Perceptual Asymmetry and Sound Change: An Articulatory, Acoustic/Perceptual, and Computational Analysis

by

Ian Calloway

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in the University of Michigan
2020

Doctoral Committee:

Professor Patrice S. Beddor, Chair
Associate Professor Steven A. Abney
Assistant Professor David Brang
Associate Professor Jelena Krivokapić

Ian Calloway
iccallow@umich.edu
ORCID iD: 0000-0002-8763-7615

ACKNOWLEDGMENTS

I am honored to have the opportunity to share this dissertation. Its completion would not have been possible without the contributions of many others.

I cannot thank Pam Beddor enough for her guidance and tireless generosity. As an instructor, she set me down the intellectual path that led me to the focus of my dissertation. As a mentor, she supported my growth as an academic and always stood in my corner, offering useful feedback and a sympathetic ear. As the chair of my dissertation committee, she helped steer this project toward completion even through the shifting landscape of my final semester. I am grateful to be able to call Pam a friend and to work alongside her as a colleague.

I would also like to thank my committee members - Steve Abney, David Brang, and Jelena Krivokapić – for their insightful feedback during the planning, analysis, and writing up of this work. I was also able to take coursework with David and Jelena; these courses exposed me to new ways of thinking about phonetics, phonology, and perception and helped refine my research approach.

Will Styler and Stephen Tobin have both offered sage advice during the conceptualizing and analysis of this project. I am grateful for the opportunity to bounce ideas off both of you.

My genuine interest in phonetics first began at the University of Chicago taking coursework and working as a research assistant for Alan Yu. I would like to thank him for first introducing me to formants, spectrograms, and sound change. My first substantial experimental project was conducted under the capable advising of Robin Queen and Pam. Through their guidance, I developed confidence in designing and conducting experiments.

I also benefited from opportunities for collaborative research. Pam, Andries Coetzee, and Nick Henriksen took me on as a research assistant in various capacities, which provided me exposure to a new experimental methodologies and analytical techniques. I have had the fortune to work with Stephen, Will, Dominique Bouavichith, Justin Craft, Tamarae Hildebrandt, and Jian Zhu on a variety of research projects, which has helped me become a better researcher.

Jen Nguyen, Sandie Petee, and Talisha Reviere-Winston have also been an essential resource during my graduate school career. I appreciate their guidance in progressing through the program and overcoming administrative obstacles.

I would like to thank the Phonetics-Phonology Discussion group for offering a platform to present research and receive constructive feedback.

Pam's bi-weekly advisee discussion group also provided invaluable support over my graduate school career. The ability to discuss small snippets of research, talk about half-formed ideas, and vent about woes helped me to make incremental progress and stay sane. Thank you to Harim Kwon, Cameron Rule, Sagan Blue, Kate Sherwood, Dave Ogden, Jiseung Kim, Dom, Justin, Fahad Alrashed, Kelly Wright, and Jian, and of course to Pam for organizing these group sessions.

I know many members of the department not just as colleagues, but as friends - hanging out at conferences, debating in the lounge, or maybe deciphering unintelligible pictures in *Drawful*. I am grateful that I got to know you all. You have enriched my time here, both inside and outside the department. I would like to give special mention to the members of my cohort - Marjorie Herbert and Alicia Stevers. I am glad to have progressed through this program with you and to share the highs and lows of this experience.

I am also grateful for my support outside the department. My close Michigan friends – Kevin, Eric, and Avin – as well as my undergraduate friends - Jonathan, Sam, Colin, Miriam and Jane - have cheered me on and kept me cognizant of the world beyond the confines of Lorch Hall. To my high school friends, I apologize for being so far west but hope that the next several years of life bring us more opportunities to reconnect. My mother, father, and Jennifer and Ray have offered unflagging encouragement since the first day of elementary school – they have always believed in my success even when I did not. I also hope to be more present in your lives in the coming years.

Finally, I am fortunate to have as caring and devoted a partner as Chris. I am thankful to be able to celebrate this achievement with him and am grateful for his patience and generosity. I hope to repay even a fraction of this kindness as he begins his own academic journey.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF EQUATIONS

EQUATION

LIST OF APPENDICES

APPENDIX

ABSTRACT

Previous experimental study of the identification of stop and fricative consonants has shown that some consonant pairs are asymmetrically confused for one another, with listeners' percepts tending to favor one member of the pair in a conditioning context. Researchers have also suggested that this phenomenon may play a conditioning role in sound change, although the mechanism by which perceptual asymmetry facilitates language change is somewhat unclear. This dissertation uses articulatory, acoustic, and perceptual data to provide insight on why perceptual asymmetry is observed among certain consonants and in specific contexts. It also uses computational modeling to generate initial predictions about the contexts in which perceptual asymmetry could contribute to stability or change in phonetic categories. Six experiments were conducted, each addressing asymmetry in the consonant pairs /k/-/t/ (before /i/), /k/-/p/ (before /i u/), /p/-/t/ (before /i/), and /θ/-/f/ (possibly unconditioned).

In the articulatory experiment, vocal tract spatial parameters were extracted from real-time MRI video of speakers producing VCV disyllables in order to address the role of vocal tract shape in the target consonants' vowel-dependent spectral similarity. The results suggest that, for consonant pairs involving /k/, CV coarticulation creates—as expected—vocal tract shapes that are most similar to one another in the environment conditioning perceptual asymmetry. However, CV coarticulation was less informative for explaining the vocalic conditioning of the /p/-/t/ asymmetry. In the second experiment, RF models were trained on acoustic samples of the target consonants from a speech corpus. Their output, which was used to identify frequency components important to the discrimination of consonant pairs, aligned well with these consonants' spectral characteristics as predicted by acoustic models. A follow-up perception experiment that examined the categorization strategies of participants listening to band-filtered CV syllables generally showed listener sensitiv-

ity to these same components, although listeners were also sensitive to band-filtering outside the predicted frequency bands.

Perceptual asymmetry is observed in CV and isolated C contexts. In the fourth experiment, a Bayesian analysis was performed to help explain why perceptual asymmetry appears when listening to isolated Cs, and a follow-up perception experiment helped to evaluate the relevance of this analysis to human perception. For /k/-/t/, for example, whose confusions favor /t/, this analysis suggested that [t] and [k] both have the highest likelihood of being generated by /t/ (relative to likelihood of /k/ generating each) in the context conditioning asymmetry. The follow-up study suggests listeners are more likely to categorize a [t] and [k] as /t/ if it has higher likelihood of being generated by /t/ (relative to /k/).

The final experiment used agent-based modeling to simulate the intergenerational transmission of phonetic categories. Its results suggest that perceptual asymmetry can affect the acquisition of categories under certain conditions. A lack of reliable access to non-phonetic information about the speaker's intended category or a tendency not to store tokens with low discriminability can both contribute to the instability of phonetic categories over time, but primarily in the contexts conditioning asymmetry.

This dissertation makes several contributions to research on perceptual asymmetry. The articulatory experiment suggests that confusability can be mirrored by gestural ambiguity. The Bayesian analysis could also be used to build and test predictions about the confusability of other sounds by context. Finally, the model simulations offer predictions of the conditions where perceptual asymmetry could condition sound change.

CHAPTER 1

Introduction

1.1 Introducing Perceptual Asymmetry

The primary focus of this dissertation is phonetic perceptual asymmetry, a phenomenon whereby speech sounds are confused for one another at unequal rates. If an individual misidentifies A as B more often than B as A, A-B as a pair show asymmetry, with B as the favored item.[1]The research undertaken in this thesis on perceptual asymmetry aims to provide further insight into the structure of phonetic units, the source of directional asymmetries in confusion rates, and the role of misperception in sound change.

Perceptual asymmetry has been documented in the auditory domain with speech sounds under laboratory settings. Table 1.1 describes cases where perceptual asymmetry has been observed among consonants and vowels. For each pair listed, listeners show higher rates of misidentification for the speech sound on the left in the specified phonetic context. Perceptual asymmetry has been observed in the visual domain as well, including in the visual identification of letters (Garner & Haun, 1978; Gilmore et al., 1979; see Ohala, 1997 for further discussion), and the categorization of animals (Quinn et al., 1993). For example, in Gilmore et al. (1979)'s study of the visual categorization of letters, participants were presented with a series of capital English letters, each appearing on a screen for a variable amount of time. The researchers observed differing rates of misidentifications for certain letter pairs – for example, 28% of 'Q' presentations were identified as 'O' (in fact 10% higher than the proportion of correctly identified 'Q's), while only 8% of 'O' presentations were identified as 'Q'. A similar effect has also been observed in the perception of handshape in American Sign Language. In Stungis (1981), signing and non-signing participants saw a video of

[1]Perceptual asymmetry is contrasted with what is described here as discriminative asymmetry, an increase in discriminative acuity for some subset of items. Discrimination asymmetry has been observed across human perceptual modalities, ranging from the auditory linguistic (e.g., Polka & Werker, 1994), visual linguistic (Best et al., 2010), visual non-linguistic (Rothkopf, 1957; Meissner & Brigham, 2001), and tactile (Williams & Julesz, 1992), among many others, but this phenomenon is distinct from asymmetry in identification patterns.

a native signer producing ASL signs, and participants identified the handshapes used in each case. The researchers found that native signers confused G and K handshapes with differing frequencies. 11% of G handshapes were misidentified as K, while only 4% of K handshapes were misidentified as G. In the auditory-acoustic domain, one of the first studies demonstrating perceptual asymmetry between speech sounds was Miller and Nicely (1955), who found that listeners tended to misidentify [θ] more often than they misidentified [f]. [θ] was misidentified as /f/ 26% of the time, but [f] was misidentified as /θ/ only 9% of the time.

