from `OpenSubtitles`. The unigram and bigram frequencies were used to calculate the probability of the word containing the tap occurring based on the previous and following words (following Bell et al., 2009; Jurafsky, Bell, Gregory, & Raymond, 2001).

We wanted to include a measure of local speech rate in our models, and previous studies have done this in varying ways. For example, Bell et al. (2009) analyzed the `Switchboard` corpus and calculated speech rate based on the total length of the conversation. Santiago and Mairano (2018) transformed all of their vowel durations to z-scores per participant to account for speech rate. These methods do not give information on how fast a speaker was talking at a particular moment, and we expected the speech rate of each participant to vary considerably throughout the 90-minute recording. We elected to get a measure of speech rate for each utterance to account for how fast the speaker was talking when they produced the tap in question. Using the utterance boundaries from the corpus and the output transcription from the forced alignment, we calculated the speech rate for each utterance in syllables per second using a script in `Praat` that divided the total number of vowels in each utterance by the total duration.

3.2.6 Statistical modeling

To analyze patterns in the variability of intervocalic taps, we fit a series of Bayesian hierarchical regression models using the `brms` package (v2.18.0, Bürkner, 2017) in `R` (v4.1.1, R Core Team, 2021). The dependent variable in these models was IntDiff. We did not statistically model tap duration for three main reasons. First, the placement of force-aligned boundaries is limited to 10ms intervals, and these were the only boundaries available for 90% of our data. Second, as mentioned by previous research, there is no reliable way to manually measure the duration for taps without visible occlusions (Amengual, 2016), which in our hand-corrected data was approximately half of all tokens, consistent with previous research (Bradley & Willis, 2012; Willis & Bradley, 2008). Third, intensity-based measurements of duration that can be applied more broadly deviate from hand-measured taps when there is a visible occlusion, which calls their validity into question (Perry et al., 2023).

We began the modeling process by fitting linear models with Gaussian likelihoods, which have been used to analyze IntDiff and other intensity-based measurements of various segments in several languages such as English (e.g., Cohen Priva & Gleason, 2020; Warner & Tucker, 2011), Spanish (Hualde et al., 2011), and Basque (Hualde et al., 2019). All models had varying intercepts by speaker, and our list of predictors is as follows, with continuous predictors being centered and scaled:

- **Unigram freq.** Log unigram frequency for word containing tap
- **Bigram freq. prev.** Log bigram frequency for word and previous word
- **Bigram freq. fol.** Log bigram frequency for word and following word
- **Cond. prob. prev.** Conditional probability of the word based on previous word

- **Cond. prob. fol.** Conditional probability of the word based on following word
- **Number of syllables** Word length in number of syllables (log-transformed)
- **Function word** Sum-coded, difference between function (0.5) vs. content words (-0.5)
- **Local speech rate** Log average speech rate of the utterance containing tap (syllables/s)
- **Prev. vowel** Preceding vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Fol. vowel** Following vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Prev. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) previous vowel
- **Fol. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) following vowel

Posterior predictive checks on our Gaussian model revealed it was badly misspecified, i.e., the model's assumptions were unambiguously violated. This can be seen in Figure 3.2 A, where data generated by the Gaussian model appear in grey and the observed hand-corrected data in black. To account for the general shape of the data, we then fit models with a skew-normal likelihood (Figure 3.2 B), which was a visual improvement during posterior checks and had better-estimated out-of-sample predictive accuracy as estimated by the leave-one-out information criterion (LOOIC; Vehtari et al., 2017). Despite these improvements, the model was still incapable of generating the observed data. We decided to fit finite mixture models, which allowed

us to model IntDiff with two distributions instead of one, accounting for the bimodal nature of the data. These two distributions can be thought of as latent pronunciation variants of the Spanish tap. Therefore, instead of modeling the mean IntDiff of all taps together, we are modeling two means: one for a 'reduced' pronunciation variant and another for an 'unreduced' variant.

Our first finite mixture model was comprised of two Gaussian distributions. This model did a poor job of capturing the behavior of the left tail due to the strong right skew created by the soft boundary of IntDiff at zero. We subsequently changed the first component distribution to a skew-normal to account for this, which was a large improvement, as evidenced by LOOIC. Our final model was a combination of two skew-normal distributions, visualized in Figure 3.2 D, which was a small improvement over the combination of a skew-normal and Gaussian (Figure 3.2 C). The details regarding these comparisons, including the fitted models and LOOIC estimates are included in the supplementary materials.

The priors for our finite mixture models were generally weakly-informative in the context of Spanish taps. An exception to this was the priors on the intercepts of our two distributions, which were chosen to fix what is referred to as a label-switching problem (Jasra et al., 2005). This occurred because our two distributions overlapped, which created bimodal posteriors as the sampler jumped between parameter values for the two distributions. To allow the model to fit properly, we needed to use domain knowledge in the form of non-exchangeable priors on the intercepts of the two distributions (Betancourt, 2017) that state we expect the means of our reduced and unreduced taps to be different. For further discussion of this issue and other aspects of the modeling process, the reader is directed to our modeling supplement.

With finite mixture models, our predictors were now associated with changes in the mean IntDiff for both reduced and unreduced taps. In addition, it is possible

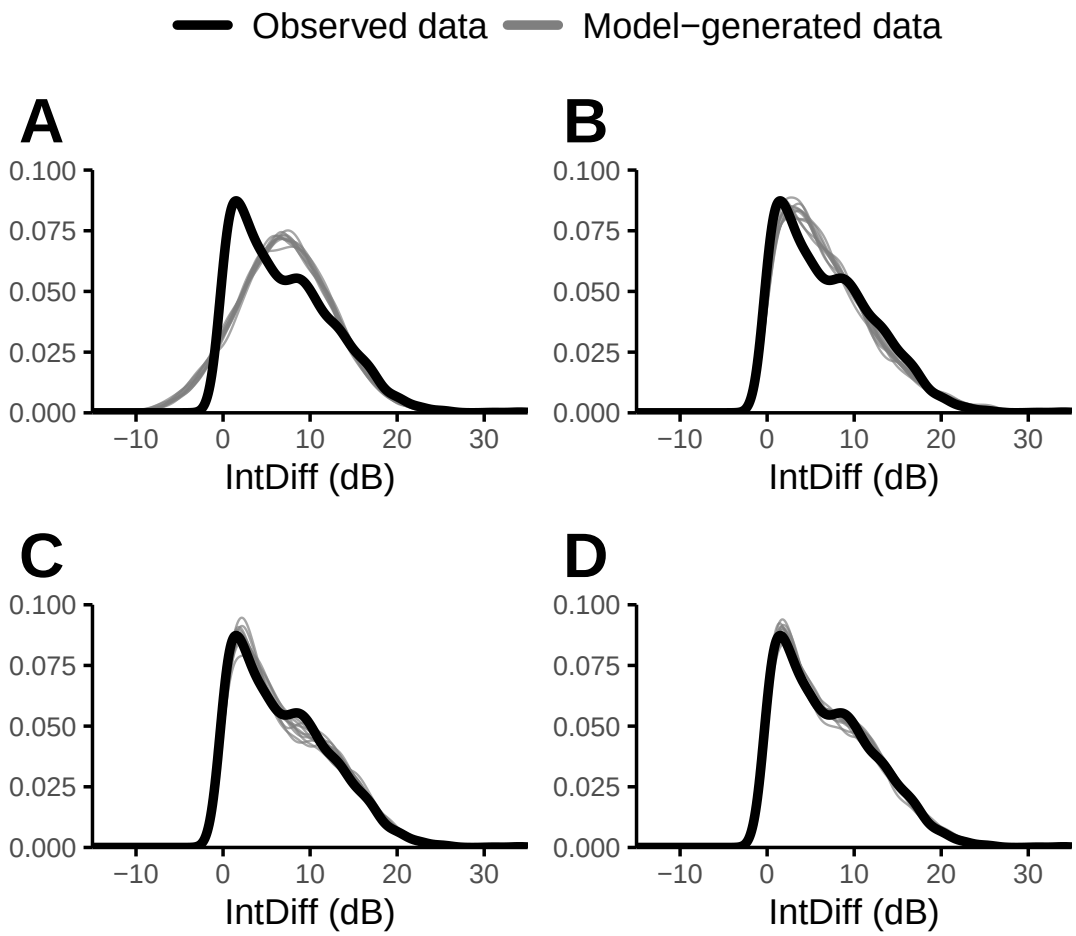


Figure 3.2: Posterior predictive checks visualized using density plots. Multiple simulated data sets from the fitted models are plotted in gray, and the observed hand-corrected data is plotted in black. Top-left is a model with a Gaussian likelihood. Top-right is a model with a skew-normal likelihood. Bottom-left is a mixture model with skew-normal and Gaussian component distributions. Bottom-right is a mixture model comprised of two skew-normal distributions.

in a finite mixture model to predict what are referred to as mixing proportions (McLachlan et al., 2019). This can be thought of conceptually as a logistic regression included as part of the overall model which estimates the probability of a tap being

unreduced according to our predictors. We included this as if there are two variants of the Spanish tap, it is plausible that our predictors could be related to the categorical alternation between variants as well as gradient changes in IntDiff.

The modeling process described above was conducted with only the hand-corrected data. As this data was responsible for choices made during the iterative model-building process, we were concerned about over-fitting. To evaluate our model, we randomly split the remaining 90% of the data into nine data sets comparable in size to the hand-corrected data. This split was done because, for mixture models exhibiting a label-switching problem, more data can overwhelm the priors used to fix the issue (Betancourt, 2017).

Model comparison on these data sets indicated that moving from the skew-normal and Gaussian mixture to a mixture of two skew-normal distributions was warranted. So we proceeded with that model, comparing it to reduced models: one that only contained phonetic variables, one that only contained lexical predictors, and a third reduced model that only contained varying intercepts by speaker. We did this to gain insight into whether Spanish tap variability was more predictable by lexical factors, phonetic factors, or both.

In order to get estimates for our predictors that were informed by all of our data, we entered the estimates from our hand-corrected model and eight cross-validation models into a Bayesian meta-analysis (excluding one cross-validation model that didn't converge). We interpret our estimates against a Region of Practical Equivalence (ROPE; Kruschke, 2018). This ROPE sets an upper and lower limit on effect sizes that we would consider meaningfully different from zero, meaning we can consider an effect to be practically null if our credible interval falls entirely within these bounds.

3.3 Results

We found a substantial amount of variability in Spanish tap production, echoing previous findings. This variability is described and depicted in Section 3.3.1. We also find that we can predict some of this variability using both lexical and phonetic predictors, indicating that Spanish, like many other languages, shows both types of patterns in speech production. Our models are reported in detail in Section 3.3.3.

3.3.1 Descriptive statistics

In the subset of our data that was hand-coded, true taps made up 21.6% of productions, approximant taps 28.8%, and perceptual taps 29.6%. Taps were deleted for 17.5% of cases and non-tap productions (mostly fricatives) made up the remaining 2.5%. The average duration of true taps was 28.8 ms (SD=8.2, range=10.0–81.3), and the average duration of approximant taps was 24.9ms (SD=7.1, range=10.3–65.7). These observations are plotted against IntDiff in Figure 3.3, where it can be seen that the two variables are correlated (r=0.35), as previous findings would predict (Cohen Priva & Gleason, 2020), but that high and low IntDiff values are possible for a large portion of duration values. The density distributions of IntDiff for the four tap types are visualized in Figure 3.4, where we see the expected ordering of IntDiff decreasing as taps become more reduced. The mean IntDiff of hand-corrected data was 6.99 dB (SD=5.44, range=-1.23–34.35), and the mean IntDiff of the force-aligned data (excluding observations that were hand-corrected) was 6.84 (SD=5.59, range=-2.24–56.55).

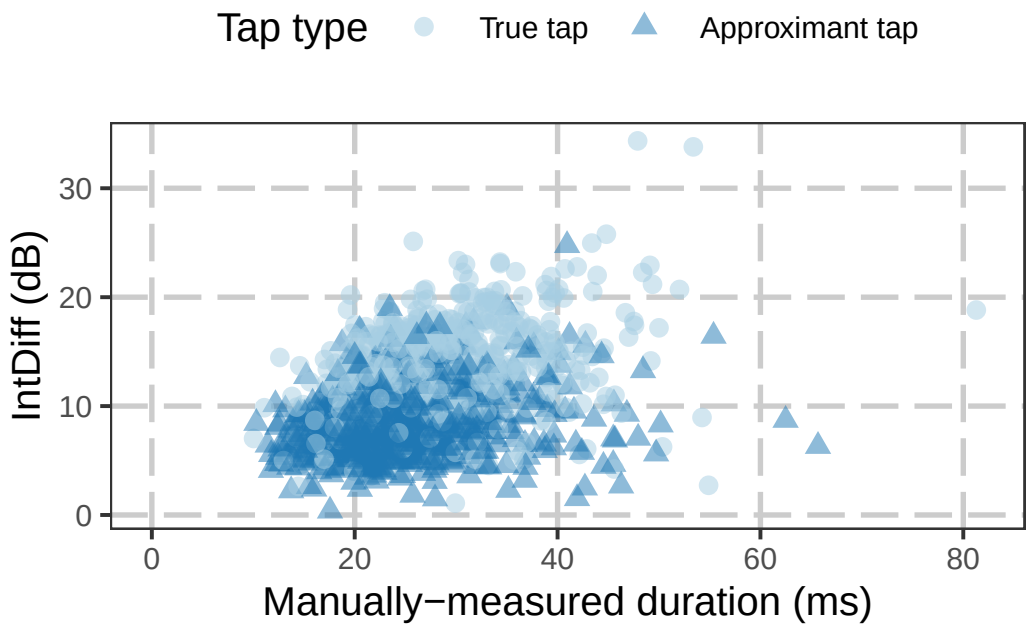


Figure 3.3: A scatter plot of manually-measured duration in milliseconds and IntDiff in decibels for taps with a visible occlusion, i.e., true taps and approximant taps.

3.3.2 Evaluation of forced-aligned data

As approximately 10% of intervocalic taps had their boundaries hand-corrected, we compared our measured values of IntDiff across the force-aligned and hand-corrected boundaries. The measurements on the two sets of boundaries were almost perfectly correlated ($r = 0.98$), and observations had the same value in 70.76% of cases. Most errors were small; the 95% quantile range for the difference between measurements spanned between -0.98 dB and 1.44 dB. Given this information, we were confident that using our force-aligned data was preferable to ignoring the 90% of intervocalic taps from the corpus we did not hand-correct. Additional visualizations of the comparison between force-aligned and hand-corrected data are available in the

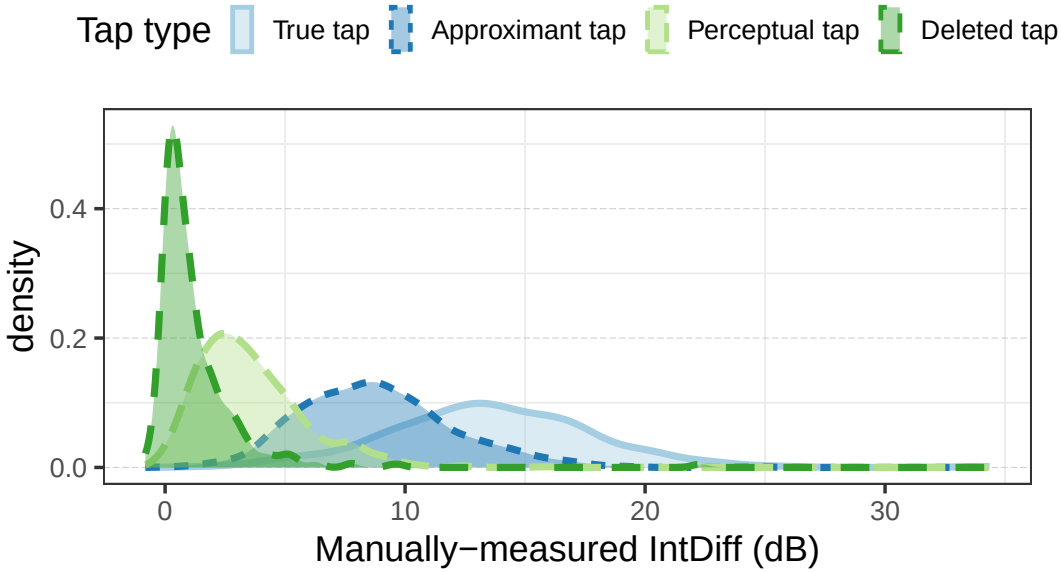


Figure 3.4: This figure shows separate density distributions of IntDiff in decibels for each of the hand-coded tap types. The tap type is denoted by color and line type.

supplementary materials.

3.3.3 Modelling results

In this section, we present the results derived from the finite mixture model with a likelihood comprised of two skew normal distributions, and its comparisons to other models. Note that these models do not contain the two measures of conditional probability. Including these variables in the more complicated models often created model fitting issues in the form of divergent transitions, possibly due to the concentration of data in the low predictability areas. In fact, the sparsity of high-predictability items may have made these effects unreliable and not worth interpreting, convergence issues aside. On conditional probability, we suggest that future investigations

Table 3.1: The percentages of tap types from the manual qualitative coding scheme, rounded to 1 decimal point, that are found above and below 5.5 dB of IntDiff.

Tap type	**% below 5.5 dB**	**% above 5.5 dB**
True tap	1.5	39.5
Approximant tap	8	47
Perceptual tap	52.5	9
Deletion	36	0.5
Non-tap production	1.5	3.5

into this variable may be more successful with experimental data, where it can be examined in a controlled way with several distinct lexical items.

Describing the two distributions

Our model's estimated predicted performance improved by modeling IntDiff using two distributions instead of one. Before we discuss model comparison and interpret parameter values, we are going to analyze these two distributions in the context of our hand-corrected data. In the density plot at the top of Figure 3.5, we plot the two theoretical distributions from our fitted model. The vertical dashed line where IntDiff equals 5.5 dB is the approximate location where observations from the two distributions are equally probable. As this data was hand-coded for tap type, we can look at the percentages of different hand-coded tap types above and below this value of IntDiff, presented in Table 3.1.

So, where the reduced distribution is more likely, we see mostly perceptual taps and deletions. To assess the idea that reduced taps were perceptual taps and deletions while unreduced taps were true taps and approximants, we removed non-tap productions from the data and then plotted the two empirical density distributions at the bottom of Figure 3.5. We grouped together deleted and perceptual taps (green) and approximant and true taps (blue). The high similarity between the two distribu-

tions estimated by our model and the empirical distributions here support the idea that there are two taps types in Spanish from the point of view of IntDiff. This is a simplification, but we believe it is a useful one to make. It is possible that we could improve our model by adding additional distributions, maybe even four, one for each tap type plotted in Figure 3.4. However, two distributions were able to explain the data adequately, and we did not want to increase model complexity without reason.

Comparison of sub-models

To learn about which types of factors were important for predicting variation in Spanish tap production, we compared the model fit to hand-corrected data reported in Section 3.3.4 to reduced models. We fit three reduced models: one with only phonetic predictors, one with only lexical predictors, and one with only an intercept that varied by speaker. Comparisons via LOOIC indicate that the full model had better estimated out-of-sample predictive accuracy in the form of the expected log predictive density (elpd) than both the phonetic model (elpd_diff = -31.6, se_diff = 9.6) and the lexical model (elpd_diff = -53.2, se_diff = 12.1), indicating that the model's predictive power is improved by considering both phonetic and lexical factors. The phonetic model was ranked second, but there was relatively weak evidence that it outperforms the lexical model (elpd_diff = -21.7, se_diff = 13.7). All models were improvements over the Intercept-only model. This is empirical support that both phonetic-level properties (e.g., speech rate, surrounding segments) and lexical information (e.g., lexical frequencies) are associated with variation in Spanish tap production.

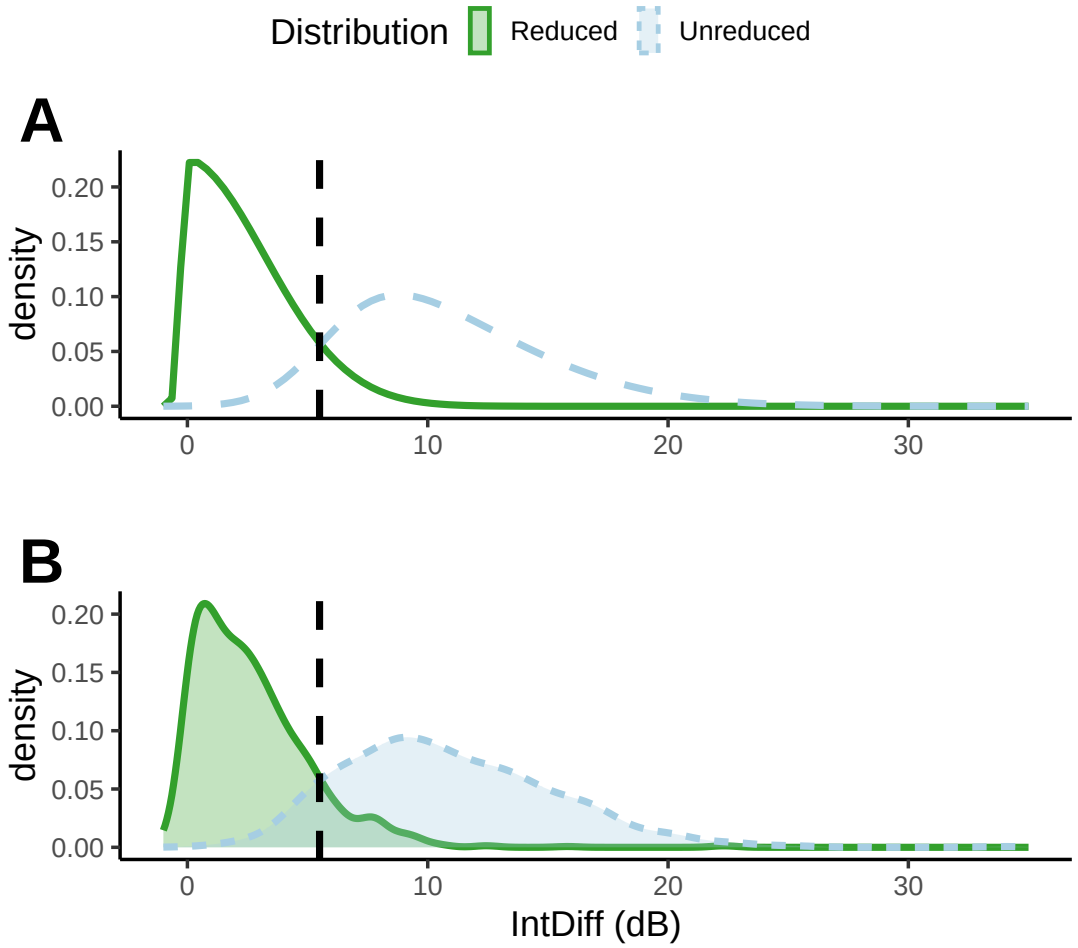


Figure 3.5: Top: plotted are the theoretical density distributions for IntDiff from our fitted finite mixture model with two skew-normal distributions. Bottom: plotted are the empirical density distributions for IntDiff from the hand-corrected data, with Deleted and Perceptual taps plotted in green and Approximant and True taps plotted in blue. The vertical dashed black line is plotted at 5.5 dB in both plots.