SOUND PAIRS	FAVOR	PHONETIC CONTEXT	OBSERVED IN. . .
/k/-/t/	/t/	_/i/	Winitz et al., 1972; Plauché et al., 1997; Guion, 1998; Plauché, 2001
/k/-/tʃ/	tʃ	_/i/	Chang et al., 2001
/k/-/t/	/t/	_/u/	Plauché, 2001; but see Winitz et al., 1972
/p/-/t/	/t/	_/i/	Winitz et al., 1972; Plauché, 2001
/p/-/t/	/t/	_/u/	Plauché, 2001; but see Winitz et al., 1972
/p/-/k/	/k/	_/i/	Winitz et al., 1972; Plauché, 2001
/p/-/k/	/p/	_/u/	Winitz et al., 1972; but see Plauché, 2001
/θ/-/f/	/f/	*context unclear*	Miller & Nicely, 1955; Wang & Bilger, 1973; Cutler et al., 2004
/ɹ/-/l/	/l/	/ɑ/_(/ɑ/)[2]	Müller, 2010
/l/-/ɹ/	/ɹ/	/i/_/i/	Müller, 2010
/æ/-/ɛ/	/ɛ/	*context unclear*	Peterson & Barney, 1952; Cutler et al., 2004; Cutler et al., 2005
/ɛ/-/ɪ/	/ɪ/	*context unclear*	Peterson & Barney, 1952; Cutler et al., 2004; Cutler et al., 2005
/ʌ/-/ɑ/	/ɑ/	*context unclear*	Peterson & Barney, 1952

Table 1.1: Sound pairs that show perceptual asymmetry

Across modalities, item pairs that demonstrate perceptual asymmetry tend to resemble one another. From the above examples, the letters 'E'-'F' tended to show asymmetry (Gilmore et al., 1979) as did the handshape pair G-K (Stungis, 1981), which differ only in whether the middle finger is extended. Similarity tends to appear among other pairs that show perceptual asymmetry: the pair of lower-case letters 'd' and 'q' show asymmetry (Garner & Haun, 1978) as do cats and dogs, at least within the context of (Quinn et al., 1993). In each case the members of each pair have strong visual or featural resemblance to one another. Likewise, pairs of speech sounds that

[2]Although Müller (2010) reports /ɹ/ asymmetries in the context of /ɑ/, this result groups intervocalic and coda contexts together.

show perceptual asymmetry also agree in articulatory parameters (e.g., manner of articulation, voicing) and have similar acoustic structure. The sounds /θ/ and /f/, for example, are both voiceless fricatives, with a consonantal constriction located at the upper incisors.

In some cases of perceptual asymmetry, the two members of a confusable pair differ minimally with respect to some feature. For the letter pair E-F, for example, where confusions favor 'F', the two can be distinguished only by the presence of a third horizontal stroke. Confusions within this pair favor the letter without this additional stroke. Similarly, the G and K handshapes of ASL can be distinguished by whether the middle finger is extended, and these confusions favor the handshape with the extended finger. For the letters 'd' and 'q', reflected variants of one another, confusions favor the orientation corresponding to 'd'. The extent to which perceptual asymmetry among consonant pairs can be described by a difference in features is unclear, though this characterization has been applied before to individual consonant pairs (Plauché et al., 1997).

Winitz et al. (1972) and Plauché (2001) both have made significant contributions to our understanding of perceptual asymmetry among voiceless stops. Despite disagreements in the directionality of some confusions, many of their findings do align and provide evidence—illustrated in Table 1—that stop asymmetries tend to emerge in specific phonetic contexts. For example, although [p] and [k] demonstrate perceptual asymmetry before high vowels, they have confusion rates comparable for one another before a low back vowel. Specifically, before [u], 30% of [k] productions were misidentified as /p/, while only 12% of [p] productions were misidentified as /k/. In contrast, before [ɑ], 14% of [p] productions were misidentified as /k/, and 18% of [k] productions were misidentified as /p/ (Winitz et al., 1972).

These studies also suggest that phonetic context can condition the direction of asymmetry. Plauché (2001) found that listeners misidentified 5% of [p] productions as /k/ and only 1% of [k] productions as /p/ when listening to consonants isolated from a /Cu/ syllable. When these same consonants were isolated from a /Ci/ syllable, listeners misidentified 4% of [k] productions as /p/ and only 1% of [p] productions as /k/ (though see Winitz et al., 1972). Though the rates of confusion in her data were low in both contexts, these findings at least suggest that [i] and [u] may affect the directionality of confusions between /p/ and /k/.

The study of perceptual asymmetries for speech sounds sits at the intersection of several phonetic sub-disciplines. This phenomenon is of course relevant to theories of speech perception. In many cases of perceptual asymmetry, the discrepancy between the production intended by the speaker and what the listener perceives also crosses articulatory boundaries. Consequently, perceptual asymmetry may be useful to better understand how phonetic units are structured. Furthermore, there is reason to believe that perceptual asymmetry may play a role in the diachronic change of phonetic categories. This dissertation pursues these lines of inquiry in Chapters 2-5. The following sections briefly describe what is known within these sub-fields in relation to perceptual asymmetry,

and the issues to be addressed in each chapter.

1.2 Overview of chapter 2: Perceptual asymmetry and the phonetic primitive

Within theories of phonetic perception, one area of disagreement has centered on the identity of the phonetic primitive. Approaches like general auditory theory (outlined in Diehl et al., 2004)[3]and DIVA (Guenther, 1994) suppose that the phonetic unit in perception is acoustic and unmediated by articulatory representations. In contrast, under gestural phonetic theories, including motor theory (Liberman & Mattingly, 1985) and direct realism (Fowler, 1986), phonetic perception is understood to be organized with respect to gestures, either those intended by the speaker, or those "physically real events that unfold during the speech production process." (Browman & Goldstein, 1992, p. 156).

Oftentimes acoustic and gestural models align closely with one another in their predictions, but one area where they can differ is when acoustic and articulatory variability mismatch. For example, when divergent productions result in acoustically similar forms, the outcome of perception in these cases can provide useful insight into how a listener structures the phonetic unit. A widely cited case where such a mismatch is observed is among American English rhotic approximants. Acoustically, this class of sounds tends to be uniformly associated with a low F3 value, and the different production styles tend to show consistent differences from one another only in higher formant values (Zhou et al., 2008). However, these productions styles are associated with a variety of articulatory settings, which can differ in tongue body constriction location and degree, whether the tongue tip is raised or lowered, and whether the tongue tip is retroflexed (Delattre & Freeman, 1968). If the English rhotic approximant can be understood as a singular phonetic unit, both auditory and gestural approaches would seem to require a single characterization that describes all the variants. An auditory approach might simply suppose that the rhotic is defined by its shared acoustic features (e.g., low F3). Researchers have also taken a gestural approach to the description of this approximant. Taking an approach consistent with Articulatory Phonology (e.g., Browman & Goldstein, 1989; Fowler, 2007) suggested that the relevant level of description for speech sounds is "constriction locations and degrees". If English rhotic approximants are all composed of constrictions at the tongue body, tip, and lips, as suggested by Fowler, all rhotic production styles could

[3]These approaches were also called 'auditory theories' in Liberman and Mattingly (1985).

share the same articulatory description.[4]

Likewise, perceptual asymmetries involving consonants often display a mismatch in perceptual and articulatory similarity. In Winitz et al. (1972), for example, when listeners were asked to categorize the initial consonant of a /Ci/ syllable, [p] was categorized as /t/ 15% of the time, while [t] was categorized as /p/ only 6% of the time. Despite gross differences in active and passive articulators, listeners frequently misidentified a labial constriction as a lingual constriction. In what way are these productions structured that might lead to perceptual and acoustic similarity despite apparent articulatory differences? Just as in the characterization of rhotic approximants, an auditory approach to perception might simply suppose that the two productions share similar acoustic characteristics. Might there also be an analogous ambiguity in the shape of the vocal tract from gestural standpoint? Like the case of rhotic, is there an articulatory description of productions showing perceptual asymmetry that mirrors their perceptual similarity?

Chapter 2 of this dissertation looks at perceptual asymmetry for consonant pairs with respect to the apparent mismatch in articulatory and acoustic similarity between members of these pairs. That chapter considers the hypothesis that there is an articulatory description of these consonant pairs that mirrors their similarity in perceptual space. Despite differences in the articulatory events involved in the production of each consonant in a pair, the respective geometries of the vocal tract in the contexts that favor confusion are predicted to take on a similar shape for both consonants. In these cases, a singular gross vocal tract shape is shown to be ambiguous between either of several articulatory events.

1.3 Overview of chapter 3: Spectral features associated with confusability

Articulatory similarity between speech sounds does not guarantee that they will be confused. For example, syllable-initial sibilant fricatives /s/ and /ʃ/, distinctive only in place of articulation, have identically low rates of confusion for one another before /ɑ/ – no [s] productions are misidentified as /ʃ/ or vice versa at +12 dB SNR in Miller and Nicely (1955)'s classic study. For consonants that are asymmetrically confused—especially voiceless stops—significant progress has been made

[4]Under the interpretation of Fowler (2007), rhotic productions are specified for tongue body, tongue tip, and lip constrictions. However, this description does not appear to be completely appropriate for 'bunched' rhotics. Unlike other production styles, where the tongue tip can be raised, this style of rhotic production is associated with a lowering of the tongue tip (e.g., Delattre & Freeman, 1968; Zhou et al., 2008). Even if tongue tip lowering is understood as being due to an explicit articulatory goal, this goal seems qualitatively different than, for example, a retroflex rhotic, which involves tongue tip raising. van Lieshout et al. (2008) offer an alternative gestural description of rhotic approximants (informed by e.g., Alwan et al., 1997), where all productions of this class are understood as possessing "an anterior lingual constriction, accompanied by a midline lowering of the tongue behind this constriction". Such a characterization would appropriately describe the tongue tip constriction of a retroflex rhotic as well as the tongue body constriction of a bunched rhotic.

toward identifying which acoustic features are relevant to that confusion. Plauché et al. (1997) made two significant findings relevant to the stop confusions in Table 1.1. First, participants always categorized a [ki] token as /t/ when a stop-band filter from 2.5 to 5 kHz was applied to the stop burst, suggesting that information in that frequency region helps to inform the listener about the place of articulation of the production. Similarly, when the stop burst of [pi] was doubled in amplitude, listeners also always categorized the consonant of the syllable as /t/. Plauché (2001) also used decision trees, a machine learning classification tool, to identify acoustic features that aided the pairwise classification of consonant pairs.

The acoustic analyses performed in these studies inform those performed here. Chapter 3 further investigates the acoustic features informing place contrasts among voiceless stops. It also expands the scope of this investigation to include the voiceless fricatives /θ/ and /f/. By using random forest models (an extension of decision trees), spectral components critical to place contrasts among these consonants are identified. Chapter 3 also tests the relevance of these regions to human perception using a design not unlike that used by Plauché et al. (1997) and Chang et al. (2001). A behavioral experiment evaluates how the perceptibility of consonant place is affected by the frequency-localized loss of spectral energy. For the frequency bands identified by the random forest classifier, where energy differences in these regions aid classification between two consonants, loss of energy in this region is predicted to increase classification error for the 'louder' consonant of the two.

1.4 Overview of chapter 4: Addressing asymmetry

Confusability also does not guarantee perceptual asymmetry. When asked to identify the initial consonant of a CV syllable, listeners misidentified the dental fricatives /θ/ and /ð/ for one another with identical likelihood – 14.6% (Cutler et al., 2004). In building an account of perceptual asymmetry, it is necessary to explain why asymmetry occurs among certain confusable consonant pairs but not others. Several accounts of asymmetry have drawn upon similarity (Ohala, 1978; Stevens & Blumstein, 1978) in one form or another. Under these approaches, consonant pairs show confusability because the acoustics of one consonant are similar to those of another consonant in the context conditioning confusion. While similarity helps to explain confusability, it does not necessarily clarify why the confusions should favor either member of a pair. If /k/ and /t/ are similar to one another before [i], why are /t/ and /k/ not simply misidentified for each other equally often? Repp and Lin (1989) suggest a bias inherent to vowels, whereby an ambiguous consonant might tend to be perceived as having a certain place according to the vowel it is adjacent to. Phonotactic probability may also play a role, as listeners show differences in processing according to the frequency of sounds or sound sequences in their language. For example, /θ/ is less common than

/f/, which may contribute to listeners' tendency to prefer /f/. Cue-based approaches to asymmetry, as in Chang et al. (2001), suggest that asymmetry may arise when a contrast within a consonant pair relies on a non-robust cue to place. If this cue is not perceived (or carries little weight for the listener), then the favored category is most consistent with the remaining cues available.

Chapter 4 uses modeling to test a probability-based hypothesis about the source of perceptual asymmetry. As noted above, confusions between /k/ and /t/ favor /t/ before high front vowels. In phonetic contexts that condition perceptual asymmetry, productions of both consonants in the pair (e.g. [k] and [t]) are predicted to be more similar to the favored consonant category (/t/) than the disfavored consonant category (/k/). While the consonant pairs that show perceptual asymmetry are most similar to one another in the context that conditions asymmetry, they are not equally similar to their respective categories. This difference is predicted to contribute to differences in confusion rate between the two consonants.

Plauché (2001) set the stage for such an analysis. Using an approach rooted in a Bayesian framework, she was interested in determining whether category structure itself was able to explain why listeners showed asymmetrical confusion patterns in certain consonant pairs. Having extracted a variety of acoustic features for a stop, she identified features that correlated with differences in listener confusion rates. Productions of [ki], for example, were found to be confused for /ti/ more often when the VOT of [k] was closer to the mean VOT of /t/ in that phonetic context. She then analyzed the degree of overlap between the two categories along that featural dimension to see if the category disfavored in the asymmetry would show greater overlap than the other in that vocalic context. Her results were mixed. In the case of /k/-/t/ confusions, although VOT significantly predicted listener confusions of [k], the two categories overlapped symmetrically with one another in this feature. Her hypothesis was confirmed, however, for some other consonant pairs, including /p/-/k/ (before /u/).

One assumption of the model built in Chapter 4 is that individual token confusability can be predicted by its acoustic similarity to the perceptually disfavored and favored categories. This assumption is tested in a perceptual study in which listeners categorize isolated consonantal tokens. Listeners are predicted to show greater classification error on the tokens identified in the model as more confusable.

1.5 Overview of chapter 5: Perceptual asymmetry and sound change

Research into sound change, the long-term change in the phonetic norms of a speech community, has taken on a variety of forms ranging from studies of how change spreads through a community

to work on the factors that can allow for a change to begin in a specific community. Researchers have also studied why only a subset of seemingly possible sound changes are observed in practice. Velar palatalization, the diachronic change of a velar obstruent (e.g., /k/ or /g/) to an alveolar or post-alveolar obstruent (/t/ or /tʃ/ or their voiced counterparts) in the environment of a palatal segment or front vowel, is a well-attested sound change, observed in proto-Slavic (Gardiner, 2008) and early Old English (Hogg, 1979), among several other languages (see Guion, 1998 for a more extensive list). In contrast, a hypothetical change whereby an alveolar or post-alveolar obstruent takes on velar place of articulation in the same phonetic environment is unattested. (Garrett & Johnson, 2013) The unidirectional character of velar palatalization invites the question of whether the factors conditioning this change also explain why the change cannot happen in the opposite direction.

Differences in listener perception are afforded a prominent role in some theories of sound change (e.g., Ohala, 1981; Beddor, 2009). If an individual fails to account for the phonetic context that a speech sound appears in (perhaps due to a noisy signal), or if they utilize a listener-specific set of cues to identify the intended production of the speaker, they may end up identifying a production unlike the one the speaker had intended, in which case this novel form could be reproduced by the listener-turned-speaker and would have the potential to contribute to community-wide change. From the perspective of these and other theoretical approaches, perceptual asymmetry is an attractive object of study for its potential relation to listener-driven sound change because here too the production recovered by the listener differs from the production intended by the speaker. For example, the listener perceives [f] even though the speaker produced /θ/.