3.3.4 Predictor estimates

For our set of population-level predictors, we can interpret how they are associated with changes in three model parameters: the gradient changes in the mean IntDiff of reduced taps and unreduced taps, as well as the changes in the probability of pro-

ducing an unreduced tap. Thus, beyond the conditional means of two distributions, we also have information about a categorical alternation. The estimates from the meta-analyis are displayed in Figures 3.6, 3.7, and 3.8. We focus on interpreting meta-analytic posteriors, as they make use of nearly an order of magnitude more data and generally overlap with plausible values from the hand-corrected model. For continuous predictors, the estimates have been multiplied by the range of the predictor in the hand-corrected data to visualize the total change.

Our posteriors are visualized together with our ROPEs, plotted as a dashed red line. Changes in mean IntDiff are visualized in decibels, and the ROPE for these predictors has been set to the range between -0.5 dB and 0.5 dB. In other words, effects resulting in a total change in IntDiff of less than half a decibel are not considered meaningful in this context. This is motivated by two reasons. First, the just noticeable difference for intensity is likely larger than 0.5 dB (e.g., Slade et al., 2022). Second, the standard deviation of IntDiff was over 5, meaning 0.5 would be considered a trivial effect size.

The ROPE for the probability of an unreduced tap is on the logit scale, and we chose values between -0.2 and 0.2. Probability is not linear with the logit scale, but this equates to saying that a change in log-odds that *at most* could equate to a 5% change in the probability of an unreduced tap is something we consider too small to be of practical significance for present purposes. If one wanted to incorporate any of this information into a cognitive model of speech production, any change that is reliably non-zero could be a meaningful component.

In our models, many predictors are associated with changes in the probability of an unreduced tap being produced, visualized in Figure 3.6. We see differences associated with the phonetic environment, lexical stress, speech rate, word length, and unigram frequency. Speech rate, in particular, is associated with large changes

in this probability, twice as large as any other predictor. If we take the posterior mean, the change in the probability of an unreduced tap changes from 0.86 when people are speaking at the slowest rates to 0.41 when people are speaking at the fastest rate observed.

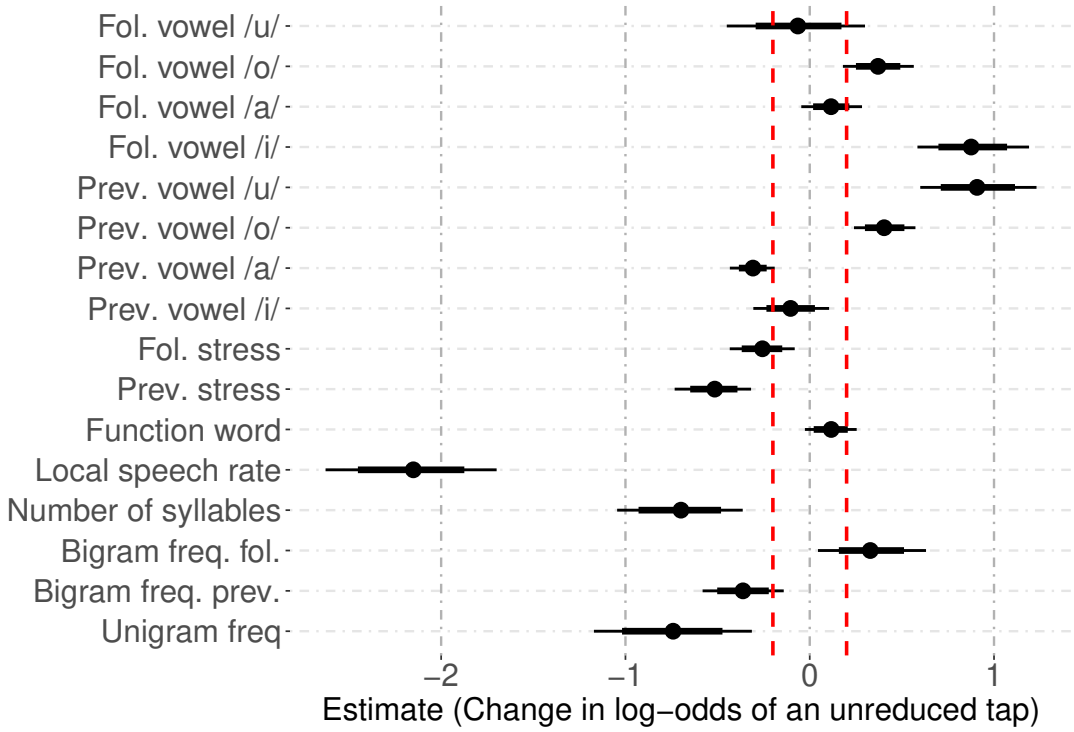


Figure 3.6: Posterior distributions from the meta-analysis for changes in the probability of producing an unreduced tap, shown in log-odds. Draws from continuous predictors were multiplied by the total range to visualize the total change between the lowest and highest values. The shape is the mean of the posterior; the thick line represents the most probable 80%, and the thin line represents the most probable 95%.

In contrast to the categorical alternation, most predictors were not associated with meaningful changes in average IntDiff of reduced taps (Figure 3.7). The main exception is speech rate, with faster speech being associated with increased reduction. The effect of the following vowel being an /i/ as compared to an /e/ straddles the

upper bound of our ROPE. This indicates that the model less than 80% confident that the effect is meaningful.

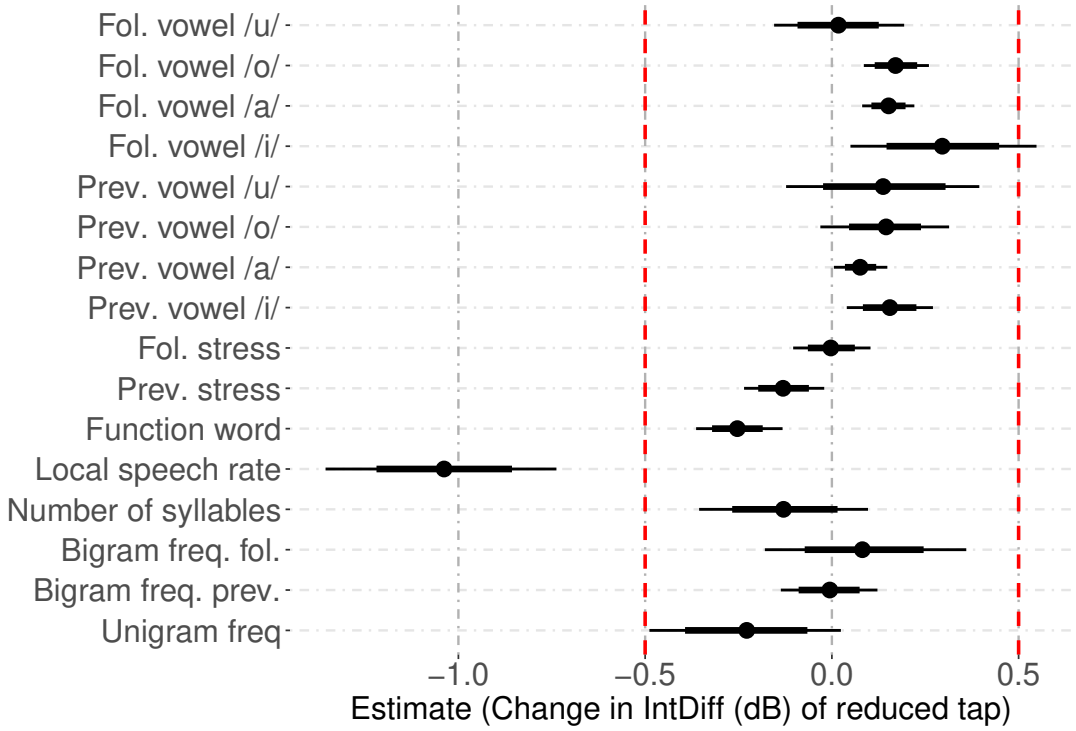


Figure 3.7: Posterior distributions from the meta-analysis for changes in average IntDiff of reduced taps. Draws from continuous predictors were multiplied by the total range to visualize the total change between the lowest and highest values. The shape is the mean of the posterior; the thick line represents the most probable 80%, and the thin line represents the most probable 95%.

Compared to the reduced taps, we are much more unsure of how our predictors are associated with changes in unreduced taps. Figure 3.8 illustrates that many posteriors from the meta-analysis overlap only partially with the ROPE. We are reasonably confident that increased unigram frequency is associated with lower IntDiff values, and that unreduced taps are more reduced before a following /u/ than before a following /e/. We are more than 80% confident that increased speech rate leads to

lower mean IntDiff values, but less than 95% confident.

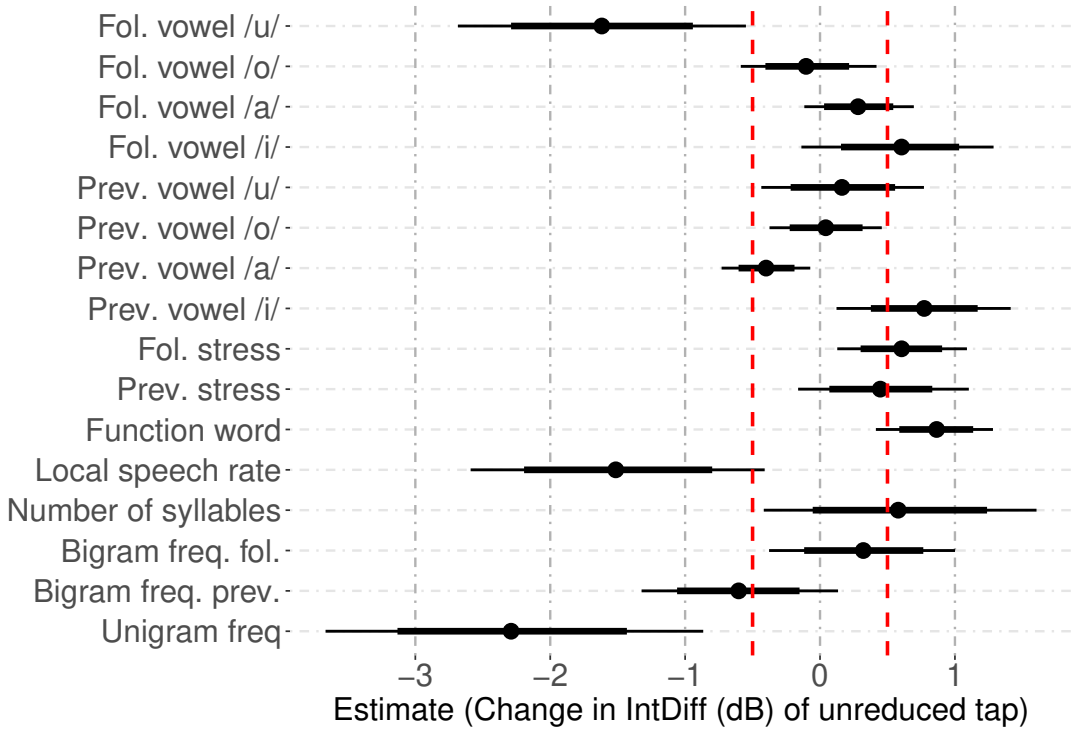


Figure 3.8: Posterior distributions from the meta-analysis for changes in average IntDiff of unreduced taps. Draws from continuous predictors were multiplied by the total range to visualize the total change between the lowest and highest values. The shape is the mean of the posterior; the thick line represents the most probable 80%, and the thin line represents the most probable 95%.

3.4 Discussion

The present study documented and modeled acoustic variation in the production of intervocalic alveolar taps in a corpus of conversational Spanish from Madrid, Spain. Similar to previous work on Spanish taps and stop-consonants in other languages, there was a high degree of variability in the acoustic realizations, with stop-like

taps, approximants, and apparent deletions all being relatively common. We found that statistical models that assumed two latent categories, which we interpret as two pronunciation variants, improved our models' predictive performance. We found that both phonetic and lexical variables can help predict variation in intensity difference (IntDiff) for Spanish taps.

Our descriptive analysis shows that our sample of Spanish taps is similar to previous work along a number of dimensions. While previous studies have not focused on conversational speech, studies using the most naturalistic elicitation methods (e.g., Bradley & Willis, 2012; Willis & Bradley, 2008) yielded percentages of different tap types similar to the present study. Manually measured durations in taps with visible occlusions are also consistent with measurements from other dialects and speaker populations (Amengual, 2016; Bradley & Willis, 2012; Willis & Bradley, 2008). Our distribution of IntDiff is also similar to the only other study we are aware of that has measured this in Spanish (J. Y. Kim & Repiso-Puigdelliura, 2020), although we did observe a larger overall range of values. This is not surprising as we analyzed more tokens and speakers as well as a more casual speech style. In general, the data in our sample is consistent with data reported from other dialects and speaker populations.

Our general result of increased reduction as word frequency increases adds to the many studies with a similar effect (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001; Pluymaekers et al., 2005). This finding is consistent with H&H Theory, as we would expect familiarity with a word to influence how it is produced and recognized. It is also consistent with claims of probabilistic reduction, as more frequent words are more likely to occur in general. The effect of word frequency is also consistent with the Smooth Signal Redundancy Hypothesis, as log word frequency was one approximation of redundancy used by Aylett and Turk (2004). Exploring other variables used as approximations of redundancy, informativity, and predictability

would be a fruitful avenue for future research.

Some studies have found that high-frequency multi-word phrases are reduced in production (Tremblay & Tucker, 2011) and processed faster (Arnon & Snider, 2010). Our effects of bigram frequency were small, and while the 95% credible intervals did not overlap with zero for predicting an unreduced tap, they did overlap with the ROPE. Higher bigram frequency with the previous word was associated with increased reduction, while higher bigram frequency with the following word was associated with decreased reduction. This could reflect the fact that these bigrams function as a multi-word unit (Arnon & Cohen Priva, 2013; Kuiper et al., 2007), and we see more reduction at the end of frequent two-word units, while enhancement occurs at the beginning. Our finding of *less* reduction with increased bigram frequency with the following word could also be a function of the level of measurement, and previous findings of decreased word duration (e.g., Bell et al., 2009) if also present in Spanish, may not entail reduction of all segments within a word as shown by findings from D. Kim and Smith (2019) for English vowels.

Among our phonetic factors, we found increased reduction at faster speech rates, consistent with articulatory undershoot (Lindblom, 1963). We also found more reduced variants following a stressed vowel than an unstressed vowel. As stress is mainly related to F1 raising in Spanish (Hernandez et al., 2023; Torreira & Ernestus, 2011), the lower position of the tongue or jaw means that, on average, a longer distance would need to be covered when the previous vowel is stressed. Other differences based on the quality of the surrounding vowels support this idea, as a following /i/ and preceding /u/ are also related to less reduction, while a preceding /a/ is associated with more reduction. As this effect of the phonetic environment was inconsistent, happening only for certain high vowels, and the evidence is acoustic, these articulatory explanations should be taken as preliminary.

A second main finding of the present study is empirical support for two pronunciation variants of the Spanish tap along the acoustic correlate IntDiff. The evidence for this comes from the fact that a model assuming two latent categories provided an improvement in the predictive accuracy of our model compared to models with a single distribution. It is possible that there are *more* than two pronunciation variants, however, our goal was to create a parsimonious model that could account for the data. Even our assumption of two pronunciation variants would be unnecessary if a predictor could account for the observed bimodal distribution with a shift in the mean of a single distribution. This effect would have to be approximately 8 dB, and as none of our predictors had estimated effects this large, we have no hypotheses as to what such a predictor could be. We make the claim of two variants along IntDiff, and believe an articulatory investigation into whether there are distinct tongue movements would be a logical next step in testing this claim more generally.

A limitation of this study, shared with almost all studies that analyze spontaneous speech, is that we did not appeal to a causal model when building and interpreting our statistical models. This is problematic, as coefficients in multiple regression may not estimate the same type of effect on the outcome (Westreich & Greenland, 2013) and therefore may not be intuitively interpretable. A shift toward causally motivated analyses of spontaneous speech, as was done by Cohen Priva and Gleason (2020), is appropriate given the observational nature of the data. Cohen Priva and Gleason highlighted the clear advantage of causal models in having explicit scientific assumptions drive analysis decisions and interpretation. Their simplified causal model, which collapsed distinct factors (e.g., speech rate and word frequency) and ignored the issues of unobserved confounds and measurement error, highlights the difficulty of building causal models given the knowledge base of the field. Indeed, even for pairs of variables with well-studied relationships, like frequency and word

length, it would not be trivial to include them in a causal model, as some current proposals have the link between them originating from an unconscious process of optimal communication from an information-theoretic perspective (e.g., Ferrer-i-Cancho et al., 2022; Kanwal et al., 2017). It is unclear whether this would be relevant to a model of speech production or is a result of long-term cognitive pressures.

Although our interpretation of individual coefficients may be problematic, the present study does provide evidence that both phonetic and lexical variables are useful in *predicting* IntDiff. While prediction and causal inference are distinct tasks (Arnold et al., 2020), we take this as evidence that some relationship exists between our variables and tap variation, even if the specifics of the system render our coefficients uninterpretable. Further research into the networks of variables used for modeling reduction would benefit from employing explicit causal models, as some variables are used to calculate others (e.g., bigram frequency and conditional probability).

Prosodic variables are one of the other limitations of this study that future studies could address. We did not prosodically annotate the corpus, and certain stressed syllables may, in fact, have been de-stressed or stressed differently in our running speech, and this was undoubtedly the case for some function words (e.g., *pero*). Regarding function words, there are only four unique function words that contain intervocalic taps, and as such the differences in tap production we find for function words in this study should be interpreted cautiously.

The results of the present study are pertinent to another discussion involving phonetic variability: the division between categorical alternations and continuous reduction (or lenition) patterns. More of our predictors were associated with categorical changes between pronunciation variants than gradient changes within variants. We might be tempted to simplify and state that Spanish tap reduction is a cate-

gorical alternation between pronunciation variants. In reality, we see both gradient and categorical effects in our model. For Spanish taps, both types of variation are happening, perhaps simultaneously. This was argued by Bürki et al. (2011) to be the case for schwa deletion in French. Future perceptual studies investigating how this type of categorical sub-phonemic variation impacts perception and word processing would help shed light on the dynamics of the sound system as a whole.

Previous analyses of stop-consonant variability employing IntDiff have, implicitly or explicitly, assumed it was a continuous reduction pattern during analysis. It is difficult to know how often this assumption is supported by the data, or if gradient and categorical patterns of stop variability occur for the same segment often cross-linguistically. Spanish taps having two variants along IntDiff may be a rarity, and a Gaussian likelihood may appropriately model variation in IntDiff for many stops across many languages. It also may be that having multiple latent pronunciation variants along continuous acoustic correlates is relatively common, and defaulting to the same statistical models is preventing us from learning this. As reporting and visualizing model diagnostics is not common, we cannot know how often IntDiff requires a non-Gaussian likelihood to be modeled appropriately. It would be fruitful to reanalyze previous data in order to see when statistical assumptions clearly depart from the observed data. Doing so could potentially resolve contrasting findings or uncover more nuanced patterns of acoustic variability, as the effects of various predictors in our models depended on the model specification.

We see general utility in our modeling approach as a flexible option for modeling acoustic correlates when distinct latent categories may apply. In hand-coded data, such categories can be qualitatively analyzed, with categories being potentially added as an independent variable to ameliorate poor model fits. However, for forced-aligned data, which is common, finite mixture models allow us to let the model decide on

the probability of an observation being in one category vs. others. This inclusion of categorical and gradient effects on acoustic measurements could be applied to diachronic sound change, looking at the shift towards one variant vs. the other and the shift in baseline rates over time. This may be directly applicable to Spanish taps in the coming years. The present study found higher lexical frequency is associated with more reduced variants, similar to recent work on other varieties of Spanish, which found a frequency effect for innovative trill variants (Pollock et al., 2023). These findings are consistent with claims that lexical frequency is a potential driver of sound change (e.g., Bybee, 2007; Bybee & Hopper, 2001; Hall et al., 2018), where high-frequency words containing reduced variants lead to the reduced variant being the default over time. In Spanish, a reduced tap could be an attempt at maintaining the phonemic contrast with reduced trills in the intervocalic position. Future data may shed light on how this potential change in progress develops.

3.5 Conclusion

The present study has documented and modeled Spanish tap variability in spontaneous, conversational speech. Our results echo previous findings of increased phonetic reduction at increased frequency and speech rates. Our modeling approach indicates that the Spanish tap has two pronunciation variants: unreduced taps that typically have a visible occlusion on a spectrogram and reduced taps without such an occlusion. Model comparisons indicate that lexical and phonetic variables help us predict the acoustic variability of Spanish taps. Overall, our results contribute to the body of knowledge regarding lexical and phonetic influences during speech production, as well as depicting the type of variation in this sound that Spanish speakers encounter on a daily basis.

3.6 Data availability

The data and materials documenting the support of the findings of this study are openly available in the University of Alberta's Education and Resource Archive at https://doi.org/10.7939/r3-66gq-mf49.

Chapter 4

Word-medial tap reduction and lexical processing in first- and second-language Spanish listeners

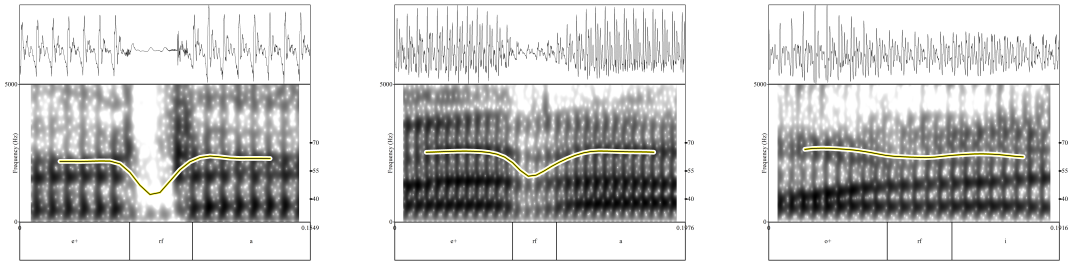
4.1 Introduction

Spoken language is highly variable (e.g., Ernestus & Warner, 2011; Greenberg, 1999; Johnson, 2004), and sounds that are characterized as stop-consonants in linguistics are often realized differently in the spontaneous, casual speech that one encounters daily. These realizations may include fricatives, approximants, and apparent deletions, and work on stops in several languages has found this process, call *reduction*, to be a common occurrence (e.g., Barry & Andreeva, 2001; Katz & Pitzanti, 2019; Mukai, 2020; Warner & Tucker, 2011). As the variability associated with reduction is ubiquitous in the speech we encounter daily (Warner, 2023), we cannot adequately understand the processes involved in spoken word recognition without incorporating

the variability we observe in spontaneous speech (Tucker & Ernestus, 2016). The present study investigates how the reduction of word-medial alveolar taps (hereafter taps) affects spoken word recognition in first-language (L1) and second-language (L2) listeners of Spanish.