Furthermore, some cases of phonetic perceptual asymmetry in the laboratory have a family resemblance to sound changes. The confusion of /f/ and /θ/ (which favors [f]) resembles what is referred to as TH-fronting in, for example, Glaswegian English (Stuart-Smith et al., 2007). For those speakers, what had historically been produced as a dental fricative (e.g., think) is now produced as a labiodental fricative ([fɪŋk]). As another example, historical velar palatalization resembles confusions between /k/ and /t/ (which favor /t/ before high front vowels). The sound changes associated with these confusion patterns appear to generally proceed in one direction, just as the asymmetries identified in Table 1.1 tend to consistently favor one sound. Perceptual asymmetry provides another opportunity to better understand how a phonetic process might contribute to diachronic change.

The precise mechanism by which perceptual asymmetry could condition sound change remains unclear. As described above, perhaps listeners recover a different production than the one the speaker intended, and this innovative form is then reproduced—for example, listeners identify /θ/ as /f/ and reproduce [f] in future communications. An explanation of how perceptual asymmetry fits into sound change should ideally explain why many speaker communities do not undergo the

appropriate sound change. English listeners generally confuse /θ/ with /f/, even in varieties where /θ/ and /f/ are stable, distinct categories. If perceptual asymmetry helps to condition sound change, it also warrants explanation for when and why these categories show stability across time within speaker communities.

Chapter 5 uses agent-based model simulations to better understand how perceptual asymmetry might influence the acquisition and transmission of phonetic categories across generations. It tests the hypothesis that the pattern of confusion characteristically observed for asymmetry can itself contribute to the instability of the two category pairs. For example, /k/ and /t/ confusions favor /t/ before high front vowels, and a subset of simulations models the long-term outcome for /k/ and /t/ in this and other contexts over several generations. These models also explore a parallel question: how can these phonetic categories remain stable across time if perceptual asymmetry is still present? This chapter explores the differing assumptions under which category stability or change can be observed across time.

1.6 Overview of chapter 6: Discussion

In Chapter 6, the combined results of each line of inquiry are addressed, and implications for the future study of perceptual asymmetry are discussed.

CHAPTER 2

Perceptual Asymmetry and the Phonetic Primitive

2.1 Background

Members of the consonant pairs identified in Chapter 1 that show perceptual asymmetry differ in active articulator and place of articulation, despite their acoustic and perceptual similarity. The goal of this chapter is to investigate whether there is an articulatory characterization of these productions that mirrors the perceptual similarity of these consonants. §2.1.1 and §2.1.2 describe the articulatory time courses of [p], [t], [k], [θ], and [f], as well as their associated spectral characteristics. §2.1.3 describes the acoustic models drawn upon in this study. The design, hypotheses, and results of this experiment are reported in subsequent sections.

2.1.1 Voiceless stops

Stop articulation consists of the formation and eventual release of a complete closure within the oral cavity. The precise location of stop closures varies, but bilabial closure for [p] is formed about the lips; alveolar closure for [t] is estimated to be formed 1.5-2.5 cm posterior to the lips , and velar [k] approximately 6 cm from the lips (Stevens, 2000). Differences in closure location lead to differential acoustic outcomes over the course of each consonant.

Upon closure release, the initial movement of compressed air across the primary constriction location creates a burst, or transient signal. The spectral acoustics of the burst is typically governed by the shape of the cavity anterior to the constriction (Stevens, 1993). In fact, despite the short respective duration of bursts, listeners can classify stop place of articulation with an accuracy greater than chance using this information alone (Repp & Lin, 1989). The front cavity is smaller for an alveolar stop than for a velar production, and the resonance frequencies for the former

(approximately 4.5 kHz) consequently tend to be higher than those of the latter (approximately 1.7 kHz and 5.1 kHz). In contrast to [t] and [k], the noise source for [p] is at the anterior end of the vocal tract, so its spectral characteristics depend on the "rate of release of the closure" and "particular anatomical details concerning the lips" (Stevens, 2000, p. 353). Despite this variability, the transient of [p] can be characterized by a spectral peak at about 1-2 kHz.

Several milliseconds into stop release, two additional acoustic events follow. Continued airflow over the consonantal constriction creates frication noise; like the transient burst, the spectral acoustics of this noise depends in part on the cavity anterior to the constriction. Soon after, turbulence generated at the glottis leads to aspiration noise. Unlike the earlier acoustic events, aspiration spectral acoustics depend on the shape of the entire vocal tract, both ahead of and behind the primary constriction.

Descriptions of voiceless stop productions vary due to coarticulation with neighboring speech sounds. The coarticulation of voiceless stops with high front [i], high back [u], and low back [ɑ] are explored in this dissertation. These three vowels can be characterized by a palatal, velar, and pharyngeal constriction, respectively (Wood, 1982). The temporal overlap of the consonantal and vocalic constrictions can have an effect on how the consonant is articulated, as well as on the acoustics of the production. In a production of [ki], for example, [k] and [i] share the tongue body as the active articulator. Under this condition, the two articulations show some degree of blending (further discussion of gestural blending can be found in Browman and Goldstein (1989)). The constriction location for [k] before [i] is velo-palatal, further forward than in a non-front vowel context. The spectral characteristics of the burst, aspiration, and frication of these productions would differ correspondingly. If the consonantal and vocalic productions require different active articulators, the two constrictions may co-occur. For a production of [p], the lower lip is the active articulator; in [pi], the labial and palatal constrictions will not show gestural blending since they require different articulators. Instead, the two articulations may overlap with one another. In this case, the acoustics of aspiration will change due to the changed shape of the vocal tract, even though the labial articulation itself did not change.

Likewise, at the start of a vowel, the consonantal constriction is typically still narrow enough to alter the spectral properties of the neighboring vowel. The overlapping consonantal and vocalic constrictions affect the formant frequencies of the vowel. After the release of [p], the continued constriction at the lips lowers the resonance frequency of the anterior cavity for a vowel. A [t] production requires the tongue body to be far enough forward for the tongue tip to contact the alveolar ridge. Consequently, after stop release, the tongue tip must be lowered, and the tongue body may need to be sufficiently retracted to achieve the proper vocalic constriction, depending on the backness of the vowel (Stevens, 2000, p. 355). The changing sizes of the anterior and posterior cavities due to the movement of the tongue body and tip will affect the formants of the vowel. A

velar constriction would divide the vocal tract into two regions where the anterior cavity is about half the length of the posterior cavity. The resonant frequency of the anterior cavity is close to the half-wavelength resonant frequency of the posterior cavity in this context. For this reason, the second and third formant frequencies of vowels are similar near a [k] production, at least when the vowel is not front. When the vowel is front, the tongue body placement for the vowel further reduces the size of the anterior cavity, so the resonant frequency associated with this cavity is much higher.

2.1.2 Voiceless dental fricatives

This section describes the articulatory and acoustic properties of the targeted voiceless fricatives, interdental [θ] and labiodental [f]. During the production of a fricative, a narrow constriction is formed in the vocal tract such that airflow across the constriction is turbulent. As a result, the spectral characteristics of the fricative are driven largely by the cavities anterior to the constriction as well as by the noise source itself. However, as a fricative constriction widens for an adjacent vowel, the anterior and posterior cavities will show increased acoustic coupling, in which case both cavities can contribute more substantially to the overall spectral characteristics of the fricative.

During a production of [f], raising the lower lip to the upper incisors directs airflow along the upper lip at a volume velocity that causes turbulence. When modeled as a monopole noise source, the turbulence along the upper lip tends to have a spectral peak at around 2.5 kHz (Stevens, 2000). The acoustic cavity anterior to the constriction is only about 0.9 cm in length; the corresponding lowest resonance frequency for this cavity is 10 kHz. Like [p], the active articulator for [f] (the lower lip) differs from those of the vowels [i], [ɑ], and [u]. Consequently, the tongue position of [f] during varies quite a bit due to vocalic context. The tongue configuration for [f] in /hVfV/ productions approximates that of the neighboring vowel (Carney & Moll, 1971). Even within the same vocalic context, the cross-sectional area of the vocal tract showed a wide degree of variability for labiodental fricatives, with speakers' productions differing in the presence of tongue tip raising and the location of linguo-palatal contact (Narayanan et al., 1995).

[θ] has a similar articulatory plan to that of [f]. This articulatory similarity leads to comparable constriction locations and cross-sectional areas (Narayanan et al., 1995) and acoustic similarities, including in spectral peak mean and variance (Jongman et al., 2000). It is also likely the case that the resonant frequency of the anterior cavity for [θ] is like that of [f]. However, the two productions do have articulatory features that distinguish them. During production of [θ], the tongue root tends to advance, and the tongue tip typically slopes downward (Narayanan et al., 1995). Analogous to [t] productions, the advancement of the tongue root may be necessary for the tongue tip to achieve the correct location for [θ]. The exact amount of tongue body raising, and the exact location of the

tongue tip varies across speakers. The two productions also use different articulators, which may create constrictions with differing cross-sectional areas, and which may also direct air through the constriction at differing velocities. Either of these factors can cause the tongue tip constriction for [θ] to have different noise source characteristics than [f]. Such a difference would be predicted to influence the spectral acoustics of the production, although the noise source of the interdental fricative remains to be modeled.

2.1.3 Vocal tract models

This section describes two models that relate speech acoustics to the shape of the vocal tract. Even models that rely on a small number of spatial variables can be used to accurately predict certain features of human speech. One such model is the Helmholtz resonator, characterized as a cavity with a narrow opening at one end, which can be used to approximate the shape of the vocal tract when one end can be treated as closed (e.g., due to a glottal constriction for voicing) while the other end is terminated by a narrowed opening (e.g., lip rounding), as in the production of [ʌp]. In such a case, the resonant frequency associated with this structure is described in Equation 2.1, where S is the cross-sectional area of the opening, L' is its effective length, and V is the volume of the chamber. This model can also be used to describe the acoustics of information relevant to burst and frication spectral acoustics.

$$f_1 = \frac{c}{2\pi}\sqrt{\frac{S}{L'V}}$$

Equation 2.1: Resonant frequency of a Helmholtz resonator (Kinsler et al., 1983)

As mentioned in §2.1.1, the tube model of the vocal tract can be used to describe vowel and stop aspiration acoustics. Under this model, the length of the vocal tract is understood as broken up into uniform tubes of varying length and cross-sectional area. A production of [i], for example, is characterized by a narrow constriction at the hard palate. A model that approximates this production would be composed of a short wide tube (corresponding to the region of the vocal tract behind the palatal constriction), a long narrow tube (corresponding to the palatal constriction itself), and a short, wide tube (corresponding to the region of the vocal tract anterior to the constriction). Each tube has its own resonant frequency, as determined by its length and cross-sectional area, and these frequencies interact with one another because the tubes are coupled as a single system. The dimensions of the narrowed tube also contribute to the spectral characteristics of the vowel, comparable to the effect of the narrowing in the Helmholtz resonator. This chapter makes use of this model to extract articulatory information in the consonant productions that are relevant to the spectral acoustics of stop aspiration.

2.2 Research question and hypothesis

The motivating research question for this study is the following:

RESEARCH QUESTION

For consonants that show perceptual asymmetry, is acoustic similarity mirrored by similarity in the geometry of the vocal tract?

Perceptual asymmetry presents a situation where dissimilar articulatory events produce similar perceptual outcomes. The pair [p] and [t] are highly confusable for one another before high front vowels despite involving a lip constriction and a tongue tip constriction, respectively. Perhaps these two events produce similar vocal tract shapes in an [i] context. In cases where the vocal tract shape is ambiguous between multiple articulatory events that could have caused it, the listener may in fact recover one of the two or more possibilities.

Differences in acoustic coupling limit which regions of the vocal tract influence speech acoustics. During stop burst and frication, for example, the characteristically narrow degree of constriction can cause the region of the vocal tract in front of the constriction to become acoustically decoupled from the region behind it (Stevens, 2000). Consequently, the spectral characteristics of both depend largely on the shape of the oral cavity anterior to the constriction, while the posterior cavity holds negligible influence. For frication and stop bursts, the dimensions of the anterior cavity are predicted to hold the greatest similarity in the phonetic context conditioning perceptual asymmetry. The respective dimensions of the anterior cavity of vocal tract for [k] and [t], for example, are predicted to show the greatest similarity before high front vowels. In this sense, the ambiguity in vocal tract shape would make it hard for listeners to figure out whether it was produced by a velar or alveolar production.

The same prediction applies to aspiration – the spectral properties of this event are sensitive to the dimensions of the anterior and posterior cavities of the vocal tract. Accordingly, when comparing consonants in a pair, their dimensions are predicted to differ the least in the vocalic context conditioning asymmetry.

The general hypothesis for Experiment 2.1 is the following:

HYPOTHESIS

Each consonant pair will show the greatest spatial similarity in the vocalic context(s) that condition perceptual asymmetry.

2.3 Experiment 2.1

2.3.1 Dataset

Data for Experiment 2.1 are from the USC Speech and Vocal Tract Morphology Database (Sorensen et al., 2017). Seventeen participants (8 male and 9 female) produced a variety of English utterances and nonsense syllable sequences. The productions were recorded using real-time MRI, which provides video of the mid-sagittal plane of the participants' oral and pharyngeal cavities during their speech. The average sampling rate of this video was approximately 23.18 frames/s. Participants were also simultaneously recorded using a fiber-optic microphone. This audio was sampled at 100 kHz at the time of recording but was down-sampled to 20 kHz. The audio also underwent noise reduction in post-processing (see Bresch et al., 2006 for methodology) to remove additional sound sources present during the MRI recording.

2.3.2 Stimuli

Participants produced /V_1CV_1/ syllables where the vowel was either [æ-ɑ] , [i], or [u]. The consonantal contexts selected for this study were [p], [t], [k], [θ], and [f], corresponding to the obstruent consonants for which perceptual asymmetry has been observed. Two repetitions of each consonant in each unique vocalic context were collected, for 510 tokens total.

2.3.3 Analyses

For stop productions, the first video frame after constriction release (i.e., up to approximately 40 ms post-release) was extracted for analysis. For fricative productions, the first video frame before constriction release was extracted for analysis. Additional frames after constriction release were extracted, but the sampling rate of the video is such that the second and third extracted frames correspond to at least about 80 and 120 ms after constriction release, respectively. While these later frames do not provide information about the spectral characteristics of stops or fricatives, they can provide a means to contrast the geometry of the vocal tract during constriction release with its shape during the vowel.

For each video frame, a vocal tract area function, corresponding to the cross-sectional of area of the vocal tract airway from the glottis to the lips, was generated using the following procedure, based on Takemoto et al. (2006).

2.3.3.1 Steps to generate vocal tract area function

Five steps were taken to determine the area function of the vocal tract:

1. The surfaces of the vocal tract were traced manually using the Matlab tool *GetContours* (Tiede & Whalen, 2015), as illustrated in Figure 2.1.

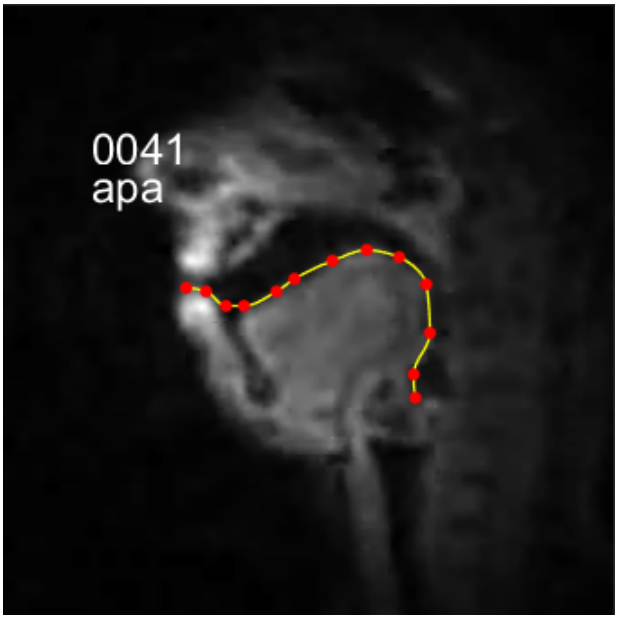

Figure 2.1: Manual trace of a participant's lower vocal tract surface during production of [ɑpɑ]

2. An active region growing algorithm (a method to segment the image into airway and non-airway portions) was implemented such that airway pixels further away from the vocal folds received a higher value if they were above the vocal folds and a lower value if they were below. Because pixels with similar values tend to be located close to one another, a roughly monotonic gradient of pixels is defined along the vocal tract.

3. A centroid was computed for each set of pixels with the same value, and a polynomial spline was fitted through these centroid values. This spline served as the vocal tract midline.

4. Thirty points evenly spaced along the vocal tract midline were identified from the glottis to the lips, and at each point, the width of the airway was calculated through a line perpendicular to the vocal tract midline.

5. An individual MRI video frame only provided information on the shape of the vocal tract along the mid-sagittal plane. If the airway were perfectly circular along the vocal tract midline, the cross-sectional area of the tract along the midline could be estimated. The cross-sectional area at a given point was calculated as $0.029\pi \left(\frac{d}{2}\right)^2$ cm^2, where d was the pixel-wise width of the airway at a given point, and 0.029 cm^2/pixel corresponds to spatial resolution of the video.