Spanish taps are highly variable across several varieties and speaker populations (Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020; Perry et al., 2024; Willis & Bradley, 2008). Realizations of taps in these studies have been qualitatively coded into discrete variants based on their visual appearance in a spectrogram, visualized in figure 4.1. Variants include stop-like 'true taps,' which have stop closures and burst releases (Figure 4.1 A), 'approximant taps' that have a visible presence but without breaks in the formant structure (Figure 4.1 B), and 'perceptual taps,' which are nearly elided tokens without a visible presence in a spectrogram (Figure 4.1 C). Some studies also code deletions and non-tap productions (e.g., J. Y. Kim & Repiso-Puigdelliura, 2020; Perry et al., 2024), which are not directly relevant to the present study. Research that has focused on analyzing elicited narration data and spontaneous speech reports that perceptual taps and apparent deletions make up approximately half of all taps that L1 Spanish speakers produce (Bradley & Willis, 2012; Perry et al., 2024; Willis & Bradley, 2008). While this sub-phonemic variation has been documented in production, we are unaware of any studies that have investigated how this variability impacts the recognition of words.

There is a wide range of phonetic variability that can be considered reduction, but reduced segments are generally less acoustically salient than their unreduced counterparts. This can come in the form of decreased duration (e.g., Bürki et al., 2011; Cohen Priva & Gleason, 2020; Warner & Tucker, 2011), more centralized vowels (e.g., Aylett & Turk, 2006; Munson & Solomon, 2004; Wright, 2004), smaller intensity drops during consonants (e.g., Perry et al., 2024; Warner & Tucker, 2011),



(A) (B) (C)

Figure 4.1: Spectrograms of spontaneously produced taps from the Nijmegen Corpus of Casual Spanish. The intensity curve is overlaid on each spectrogram using the same scale, with decibel values indicated on the right vertical axis. Panel A is a true tap, panel B is an approximant tap, and panel C is a perceptual tap.

and higher probabilities of being deleted (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001). Many previous studies that have investigated how we process and recognize reduced speech have found that reduction is an inhibitory factor (e.g., Ernestus et al., 2002; Ernestus & Baayen, 2007; Mukai, 2020; Pitt, 2009; Ranbom & Connine, 2007; Tucker, 2011; Wanrooij & Raijmakers, 2020). This can come in the form of slower and less accurate responses in lexical decision (e.g., Ernestus & Baayen, 2007; Tucker, 2011), increased pupil dilation during lexical retrieval (Mukai et al., 2023), and less accurate orthographic transcriptions of recorded audio (e.g., Ernestus et al., 2002; Wanrooij & Raijmakers, 2020).

In contrast to the research findings on the inhibitory effects of reduction, some research has also documented a *facilitatory* effect of reduction (McLennan et al., 2003; van de Ven & Ernestus, 2018). van de Ven and Ernestus (2018) explain their results through methodological differences, pointing out that studies finding an inhibitory effect of reduction generally investigate the recognition of isolated words, while their study provided some of the surrounding context, which is known to aid

the recognition of reduced forms (Ernestus et al., 2002). In contrast, the apparent advantage of 'casual' over 'careful' pronunciations found by McLennan et al. (2003) can be attributed to a terminological difference when looking at the phonetic properties of the stimuli used (Tucker & Mukai, 2023). Inhibitory effects of reduction for the American English flap were reported by Tucker (2011) while McLennan et al. (2003) found a facilitatory effect of casual (vs. careful) pronunciation in the same phoneme. These two studies are not, in fact, in disagreement, because all items in Tucker (2011) were flaps while McLennan et al. (2003) compared flaps to hyperarticulated /t/ productions. This demonstrates the need for researchers to be clear about the phonetic nature of the stimuli they are comparing, as determining what constitutes a 'reduced' production of a segment is not always clear. For related reasons, an acoustic analysis of the stimuli used in perception experiments helps contextualize results and facilitate comparisons across different studies, as the same phonetic 'label' can correspond to a wide variety of acoustic measurements.

While there are several studies documenting the variability of Spanish taps, it is not clear as to what should constitute a 'reduced' tap in Spanish. If we consider that the tap *can* be produced with a stop closure and burst release, meaning the vocal tract is completely obstructed, then any variant of this sound where that does not happen could be considered to be reduced. A point in favour of this is that Amengual (2016) found Spanish-dominant heritage speakers overwhelmingly produced true taps in read speech elicited with carrier sentences. This shows that in a situation where one would be most likely to enunciate carefully, a true tap is much more likely than any other variant. A counterargument can be found in work on more natural speech styles, which finds that true taps constitute the minority of productions (e.g., Bradley & Willis, 2012; J. Y. Kim & Repiso-Puigdelliura, 2020; Perry et al., 2024; Willis & Bradley, 2008). Perry et al. (2024) found that true taps and approximant

taps pattern together in terms of their intensity difference (IntDiff), while perceptual taps and deletions were best characterized by a separate distribution, and interpreted this as evidence that there are two pronunciation variants of the Spanish tap, at least along the acoustic correlate of IntDiff. Based on this production data, perceptual taps would count as reduced in Spanish, but approximant taps would not. Additional support for the idea that approximant taps may not count as reduced is that the duration of true and approximant taps are well-characterized by a single lognormal distribution (Perry et al., 2023) and that J. Y. Kim and Repiso-Puigdelliura (2020) coded both true and approximant taps as 'target-like' when looking at variation in tap production by heritage speakers of Spanish from southern California.

Another factor related to phonetic variability that has been shown to impact spoken word recognition is the frequency of occurrence of a pronunciation variant. In spontaneous speech, and therefore in the input that language users are exposed to, different possible realizations of segments occur at different frequencies (see Tucker & Mukai, 2023, and references therein). The frequency of occurrence of a pronunciation variant has been shown to be a facilitatory factor in a variety of experimental tasks (e.g., Brand & Ernestus, 2018; Bürki & Frauenfelder, 2012; Bürki et al., 2018; Connine et al., 2008; Pitt et al., 2011), with the general findings being faster and more accurate responses for more frequent variants. Thus, there are two separate issues concerning phonetic variability that we must consider when looking into spoken word recognition. As researchers have observed reduction in all languages studied (Warner, 2023), and frequencies of pronunciation variants for different segments occur at different rates for different languages (e.g., Broś et al., 2021; Warner & Tucker, 2011), we need to gather empirical and experimental evidence from as many languages as possible in order to tease apart how these two factors impact spoken word recognition.

So far, we have focused on how reduction is perceived and processed by L1 speakers of the language. These are language users who have had substantial exposure to the language that began early in life, where input may have differed (Dilley et al., 2019). Compared to work on L1 listeners, there has been much less work done on how L2 listeners perceive and process reduced speech. Findings generally indicate that L2 listeners have a harder time recognizing reduced speech than L1 listeners (e.g., Ernestus, Dikmans, & Giezenaar, 2017; Ernestus, Kouwenhoven, & van Mulken, 2017; Wanrooij & Raijmakers, 2021). This is generally assumed to reflect a lack of sufficient experience with the language (e.g., Morano et al., 2023; Wanrooij & Raijmakers, 2021). Wanrooij and Raijmakers (2020) put forward evidence that even L1 German adolescents perceived reduced speech less accurately than L1 German adults, indicating that it can take a considerable amount of input and practice to process reduced forms. When we consider the fact that classroom L2 learners may be exposed to less reduced variants of words, the findings of a larger inhibitory effect of reduction in L2 learners are not surprising.

Compared to work on reduction and L2 listeners, L2 and non-native spoken word recognition, in general, has a much larger body of research (Warner, 2023). Studies typically report that L2 listeners are slower and less accurate overall as compared to L1 listeners (e.g., Díaz et al., 2012; Nijveld et al., 2022). Within L2 listeners, studies on individual differences have shown that the speed and accuracy of recognition of spoken language improve as their proficiency increases and their amount or length of exposure to the L2 increases (see Grosjean, 2018, and references therein). Beyond these overall differences, cross-linguistic effects on spoken word recognition are well documented (e.g., Llompart et al., 2021; Weber & Cutler, 2004), which show that properties of the L1 sound system impact how words are recognized in the L2.

The goal of the present study is to add to the body of literature regarding reduc-

tion and spoken word recognition, offering data from Spanish, which is a language that has received less attention in the literature. We also included L2 listeners in our study, in order to add to the body of research on how reduced speech impacts lexical access in L2 listeners and evaluate whether differences between L1 and L2 learners in Spanish may be able to inform models of bilingual spoken word recognition. The research questions that guide the present study are:

- RQ1: How does tap reduction impact the speed and accuracy of spoken word recognition in L1 and L2 Spanish listeners?
- RQ2: Does accounting for specific tap realizations based on previous literature lead to different patterns in reaction times and accuracy compared to assuming all lenited variants are reduced?
- RQ3: Does predicting responses by a continuous measure of reduction (as in Tucker, 2011) work better than a categorical variable?

Considering the frequency of reduced taps in Spanish, both L1 and L2 speakers must learn to recognize words containing reduced taps in order to communicate successfully. However, this doesn't mean that words with reduced taps won't take longer to recognize or are not identified less accurately when heard in isolation, as demonstrated by the research on the effect for L1 speakers (e.g., Ernestus & Baayen, 2007; Tucker, 2011). Based on previous research, we expect words containing reduced taps to be recognized more slowly and less accurately than words containing unreduced taps (Ernestus et al., 2002; Ernestus & Baayen, 2007; Mukai, 2020; Pitt, 2009; Ranbom & Connine, 2007; Tucker, 2011; Wanrooij & Raijmakers, 2020), and this difference will be larger for L2 listeners (Ernestus, Dikmans, & Giezenaar, 2017; Ernestus, Kouwenhoven, & van Mulken, 2017; Wanrooij & Raijmakers, 2021). Related to the

second research question, we believe that it may be the case that we only see differences with perceptual taps, which are more highly reduced than approximant taps and have been shown to pattern differently than true taps and approximant taps in terms of their intensity difference (Perry et al., 2024). Regarding the third and final research question, we have reason to believe that using a continuous measurement of reduction, which is based on the acoustic signal, could be able to provide a more fine-grained measurement of the stimuli, allowing us to predict responses better.

4.2 Methods

4.2.1 Participants

This study collected data from 168 participants who reported being L1 speakers of either Spanish (n=83) or English (n=85). Participant recruitment began with direct recruitment via email and moved to Prolific shortly thereafter. Four English L1 participants were removed for not indicating handedness properly at the beginning of the study. We removed 4 Spanish L1 and 8 English L1 participants due to having more than 16 non-responses in the data, a cutoff determined by examining the overall distribution of missing responses per participant. We removed participants that had an overall accuracy rate below 66%, which was the case for 18 English and 2 Spanish L1 participants. This accuracy cutoff was decided upon by looking at the overall distribution of accuracy rates and wanting participants who were, in general, responding correctly more often than not. It is also similar to the exclusion criterion of 60% employed by Tucker et al. (2019). Employing more strict cutoffs would exclude the majority of L2 listeners. Regarding the percentage of English participants who were removed for low accuracy rates, the ethical protocol stipulated

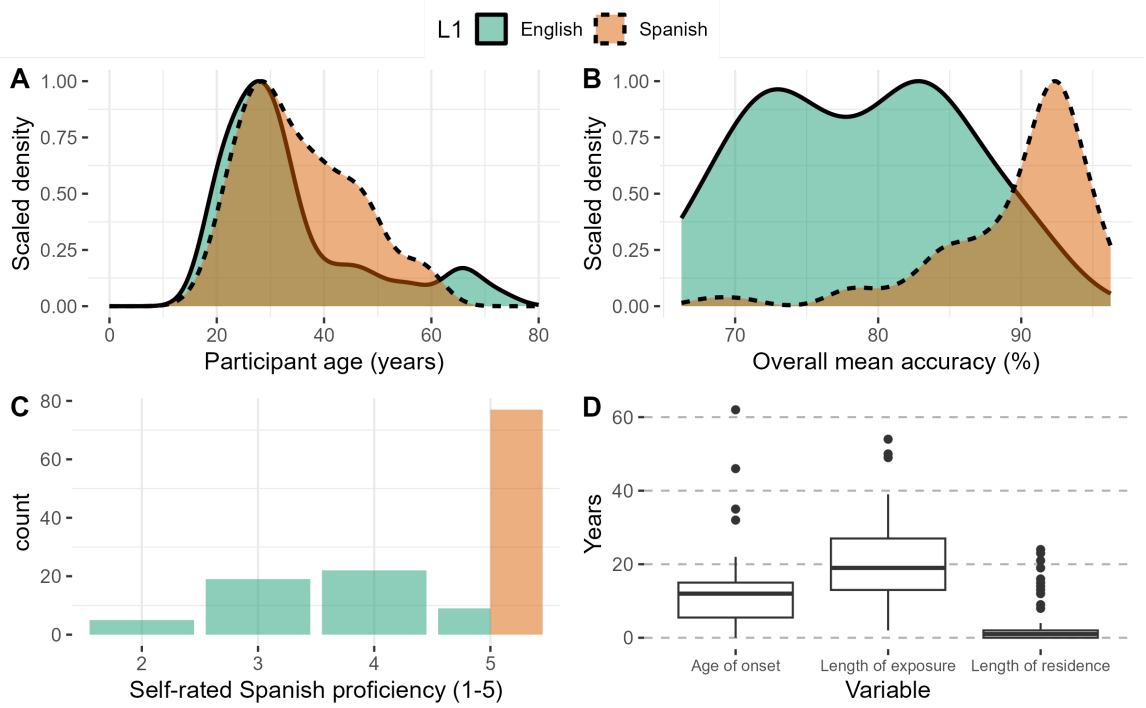


Figure 4.2: Plots of participant-level information regarding age at the time of the experiment (A), overall accuracy for all items, including fillers (B), self-rated proficiency in Spanish on a scale of 1-5 (C). D visualizes only L1 English participants' Spanish age of onset, their length of exposure to Spanish, and their length of residence in a Spanish-speaking country.

clearly that participants would receive compensation regardless of their performance or completion of the experiment, which we believe led to some participants either not responding or guessing, that is to say, that they were not truly engaging in the experimental task.

The participants whose data we analyzed (N=132) were Spanish L1 speakers (n=77, 62 from Spain) and English L1 speakers (n=55, 38 from USA). Participant age, overall mean accuracy, and self-rated proficiency in Spanish are visualized in Figure 4.2 for all participants. Figure 4.2 D visualizes Spanish Age of onset, length

of exposure, and length of residence in a Spanish-speaking country for the English L1 participants only. L1 Spanish participants were between 20 and 59 years of age at the time of testing (mean = 35.6, SD = 10.4), while L1 English participants were between 19 and 73 years of age (mean = 33.7, SD = 13.8). Overall, in all items, L1 Spanish participants had an average mean accuracy of 89.7% (SD = 5.4%), and L1 English participants had an average mean accuracy of 78.7% (SD = 7.1%). All L1 Spanish participants reported a self-rated proficiency of 5, while L1 English participants were much more variable, with 74% of participants responding either *3* or *4*, 9.5% with a *2*, and 16.4% with a *5*. English L1 participants started learning Spanish between birth and 62 years (mean = 13.0, SD = 10.6) and had been speaking Spanish between 2 and 54 years (mean = 20.7, SD = 11.5). Forty percent of L1 English participants had never lived in a Spanish-speaking country, and the maximum number of years was 24 (mean = 3.7, SD = 6.6). For additional details regarding the participants whose data we analyzed, including handedness, country of origin, and whether that participant was born or lived in Madrid, the reader is directed to supplementary materials, where a participant-level data file is available.

4.2.2 Materials

Stimuli creation

The selection of target words began by taking bisyllabic words containing a word-medial tap from the Nijmegen Corpus of Casual Spanish (Torreira & Ernestus, 2010) in which the range of values for the intensity difference (IntDiff) between the average of the surrounding vowels and the minimum during the tap was at least 5dB. This meant we were sure these words were produced in casual speech with a fair amount of variability. If two words came from the same lemma, one was randomly chosen

to be included as a target word, although this did leave homophones in the list that *could* have come from the same lemma. This process left us with 40 words we were certain were produced variably, which we supplemented with an additional 13 words containing word-medial taps. The target experimental items (n=48) for the present study were those that were produced by our speaker during recording with a token of a true tap and a token of either an approximant or a perceptual tap. Two words from the initial list were not produced with a true tap, and three were never produced as anything other than a true tap. Our real-word fillers were chosen by taking a random subset of two-syllable words from the same corpus that did not contain an intervocalic tap.

We generated our pseudowords using the Wuggy software (Keuleers & Brysbaert, 2010). We provided the software with the list of the two-syllable words from the Nijmegen Corpus of Casual Spanish that were not used as fillers in order to generate two-syllable pseudowords. We randomly chose a subset of the candidate pseudowords generated by Wuggy after confirming that they were not in the dictionary. In addition to the pseudoword fillers, we flagged generated pseudowords that contained word-medial taps. We included the same number of pseudoword distractors containing word-medial taps to prevent the presence of the tap from being a clue to the item being a word. Half of these distractors contained reduced taps (approximant or perceptual taps), and the other contained unreduced taps (true taps). The final counts in each list were 48 real target words containing a word-medial tap, 48 pseudoword distractors containing a word-medial tap, 152 real-word fillers and 152 pseudoword fillers for a total of 400 trials.

Stimuli recording and preparation

All stimuli for the auditory lexical decision task were recorded by a male L1 Spanish speaker from Madrid residing in Canada at the time of recording. He was presented with all words and pseudowords in orthographic form and recorded using a head-mounted microphone in a sound-attenuated booth. He was presented with randomized lists of either words or pseudowords. For some lists, he was asked to read out loud while clearly enunciating, while for others he was asked to read as naturally as possible in order to get variation in his productions. Our filler items included actual words carefully enunciated (n=76) and produced naturally (n=76), as well as pseudowords produced carefully (n=81) and produced as naturally as possible (n=71). This was done to further obscure the variation in the pronunciation of our target items by making both careful and natural pronunciations something that varied constantly throughout all items heard by participants.

After the recording of our speaker, we had twenty target words where the tokens produced included all three pronunciation variants from previous production research: true taps, approximant taps, and perceptual taps. As we considered the possibility that differences might lie between true/approximant taps and perceptual taps based on recent production studies (J. Y. Kim & Repiso-Puigdelliura, 2020; Perry et al., 2024), we randomly assigned 10 of these to have approximants for the reduced category while the other ten had perceptual taps for the reduced category. This subset of 20 items was used for an additional planned analysis where we modelled the three categories separately.

To create our experimental lists, we first placed all non-target items in random order. Then, target items were placed between two pseudowords so that, no matter the order of presentation, each target word would be preceded by a pseudoword.

The Condition (Reduced vs. Unreduced) of each target item was counterbalanced across lists 1 and 2, with each list having 24 reduced and 24 unreduced items. Lists 1 and 2 then had their order of presentation reversed to create lists 3 and 4. For each participant, an experimental list was randomly assigned. For a full listing of the words of each experimental list, the reader is directed to the supplementary materials. The target items are listed in the appendix.

Acoustic analysis of target items

We visualize the results of an acoustic analysis of the stimuli in Figure 4.3 along with a word frequency comparison between our target items and comparable words from the Nijmegen Corpus of Casual Spanish. This included a qualitative coding of each realized tap as either a true tap, approximant tap, or perceptual tap. After manual coding and boundary placement were completed by the first author, acoustic measurements were taken automatically using a custom script in Praat (Boersma & Weenink, 2022). The IntDiff according to *Condition* and *tap type* are visualized in Figure 4.3 A and B. Also plotted in these two panels is the density distribution of IntDiff for two-syllable words from the Nijmegen Corpus of Casual Spanish, which was taken from the data included in the supplementary materials from Perry et al. (2024). In Figure 4.3 A, we see the mode of both the 'reduced' and 'unreduced' stimuli is larger than the mode from spontaneous speech data, while in Figure 4.3 B the perceptual taps more closely align with this peak. In comparison to IntDiff, where we see differences by Condition and tap type, the overall word duration split by these variables is quite similar (visualized in Figure 4.3 C/D). We also see the lexical (unigram) frequency of our target items plotted against the lexical frequency of all two-syllable words in the Nijmegen Corpus of Casual Spanish in Figure 4.3 E,

where we see that the spread of our items covers a similar range compared to the general distribution of words of this length.

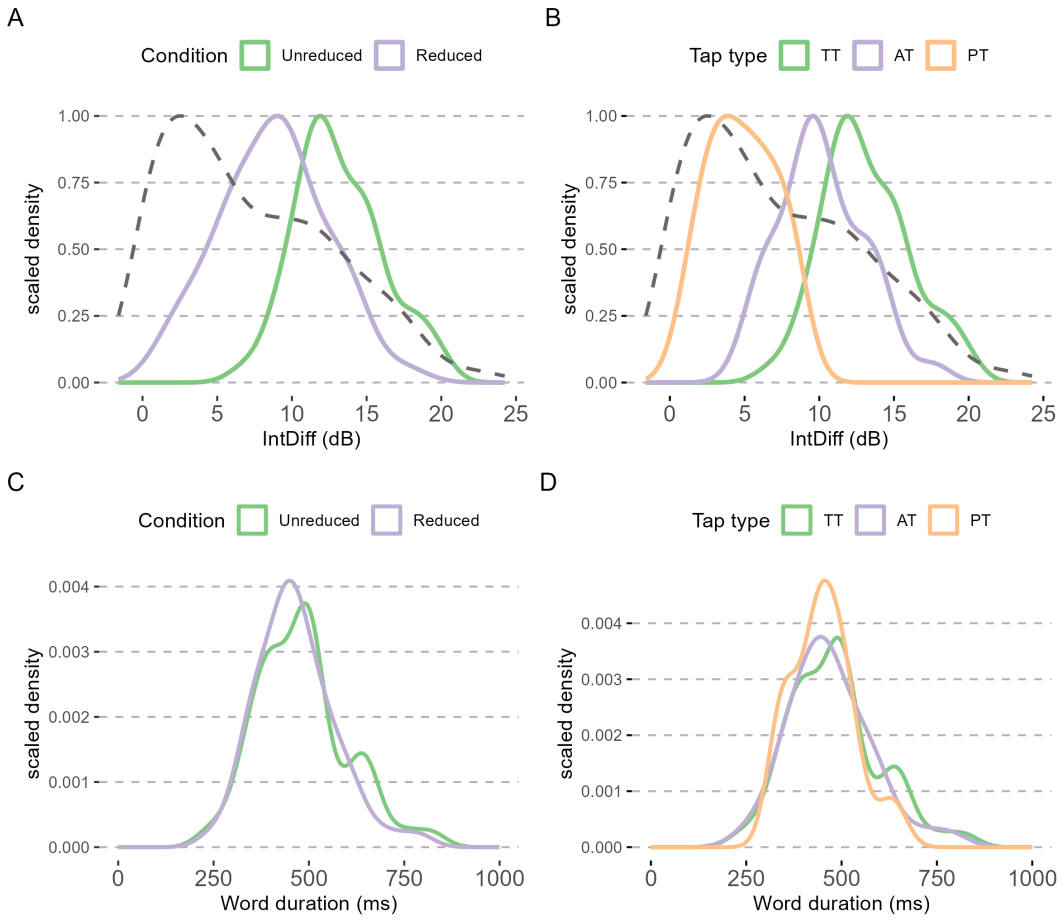


Figure 4.3: Visualizations of item-level properties across conditions, with comparisons to corpus data when applicable. Panel A is a density plot split by Condition: Reduced vs Unreduced taps. The gray dashed line in the density distribution of IntDiff from two-syllable words from the corpus data. Panel B is the IntDiff from the taps in our target stimuli split by the type of tap. TT are true taps, AT are approximant taps, and PT are perceptual taps. The gray dashed line in the density distribution of IntDiff from two-syllable words from the corpus data. Panels C and D show the total word duration of the target stimuli split by Condition and tap type, respectively.