2.3.3.2 Steps to calculate anterior and posterior cavity dimensions

As discussed in §2.1.3, the spectral characteristics of stop burst and frication noise are primarily associated with the geometry of the vocal tract anterior to the consonantal constriction. In contrast, the spectral characteristics of aspiration are sensitive to the geometry of cavities ahead of and behind the vocalic constriction. Accordingly, the average lengths and cross-sectional areas of the anterior and posterior cavities were extracted using the following process:

1. The primary consonantal and vocalic constrictions were identified as the point along the vocal tract midline with the smallest cross-sectional area appropriate to that of the consonant or vowel, respectively.

2. Due to reverberation in the audio recording, it was not possible to acoustically verify which video frame corresponds to burst release or aspiration. Instead, under the assumption that a wide primary constriction was unlikely to be associated with frication, a maximum primary constriction cross-sectional area (0.5 cm^2) was selected; tokens with higher values at all points were excluded. The interval from the front of the mouth to the first point with cross-sectional area less than 0.5 cm^2 was identified as the anterior cavity. Analogously, the interval from the back of the mouth to the last point with cross-sectional area less than 0.5 cm^2 was identified as the posterior cavity.

   Figure 2.2 gives the vocal tract area function of a participant's production of [t] before [ɑ]. The region in green corresponds to the location of the consonantal constriction. The region in red corresponds to the location of the anterior cavity.

Figure 2.2: Vocal tract area function of [t]. Boundary of cavity anterior to consonantal constriction marked in red. The region corresponding to the alveolar constriction is marked in green.

Figure 2.3 plots the same production of [t] before [ɑ]. In this figure, however, the region in green corresponds to the location of the vocalic constriction, and the regions in red and blue correspond to the locations of the constrictions anterior and posterior, respectively, to the vocalic constriction.

Figure 2.3: Vocal tract area function of [t]. Boundaries of cavities anterior and posterior to primary vocalic constriction marked in red and blue, respectively.

For most consonants addressed in this study, the vocalic and consonantal constrictions tend to occur at dissimilar locations. In contrast, for [k], the vocalic and consonantal constrictions tend to be located in similar places, which can affect the dimensions of the vocal tract in ways relevant to the acoustics of aspiration and burst release.

3. As noted earlier, the frequency characteristics of resonating chambers are sensitive to the cross-sectional areas and lengths of the corresponding cavities as separate variables. Despite the theoretical independence of length and cross-sectional area, the vocal tract is limited in how it can change shape. As a result, patterns emerge when considering cross-sectional area as a function of cavity length. For the anterior cavity, the cross-sectional area is significantly correlated with cavity length $[r = 0.61, t(1387) = 28.3, p < 0.001]$ – as the length of the anterior cavity increases, the cross-sectional area of the cavity also increases. The cross-sectional area of the posterior cavity is also significantly correlated with its length $[r = -0.22, t(1387) = 8.35, p < 0.001]$ – as the length of the posterior cavity increases, its cross sectional area decreases. Length is the primary spatial measure reported in this study, and this analysis describes how these two length measures can interact in consonants by vocalic context.

4. The total distance was computed for points along the vocal-tract midline corresponding to the anterior or posterior cavity.

2.3.4 Measures

Two measures were taken for each vocal tract frame extracted for this study:

PAIRWISE ABSOLUTE DIFFERENCE IN LENGTH

For each speaker, consonant pair, and vocalic context, I calculated the absolute value of the difference in length of the anterior cavity (relevant to the consonantal constriction) between one consonant and another in the same vocalic context (e.g., [k] and [t] before [i]). This measure serves to capture the relevant spatial characteristics of the vocal tract for stop burst noise and frication noise.

EUCLIDIAN DISTANCE OF CAVITIES' LENGTHS

Because the spectral acoustics of aspiration are influenced by the shape of the anterior and posterior cavities (relevant to the vocalic constriction), the dimensions of both regions must be considered for stops (though not for fricatives). To compare both cavity dimensions for aspiration, the Euclidean distance between two stop productions was computed with respect to these two measures, generating a measure for each speaker, consonant, and vocalic context.

2.3.5 Predictions

The following consonant pairs are analyzed in this study:

[k]-[t]: Before [i], [k] tends to be misidentified as /t/ more often than [t] as /k/ (e.g., Plauché et al., 1997; Guion, 1998). Accordingly, [k] and [t] are predicted to show greater spatial similarity in the context of [i] than in the context of [u] and [ɑ]. The pairwise absolute difference and Euclidian distance measures between [k] and [t] should be smallest in the context of [i].

[k]-[p]: [k] and [p] tend to be misidentified for one another more often in the context of a high vowel than in the context of other vowels (Winitz et al., 1972; Plauché, 2001), though the exact directionality of the asymmetry according to each vocalic context ([i], [u]) is not clear. [k] and [t] are predicted to show a greater spatial similarity before a high vowel than before [ɑ]. The pairwise absolute difference and Euclidian distance measures between [k] and [t] should be smaller before high vowels.

[p]-[t]: Like the consonant pair [k]-[t], before [i], [p] tends to be misidentified as /t/ more often than [t] as /p/ (e.g., Winitz et al., 1972). Accordingly, [p] and [t] are predicted to show a greater spatial similarity, and the smallest pairwise absolute difference and Euclidian distance measures, in the context of [i] compared to the context of [u] and [ɑ].

[θ]-[f]: Listeners tend to misidentify [θ] as /f/ (e.g., Miller & Nicely, 1955 Cutler et al., 2004), but the existing literature demonstrating this result does not also explore the effect of phonetic environment. This gap in the literature could in fact be indicative of a lack of context conditioning this confusion, but it is unclear how to interpret it without explicit perceptual study.

2.3.6 Results

The results for this experiment are organized by consonant pair. Pairwise t-tests with Holm correction were taken for each measure and consonant pair, with the measure as the dependent variable and vocalic context as the independent variable.

2.3.6.1 [k]-[t]

As predicted based on the context-dependent perceptual asymmetry of [k] and [t], and as visualized in Figure 2.4, the pairwise absolute difference for these stops was smaller in the context of [i] than in the contexts of [ɑ] $[t(27.38) = 3.44, p = 0.006]$ and [u] $[t(24.41) = 2.71, p = 0.02]$. The pairwise absolute difference for [k] and [t] was also not significantly different in the context of [u] than in the context of [ɑ] $[t(20.40) = 1.79, p = 0.08]$.

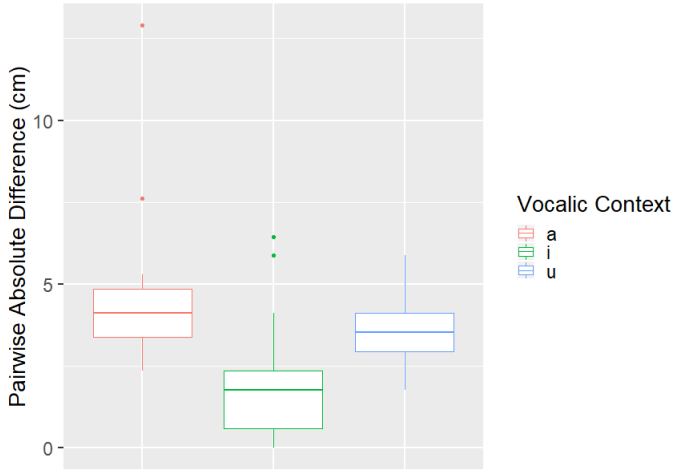

Figure 2.4: Pairwise absolute difference for [k] and [t] by vocalic context

Also as predicted, the Euclidean distance measure for [k]-[t] was smaller in the context of [i] than in the contexts of [ɑ] $[t(25.01) = 8.60, p < 0.0001]$ and [u] $[t(24.53) = 2.62, p = 0.02]$. The Euclidean distance measure was also smaller context of [u] than [ɑ] $[t(29.97) = 4.84, p < 0.0001]$, though this outcome was not predicted. These results are plotted in Figure 2.5.