4.2.3 Procedure

Our auditory lexical decision experiment was run on a university server using `jsPsych` (de Leeuw, 2015). Data collection began by contacting university departments in Canada, the United States, and Spain, and asking them to forward the recruitment email. Due to slow data collection, participant recruitment was moved to Prolific, where data collection was completed. Participants first completed a short questionnaire that was designed to gather basic information regarding their Spanish language background and self-reported proficiency.

4.2.4 Statistical analysis

In this section, we outline the planned analyses that were chosen before collecting the data. The exploratory analyses that were conducted after seeing the results from the analyses described below are described along with the results in Section 4.3.4. Additional information on all analyses, including the data and code used to produce them, are available in the supplementary materials.

The primary analysis involved analyzing reaction times and accuracy based on Condition (unreduced vs. reduced) and L1 (Spanish vs. English) and included all of our target words (N=48). For this first set of models, the unreduced taps were all true taps, while the reduced category was comprised of approximant taps and perceptual taps. We fit hierarchical generalized linear models in a Bayesian framework using package `brms` (Bürkner, 2017) in `R` (R Core Team, 2021). Models of reaction times for correct responses used a shifted lognormal distribution as the likelihood as in Ciaccio and Veríssimo (2022), as this more closely resembles our assumptions about human reaction times. Models of accuracy used a Bernoulli distribution with a logit link as the likelihood in order to model the probability of a correct response. Priors

for our models were based on prior predictive simulations that began with values from similar models fit to data taken from the Massive Auditory Lexical Decision database (Tucker et al., 2019).

The model predicting reaction times predicted the conditional mean by the interaction of Condition and L1 as the main predictors of interest, which were sum-coded into the model. We also included word duration in milliseconds and word frequency as control variables, which were both log-transformed. Group-level effects included varying intercepts for listener and word, as well as correlated random slopes for Condition by Participant and Condition and L1 by Word. In addition to predicting changes in the conditional mean, we also predicted changes in the σ parameter, allowing us to model non-constant variance. We allowed the variance of our lognormal distribution to vary by L1, as we were not willing to assume *a priori* that the variability of L1 and L2 listeners was equal. We also included varying intercepts on σ by Participant and Word, as we expected that the amount of variation in reaction times would vary for different items and listeners. The model predicting accuracy included the same specification as the model for the conditional mean of reaction time. Priors for our model were weakly informative in the context of lexical decision, serving to provide a small amount of regularization. The priors for the reaction time models constrained the majority of expected reaction times to a reasonable range between 250 and 4000ms, while the priors for the accuracy model served to limit the model to mostly being above chance. Both of the models described above were also fit with an alternative set of priors that provided less information to the model, with the resulting posteriors being highly similar across different sets of prior assumptions.

While we will include both L1 and L2 listeners in the same statistical models during analysis, the primary function of the Spanish L1 group is not a control group. We currently do not have empirical data on how phonetic variability of this sound

impacts spoken word recognition for *any* group of listeners, so providing empirical evidence for both groups is needed. As such, we include the interaction between the L1 group and our experimental variable to allow for the interpretation of how reduction impacts each group separately. This is not to say we won't look at comparisons between groups in our model, as we do believe that the differences in language experience between L1 and L2 listeners is a valuable source of information for learning about the general process of spoken word recognition.

Tap type analysis

As mentioned above, for twenty of our target words, our speaker produced tokens of all three tap types: true taps, approximant taps, and perceptual taps. For these twenty words, we randomly selected 10 of them to have approximant taps as the reduced variant and the other 10 to have perceptual taps. We ran new versions of the models from the previous section, swapping out Condition for tap type. We had two reasons for conducting this analysis. The first was recent work on tap production where true taps and approximant taps patterned together along the acoustic correlate of IntDiff (Perry et al., 2024), in which the more reduced perceptual taps and deletions were characterized as a separate pronunciation variant. The second is that spontaneous productions have a peak in the distribution with low IntDiff values, which our perceptual taps approximate but which our 'reduced' items do not. For these reasons, we thought that differences in lexical processing may only be apparent for highly-reduced perceptual taps.

A continuous measure of reduction

The last facet of our planned analysis was removing our categorical coding of 'Condition' for indicating whether a tap was reduced or not and replacing it with a continuous measurement of reduction as was done by Tucker (2011). For this model, the same models described in Section 4.2.4 were run, but with IntDiff replacing Condition. IntDiff was the difference in intensity in decibels between the average maximum of the two surrounding vowels and the minimum intensity during the tap. This variable was scaled and centered before modelling.

4.3 Results

In this section, we provide the interpretation of the models that were fit to examine how tap reduction affects spoken word recognition in L1 and L2 Spanish listeners. To aid in this endeavour, we provide visualizations of model-based predictions that have been back-transformed to more intuitively understandable units. We also report the conditional effects from the model summaries, which are the effects for an average item and listener. Where appropriate, we further probe interactions with the use of package `emmeans` (Lenth, 2021), and provide the 95% highest posterior density intervals for relevant contrasts as calculated by the `emmeans()` function. Many additional details regarding the interpretation of these models, including visualizations of posterior distributions, can be found in the modelling supplement.

4.3.1 Reduced vs unreduced

The initial models fit to analyze Condition contained 6,303 responses (accuracy model) and 5,483 responses (reaction time model). The model-based predictions

in Figure 4.4 are for an average item and participant. The predicted reaction times in milliseconds by L1 and Condition are visualized in Figure 4.4 A. The predicted probabilities of a correct response by L1 and Condition are visualized in Figure 4.4 B. We have a main effect of L1 for both reaction time and accuracy of spoken word recognition. L1 English speakers respond more slowly than L1 Spanish listeners ($\hat{\beta}$ = 0.12, 89% CI = 0.07–0.18) and with a lower probability of responding correctly ($\hat{\beta}$ = -0.86, 89% CI = -1.19 – -0.52). The main effect of Condition was uncertain in both models, leaving us uncertain as to whether there was an effect in terms of reaction time ($\hat{\beta}$ = 0.00, 89% CI = -0.01–0.02) or accuracy ($\hat{\beta}$ = -0.19, 89% CI = -0.47–0.11).

The interaction between Condition and L1 in the reaction time model was uncertain and centered at zero ($\hat{\beta}$ = -0.00, 89% CI = -0.03–0.02), indicating we do not have evidence that the effect of reduction was different for the two L1 groups. In contrast, we are confident that the effect of reduction on accuracy was different for the two groups ($\hat{\beta}$ = 0.50, 89% CI = 0.19–0.81), with more than 99.9% of the posterior being over zero. Further exploration of the interaction using emmeans showed that the 95% highest posterior density interval for the effect of Condition does not include zero for the Spanish L1 group (lower HPD = -0.842, upper HPD = -0.0299) while the same interval for the L1 English group straddles zero (lower HPD = -0.354, upper HPD = 0.4727). We, therefore, have evidence that words containing unreduced taps are recognized more accurately by L1 listeners and that their behaviour differs from L2 listeners who do not show this effect.

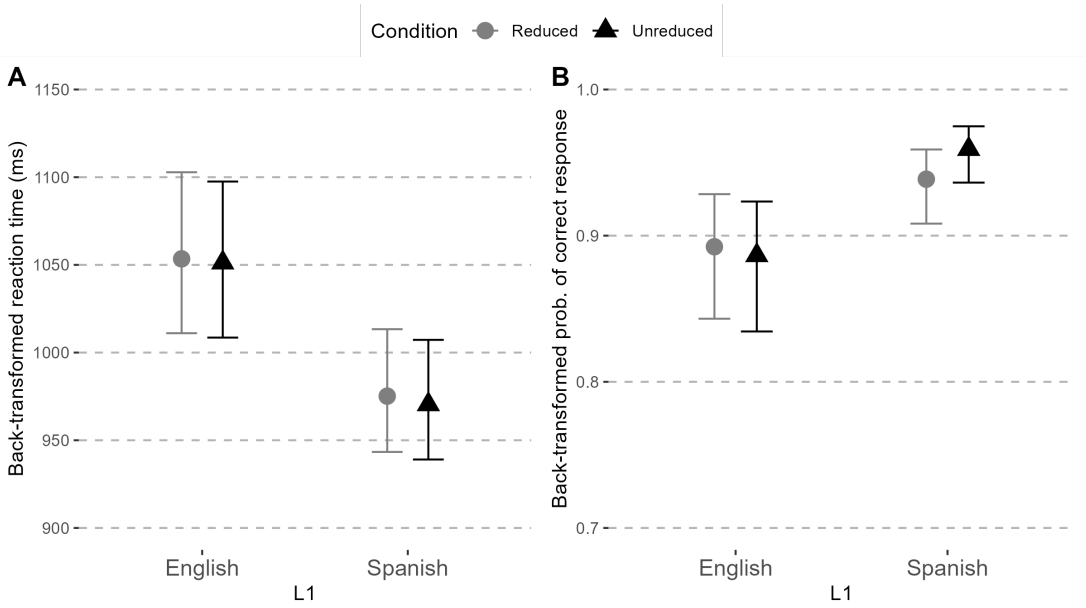


Figure 4.4: Model-based predictions for the interaction between Condition and L1 group from the full model with all items. Values visualized the average effects. Plot A reaction times back-transformed from log scale to milliseconds, visualizing model predictions for an average word and participant. Plot B shows the probability of a correct response back-transformed from log-odds to probability for an average word and participant.

4.3.2 Tap type

The models fit to analyze tap type contained 2,627 responses (accuracy model) and 2,252 responses (reaction time model). Predicted reaction times by tap type and L1 are visualized in Figure 4.5 A. In this model, as with the initial model, we have a main effect of L1, with L1 English listeners being slower to respond than L1 Spanish listeners ($\hat{\beta} = 0.19$, 89% CI = 0.10–0.28). Pairwise comparisons between the three levels of tap type, split by L1 group, are visualized in Figure 4.5 B, where we can see that we are more than 95% confident that both L1 and L2 listeners are faster to process words containing true taps as compared to perceptual taps. Based on the

interaction terms in our model, we do not have evidence that the differences between tap types differ according to L1 group ($\hat{\beta}_{Tap\_type1:L11}$ = -0.03, 89% CI = -0.12–0.06, $\hat{\beta}_{Tap\_type2:L11}$ = 0.02, 89% CI = -0.09–0.13).

Predicted probabilities of a correct response by tap type and L1 group are visualized in Figure 4.6 A. Echoing the results of the model with all items, we have a main effect of L1, with L1 English listeners having a lower probability of a correct response overall compared to L1 Spanish listeners ($\hat{\beta}$ = -0.80, 89% CI = -1.23– -0.36).

For this model, we did have some weaker evidence that the differences between tap type may be different for the two L1 groups ($\hat{\beta}_{Tap\_type1:L11}$ = -0.56, 89% CI = -1.14– -0.06, $\hat{\beta}_{Tap\_type2:L11}$ = 0.66, 89% CI = -0.08–1.41). This can be seen in the pairwise differences between the three tap types visualized in Figure 4.6 B, split by L1 group. We are more than 95% confident that there is a difference between true taps and perceptual taps in terms of recognition accuracy by both L1 and L2 listeners, and that L1 English listeners are more heavily inhibited by the presence of a perceptual tap compared to an approximant tap.

4.3.3 A continuous measurement of reduction

As part of our planned analysis, we replaced Condition with a continuous measurement of reduction: IntDiff, based on Tucker (2011), and fit the model to the full data set. The main effect of IntDiff was uncertain and had the bulk of the posterior near zero for both the reaction time model ($\hat{\beta}$ = -0.00, 89% CI = -0.02–0.01) and the accuracy model ($\hat{\beta}$ = -0.02, 89% CI = -0.41–0.38). Additionally, the interaction of IntDiff with L1 group was uncertain in both the reaction time model ($\hat{\beta}$ = -0.00, 89% CI = -0.02–0.01) and the accuracy model ($\hat{\beta}$ = -0.13, 89% CI = -0.36–0.09).

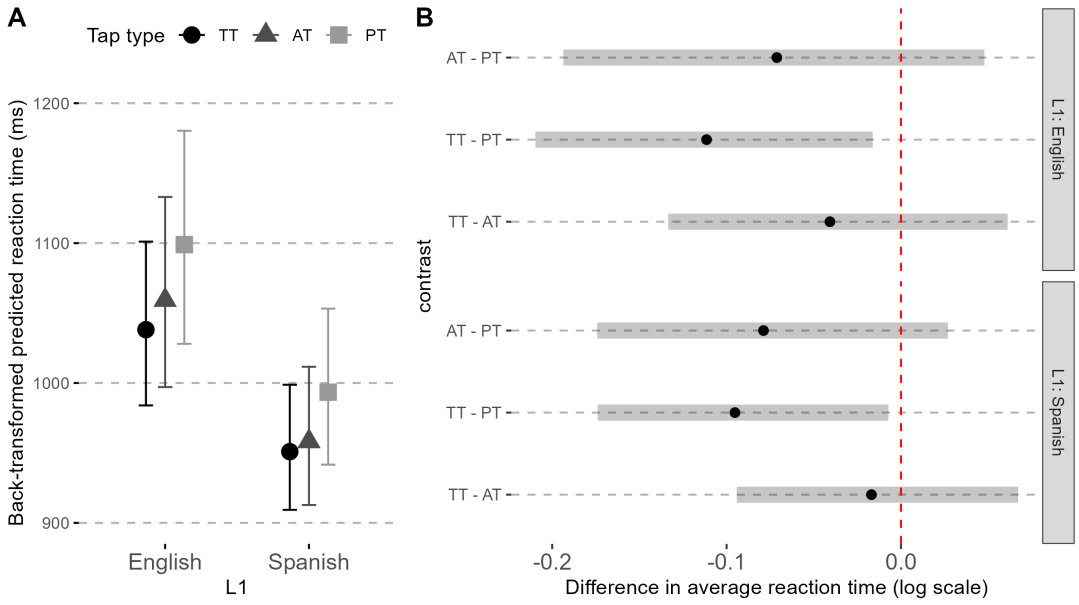


Figure 4.5: Model-based predictions for the interaction between Condition and L1 group from the subset model with the 20 items. Values visualized are model predictions for an average item and participant. Plot (a) shows average reaction times back-transformed from log scale to milliseconds. Plot (b) shows the probability of a correct response back-transformed from log-odds to probability. In the legend, TT is True tap, AT is approximant tap, and PT is perceptual tap.

4.3.4 Exploratory analysis

After conducting the planned stages of our analysis, we were left with two potentially conflicting pieces of information that we attempted to delve further into with exploratory models. The first was that we did not observe an effect of IntDiff, contrary to our expectations. The second was that we did observe an effect of tap type in a subset of our items, and our acoustic analysis indicated that these taps differed considerably in terms of IntDiff - an acoustic correlate that has been demonstrated to be relevant to the perception of the English flap (Warner et al., 2009), which is a similar sound in many respects. We fit two additional sets of models in order to

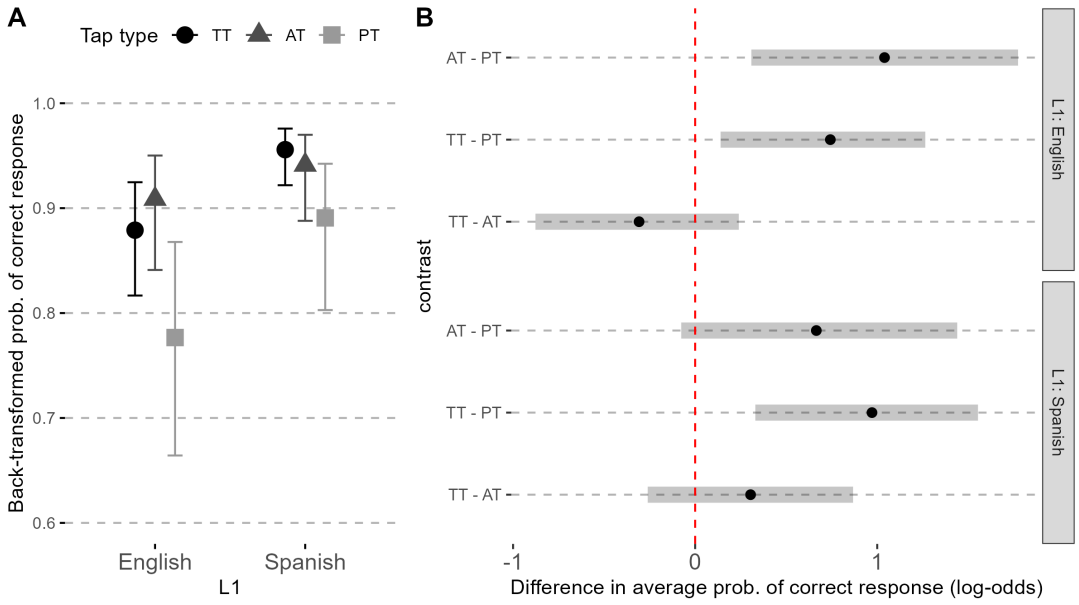


Figure 4.6: A: Model-based predictions for the probability of a correct response between tap type and L1 from the subset model with the 20 items. Values visualized are marginal effects back-transformed onto the probability scale. B: Pairwise comparisons between the three levels of tap type for each L1 group. In the legend, TT is True tap, AT is approximant tap, and PT is perceptual tap.

explore these results further and attempt to provide additional focus for the direction of future studies.

The first set of models explores the effect of what we will refer to as an 'expected production', which can be thought of as a continuous variable related to variant frequency. The expected production is based on the measurements of IntDiff from spontaneous speech data and is calculated on a per-word basis. The idea for this comes from the previous studies that have found effects of the frequency of occurrence for pronunciation variants on word recognition (Bürki et al., 2018; Pitt, 2009). We thought that if raw IntDiff values were not associated with changes in lexical access, what may influence the recognition of these words is how frequent or typical the

pronunciation of the tap was for each word in our stimuli. To evaluate this idea, we compared the IntDiff value of each tap in our stimuli to measures of central tendency for IntDiff from the same word in spontaneous speech using the publicly-available corpus data from Perry et al. (2024).

In Figure 4.7, we see the distribution of IntDiff values from the Nijmegen Corpus of Causal Spanish for two words from our stimuli: 'caro' (expensive) and 'era' (it was). We also visualize three different measures of central tendency along with the density distributions: the mean IntDiff, the median IntDiff, and the mode IntDiff as estimated using the meanshift method as implemented in package `modeest` (Poncet, 2019). For relatively symmetrical distributions, like the one for the word 'caro', all three measures are similar. For skewed distributions, the three different measures can diverge, as is the case for the word 'era'.

We calculated the absolute difference in decibels between the IntDiff in our stimuli and the mean, median, and mode for their respective words. We then fit three models predicting reaction time and three predicting accuracy, replacing Condition with the measure of distance from the central tendency measure from the corpus. We included only words for which we had at least 10 measurements from the corpus data, which was 21 distinct words that are naturally skewed toward higher-frequency items (see Appendix for a visualization of word frequency in different subsets). The accuracy models were fit to a total of 2,758 observations and the reaction time models to 2,419 observations. As these models were exploratory in nature, we did not include varying slopes.

The effect of distance from an expected production on reaction times and accuracy according to which measure of central tendency was used are presented in Figure 4.8. Across all three measures of central tendency, both English and Spanish L1 listeners were slower to recognize productions that were farther away from the

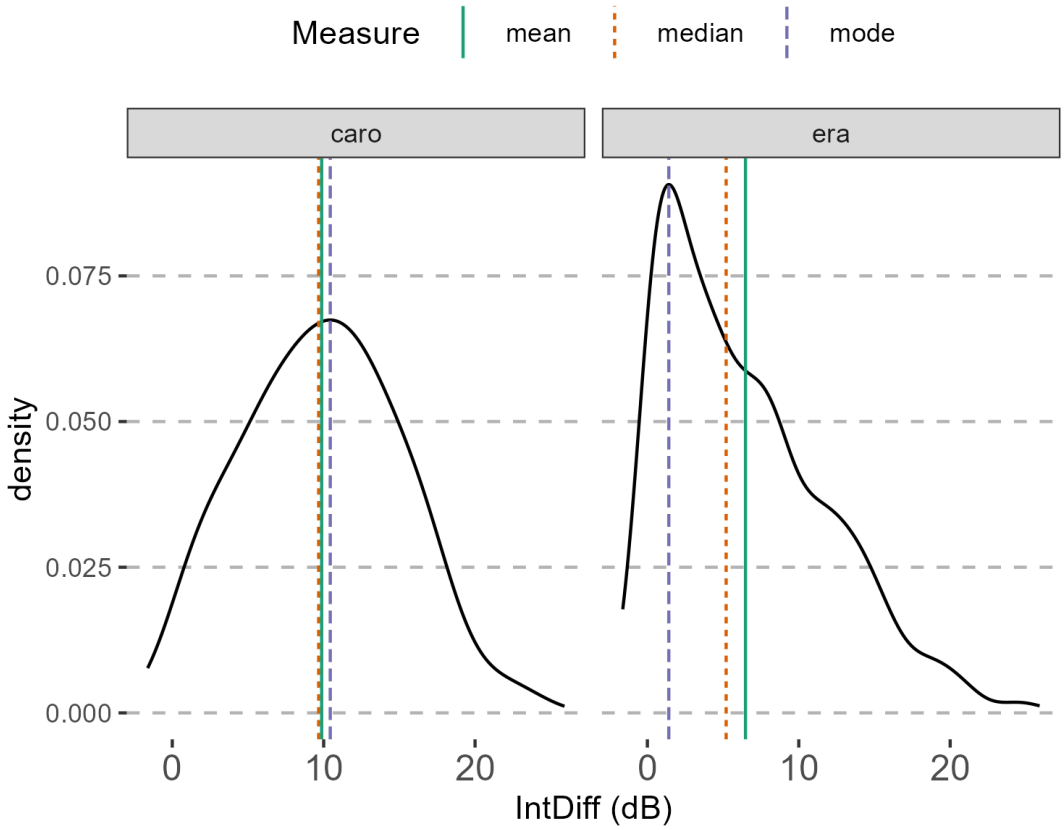


Figure 4.7: Two examples of the distribution of IntDiff values from intervocalic taps in the Nijmegen Corpus of Casual Spanish. The mean, median, and estimated model are plotted on the density distribution by color and line type.

expected production from the corpus. For each of the three models, the main effect of distance from an expected production had more than 99.9% of the posterior on the positive side of zero. Split by L1 group, the 95% highest posterior intervals estimated by `emmeans()` did not include zero for either L1 English or L1 Spanish listeners. The credible intervals for the interaction terms between the distance measure and the L1 group in the three models were also highly uncertain, indicating that we do not have evidence that this effect differs across our two groups.