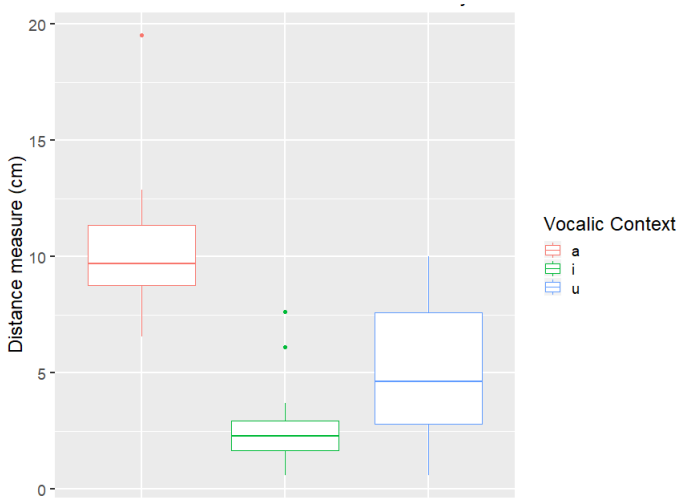

Figure 2.5: Euclidean distance measure for [k] and [t] by vocalic context

2.3.6.2 [k]-[p]

All results of the pairwise difference measure followed predictions for [k] and [p]. This measure was larger in the context of [ɑ] than in the contexts of [i] $[t(27.42) = 3.35, p = 0.007]$ and [u] $[t(23.37) = 2.80, p = 0.02]$. This measure did not differ between the contexts [u] and [i] [t(27.96)=1.04, p=0.30]. These results are plotted below in Figure 2.6.

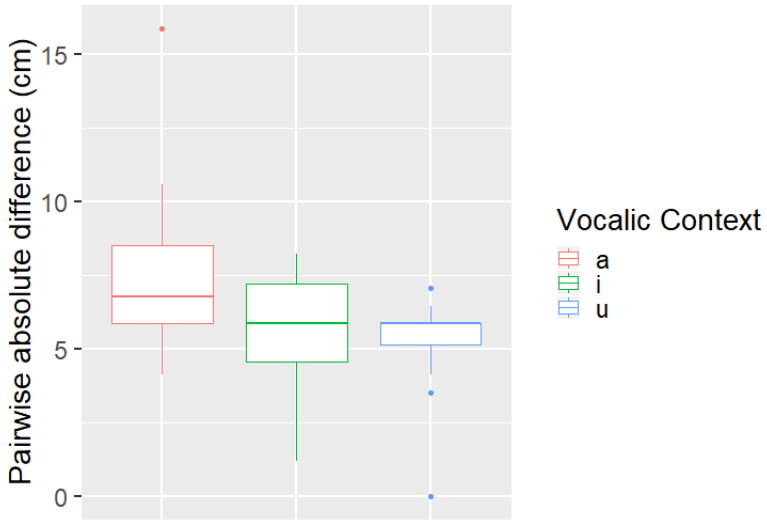

Figure 2.6: Pairwise absolute difference for [k] and [p] by vocalic context

Similarly, the Euclidean distance measures were smaller in the context of [u] than [ɑ] $[t(16.67) = 4.18, p = 0.001]$ and smaller in the context of [i] than [ɑ] $[t(16.16) = 4.61, p = 0.001]$. These measures did not differ between the contexts [u] and [i] $[t(29.02) = 1.33, p = 0.19]$.

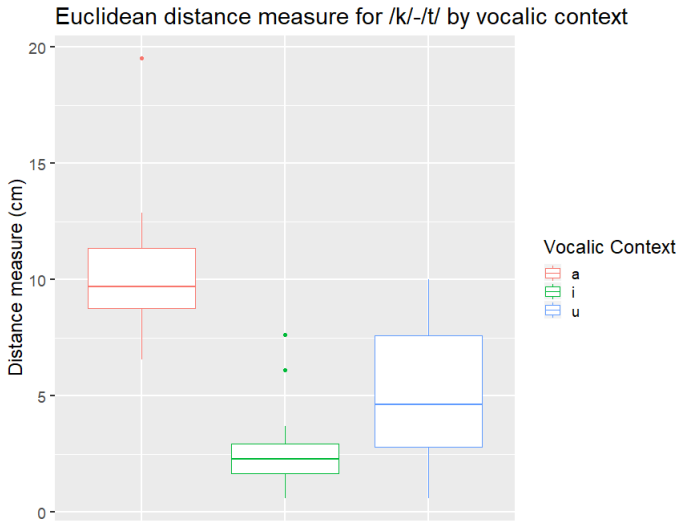

Figure 2.7: Euclidean distance measure for [k] and [p] by vocalic context

2.3.6.3 [p]-[t]

Figure 2.8 illustrates that, contrary to predicted effects, the pairwise absolute difference measures for [p] and [t] did not differ between [i] and [ɑ] $[t(29.82) = 1.07, p = 0.29]$, and was larger before [i] than before [u] $[t(20.62) = 2.95, p = 0.02]$. This measure did not differ between the contexts of [ɑ] and [u] $[t(19.86) = 2.39, p = 0.05]$.

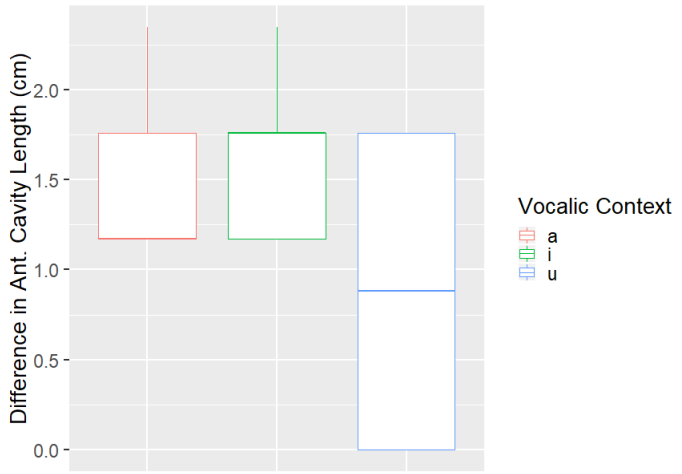

Figure 2.8: Pairwise absolute difference for [p] and [t] by vocalic context

The Euclidean distance measure for [p]-[t] was smaller in the context of [i] than in the contexts of [u] $[t(29.84) = 2.46, p = 0.02]$ and [ɑ] $[t(18.70) = 4.32, p = 0.001]$, as predicted. The Euclidean distance measure was also smaller context of [u] than [ɑ] $[t(19.26) = 3.08, p = 0.01]$,

though this outcome was not predicted. These results are plotted in Figure 2.9.

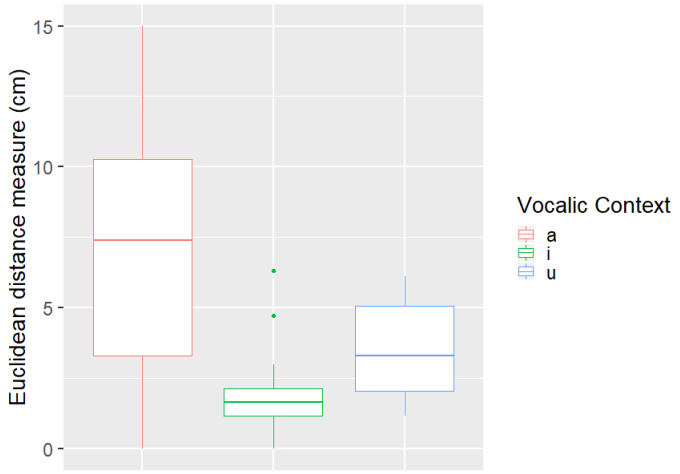

Figure 2.9: Euclidean distance measure for [p] and [t] by vocalic context

2.3.6.4 [θ]-[f]

Unlike for the previous pairs, there are no perceptual results from which to make a prediction about the vocalic context where [f] and [θ] should show the smallest difference. The pairwise absolute distance measure for [f] and [θ] was smaller before [u] than [i] $[t(20.16) = 4.22, p = 0.001]$. This measure did not differ between the contexts [u] and [ɑ] $[t(23.40) = 2.17, p = 0.08]$ or [i] and [ɑ] [t(24.47)=2.17, p=0.08]. Results are plotted in Figure 2.10.

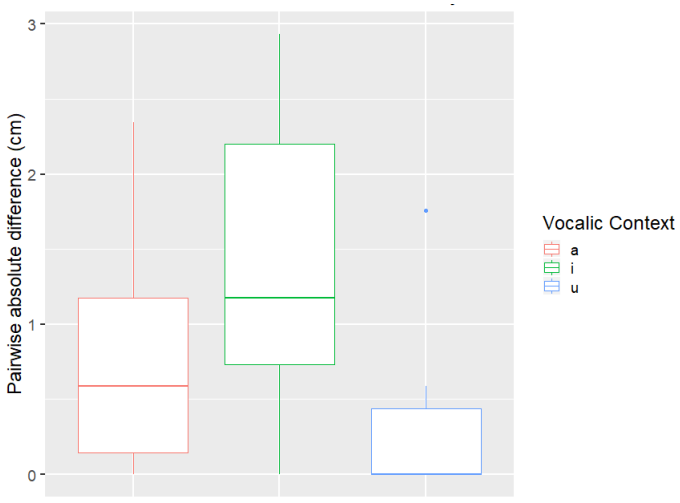

Figure 2.10: Pairwise absolute difference for [θ] and [f] by vocalic context

2.4 Discussion

Consonant pairs that show perceptual asymmetry tend to be acoustically similar despite differences in the exact articulatory events that take place during their production. The goal of this study was to see if there was an articulatory description of members of four targeted consonant pairs whose similarity mirrored their (context-dependent) acoustic similarity. The results of this study are summarized below in Table 2.1.