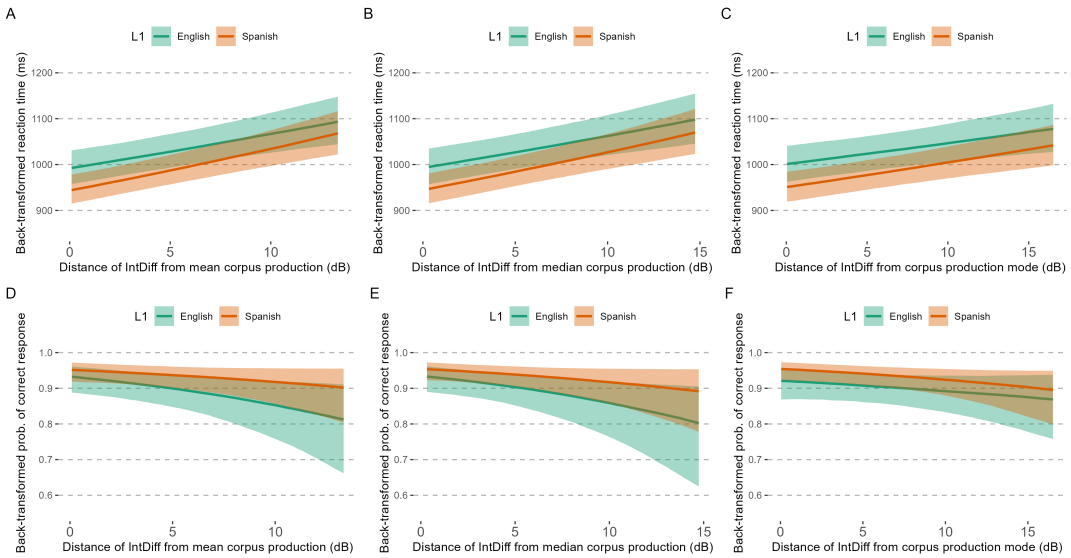


Figure 4.8: The predicted effect of distance from an expected production on reaction times (A, B, C) and accuracy (D, E, F) for an average item and listener. Panels A and D show the effect of distance from the mean IntDiff, panels B and E from the median IntDiff, and panels C and F from the mode.

In contrast, the evidence for the effect of these distance measures on accuracy is weaker. The posteriors for the main effect of the distance from an expected production do have a small amount of overlap with zero. Additionally, when we split by L1 group and estimate the effect of distance from an expected production separately, almost all of the 95% HPD intervals include zero, with the exception of the distance as calculated from the median for the English L1 group (lower HPD = -0.102, upper HPD = -0.003).

After this set of models, we had some evidence that listeners responded to expected productions of a word faster. This provided a potential explanation for why we did not see an effect of reduction in our first set of models, which included Condition. If higher frequency pronunciation variants in specific words facilitate lexical access, it may have been that our speaker produced many expected productions that

we used for our reduced stimuli. Supporting this idea was the fact that, during recording, we were only able to elicit an approximant tap or a perceptual tap from our speaker, but not both, for the majority of words (28/48). This would mean that there was a confound in our original sample of 48 items, as our speaker was producing more expected productions for all words, including what we originally referred to as reduced variants. If this were the case, our random assignment of tap type for the 20 words that were produced with a wider range of variability would have also blocked the influence of the expected production on our stimuli. To evaluate this idea, we fit another set of models that used raw IntDiff values of our stimuli, this time only using the subset of items from our tap type analysis, replacing the categorical variable tap type with the continuous measurement of IntDiff.

The predicted reaction times for an average participant and item are visualized in Figure 4.9 A, and the predicted probabilities of accuracy in Figure 4.9 B. In our subset of items, we do find a main effect of IntDiff, with more stop-like taps being recognized faster ($\hat{\beta}$ = -0.03, 89% CI = -0.05–0.02), with more than 99.9% of the posterior being negative. The model is also more than 95% confident of the effect being present for both L1 English and L1 Spanish listeners. The effect of IntDiff is also present in the accuracy model, with words containing more stop-like taps being recognized more accurately ($\hat{\beta}$ = 0.27, 89% CI = 0.16–0.38), with more than 99.9% of the posterior being positive. Like the reaction time model, the model is also more than 95% confident of the presence of this effect for the two L1 groups separately.

4.4 Discussion

The present study investigated how phonetic reduction in word-medial alveolar taps impacts spoken word recognition in Spanish for L1 and L2 listeners using an auditory

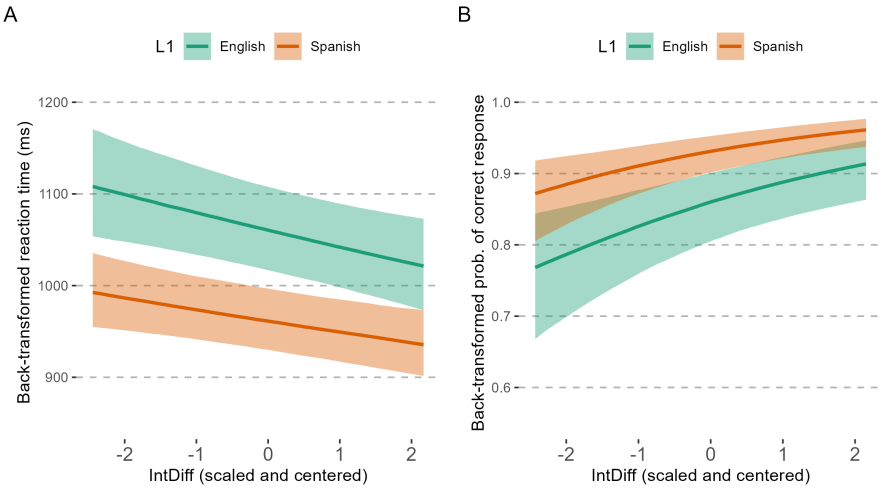


Figure 4.9: The predicted effect of IntDiff for an average item and listener on reaction times (A) and probability of a correct response (B) from the exploratory follow-up analysis on the subset of items used for the tap type analysis. Predictions for L1 English listeners are plotted in green and L1 Spanish listeners are plotted in orange.

lexical decision task. We find that whether or not this phonetic variability impacts spoken word recognition in Spanish partially depends on how we operationalize phonetic variability. We also find only limited evidence showing that the effect of this variability differs between L1 and L2 listeners. In addition to our planned analyses, we also communicate the results of an exploratory analysis that complements the findings of the planned analyses and provides direction for future work.

Our first research question concerned how reduction impacts reaction times and accuracy for both L1 and L2 listeners. In our initial analysis, we coded all taps that were not strictly stop-like the same – as reduced variants of the tap. Under this definition of reduction, we found no evidence for an effect of reduction on reaction times for L1 or L2 listeners or for the accuracy of our L2 listeners. We did find an effect of reduction for our L1 Spanish group, who showed a small inhibitory effect of reduction on accuracy. This is consistent with previous studies that have found an

inhibitory effect of reduction (e.g., Ernestus & Baayen, 2007; Tucker, 2011). While these findings are consistent with an inhibitory effect of reduction, we believe it is useful to contextualize these findings with spontaneous and semi-spontaneous speech data, where true taps may be a less common pronunciation variant overall (Bradley & Willis, 2012; Perry et al., 2024; Willis & Bradley, 2008). That information, as well as the fact that our 'reduced' category appeared less reduced on average than we see in spontaneous productions, leads us to propose that we could interpret these same results not as an inhibitory effect of reduction but as a facilitatory effect of hyper-articulation, which the L1 listeners were able to benefit from, in contrast to the L2 listeners.

For our second research question, we wanted to see if the qualitative coding scheme employed by production studies of the Spanish tap provided a different interpretation of our participant's behaviour. In recording our speaker, we had the option of choosing what variant (approximant vs. perceptual) would be used for a 'reduced' tap for less than half of our target words. When we did have a choice, the reduced variant was chosen at random. The model that analyzed this subset of items eschewed the binary coding from the first analysis in favour of using the qualitative coding scheme that has been used by several production studies looking into the Spanish tap (Bradley & Willis, 2012; Henriksen, 2015; J. Y. Kim & Repiso-Puigdelliura, 2020; Willis & Bradley, 2008). Under this view of phonetic variability, we see an effect of reduction on reaction times for both L1 and L2 listeners, but only between the two ends of the spectrum (i.e., true taps and perceptual taps), and no convincing evidence that the differences between tap types were different for the two groups. For the accuracy model looking at tap type, we see a different qualitative pattern emerge between the L1 and L2 groups, with some weak evidence from the interactions that they might behave differently. L1 Spanish listeners showed

the expected ordering of true tap > approximant tap > perceptual tap in terms of predicted accuracy, and we were confident of the estimated difference between true taps and perceptual taps. L2 English listeners instead were predicted to be most accurate on approximant taps, then true taps, and then perceptual taps. A possible explanation for this is related to variant frequency, as approximant taps are more common than true taps (Perry et al., 2024).

Both of these analyses, taken together, indicate that L1 listeners are able to process true taps quicker than perceptual taps, even though the reduced variant may be more frequent (Bradley & Willis, 2012; Perry et al., 2024; Willis & Bradley, 2008), and that there may be differences between L1 and L2 listeners regarding the different pronunciation variants. This may stem from the fact that the lexical representations of L1 and L2 listeners are most likely based on input that is both qualitatively and quantitatively different. Our findings of a facilitatory effect for the true tap are consistent with studies that have found a benefit for the 'canonical' pronunciation variant (e.g., Pitt et al., 2011). The fact that we see it only for L1 listeners is consistent with claims that lexical representations formed during early childhood may have been based on input that contains more 'canonical' pronunciations than typical adult-directed speech (Dilley et al., 2019). It is also consistent with the proposal from Sumner (2013), who posited that when words are stored in multiple forms within long-term memory, canonical forms may be stored with increased acoustic detail, leading to faster lexical processing. It could be that L2 listeners do not store this increased acoustic detail or that aspects of their shared bilingual sound system prevent them from exploiting this detail for faster lexical retrieval.

We have seen that our decisions on the categorical coding of phonetic variability determine what results we see in our items. Our final research question was related to a desire to bypass this potentially false categorization of phonetic variability by

predicting reaction times and accuracy using a continuous measurement of tap reduction. We did not find evidence that IntDiff predicts either aspect of spoken word recognition in Spanish when considering our full set of words, but we did see the effect of IntDiff in the subset of our items that were produced with more variability by our speaker and used for our tap type analysis, with more carefully articulated tap being recognized faster and more accurately for both L1 and L2 listeners. As we see it, there are two potential explanations for why we see this effect in the subset of words but not for all items, and that these explanations are not mutually exclusive.

The first explanation for these differences appeals to the experimental control of a potentially confounding variable. Randomizing which words were presented to participants containing a perceptual vs. approximant tap may have controlled for variant frequency. For the majority of our items, the 'reduced' variant was the only variant produced by our speaker, other than a true tap, for that specific word. This means that variant frequency may have impacted our study design indirectly and that the randomization of variant type in this subset blocked the effect from confounding with IntDiff. The second explanation has to do with the subset of words themselves. As these words were produced in a wider amount of variability by our speaker, they may share some property that makes them ideal for showing an effect of IntDiff compared to other words.

The last part of our exploratory analysis was analogous to an analysis of variant frequency but with a continuous acoustic correlate. Looking at the production of taps through the lens of IntDiff, we looked at the distribution of this acoustic correlate of tap variation in spontaneous speech and calculated what we refer to as the expected production of the tap for each word. This is the value of IntDiff that is estimated to be the most common (mode), or expected based on some other measure of central tendency (mean or median). We find evidence that increased distance

from this expected production is associated with slower reaction times but only weak evidence that it is associated with decreased accuracy. The predicted effects of these variables are consistent with research that has found an advantage for processing more frequent variants of words (Bürki et al., 2018; Pitt, 2009). The presence of this effect also provides additional context to our analysis of raw IntDiff values from both the planned and exploratory models, indicating that lower IntDiff values may not predict changes in the process of spoken word recognition by themselves, as those values may be close to expected productions, which facilitates word recognition.

Taken together, our planned and exploratory analysis indicates that reduction impacts the speed and accuracy of spoken word recognition only after experimentally controlling for expected productions. Both of these effects being present simultaneously support the idea that models of spoken word recognition need to incorporate detailed acoustic information that can be stored in long-term memory in relation to specific lexical entries (e.g., Pierrehumbert, 2002). A limitation of the present study is that it was not initially designed to test the effect of high-frequency or expected productions of words, but we believe these results warrant further attention. A study targeting expected productions would start with a corpus analysis to find a set of words with a large range of expected productions and then systematically vary the IntDiff presented to participants. In such a design, the distance from an expected production could be changed to a non-linear interaction between the value of IntDiff in an expected production and the IntDiff value of the stimuli, allowing for the distance to be evaluated visually and in both directions.

We turn now to a more focused discussion of the results of our L2 listeners. While we found some specific differences between this group and the L1 listeners, which we have discussed above, for many of the effects we observe similar patterns for both L1 and L2 listeners. There are a few potential reasons for this, one of which

is that the vast majority of our L2 listeners are L1 speakers of North American English and, therefore, have the alveolar flap as a position-sensitive allophone in their L1 (Warner & Tucker, 2011). In models of L2 speech learning like the revised Speech Learning Model (Flege & Bohn, 2021) and the Speech Learning Model (Flege, 1995), the position-sensitive allophone is appealed to as the phonetic entity that interacts between the two languages. While the similarity between to two sounds should make learning the new sound difficult, it is possible that the English flap and the Spanish tap are similar enough that any automatic perceptual routines English speakers have in their L1 can be used in the L2 without creating problems for word recognition. As flap reduction in English is ubiquitous (Warner & Tucker, 2011), any such perceptual routine would already excel at processing and recognizing reduced forms. Any follow-up study attempting to replicate these findings would benefit from another L2 Spanish listener group whose L1 does not contain any segment closely resembling the Spanish alveolar tap. Other aspects to consider would be providing a detailed language background questionnaire to L2 listeners. As we argue here that variation in language backgrounds between our L1 and L2 listeners is responsible for some of the differences that we see, we should also see variation within the L2 group based on similar differences, such as the age of onset of acquisition.

4.5 Conclusion

The present study has found evidence that tap reduction impacts both L1 and L2 spoken word recognition in Spanish as long as stimuli are sufficiently experimentally controlled. We find inconsistent evidence of differences in our effects for L1 and L2 listeners. We argue that the differences that do exist may be due to qualitative and quantitative differences in the input that result in different long-term representations

of words but that English L1 listeners may be exploiting their L1 perceptual routines to process phonetic variability in their L2. We also find a processing advantage for pronunciation variants that are more likely to occur in spontaneous speech, indicating the lexical representations are influenced by phonetic variability.

4.6 Appendix

Target words:

barón, cara, caro, cera, cero, clara, claros, coro, cuero, cura, dará, dirá, dura, era, flores, foro, fueran, gira, giran, hora, iris, juro, llora, loro, mira, moral, moro, muera, muro, nariz, oral, oro, oros, paran, pared, pira, pura, quiera, rara, rieron, será, tira, tirón, tiros, toro, vara, verás, virus

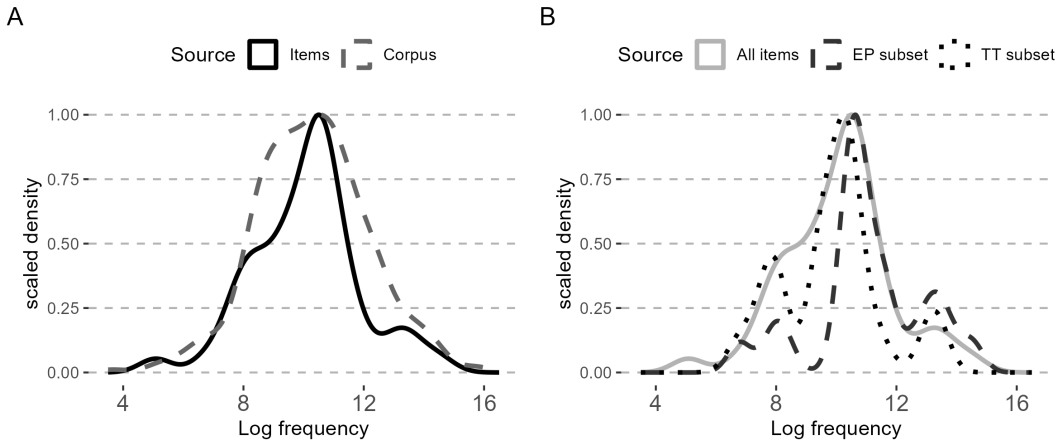


Figure 4.10: Panels A and B show estimated density distributions of word frequency for different sets of words. Panel A shows a comparison between our target words and all two-syllable words containing taps from the Nijmegen Corpus of Casual Spanish. Panel B shows a comparison between all of our items and the two subsets of items used for different analyses in the paper: the tap type analysis and the analysis of the distance from an expected production.

Chapter 5

General discussion and conclusions

The primary goals of this dissertation were to study patterns of phonetic reduction in spontaneous productions of the Spanish alveolar tap and to experimentally investigate how such patterns impact the process of spoken word recognition in L1 and L2 Spanish listeners. The production side of this dissertation involved a corpus analysis of spontaneous speech that qualitatively documented the variability through spectrographic analysis and quantified this variability by measuring the acoustic correlates of duration and the intensity difference (IntDiff) between the tap and the surrounding vowels. In our quantitative analysis, we were interested in how we could predict variability in our acoustic correlates with a combination of phonetic variables related to articulation (e.g., speech rate, phonetic environment, surrounding lexical stress) and to variables at the lexical level (e.g., unigram and bigram frequencies, content vs. function word).

After documenting the kind of variability that Spanish speakers encounter in their daily interactions, we designed an auditory lexical decision experiment to investigate how that kind of variability impacts spoken word recognition. We collected

this lexical decision data from L1 Spanish listeners and L2 Spanish listeners who spoke English as an L1. Our planned analyses included comparing canonical 'true taps' to other reduced variants, comparing all variants separately in a subset of our items, and replacing categorical measures of phonetic variation with the continuous measurement of IntDiff.

In the present chapter, I summarize the main findings of the corpus analysis and the lexical decision experiment, after which we discuss how these findings inform theories of phonetic variability in production and how it impacts spoken word recognition. We will also discuss the methodological considerations raised by our studies, including current practices in analyzing large amounts of acoustic data and interpreting regression coefficients from models fit to observational data. Throughout this chapter, we will highlight the limitations of the current work and discuss how future work may build on the research from the present dissertation.

A secondary goal of the present dissertation was to set the stage for future work investigating the L2 production of the Spanish tap in spontaneous speech. Studies of spontaneous speech and the patterns of variability we find are relevant to studies of L2 production because this type of speech is likely to constitute the majority of the input that L2 learners receive during naturalistic L2 acquisition. Furthermore, many studies of L2 segmental production often analyze data in a way that is at odds with the variability present in spontaneous speech. The implications of our studies on future work on the Spanish tap in L2 contexts are discussed in depth in this chapter, including recommendations for future research.

5.1 Summary of findings

The corpus analysis from the present dissertation (chapters 2 & 3, published as Perry et al., 2023, 2024, respectively) uncovered patterns of phonetic variability in the production of Spanish taps that have not previously been reported. Our duration modelling was limited to a subset of our hand-corrected data set due to methodological limitations. Namely, we learned that we could not rely on automatic measurements of duration, and those were the only measurements of duration present for 90% of our data. Thus, our findings of decreased duration as speech rate increases, shorter taps in words that are predictable based on the following word, and longer taps when following or preceding an /i/ vowel can only be generalized to taps that have a visible occlusion on a spectrogram that a phonetician could consistently measure. This hurts the generalizability of our results, as this was approximately half of our data, and we know that this missing data is more likely to be missing at certain frequencies and speech rates based on our subsequent analyses.

As part of modelling the duration of our taps, a methodological contribution was also reported by Perry et al. (2023). We compared several methods of measuring tap duration to hand-corrected boundaries. These automated methods included force-aligned boundaries and two different methods for approximating the duration of a segment based on intensity information. The automated forms of measuring duration did a poor job of approximating hand measurements, with two of the methods drastically overestimating tap duration. Versions of our model fit to these automated measurements differed substantially in their estimated effects. Based on our findings, we were forced to make our inferences about variability in tap duration based only on hand-corrected data.

In our modelling of the intensity difference of our taps (Perry et al., 2024), we

were forced to abandon linear models for finite mixture models after the former could not account for the observed data. The two distributions we estimated in our model, which we referred to as the reduced and unreduced distributions, map closely to our hand-corrected data in a way that indicates 'reduced' taps do not have a visible presence on a spectrogram while 'unreduced' taps do.

We found that lexical frequency was related to changes in the IntDiff of Spanish taps and that speech rate and the quality of the surrounding vowels were associated with changes in both duration and IntDiff. Furthermore, our modelling of IntDiff required two distributions to properly account for the data, and we found that more of our predictors were related to changes in the categorical alternation between the two distributions than to gradient changes within them.

We found patterns of variability in IntDiff that, in some ways, were similar to the factors that affected the tap duration when a visible occlusion was present. This is the case for the estimated effects of speech rate and the surrounding vowels. We also had several predictors with non-negligible effects that predicted changes in IntDiff, but that had highly uncertain effects on duration. Lexical frequency predicted gradient reduction within unreduced taps in IntDiff, but we found an uncertain effect on the duration of taps with visible occlusions.

Based on the set of phonetic and lexical predictors we used to model IntDiff, we were unable to account for the bimodal nature of the data without the assumption of two latent categories, which we interpret as two pronunciation variants of the tap. This assumption drastically improved the predictive performance of our model of tap variability. In interpreting our models, we found that phonetic variables like speech rate and surrounding vowels were associated with changes in the probability of realizing a reduced tap and gradient changes within each pronunciation variant. While these patterns of variability in IntDiff had not previously been documented, the

qualitative side of the corpus analysis revealed average rates of different pronunciation variants to be relatively similar to previous work on the Spanish tap in other varieties. This suggests that taps across different varieties of Spanish may be more similar than they are different and that the findings of the present dissertation may apply to other varieties as well.

Complementing the analysis of spontaneous speech in chapters 2 and 3, the fourth chapter investigated how the types of tap variability we documented in our corpus analysis impacted spoken word recognition using an auditory lexical decision experiment. We gathered this experimental data from both first-language (L1) and second-language (L2) speakers of Spanish, as research on reduction for non-native speakers is an under-researched area of study. Our experimental stimuli were all real, two-syllable Spanish words containing an intervocalic, word-medial alveolar tap. Our initial analysis compared taps with a stop closure to those without, while a second analysis on a subset of our items used the qualitative coding from previous production studies. We also explored a continuous measurement of reduction, using the IntDiff of the taps in our stimuli to try to predict responses.

Results from our planned analyses indicate that reduction is likely an inhibitory factor in Spanish spoken word recognition when the words are presented in isolation. We see differences between true and perceptual taps, the most and least reduced versions, for both L1 and L2 listeners. In most of our models, we do not have evidence that L2 listeners differed from L1 listeners regarding the effect of reduction on how long it takes to recognize a word correctly. Still, we find some differences in how reduction impacts accuracy, with L1 listeners benefiting more from unreduced productions. When we replaced our reduced/unreduced variable with a continuous measure of IntDiff, we saw no effect on recognition in either group.

In our exploratory analysis, we used the data from our corpus analysis in the

previous chapters to look at the distribution of the intensity difference for the words used as stimuli in our lexical decision experiment. We then calculated the distance between the taps in our stimuli and the 'expected production' from the corpus. We find that L1 and L2 listeners are slower to recognize words when they are produced far away from an average or typical production. This indicated that our initial set of 48 items may have been influenced by this, which prevented us from seeing the effect of IntDiff. A model of continuous reduction on the more controlled subset of words that were produced more variably by our speaker showed that continuous reduction is an inhibitory factor.

5.2 Probabilistic variation in speech production

The phonetic variability we documented in the Spanish taps is consistent with H&H Theory along a number of important dimensions, but perhaps the most important one is that we see modulations in the acoustic signal when there are changes in the nature of the signal-independent information that is present. This signal-independent information includes lexical frequency and the predictability of the word based on surrounding words. In this section, these findings are discussed in terms of current theoretical discussions regarding phonetic variability in speech production and how it relates to phonological theory.

In our model of IntDiff variation (Perry et al., 2024), we have an effect of lexical frequency on gradient changes in unreduced taps, with taps being less stop-like at higher frequencies. In duration, however, we see either similar or longer taps as frequency increases (Perry et al., 2023). If we accept both these things as true, then it would be an example of phonetic reduction that is not due to articulatory undershoot, contrary to the claim that durational shortening causes intensity changes

(e.g., Cohen Priva & Gleason, 2020). If speakers put less effort into articulation even though they are not producing the sound more quickly, we could take this as evidence that the economy of effort is sensitive to changes in signal-independent information, that the implicit knowledge that lexical discrimination will be easier can affect intensity directly, and that this is not undershoot in the classical sense.

Articulatory undershoot, as first described by Lindblom (1963), is a phenomenon whereby the acoustic properties of a sound are changed at increased speech rates due to the speaker having less time to reach an articulatory target. It assumes that the same target is aimed for but not reached due to temporal constraints, as the speaker would have to articulate faster to reach the target in less time. In contrast, tap reduction based on increased frequency may not be due to temporal constraints. While this is not consistent with a single target, it has been argued that a 'window' of possible articulations may be a better analogy to viewing coarticulation (Keating et al., 1990) and that this may be a better lens through which to view reduction (Warner & Tucker, 2011). If there is a window containing an infinite amount of possible articulations, then all you would need to incorporate our findings into this claim is a mechanism where lexical frequency can modulate which parts of the window are most probable.

There are two main problems with the evidence for this claim as it stands. First, the reported effects of frequency on duration are from a modest amount of data that was hand-measured, while the effect on the IntDiff on 'unreduced' taps is estimated using an order of magnitude more data and based on a model-based inference of what an unreduced tap is. While Perry et al. (2024) showed that the 'unreduced' distribution from the model was visually similar to the taps for which we measured duration (Perry et al., 2023), we don't know for sure that we are comparing the same types of taps to each other. We did not explicitly model duration and IntDiff

together at any point due to methodological constraints on measuring duration. A future study of Spanish taps could be set up to test the relationships between frequency, duration, and IntDiff explicitly. However, it will have to find a way to deal with the missing duration information for a substantial portion of taps.

While we found lexical effects on tap production, there is some evidence that phonetic variables may be more important in predicting tap reduction. Speech rate had the largest effect on tap production, both in terms of intensity and duration (Perry et al., 2023, 2024). This was unsurprising, as it is perhaps the most robust predictor of reduction overall (Ernestus, 2014). In interpreting this effect and the effect of the surrounding vowels, we proposed an explanation based on articulatory factors. However, we did not measure articulation in our studies, and some acoustic measurements map better to articulation than others (e.g., see Noiray et al., 2014, and references therein). Future articulatory work could easily corroborate or refute certain claims, such as our claim of two pronunciation variants of the tap in production. I would consider this claim falsified, for example, if a smooth and continuous increase in the minimum distance of the tongue blade from the alveolar ridge during tap production leads to a bimodal distribution of IntDiff. This would indicate that there are not two variants, and we observed an artifact stemming from acoustic measurement.

I have just discussed our lexical frequency effect and physiological effects as if they were separate sources of variation, and indeed, our study did control for various phonetic constructs that are more closely related to articulation, such as speech rate and the surrounding vowels. After controlling for these factors, we still see an effect of lexical frequency on tap production. Exactly how frequency impacts the cognitive processes of speech production is unclear. Still, in exploring the effects of cognitive factors on speech production, it is important to acknowledge that they must affect

acoustics *through* articulation (Tomaschek & Tucker, 2023), so controlling for enough articulatory variables should always leave the effects of cognitive variables at zero.

Lexical frequency effects are one of psycholinguistics' most studied and consistent effects, but what mechanisms lead to these effects is still under debate. Older models assume that the architecture of frequency effects boils down to counting, with frequency effects accounted for by different resting activation levels (e.g., McClelland & Rumelhart, 1981; Pierrehumbert, 2002; Van Heuven et al., 1998). In contrast, Baayen (2010) argues that lexical frequency effects arise instead from the cognitive process of mapping form to meaning, demonstrating that the effect of frequency on visual word recognition can be broken down into several distinct facets. Many more recent investigations have replaced frequency measures with variables taken from discriminative models of the lexicon, showing frequency-like effects that are based on a morphological model of mapping form to meaning (Schmitz et al., 2021; Stein & Plag, 2021; Tomaschek et al., 2021). As far as I know, no work on variables derived from discriminative modelling has been applied to Spanish production data, so expanding on the present work with this approach could offer alternative interpretations of our frequency effect.

While the effect of frequency and morphology on speech production has typically been concerned with cognitive processes (Bell et al., 2009; Tomaschek & Tucker, 2023), we have to consider that frequency effects afford the advantage of practice, as high-frequency words have been spoken aloud many, many times (Tomaschek et al., 2018). This should afford advantages to the motor planning and execution that are separate from whatever effects frequency has on the cognitive aspects. In building a more complete model of speech production, one would have to disentangle frequency-related effects at the cognitive level with practice effects at the motor level, as well as how these levels could interact over time.

The data from our corpus study has ramifications for phonological theories, as corpus-based studies contribute crucial information to our collective knowledge of how the sound systems of different languages actually function (Ernestus & Baayen, 2011). The third chapter of this dissertation focused on phonetic variation in the Spanish tap, and an in-depth discussion of the potential implications of our findings to phonological theory was therefore not included. Now, however, we will discuss how our findings of lexically driven subphonemic variation fit into current theories of phonology and what future work on the Spanish tap we think should follow the present dissertation.

Our finding of initial evidence for categorical subphonemic variation in the Spanish tap is consistent with the consensus that probabilistic variation must play a role in phonological systems (Alderete & Finley, 2023) and that purely symbolic or abstract accounts of phonology do not align with empirical speech data. It is similar to findings from other languages that information related to the word (or lemma) is active in the speech production system (e.g., Bell et al., 2009; Drager, 2011). A theoretical question raised by our production findings concerns whether the categorical alternation between reduced and unreduced taps is an online phenomenon based on lexical retrieval (e.g., Bell et al., 2009; Jurafsky, Bell, Gregory, & Raymond, 2001) or whether there are two versions of the tap stored in long-term memory. These two allophones would not be position-sensitive and are driven by word-level properties, so it is unclear how they would be incorporated into existing models of phonology that specify phonemic encoding at any level. In contrast, our data are easily explained by one of the models proposed by Pierrehumbert (2002), which includes frequency distributions of allophonic variation stored on a per-word basis. We can, therefore, explain the categorical effects of our model as stemming from frequency-based allophonic variation at the word level and the gradient effect of lexical frequency on

IntDiff as stemming from the speed of lexical retrieval. This would indicate that there is both phonetic and phonological variation in the Spanish tap attributable to lexical frequency.

In building upon the current findings in Spanish tap production, the first steps that should be taken are to replicate and consolidate the statistical model proposed by Perry et al. (2024). I see two potential avenues in this regard. The first is to replicate the findings of Perry et al. (2024) with another corpus of spontaneous Spanish, and the second is to see if the findings hold up to a multi-verse style analysis which explores other methods of measuring stop reduction (e.g., those used in Hualde et al., 2011). An analysis of a different corpus could use the data from Perry et al. (2024) as priors to identify models with more than two distributions and compare their predictive accuracy. This would let us explore the idea of multiple pronunciation variants along acoustic correlates and see if two is the optimal number. A multi-verse style analysis showing that the pronunciation variants are present along other acoustic correlates that purport to measure that same underlying construct would provide clear support or highlight how researcher degrees of freedom impact the interpretation of results (Coretta et al., 2023).

In any attempt to replicate these findings using another corpus of spontaneous Spanish, we must consider that varieties of Spanish may differ along this dimension. There is no guarantee that the phonological system of Madrilenian Spanish is consistent with the Spanish spoken elsewhere. The fact that corpora made up of various speaker populations prevent the generalizability of analysis (Newmeyer, 2003) is important to consider, both for the present dissertation and for corpus phonetics more generally. What are we learning in comparing phonetic patterns across populations that see only limited or unidirectional contact? Ernestus and Baayen (2011) addresses the concerns of analyzing large corpora containing multiple varieties,

claiming that mixed-effects models allow us to find what is common to all dialects and speakers, which is misleading. While it is true that mixed-effects models can be set up to estimate variation in an effect according to multiple dimensions, such as dialect and individual speakers, their claim that the main slope of the fixed effects allows for the generalizability of an effect across speakers and dialects is incorrect. As most commonly reported, this effect will be for an *average* dialect and speaker (i.e., the effect when all random effects are set to zero). While one *could* examine the generalizability of an effect across speakers and dialects using these models, it would require a much more careful inspection of the model. Relying on the estimates from the fixed effects is insufficient. Given the by-speaker variability in the models reported by Perry et al. (2024), a more in-depth inspection of statistical models used in corpus phonetics may be warranted, allowing for a more complete look at the generalizability of phonetic and phonological patterns.

5.3 Phonetic variability and spoken word recognition

Our auditory lexical decision experiment contributes experimental data from Spanish to the research investigating how reduction and variant frequency influence spoken word recognition. In contrast to studies that report an advantage of the unreduced or canonical variant (e.g., Tucker, 2011) or studies that find no effect of reduction after controlling for variant frequency (e.g., Bürki et al., 2018), the present study mirrors the results of Pitt (2009), with findings consistent with a variant frequency effect appearing alongside an advantage for unreduced taps. While intuitively, reduction and variant frequency are different effects, it is important that we discuss how to

tease apart these variables and how disparate findings may work together to inform theories of speech perception and spoken word recognition.

As discussed in Chapter 4 of the present dissertation, the canonical variant advantage in lexical access, which in many cases is the same as what we would consider the unreduced variant, has inconsistently been shown to have a facilitatory effect on lexical access in the auditory modality. Arguments for why this is the case are varied, with Sumner (2013) arguing for multiple lexical representations, with the canonical variant containing more nuanced representations of acoustic information that more heavily relies on bottom-up processing, while more casual speech processing is more reliant on top-down information (note that this is consistent with H&H Theory). A potentially complementary proposal by Dilley et al. (2019) raises the issue that initial lexical representations are created during childhood for L1 speakers of a language, and that child-directed speech contains many more canonical variants than adult-directed speech.

H&H Theory proposes that speech perception is *discriminative* in nature. This is consistent with recent computational work from Baayen et al. (2019), who propose that the lexicon is fundamentally based on discrimination. While frequency effects are easily accounted for by this model, it remains unclear exactly how it would account for a co-existing effect of the advantage of unreduced variants. We would have to assume that those differences in the acoustic signal allow for easier discrimination, which, as far as I am aware, is an unanswered empirical issue. Work from Warner et al. (2009) seems to contradict this, as they found different levels of intensity reduction lead to similar discrimination between a VV sequence and a sequence with an intervocalic tap.

How to answer the issue of whether or not canonical productions of words are more easily discriminated from other candidates needs to be based on something

other than listener behaviour if we are to avoid circularity in our argumentation, something that has historically been an issue in phonological theory (Ohala, 1990). As phonemes are an implausible entity in speech perception (Goldinger, 1998; Mitterer et al., 2018), this rules out using phonological neighbourhood density, as it relies on abstract phonological units. While we could calculate a similar metric based on allophones, this is also removed from the acoustic signal that listeners are discriminating. Additionally, it is perhaps fruitless to argue for any linguistic unit being cognitively 'real' (Samuel, 2020). As such, we should base any measurement of discriminability on the acoustic signal. One option is adapting the distance measure proposed by Kelley and Tucker (2022), which relies more directly on the acoustic signal. However, if we are to use it to calculate a measure of discrimination that works as a reasonable approximation to the knowledge an adult language speaker has, we need to calculate the acoustic distance using a large amount of acoustic data that includes many varying productions of high-frequency words from several speakers. If, after doing this, we see that canonical productions have a higher average acoustic distance from all other words in the lexicon, this could be taken as evidence that canonical productions provide more signal-dependent information to listeners that aid in the discrimination task.

Turning back to our lexical decision data, we see that the distance from an expected production (a continuous version of variant frequency) impacts the speed of lexical retrieval. Still, our results regarding the accuracy of identification were less robust to researcher degrees of freedom. Perceptual work on the English alveolar flap by Warner et al. (2009) may provide an explanation here. Warner et al. (2009) found that introducing even the minimal amount of acoustic evidence that a flap was present between two vowels led to almost perfect discrimination. Listeners only needed small amounts of information that there was a tap present. Warner et al.

(2009) collected but did not report reaction time data, but if it took longer for identification when there was less acoustic information, that would be consistent with our data from the lexical decision, as well as consistent with the claim from Sumner (2013) that different pronunciation variants are all equally able to activate the relevant target. While this is clearly the case to some extent, as speakers are not generally aware of pronunciation variants nor their impact on spoken word recognition, it may be the case that the speed of lexical retrieval of different variants is not the same, even though the word will be recognized regardless.

The presence of both an advantage of unreduced taps as well as a continuous version of variant frequency effects are interesting from a theoretical perspective, but our lexical decision study was not intended to examine word-specific phonetic properties, and this was a limitation that we will discuss in further depth here, along with a proposal for a follow-up experiment. Our full set of items in our lexical decision study was likely confounded due to a variant frequency effect stemming from the fact that our recordings came from a speaker of the language and were not synthetic in nature. Of the initial 53 tap-containing words that we presented our speaker with, he produced 51 of them with at least one instance of a true tap (the canonical variant). We also needed a 'reduced' category for our study of reduction, but for 28 of the 48 words that were produced with a true tap and one reduced variant, all reduced taps over several hours of recording and several repetitions of the word were the same, and we had recordings of either an approximant tap *or* a perceptual tap for that word. We only got all three variants for 20 words. Given our production data, which shows that there may be probabilistic variation occurring between two pronunciation variants based on word-specific properties, it is easy to see how either abstract representations of the word or fine-phonetic detail in the form of exemplars could impact our stimuli. A formalization of this confounding structure

will be discussed in Section 5.5.

I propose here the outline of a study that could, in theory, tease apart reduction effects and variant frequency effects. Any such study needs to start with a corpus analysis. This is unavoidable, as the variant frequency effect itself stems from the experience of the language user and must be based on data that the researcher believes to be a representative sample of that experience. A further limitation of studying this phenomenon is that it will always be biased towards investigating words that are not low-frequency, as getting enough data to consider what an expected production of a word is requires multiple tokens. These caveats aside, the corpus study should begin by quantifying the variation at the word level for as many words as possible so that experimental items can be selected that cover as large a range as possible in the acoustic parameters of interest. To give a concrete example based on the previous dissertation, this would mean finding words that have an expected production of IntDiff ranging from approximately zero to as high as possible while still having continuous coverage, as it would be undesirable to have a gap in the coverage where one side of the gap is supported by very few lexical items.

After the corpus analysis is complete and the words are selected, the experimental stimuli would be designed to vary along a grid, covering a greater amount of variation in IntDiff than the items for the expected productions (so that each item can have tokens both above and below an expected production). Using pseudo-spontaneous speech from a real speaker would provide the benefit of ecological validity, but it also may be exceedingly difficult to get an adequate range of productions. Synthesized stimuli may be an acceptable alternative here, as it would allow for increased control over the other aspects of the production. The analysis of such items, carefully controlled in a two-dimensional grid, with expected production of IntDiff for the word on the x-axis and the IntDiff of the stimuli on the y-axis, would be analyzed using

a non-linear interaction term in a generalized additive mixed model (Wood, 2017). Plotting the line where $IntDiff_{expected} = IntDiff_{actual}$ would allow for the effect of the distance from an expected production to be visualized by the model instead of calculated like was done in Chapter 4. This would also allow for a potential asymmetry to be explored, as it is possible that being more reduced than an expected production is not the same as being less reduced than an expected production.

We have so far focused on discussing the implications of our general findings regarding how the variability of the tap impacts spoken word recognition, and now we turn to a discussion of our L2 listeners. While most published studies find a difference in how reduction impacts L1 and L2 listeners (Ernestus, Dikmans, & Giezenaar, 2017; Ernestus, Kouwenhoven, & van Mulken, 2017; Wanrooij & Raijmakers, 2021), our study found this only for some of our analyses. For others, the interaction terms between L1 groups and other variables were centred near zero and highly uncertain. We never found evidence that reduction impacts both groups in the same way, as this would have required a much larger sample size. In our analysis of tap variant instead of reduction, we saw that this difference may be due to a variant frequency effect being different for L2 listeners. However, this was not borne out in the analysis using expected productions from the corpus. As this should be similar in theory, it suggests that further study into variant frequency effects in L2 spoken word recognition is needed. Research into variant frequency effects in L2 learners is not common, as pointed out by Llompart et al. (2021), who found that L1 variants of corresponding sounds were important for variant frequency effects in L2 spoken word recognition.

If the effects of tap reduction are, in fact, similar in L1 and L2 listeners, why would that be the case? The answer may lie in the similarity between the English flap and the Spanish tap. The majority of our L2 listeners were L1 speakers of North American English, which has the flap as a position-sensitive allophone. These are

the units proposed to be active in L2 acquisition (Flege & Bohn, 2021), and as we know, flap reduction is common in American English (Warner & Tucker, 2011), so it may be that any automatic perceptual routines from the L1 can apply successfully in the L2. While models of L2 speech learning would likely predict that these two categories would become merged - distinguishing them would be highly difficult due to the similarity - it might not matter for the accurate perception of this sound in either language. This could be explored by including an additional group whose L1 does not contain the alveolar tap or flap as a position-sensitive allophone in a future version of this study.

The presence/absence of the tap/flap in the L1 sound system should be a sufficient control in terms of the assumptions of theoretical models which assume allophones are interacting in a common phonetic space (Flege, 1995; Flege & Bohn, 2021). However, there is another more general language factor that deserves to be considered, which is the general prevalence and degree of stop reduction in the language. There is evidence that languages can differ in terms of general tendencies of stop reduction (Torreira & Ernestus, 2011), and it is plausible that general patterns of stop reduction in the L1 could also influence patterns of word recognition in the L2. Ideally, then, the two L1 groups would have languages that behaved similarly in terms of overall levels of consonant reduction but with one language missing the tap/flap in the phonological inventory. This would allow for a direct evaluation of theoretical models of L2 speech, which assume that the phonetic categories that interact are allophones. It would also be of theoretical interest to have two languages that differed greatly in terms of voiced stop reduction, neither of which had a similar sound to the Spanish tap, in order to explore how general processing adaptations from the L1 can be exploited (or not) in the L2.

As this was an initial examination of how tap variability impacts spoken word

recognition in L2 speakers, we focused on the broad population of self-identified L2 speakers of Spanish and did not explore variation in this group according to factors related to their experience with the language, which are known to affect virtually every aspect of production and perception (see Flege & Bohn, 2021, and the references therein). The data from our lexical decision experiment, which includes information about when the participants started learning Spanish, how long they have been speaking Spanish, their self-rated proficiency, and their time spent living in a country where Spanish was the societal language, can be used as pilot data for future studies of the phenomenon, providing a jumping off point that can be used in simulations to investigate required sample sizes as well as for providing informative priors for future investigations into this topic.

5.4 Implications for L2 speech learning

Although we did not investigate the L2 production of the Spanish tap, the findings of our corpus analysis have implications for future studies of this sound, as well as for theories of L2 speech learning more broadly. Models of L2 speech learning place critical importance on the input received during the acquisition process (e.g., Flege, 1995; Flege & Bohn, 2021; Van Leussen & Escudero, 2015). Following these models, the theoretical framing of many studies assumes that the input received by learners interacts with pre-established phonetic categories to produce the various patterns we observe in L2 production and perception. As the acquisition process is dependent on the input, empirical evaluations of theoretical claims are diminished when there is a misalignment between the properties of the input and our quantitative analysis of L2 segment productions. I believe there is such a misalignment in some of the literature investigating the acquisition of the Spanish tap by L2 learners.

Several studies of the acquisition of the Spanish tap by L2 learners have analyzed productions as either 'target-like' or not, with the most common definition of the target being a single visible occlusion on a spectrogram (e.g., Face, 2006; Herd, 2011; Patience, 2018), although J. Y. Kim and Repiso-Puigdelliura (2020) coded both true and approximant taps as the target. Using the canonical variant as the target appears reasonable given the textbook description of this sound (Hualde, 2005), but is at odds with studies of semi-spontaneous speech from different varieties of Spanish (Bradley & Willis, 2012; Willis & Bradley, 2008) as well as the data analyzed in the present dissertation (Perry et al., 2023, 2024), where we only find this occlusion present for half of all taps. We wouldn't claim that L1 speakers only produce target-like sounds half the time, and Chapter 4 of this dissertation provided evidence that L2 Spanish listeners can recognize reduced forms fairly accurately. As models of L2 speech generally assume that perception precedes and drives production (e.g., Flege & Bohn, 2021), it is unclear how any analysis of L2 tap production can provide meaningful results if reductions that are present in the input and accurately perceived by L2 learners are deemed not target-like by researchers.

This issue is not restricted to Spanish, taps, or studies of L2 production that categorize productions into target vs. non-target. Many studies of L2 segment production are based on acoustic correlates, which can run into the same general issue of making comparisons using laboratory speech without the added context of how the segment's acoustic correlates may differ in spontaneous speech and, therefore, the input that L2 learners receive. If a comparison of laboratory speech between L1 and L2 speakers finds a difference, but L2 speakers pattern closely to how that acoustic correlate is produced in the spontaneous input they receive, this is simply evidence that they differ in terms of producing the canonical variant in a situation where L1 speakers produce more careful speech. This would amount to a task effect and may

not tell the whole story of the processes taking place during L2 speech learning.

A reasonable rebuttal to the points I've raised here would be that those who are interlocutors with L2 speakers may adapt their speech (e.g., as seen in Lorge & Katsos, 2019), leading to differences compared to spontaneous productions within a group on L1 speakers. This is a fair criticism but does not escape the central recommendation, which is to empirically quantify the input that your L2 learners receive in order to properly evaluate how the properties of that input lead to their patterns of perception and production on an individual level. Instead of generally looking at L1 speech in the surrounding community, it may be more relevant to look at the spontaneous speech of the L1 speaker with which your L2 participants speak most often while interacting with your participant. While this adds practical challenges, research on L2 speech will not be able to progress if the properties of the input are not better defined than they currently are. This recommendation, to empirically quantify the input received by participants, is similar to a recent proposal from Flege (2021), which proposed sampling the input of interlocutors in order to quantify the proportion of the input that may be received from other L2 speakers, affecting the formation and/or composition of phonetic categories. This general argument also applies to spontaneous speech, as several acoustic correlates are known to vary systematically between lab speech and spontaneous speech (e.g., Laan, 1997; Warner & Tucker, 2011).

Another finding of the present dissertation which has implications for L2 speech learning is seeing the effect of distance from an expected production for specific words in L2 listeners of Spanish. Current theoretical models of L2 speech either ignore word-specific phonetics entirely (Flege, 1995; Flege & Bohn, 2021) or only appeal to word recognition as a mechanism for driving error-driven learning (Van Leussen & Escudero, 2015). While the evidence from our exploratory analysis still needs to be

replicated, doing so would indicate that models of L2 speech learning which ignore this acoustics-to-meaning mapping at the word level need to be updated, a position supported by the fact that L2 phonological processing involving lexical information diverges from the auditory processing of isolated sounds (Llompart, 2023).

5.5 Methodological considerations for research into phonetic reduction

In addition to the theoretical contribution of the present dissertation, our work brings up some methodological issues that the field as a whole needs to consider and address if we are to improve the quality of the empirical evidence of reduction phenomena. Ideally, these improvements would be paired with increased methodological transparency, sharing the code and data that support the claims of a paper when it is possible to do so. The rest of this section outlines the three main methodological issues raised during the present dissertation and provides concrete suggestions on how to address them. The first issue is the use of forced aligners in large-scale phonetic analysis, the second is defaulting to similar statistical models that don't align with theoretical knowledge, and the third is making inferences based on regression coefficients estimated from observational data.

Perhaps the most important issue research on phonetic variability has to reckon with is the widespread use of force-aligned boundaries for phonetic segmentation. While the fact that this could skew results is not a new idea (e.g., see a discussion of this issue by Watson & Evans, 2016), many large-scale analyses of phonetic data use a scripting language to take acoustic measurements based on force-aligned boundaries which have not been corrected (e.g., 13 of the 21 corpus studies published in the

proceedings of the International Congress of the Phonetic Sciences 2023 that cited the Montreal Forced Aligner did not hand-correct the output). While forced alignment allows for fast analyses of large amounts of data, we need to exercise caution as a field in trusting the results of analyses that do not provide empirical evidence that this crucial stage is functioning as intended.

Evaluation of the performance of force-aligned software is completed by those developing the software (Kelley et al., 2023; McAuliffe et al., 2017; Yuan & Liberman, 2008) and by end-users wanting to verify performance or compare different aligners (e.g., Gonzalez et al., 2020; Watson & Evans, 2016). The reality is that any single investigation of an aligner's performance may not provide the information needed for a researcher to know if it will work well *for the analysis at hand*. The typical evaluation method is to calculate and report distances between automatically-placed boundaries and human-placed boundaries for a speech sample. Some work is more in-depth, such as the comparison between aligners from Gonzalez et al. (2020), which provides specific circumstances where certain aligners may be more likely to fail. These analyses of boundary placement do not consider how measurements of formants, intensity, or pitch systematically change after hand correction. The fact that an aligner has worked better than others based on one set of criteria does not mean it will be reliable for a given analysis. The evaluations of force-aligned boundaries in the present dissertation provide a concrete example of the same boundaries, which were unreliable for measuring duration, being highly reliable for measuring IntDiff.

My proposed solution is simple, but not easy. Every study that uses force-aligned boundaries to take acoustic measurements should hand-correct a random subset of the data and evaluate how their specific measurements of interest change when using the force-aligned vs. the hand-corrected boundaries. This information should be reported either in the paper itself or in an appendix so that readers are aware of how

much additional noise the forced aligner was responsible for and whether it seems likely to have *biased* the measurements. Unfortunately for researchers who deal with large datasets, the increased sample size does not allow us to ignore the noise introduced by forced alignment, and the increased sample size may only exacerbate the issue of bias (for a discussion of this topic, see Kaplan et al., 2014). Speech science should continue to use forced-aligned software as a tool for speeding up research and leveraging big data, but we should not discount the benefit of smaller, high-quality data sets that allow for increased confidence that the results are not due to measurement error.

In our modelling of the duration and IntDiff of Spanish taps, we started with a log-normal model (duration) and a linear (Gaussian) model consistent with previous studies of these correlates. The log-normal model of duration was a good fit to the hand-corrected data but showed severe misspecification when fit to the automated measurements. As discussed in detail in Chapter 3, the linear model based on the Gaussian distribution did not fit the data well, but the estimates of several effects were consistent with previous literature. Defaulting to models often used by the field for specific purposes, without verifying that the assumptions of that model are reasonable, is related to a more general problem of not allowing theoretical assumptions to dictate the model chosen to analyze the data.

An easy example to illustrate a disconnect between the model and the data-generating process is linear or log-normal models of duration that are based on force-aligned data. Let's give the example of the log-normal model, which better approximates duration in general (as duration can't be negative and the variance and the mean are often correlated). A log-normal model of segment duration assumes that duration can be any non-negative number and that the duration entered into the model is an accurate measurement. Most force-aligned duration data can only

be a multiple of 10ms due to window advancement (Kelley et al., 2023; McAuliffe et al., 2017; Tucker & Mukai, 2023), which means that, in the best case scenario, the true duration is being rounded to the nearest 10ms value. There are models that can deal with the issue of rounded data (e.g., Zhao & Bai, 2020), but as far as I'm aware, they have never been used to address this issue.

Another ubiquitous example of scientific and statistical assumptions clashing is analyses of reaction times where responses are nested within different words and listeners. While there is currently a heavy emphasis placed on accounting for differences in the conditional mean by word and by participant through the use of mixed-effects models, the same models estimate a single number for the standard deviation of the residuals. This means that all words and participants are assumed to be equally variable with respect to reaction times, which is highly implausible. For example, the density distributions for different words visualized in Baayen and Milin (2010) show differences in the variability of reaction times, and visualizing reaction time distributions per participant in the Massive Auditory Lexical Decision database (Tucker et al., 2019) shows that the variability across different listeners is far from constant.

Unlike the issue of not validating results derived from force-aligned data, this issue does not have a simple solution. Building statistical models that are reasonable approximations of the data-generating process, that have been informed by expert knowledge, and that align theoretical assumptions of the phenomenon in question with the mathematical structure of the model takes a substantial amount of time and thought. There is no one-size-fits-all solution. However, given that the use of complex statistical models in growing in the phonetic sciences (e.g., see papers from special issue Roettger et al., 2019), it is important to remember that, in the absence of theoretical assumptions that warrant increased complexity, models should

be simple (e.g., Bates et al., 2015; Matuschek et al., 2017), although what constitutes a simpler model also deserves greater thought that we currently give it (e.g., Falk & Muthukrishna, 2023). In an ideal world, the data would be available in an online repository along with the processing and analysis scripts so that other researchers could re-run the models supporting a paper's conclusion and explore how robust the findings were to what they believed was a more appropriate set of assumptions. In practice, this may not always be possible or feasible, and in some cases, making this mandatory in all cases could propagate inequality along a number of dimensions (see Whitaker & Guest, 2020, for a discussion of this). A more equitable starting point in the absence of open data would be to mandate the sharing of model diagnostic plots in an appendix so that readers can evaluate their own subjective belief in the results of that model. The additional burden of such a requirement is minimal, as a single line of code can often complete this.

The third systemic issue that needs to be addressed is the widespread use of multiple regression models to analyze observational data in studies on phonetic reduction. This issue was briefly discussed by Perry et al. (2024), but this is a fairly serious limitation that deserves an expanded discussion. What follows is not a proper introduction to modern approaches to causal inference (the reader is directed to Pearl, 2009; Pearl et al., 2016), but rather a brief discussion of how one type of causal model, the Directed Acyclic Graph (DAG), can be used to motivate variables included or omitted from statistical models of observational data, as was done by Cohen Priva and Gleason (2020) for English consonants.

As the field of linguistics shifted to the use of multiple regression models from ANOVA, it became general knowledge that adding additional variables to a regression could statistically control for potential confounds. What is virtually never discussed is that adding additional variables into a regression can also *create* confounds (for a

thorough explanation of this, see McElreath, 2020). Whether a variable should be added depends on the specifics of the system under study and for which variable of the system you want to estimate the causal effect. A DAG is a graphical model of the relationships between variables relevant to a certain system under study, which has the key property that edges (arrows) connecting one variable to another can only go in one direction.

To illustrate a DAG for phonetic reduction where we *know* that there are causal relationships between variables (because we use them to calculate each other), we take a closer look at variables related to frequency and contextual probability, which have been used to study reduction (Aylett & Turk, 2004; Bell et al., 2009; Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, Bell, Gregory, & Raymond, 2001). The variables are word frequency (Freq), bigram frequency with the previous word (Bigram), conditional probability based on the previous word (CondProb), and mutual information with the previous word (MuInf). We also have our outcome construct Reduction, which is a deliberate simplification only for illustration purposes, as well as the frequency of the previous word (PrevWord), which is needed to calculate several of the other variables. U is an unobserved confound between word frequency and bigram frequency, as there is a strong relationship between the two variables, the nature of which is unclear.

We now outline the individual relationships that underlie our DAG, which is visualized in Figure 5.1. First, we have an unobserved variable U that is a joint cause of both frequency and bigram frequency ($Freq \leftarrow U \rightarrow Bigram$). We then have mutual information with the previous word, which is calculated using bigram information ($Bigram \rightarrow MuInf$) along with frequency information of the word ($Freq \rightarrow MuInf$) and previous word ($PrevFreq \rightarrow MuInf$). We know there is a causal relationship present, as changes in any of these variables will result in

changes in mutual information. Conditional probability is calculated using bigram information and the frequency of the previous word ($PrevFreq \rightarrow CondProb \leftarrow Bigram$). As all of these variables have been claimed to be related to reduction, we draw an arrow between all of them and Reduction (with the exception of the previous word).

A DAG guides the variable selection process by providing sets of variables that close the additional paths between the variable of interest and the outcome, isolating the effect that we care about. While cause flows in one direction down the arrows, information flows both ways. To isolate the *direct* causal effect of a variable, one needs to block the paths, which can be done by controlling for the variable in your model. Let's say we wanted to examine the effect of lexical frequency on reduction. To estimate this, we would need to control for Bigram, MuInf, and one of either CondProb or PrevFreq, as either of those would close the path. If, however, we only wanted to estimate the total (not direct) effect, any 'downstream' variables between frequency and reduction can be left open. In this case, we could just control for bigram frequency.

It is possible, given theoretical assumptions, to realize that a given causal effect cannot be estimated from a given set of data. While causal models are something that should be discussed openly in order to unite scientific theory and statistical analyses, we may end up acknowledging that causal hypotheses (e.g., the Probabilistic Reduction Hypothesis) are impossible to evaluate using spontaneous speech data, as there are too many interconnected variables that have relationships that we do not fully understand. If our goal is instead *predicting* phonetic variability, researchers should stop reporting and interpreting individual regression coefficients altogether and focus on comparing how models derived from different theories predict the data we observe.

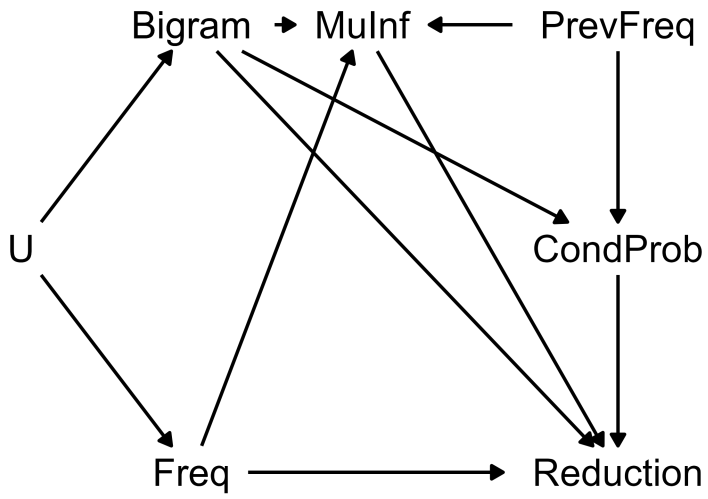


Figure 5.1: A directed acyclic graph depicting the proposed relationships between information-theoretic factors and reduction, as well as the relationships between them.

The advantage of explicit causal models is not limited to observational data, but can be useful to examining any system under study. The confound of variant frequency in our investigation into how reduction impacts spoken word recognition, discussed in Chapter 4, is visualized as a DAG in Figure 5.2. We wanted to study the causal effect of reduction on spoken word recognition ($Reduction \rightarrow SWR$), but they had a common cause as variant frequency (expected production) impacted both our reduction variable and the process of spoken word recognition ($SWR \leftarrow Variant\_frequency \rightarrow Reduction$).

In our study, the randomizing of which variant was selected in a subset of our items broke the causal link between variant frequency and reduction, leading to an effect of IntDiff in that subset of words when there was none present for all items together. This is a good example of how randomization allows for the estimation of

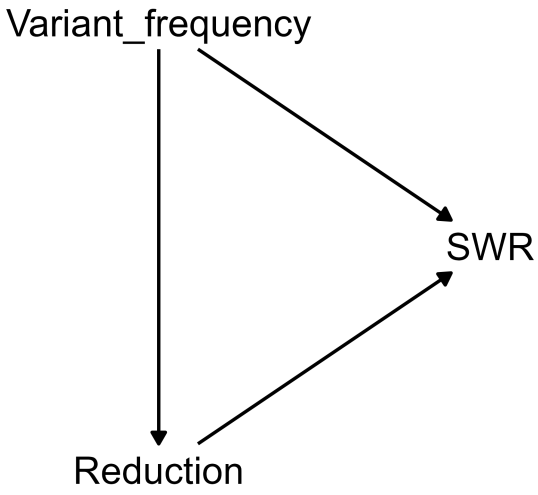


Figure 5.2: A simplified causal model of reduction and variant frequency effects in spoken word recognition.

causal effects in experimental settings. As an illustration of statistical control, we present the estimated effect of IntDiff for the subset of words used in the analysis of expected productions, both with and without the distance from an expected production included in the model. In Figure 5.3 A, the estimated effect of IntDiff without controlling for variant frequency is a small positive effect of which we are slightly uncertain. Figure 5.3 B shows the estimated effect of IntDiff from the same data after controlling for variant frequency by including the distance from an expected mean production in the model. After controlling for an expected production of the word, we have a large negative effect of IntDiff, of which we are more than 99.9% confident. It is important to point out that the 'true' direction of the effect is impossible to discern based on the data; it has to be interpreted through our theoretical assumptions.

Given the difficulties of making causal inferences from observational data, there

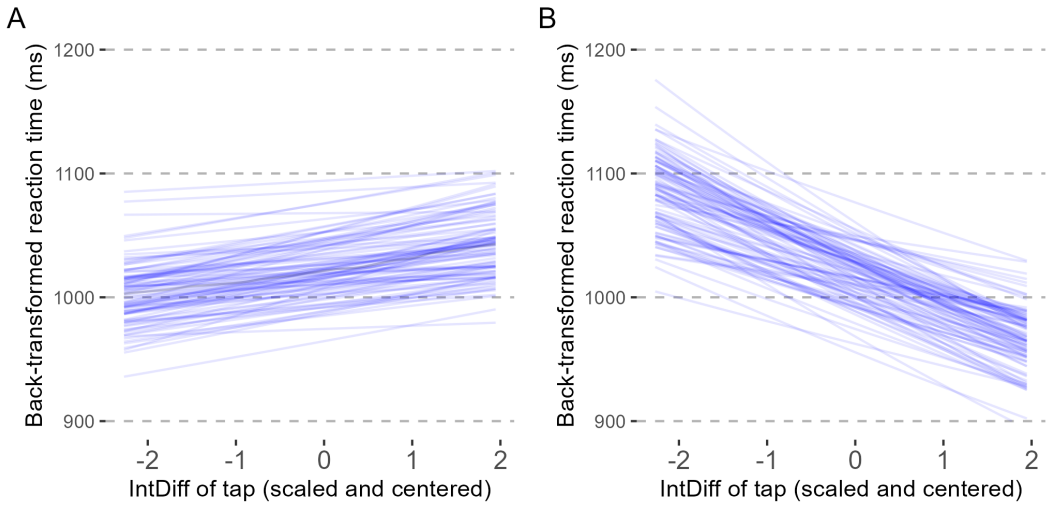


Figure 5.3: Panel A shows the effect of IntDiff on reaction time without controlling for the distance from an expected production of the word. Panel B shows the estimated effect of IntDiff in an identical model where the distance from the mean IntDiff for the word from the Nijmegen Corpus of Casual Spanish has been added. Lines in both panels represent random draws from the 89% credible interval.

is a need for more studies of phonetic variation that follow the procedure of a randomized controlled trial, randomly assigning participants from the same speaker population into different groups, one or more of which are exposed to experimental manipulation. Experimental elicitation of speech forms a spectrum from read speech using carrier phrases to situations with unguided speech, such as map tasks and sociolinguistic interviews. While these differ from conversational speech, I contend that the most important aspects from the point of view of ecological validity are that the talker is not reading nor repeating recorded stimuli and that they have not been coached or trained to produce speech following specific phonological or syntactic patterns. They must also be actively communicating with another person. Within these parameters, there is clear room for experimental manipulation.

I propose that solutions-based games, such as the map task (Anderson et al., 1991; Van Engen et al., 2010), in which participants must communicate information back and forth in service of a common goal, should be sufficiently spontaneous to allow for any mechanism that is responsible for reductions in daily interactions to be present in an experimental setting. The key components here would be initial randomization into a control or experimental group and the properties of the objects that are involved in the map task. As an example, let's imagine the simplest case of having only two descriptors and two objects for a total of four items. The descriptors could be 'red' and 'blue' while the objects are 'bike' and 'car'. In the control group, there would be an equal amount of red and blue versions of each object, so the colour imparts no information about the object itself. In the experimental group, there would be a manipulated range so that, for example, red is 25% more likely for a bike and blue is 25% more likely for a car. Long-term and repeated habituation into this context over repeated experimental sessions could feasibly allow for an experimental estimate of whether changes in predictability impact word duration. Any estimates from such a paradigm could then be compared to estimates based on different causal assumptions investigating spontaneous speech.

5.6 Conclusion

As was noted at the beginning of this dissertation, speech is highly variable. The data and analyses presented here add to the body of research that views this variability not as noise but as a window into the structure of language in the mind. We find lexical information is responsible for variation in the production of the Spanish alveolar tap in spontaneous productions, in addition to articulatory variables such as speech rate and phonetic environment. In perception, we see that phonetic reduction can impact

how both L1 and L2 listeners recognize words, as well as finding evidence for a version of the variant frequency effect based on a continuous acoustic correlate, indicating that the way that words are produced in everyday speech shapes how these words are recognized, regardless of whether Spanish is your first or second language.

References

Alderete, J., & Finley, S. (2023). Probabilistic phonology: A review of theoretical perspectives, applications, and problems [Publisher: John Benjamins]. *Language and Linguistics*, *24*(4), 565–610. https://doi.org/10.1075/lali.00141.ald

Amengual, M. (2016). Acoustic correlates of the Spanish tap-trill contrast: Heritage and L2 Spanish speakers. *Heritage Language Journal*, *13*(2), 88–112. https://doi.org/10.46538/hlj.13.2.2

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus [Publisher: Sage Publications Sage UK: London, England]. *Language and speech*, *34*(4), 351–366.

Arnold, K. F., Davies, V., de Kamps, M., Tennant, P. W. G., Mbotwa, J., & Gilthorpe, M. S. (2020). Reflection on modern methods: Generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology*, *49*(6), 2074–2082. https://doi.org/10.1093/ije/dyaa049

Arnon, I., & Cohen Priva, U. (2013). More than Words: The Effect of Multi-word Frequency and Constituency on Phonetic Duration. *Language and Speech*, *56*(3), 349–371. https://doi.org/10.1177/0023830913484891

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058. https://doi.org/10.1121/1.2188331

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective [Publisher: John Benjamins]. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, *2019*, 1–39. https://doi.org/10.1155/2019/4895891

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times [Publisher: Universidad de San Buenaventura (USB), Medellín]. *International journal of psychological research*, *3*(2), 12–28.

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Barry, W. J., & Andreeva, B. (2001). Cross-language similarities and differ-

ences in spontaneous speech patterns. *Journal of the International Phonetic Association*, *31*(1), 51–66. Retrieved November 14, 2020, from https://www.jstor.org/stable/44645150

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models [arXiv: 1506.04967]. *arXiv:1506.04967 [stat]*. Retrieved September 25, 2019, from http://arxiv.org/abs/1506.04967

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*. John Benjamins Publishing Company. Retrieved January 24, 2024, from http://ebookcentral.proquest.com/lib/ualberta/detail.action?docID=622748

Betancourt, M. (2017). Identifying Bayesian Mixture Models. https://betanalpha.github.io/assets/case_studies/identifying_mixture_models.html

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer [Computer program]. http://www.praat.org/

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607–624. https://doi.org/10.3758/s13415-015-0340-0

Bradley, T. G., & Willis, E. W. (2012). Rhotic variation and contrast in Veracruz Mexican Spanish. *Estudios de fonética experimental*, *21*, 43–74.

Retrieved November 14, 2020, from https://www.raco.cat/index.php/EFE/article/view/260320

Bradlow, A. R. (2022). Information encoding and transmission profiles of first-language (L1) and second-language (L2) speech. *Bilingualism: Language and Cognition*, *25*(1), 148–162. https://doi.org/10.1017/S1366728921000717

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., & Mok, K. (2011). Language-and talker-dependent variation in global features of native and non-native speech. *Proceedings of the... International Congress of Phonetic Sciences. International Congress of Phonetic Sciences*, 356.

Brand, S., & Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of French [Publisher: SAGE Publications]. *Quarterly Journal of Experimental Psychology*, *71*(5), 1240–1259. https://doi.org/10.1080/17470218.2017.1313282

Broś, K., Żygis, M., Sikorski, A., & Wołłejko, J. (2021). Phonological contrasts and gradient effects in ongoing lenition in the Spanish of Gran Canaria. *Phonology*, *38*(1), 1–40. https://doi.org/10.1017/S0952675721000038

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect [Publisher: Hogrefe Publishing]. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Bürki, A., Fougeron, C., Gendrot, C., & Frauenfelder, U. H. (2011). Phonetic reduction versus phonological deletion of French schwa: Some methodological issues. *Journal of Phonetics*, *39*(3), 279–288. https://doi.org/10.1016/j.wocn.2010.07.003

Bürki, A., & Frauenfelder, U. H. (2012). Producing and recognizing words with two

pronunciation variants: Evidence from novel schwa words [Publisher: SAGE Publications]. *Quarterly Journal of Experimental Psychology*, *65*(4), 796–824. https://doi.org/10.1080/17470218.2011.634915

Bürki, A., Viebahn, M. C., Racine, I., Mabut, C., & Spinelli, E. (2018). Intrinsic advantage for canonical forms in spoken word recognition: Myth or reality? [Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2017.1388412]. *Language, Cognition and Neuroscience*, *33*(4), 494–511. https://doi.org/10.1080/23273798.2017.1388412

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bybee, J. (2007). *Frequency of use and the organization of language. [electronic resource]*. Oxford University Press.

Bybee, J., & Hopper, P. J. (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins Publishing Company. Retrieved November 14, 2020, from http://ebookcentral.proquest.com/lib/ualberta/detail.action?docID=622579

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.63]. *WIREs Cognitive Science*, *1*(5), 629–647. https://doi.org/10.1002/wcs.63

Chang, C. B., & Yao, Y. (2016). Toward an Understanding of Heritage Prosody: Acoustic and Perceptual Properties of Tone Produced by Heritage, Native, and Second Language Speakers of Mandarin [Publisher: Brill]. *Heritage Language Journal*, *13*(2), 134–160. https://doi.org/10.46538/hlj.13.2.4

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production

in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001

Ciaccio, L. A., & Veríssimo, J. (2022). Investigating variability in morphological processing with Bayesian distributional models. *Psychonomic Bulletin & Review*, *29*(6), 2264–2274. https://doi.org/10.3758/s13423-022-02109-w

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278. Retrieved December 21, 2022, from https://www.degruyter.com/document/doi/10.1515/lp-2015-0008/html

Cohen Priva, U., & Gleason, E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language*, *96*(2), 413–448.

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics*, *70*(3), 403–411. https://doi.org/10.3758/PP.70.3.403

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., et al. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses [Publisher: SAGE Publications Sage CA: Los Angeles, CA]. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162567.

Davidson, L. (2011). Characteristics of stop releases in American English spontaneous speech. *Speech Communication*, *53*(8), 1042–1058. https://doi.org/10.1016/j.specom.2011.05.010

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, *54*, 35–50. https://doi.org/10.1016/j.wocn.2015.09.003

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*(6), 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

DiCanio, C., Chen, W.-R., Benn, J., Amith, J. D., & Castillo García, R. (2022). Extreme stop allophony in Mixtec spontaneous speech: Data, word prosody, and modelling. *Journal of Phonetics*, *92*, 101147. https://doi.org/10.1016/j.wocn.2022.101147

Dilley, L., Gamache, J., Wang, Y., Houston, D. M., & Bergeson, T. R. (2019). Statistical distributions of consonant variants in infant-directed speech: Evidence that /t/ may be exceptional. *Journal of Phonetics*, *75*, 73–87. https://doi.org/10.1016/j.wocn.2019.05.004

Drager, K. K. (2011). Sociophonetic variation and the lemma [Publisher: Elsevier]. *Journal of Phonetics*, *39*(4), 694–707.

Dufour, S., & Frauenfelder, U. H. (2010). Phonological neighbourhood effects in French spoken-word recognition [Publisher: SAGE Publications]. *Quarterly Journal of Experimental Psychology*, *63*(2), 226–238. https://doi.org/10.1080/17470210903308336

Ennever, T., Meakins, F., & Round, E. R. (2017). A replicable acoustic measure of lenition and the nature of variability in Gurindji stops [Number: 1]. *Laboratory Phonology*, *8*(1). https://doi.org/10.5334/labphon.18

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27–41. https://doi.org/10.1016/j.lingua.

2012.12.006

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The Recognition of Reduced Word Forms. *Brain and Language*, *81*(1), 162–173. https://doi.org/10.1006/brln.2001.2514

Ernestus, M., & Baayen, R. H. (2011). Corpora and Exemplars in Phonology. In *The Handbook of Phonological Theory* (pp. 374–400). John Wiley & Sons, Ltd. Retrieved January 30, 2024, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444343069.ch12

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions [Publisher: SAGE Publications]. *Quarterly Journal of Experimental Psychology*, *68*(8), 1469–1488. https://doi.org/10.1080/17470218.2014.984730

Ernestus, M., Dikmans, M. E., & Giezenaar, G. (2017). Advanced second language learners experience difficulties processing reduced word pronunciation variants. *Dutch Journal of Applied Linguistics*, *6*(1), 1–20. https://doi.org/10.1075/dujal.6.1.01ern

Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics*, *48*, 60–75. https://doi.org/10.1016/j.wocn.2014.08.001

Ernestus, M., Kouwenhoven, H., & van Mulken, M. (2017). The direct and indirect effects of the phonotactic constraints in the listener's native language on the comprehension of reduced and unreduced word pronunciation variants in a foreign language. *Journal of Phonetics*, *62*, 50–64. https://doi.org/10.1016/j.wocn.2017.02.003

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants [Editorial]. *Journal of Phonetics*, *39*(SI), 253–260. https://doi.org/10.

1016/S0095-4470(11)00055-6

Ernestus, M., & Baayen, R. H. (2007). The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration, and frequency of occurrence [Publisher: Saarbrücken:[Sn]].

Face, T. L. (2006). Intervocalic rhotic pronunciation by adult learners of Spanish as a second language. *Selected proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, 47–58.

Falk, C. F., & Muthukrishna, M. (2023). Parsimony in model selection: Tools for assessing fit propensity [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *28*(1), 123–136. https://doi.org/10.1037/met0000422

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2022). Optimal Coding and the Origins of Zipfian Laws. *Journal of Quantitative Linguistics*, *29*(2), 165–194. https://doi.org/10.1080/09296174.2020.1778387

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). York Press.

Flege, J. E. (2021). New Methods for Second Language (L2) Speech Research. In R. Wayland (Ed.), *Second Language Speech Learning, Theoretical and Empirical Progress* (pp. 119–156). Cambridge University Press.

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

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech [Publisher: Linguistic Society of America]. *Language*, *84*(3), 474–496.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178. https://doi.org/10.1016/j.jml.2015.12.006

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. [Publisher: American Psychological Association]. *Psychological review*, *105*(2), 251.

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research [Publisher: De Gruyter Mouton]. *Linguistics Vanguard*, *6*(1). https://doi.org/10.1515/lingvan-2019-0058

González-Alvarez, J., & Palomar-García, M.-A. (2016). Syllable frequency and spoken word recognition: An inhibitory effect [Publisher: Sage Publications Sage CA: Los Angeles, CA]. *Psychological Reports*, *119*(1), 263–275.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*(2), 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3

Grosjean, F. (2018). Spoken Word Recognition. In *The Listening Bilingual* (pp. 65–85). John Wiley & Sons, Ltd. Retrieved January 15, 2024, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118835722.ch4

Grüter, T., Lau, E., & Ling, W. (2020). How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues [Publisher: Taylor & Francis]. *Language, Cognition and Neuroscience*, *35*(2), 221–234.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in

shaping phonological patterns. *Linguistics Vanguard*, *4*(s2), 20170027. https://doi.org/10.1515/lingvan-2017-0027

Henriksen, N. C. (2015). Acoustic analysis of the rhotic contrast in Chicagoland Spanish: An intergenerational study. *Linguistic Approaches to Bilingualism*, *5*(3), 285–321. https://doi.org/10.1075/lab.5.3.01hen

Herd, W. (2011). *The Perceptual and Production Training of/d, tap, r/in L2 Spanish: Behavioral, Psycholinguistic, and Neurolinguistic Evidence* [PhD Thesis]. University of Kansas.

Hernandez, L. A., Perry, S. J., & Tucker, B. V. (2023). The effect of stress and speech rate on vowel quality in spontaneous Central Mexican Spanish. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 694–698). Guarant International.

Hilton, N. H., Schüppert, A., & Gooskens, C. (2011). Syllable reduction and articulation rates in Danish, Norwegian and Swedish [Publisher: Cambridge University Press]. *Nordic Journal of Linguistics*, *34*(2), 215–237.

Hualde, J. I. (2005). *The Sounds of Spanish*. Cambridge University Press.

Hualde, J. I., Beristain, A., Isasa, A. I., & Zhang, J. (2019). Lenition of word-final plosives in Basque. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 642–646). Australasian Speech Science; Technology Association Inc.

Hualde, J. I., Shosted, R., & Scarpace, D. (2011). Acoustics and Articulation of Spanish /d/ Spirantization. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 906–909).

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic in-

formation density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, *20*(1), 50–67. https://doi.org/10.1214/088342305000000016

Johnson, K. (2004). Massive reduction in conversational American English. *Proceedings Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). The effect of language model probability on pronunciation reduction. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, *2*, 801–804. https://doi.org/10.1109/ICASSP.2001.941036

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W., D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 229–254, Vol. 45).

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52. https://doi.org/10.1016/j.cognition.2017.05.001

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clinical and Translational Science*, *7*(4), 342–346. https://doi.org/10.1111/cts.12178

Katz, J. (2021). Intervocalic lenition is not phonological: Evidence from Campidanese Sardinian [Publisher: Cambridge University Press]. *Phonology*, *38*(4),

651–692. https://doi.org/10.1017/S095267572100035X

Katz, J., & Pitzanti, G. (2019). The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 16. https://doi.org/10.5334/labphon.184

Keating, P. A., Kingston, J., & Beckman, M. E. (1990). The window model of coarticulation: Articulatory evidence. *UCLA Working papers in Phonetics*, *69*, 3–29.

Kelley, M. C., Perry, S. J., & Tucker, B. V. (2023). The Mason-Alberta Phonetic Segmenter: A forced alignment system based on deep neural networks and interpolation. *arXiv preprint arXiv:2310.15425*.

Kelley, M. C., & Tucker, B. V. (2022). Using acoustic distance and acoustic absement to quantify lexical competition [Publisher: Acoustical Society of America]. *The Journal of the Acoustical Society of America*, *151*(2), 1367–1379.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Kim, D., & Smith, C. L. (2019). Phonetic conditioning of word frequency and contextual predictability effects in American English conversational speech. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 2836–2840). Australasian Speech Science; Technology Association Inc. Retrieved November 20, 2023, from https://www.semanticscholar.org/paper/Phonetic-conditioning-of-word-frequency-and-effects-Kim-Smith/9fc6905582fabfce086dcca06b07186a7de8de76

Kim, J. Y., & Repiso-Puigdelliura, G. (2020). Deconstructing heritage language

dominance: Effects of proficiency, use, and input on heritage speakers' production of the Spanish alveolar tap. *Phonetica*, *77*(1), 55–80. https://doi.org/10.1159/000501188

Kingston, J. (2008). Lenition. In L. Colantoni & J. Steele (Eds.), *Proceedings of the 3rd conference on laboratory approaches to Spanish phonology* (pp. 1–31). Cascadilla.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in methods and practices in psychological science*, *1*(2), 270–280.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Kuiper, K., Egmond, M.-E. v., Kempen, G., & Sprenger, S. (2007). Slipping on superlemmas: Multi-word lexical items in speech production. *The Mental Lexicon*, *2*(3), 313–357. https://doi.org/10.1075/ml.2.3.03kui

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? [Publisher: Taylor & Francis]. *Language, cognition and neuroscience*, *31*(1), 32–59.

Laan, G. P. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style [Publisher: Elsevier]. *Speech Communication*, *22*(1), 43–65.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Lenth, R. V. (2021). Emmeans: Estimated marginal means, aka least-squares means.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech*

Modelling (pp. 403–439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 923–929. Retrieved December 21, 2022, from https://www.duo.uio.no/handle/10852/50459

Llompart, M. (2023). On the effects of task focus and processing level on the perception-production link in second-language speech learning. *Studies in Second Language Acquisition*, 1–13. https://doi.org/https://doi.org/10.1017/S0272263123000414

Llompart, M., Eger, N. A., & Reinisch, E. (2021). Free Allophonic Variation in Native and Second Language Spoken Word Recognition: The Case of the German Rhotic. *Frontiers in Psychology*, *12*. Retrieved January 15, 2024, from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.711230

Lõo, K., Järvikivi, J., & Baayen, R. H. (2018). Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing. *Cognition*, *175*, 20–25. https://doi.org/10.1016/j.cognition.2018.02.002

Lorge, I., & Katsos, N. (2019). Listener-adapted speech: Bilinguals adapt in a more sensitive way [Publisher: John Benjamins Publishing Company Amsterdam/Philadelphia]. *Linguistic Approaches to Bilingualism*, *9*(3), 376–397.

Luce, P. A., & Cluff, M. S. (1998). Delayed commitment in spoken word recognition: Evidence from cross-modal priming. *Perception & Psychophysics*, *60*(3), 484–490. https://doi.org/10.3758/BF03206868

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing*, *19*(1), 1–36. Retrieved January 27, 2024,

from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467695/

MacKay, I. R., & Flege, J. E. (2004). Effects of the age of second language learning on the duration of first and second language sentences: The role of suppression [Publisher: Cambridge University Press]. *Applied Psycholinguistics*, *25*(3), 373–396.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models [ISBN: 0027-8874 (Print) 0027-8874 (Linking) _eprint: 1511.01864]. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the annual conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.21437/interspeech.2017-1386

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. [Publisher: American Psychological Association]. *Psychological review*, *88*(5), 375.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition* (2nd ed.). CRC Press. http://xcelab.net/rm/statistical-rethinking/

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 539–553. https://doi.org/10.1037/0278-7393.29.4.539

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition [Publisher: Elsevier]. *Journal of Memory and Language*, *98*, 77–92.

Morano, L., Bosch, L. t., & Ernestus, M. (2023). Second language learners acquire reduced word forms just like they acquire full forms: From exposure [Publisher: John Benjamins]. *Linguistic Approaches to Bilingualism*. https://doi.org/10.1075/lab.22043.mor

Mukai, Y. (2020). *Production and perception of reduced speech and the role of phonological-orthographic consistency* [Doctoral dissertation]. University of Alberta.

Mukai, Y., Järvikivi, J., & Tucker, B. V. (2023). The role of phonology-to-orthography consistency in predicting the degree of pupil dilation induced in processing reduced and unreduced speech [Publisher: Cambridge University Press]. *Applied Psycholinguistics*, 1–32. https://doi.org/10.1017/S0142716423000279

Munson, B., & Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1048–1058. https://doi.org/10.1044/1092-4388(2004/078)

Narayan, C. R. (2023). Speaking Rate, Oro-Laryngeal Timing, and Place of Articulation Effects on Burst Amplitude: Evidence From English and Tamil [Publisher: SAGE Publications Ltd]. *Language and Speech*, *66*(4), 851–869. https://doi.org/10.1177/00238309221133836

Newmeyer, F. J. (2003). Grammar Is Grammar and Usage Is Usage [Publisher:

Linguistic Society of America]. *Language*, *79*(4), 682–707. Retrieved January 30, 2024, from https://www.jstor.org/stable/4489522

Nijveld, A., Bosch, L. t., & Ernestus, M. (2022). The use of exemplars differs between native and non-native listening [Publisher: Cambridge University Press]. *Bilingualism: Language and Cognition*, *25*(5), 841–855. https://doi.org/10.1017/S1366728922000116

Noiray, A., Iskarous, K., & Whalen, D. H. (2014). Variability in English vowels is comparable in articulation and acoustics [Publisher: De Gruyter Mouton]. *Laboratory Phonology*, *5*(2), 271–288. https://doi.org/10.1515/lp-2014-0010

Ohala, J. J. (1990). There is no interface between phonology and phonetics: A personal view. *Journal of Phonetics*, *18*(2), 153–171. https://doi.org/10.1016/S0095-4470(19)30399-7

Patience, M. (2018). Acquisition of the Tap-Trill Contrast by L1 Mandarin–L2 English–L3 Spanish Speakers [Number: 4 Publisher: Multidisciplinary Digital Publishing Institute]. *Languages*, *3*(4), 42. https://doi.org/10.3390/languages3040042

Pearl, J. (2009). Causal inference in statistics: An overview.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Perry, S. J., Kelley, M. C., & Tucker, B. V. (2023). Measuring and modelling the duration of intervocalic alveolar taps in Peninsular Spanish. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 699–703). Guarant International.

Perry, S. J., Kelley, M. C., & Tucker, B. V. (2024). Documenting and modeling the acoustic variability of intervocalic alveolar taps in conversational Peninsular Spanish. *The Journal of the Acoustical Society of America*.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory phonology*, *7*.

Pitt, M. A. (2009). The strength and time course of lexical activation of pronunciation variants. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 896–910. https://doi.org/10.1037/a0013160

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics*, *39*(3), 304–311. https://doi.org/10.1016/j.wocn.2010.07.004

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569. https://doi.org/10.1121/1.2011150

Pollock, M., Delgado-Díaz, G., Galarza, I., Díaz-Campos, M., & Willis, E. W. (2023). The emergence of sound change in two varieties of Spanish. In S. Fernández Cuenca, T. Judy, & L. Miller (Eds.), *Innovative Approaches to Research in Hispanic Linguistics: Regional, diachronic, and learner profile variation* (p. 106, Vol. 38). John Benjamins Publishing Company.

Poncet, P. (2019). *Modeest: Mode Estimation*. https://CRAN.R-project.org/package=modeest

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ranbom, L., & Connine, C. (2007). Lexical representation of phonological variation. *Journal of Memory and Language*, *57*, 273–298. https://doi.org/10.1016/j.jml.2007.04.001

Roettger, T., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, 1–7. https://doi.org/10.1016/j.wocn.2018.12.001

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. https://doi.org/10.1016/j.jml.2019.104070

Sanchez, K., Hay, J., & Nilson, E. (2015). Contextual activation of Australia can affect New Zealanders' vowel productions. *Journal of Phonetics*, *48*, 76–95. https://doi.org/10.1016/j.wocn.2014.10.004

Santiago, F., & Mairano, P. (2018). The role of lexical stress on vowel space and duration in two varieties of Spanish. *9th International Conference on Speech Prosody*, 453–457. https://doi.org/10.21437/SpeechProsody.2018-92

Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. D. (2021). Durational Differences of Word-Final /s/ Emerge From the Lexicon: Modelling Morpho-Phonetic Effects in Pseudowords With Linear Discriminative Learning. *Frontiers in Psychology*, *12*. Retrieved February 10, 2024, from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.680889

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. https://doi.org/10.1016/j.cognition.2014.06.013

Slade, T., Gascon, A., Comeau, G., & Russell, D. (2022). Just noticeable differences in sound intensity of piano tones in non-musicians and experienced pianists. *Psychology of Music*, *51*(3), 924–937. https://doi.org/10.1177/03057356221126203

Spilková, H. (2014). *Phonetic reduction in spontaneous speech:an investigation of native and non-native production* [Doctoral thesis]. Norges teknisk-naturvitenskapelige universitet, Det humanistiske fakultet, Institutt for språk og litteratur [Accepted: 2014-12-19T13:09:48Z ISBN: 9788247149454]. Retrieved January 28, 2024, from https://ntnuopen.ntnu.no/ntnu-xmlui/

handle/11250/244283

Stein, S. D., & Plag, I. (2021). Morpho-Phonetic Effects in Speech Production: Modeling the Acoustic Duration of English Derived Words With Linear Discriminative Learning [Publisher: Frontiers]. *Frontiers in Psychology*, *12*, 678712. https://doi.org/10.3389/fpsyg.2021.678712

Sumner, M. (2013). A phonetic explanation of pronunciation variant effects. *The Journal of the Acoustical Society of America*, *134*(1), EL26–EL32. https://doi.org/10.1121/1.4807432

Tang, K., & Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (Kaqchikel Mayan). *The Journal of the Acoustical Society of America*, *144*(2), 997–1017. https://doi.org/10.1121/1.5046095

Tily, H., & Kuperman, V. (2012). Rational phonological lengthening in spoken Dutch. *The Journal of the Acoustical Society of America*, *132*(6), 3935–3940. https://doi.org/10.1121/1.4765071

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2021). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning [Publisher: Cambridge University Press]. *Journal of Linguistics*, *57*(1), 123–161.

Tomaschek, F., & Tucker, B. V. (2023). Speech Production: Where Does Morphology Fit? In D. Crepaldi (Ed.), *Linguistic Morphology in the Mind and Brain* (pp. 80–95). Routledge.

Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, *4*(s2), 20170018. https://doi.org/10.1515/lingvan-2017-0018

Torreira, F., & Ernestus, M. (2010). The Nijmegen corpus of casual Spanish. In

N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 2981–2985). European Language Resources Association (ELRA).

Torreira, F., & Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *2*(2), 331–353. https://doi.org/10.1515/labphon.2011.012

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302–324. https://doi.org/10.1075/ml.6.2.04tre

Tucker, B. V. (2011). The effect of reduction on the processing of flaps and /g/ in isolated words. *Journal of Phonetics*, *39*(3), 312–318. https://doi.org/10.1016/j.wocn.2010.12.001

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375–400. https://doi.org/10.1075/ml.11.3.03tuc

Tucker, B. V., & Mukai, Y. (2023). *Spontaneous Speech*. Cambidge University Press.

van de Ven, M., & Ernestus, M. (2018). The role of segmental and durational cues in the processing of reduced words [Publisher: SAGE Publications Ltd]. *Language and Speech*, *61*(3), 358–383. https://doi.org/10.1177/0023830917727774

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles [Publisher: Sage Publications Sage UK: London, England]. *Language and speech*, *53*(4), 510–540.

Van Heuven, W. J., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition [Publisher: Elsevier]. *Journal of memory and language*, *39*(3), 458–483.

Van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*. Retrieved May 2, 2022, from https://www.frontiersin.org/article/10.3389/fpsyg.2015.01000

van Dommelen, W. A. (2018). Reduction in native and non-native read and spontaneous speech. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, M. Zellers, & A. Lahiri (Eds.), *Rethinking Reduction* (Vol. 25).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413–1432.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish [Publisher: Taylor & Francis]. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73.

Vroomen, J., & De Gelder, B. (1997). Activation of embedded words in spoken word recognition. [Publisher: American Psychological Association]. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 710.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. https://doi.org/10.1016/j.

wocn.2014.11.001

Wanrooij, K., & Raijmakers, M. E. J. (2020). Evidence for immature perception in adolescents: Adults process reduced speech better and faster than 16-year olds. *Language Acquisition*, *27*(4), 434–459. https://doi.org/10.1080/10489223.2020.1769627

Wanrooij, K., & Raijmakers, M. E. J. (2021). "Hama"? Reduced pronunciations in non-native natural speech obstruct high-school students' comprehension at lower processing levels. *Journal of Phonetics*, *88*, 101082. https://doi.org/10.1016/j.wocn.2021.101082

Warner, N. (2019). Reduced speech: All is variability. *WIREs Cognitive Science*, *10*(4), e1496. https://doi.org/https://doi.org/10.1002/wcs.1496

Warner, N. (2023). Advancements of phonetics in the 21st century: Theoretical and empirical issues of spoken word recognition in phonetic research. *Journal of Phonetics*, *101*, 101275. https://doi.org/10.1016/j.wocn.2023.101275

Warner, N., Fountain, A., & Tucker, B. V. (2009). Cues to perception of reduced flaps. *The Journal of the Acoustical Society of America*, *125*(5), 3317–3327. https://doi.org/10.1121/1.3097773

Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustical Society of America*, *130*(3), 1606–1617. https://doi.org/10.1121/1.3621306

Watson, C., & Evans, Z. (2016, December). *Sound Change or Experimental Artifact?: A study on the impact of data preparation on measuring sound change.*

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition [Place: Netherlands Publisher: Elsevier Science]. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition [Publisher: Wiley Online Library]. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401.

Westreich, D., & Greenland, S. (2013). The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. https://doi.org/10.1093/aje/kws412

Whitaker, K., & Guest, O. (2020). #Bropenscience is broken science. Retrieved January 26, 2024, from https://www.bps.org.uk/psychologist/bropenscience-broken-science

Willis, E., & Bradley, T. (2008). Contrast maintenance of taps and trills in Dominican Spanish: Data and analysis. In L. Colantoni & J. Steele (Eds.), *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology* (pp. 87–100). Cascadilla Proceedings Project.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (Second edition). CRC Press/Taylor & Francis Group.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, *123*(5_Supplement), 3878–3878. https://doi.org/10.1121/1.2935783

Zampini, M. L. (2008). L2 speech production research [Publisher: John Benjamins Philadelphia]. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219–249, Vol. 36).

Zhao, N., & Bai, Z. (2020). Bayesian statistical inference based on rounded data [Publisher: Taylor & Francis _eprint:

https://doi.org/10.1080/03610918.2018.1476701]. *Communications in Statistics - Simulation and Computation*, *49*(1), 135–146. https://doi.org/10.1080/03610918.2018.1476701