The spectral characteristics of stop burst noise are sensitive to the geometry of the cavity anterior to the primary constriction. Consequently, the three targeted stop pairs were predicted to show greatest similarity in the spatial features relevant to burst spectral characteristics in the vocalic environments conditioning confusion. This prediction was partially supported by the results: the pairwise absolute difference in anterior cavity length was smaller for [k]-[t] and [k]-[p] in the vocalic environments that favored confusion. For [p]-[t], however, the context of [u] actually showed greater similarity than that of [i].

The spectral characteristics of aspiration are sensitive to the shapes of the cavities both ahead of and behind the constriction. Stop pairs were predicted to show the greatest similarity in the spatial features relevant to aspiration in the vocalic contexts that favor confusability. This prediction was supported by the Euclidean distance results for all consonant pairs.

For [k]-[t] and [p]-[t], the Euclidean distance measure shows greater similarity in the context of [u] than in the context of [ɑ], which, while not inconsistent with predictions, was unexpected. This result may be suggestive of perceptual asymmetry of these consonants before [u], but an increase in [k]/[p] and [t] confusions in this context has only been observed in Plauché (2001). Most of her participant population were in their 20s and from California, a region characteristically associated with [u]-fronting. Compared to a back [u], a fronted [u] has a constriction location closer to that of [i], and so its coarticulatory effect on [k] and [t] might be expected to resemble [i] in articulation and perception. While the population recorded for this study was not majority Californian, [u] fronting might also help to explain such a result. The audio associated with the MRI video in this study was unfortunately noisy due to the activity of the MRI scanner. In future work where audio quality presents less of an issue, it would be worth investigating the effect of gradient /u/-fronting on the articulation of neighboring consonants.

CONSONANT PAIR	PREDICTED CONTEXT FOR SMALLEST DIFFERENCE	ANTERIOR CAVITY LENGTH DIFFERENCE (FOR BURST AND FRICATION)	EUCLIDEAN DISTANCE MEASURE (FOR ASPIRATION)
[k]-[t]	[i]<[u],[ɑ]	[i]<[u,ɑ]	[i]<[u]<[ɑ]
[k]-[p]	[i,u] <[ɑ]	[i,u] <[ɑ]	[i,u] <[ɑ]
[p]-[t]	[i] <[u],[ɑ]	[u] <[i]	[i] <[u] <[ɑ]
[θ]-[f]	*no context attributed*	[u] <[i]	-

Table 2.1: Predictions and results for each consonant pair; green highlighting indicates the results were consistent with predictions; red highlighting indicates the results were inconsistent with predictions

2.4.1 Delving deeper into the dental fricative results

Like stop burst noise, the spectral characteristics of dental frication are sensitive to the shape of the cavity ahead of it. Unlike the stop pairs, however, no conditioning phonetic environment for [θ]-[f] confusion has been identified in the literature. As a result, it was not possible to use perceptual results to predict which phonetic environment should show the smallest pairwise absolute difference measure.

For the dental fricatives, the pairwise absolute difference measure in the context of [u] was smaller than in the context of [i]. It may be difficult to tease apart this result from the effect of how constriction location and cavity boundaries are identified. In the context of [u], a production of [θ] involves a lingual and labial constriction that are located close to one another in the vocal tract. Because the boundary of the anterior cavity is identified as the location where the cross-sectional area achieves above a minimal cross-sectional area, the presence of an additional constriction nearby may prevent the location of the boundary from being consistently identified.

The location of the posterior edge of the consonantal constriction can provide information about the location of consonantal constriction with needing to deal with the presence of a complicating coarticulatory labial constriction. The method of identifying this boundary is identical to the method described in §2.3.3.2, except for one change. Figure 2.11 includes two vocal tract area functions corresponding to productions of [θ] in the context of [i]. The gray dashed line corresponds to the threshold measure used in §2.3.3.2 to calculate the boundary of anterior and posterior cavities of the vocal tract. In both panels, a dental and palatal constriction can be observed, but in the second panel, these two constrictions blend into one long constriction if its bounds are defined at 0.5 cm^2. For this area function (and other productions in the corpus), the dorsal and coronal constrictions overlap enough that a narrower cross-sectional area criterion is necessary in order to specifically identify the posterior boundary of the consonantal constriction.

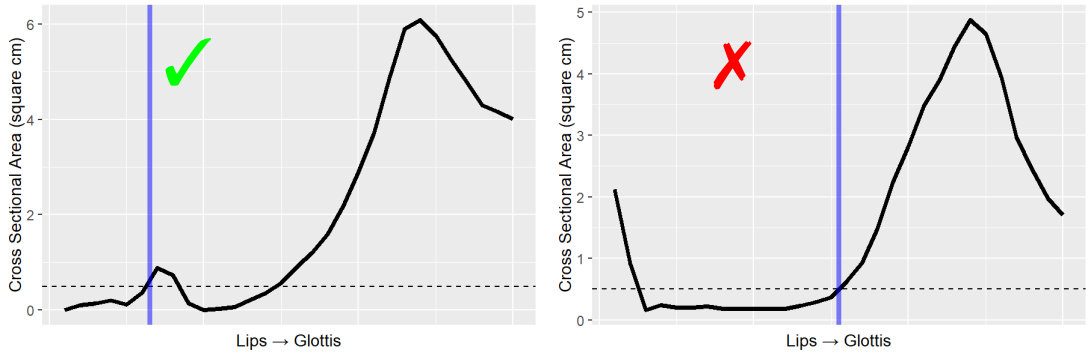

Figure 2.11: Estimated posterior boundary of constriction for [θ] at a threshold cross-sectional area of 0.5 cm^2

The black dotted line in both panels of Figure 2.12 represent this choice – 0.25 cm^2. At this value, the posterior boundary of the vocalic constriction is not misidentified as part of the consonant, nor does the labial constriction factor in meaningfully into the identification of the dental constriction location.

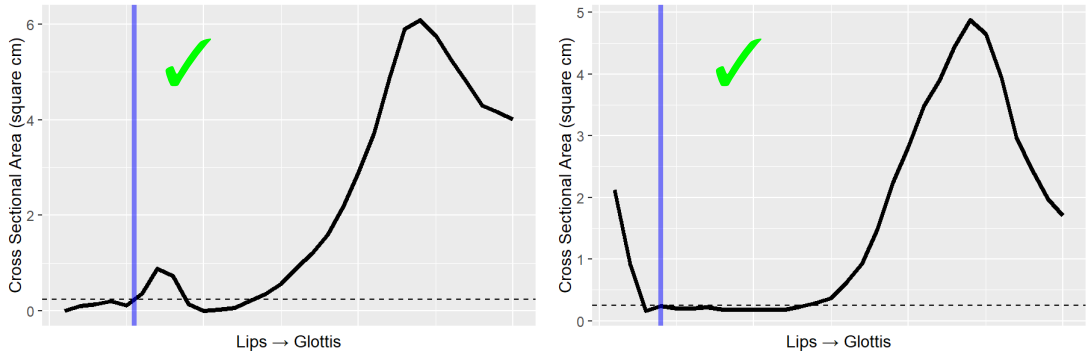

Figure 2.12: Estimated posterior boundary of constriction for [θ] at a threshold cross-sectional area of 0.25 cm^2

In fact, Figure 2.13 shows that [θ] has a significantly smaller posterior cavity length (indicating retraction) before [i] than before [u] $[t(20.94) = 2.40, p = 0.03]$, but no significant differences between the other vocalic contexts. In contrast, there is no significant difference in posterior cavity length according to vocalic context for productions of [f]. Although the anterior cavity length measure is somewhat complicated by the presence of a labial constriction, productions of [θ] do appear to have a slightly retracted place of articulation before [i] relative to [u]. The dental fricatives differ in their active articulator. The labial articulation of [f] does not constrain the movement of a tongue for any of the vocalic constrictions. In contrast, the tongue tip constriction of [θ] may constrain the movement of the rest of the tongue, which is used to articulate the vowels included in this study. Although [i] and [u] both involve movement of the tongue dorsum, the two vowels appear to show

differing degrees of variability according to phonetic context. The Degree of Coarticulatory Resistance model (Recasens et al., 1997) makes explicit predictions about how much variability a speech sound will show due to coarticulation. Within this model, [i] is highly constrained, relative to the back vowels [ɑ] and [u], and so shows little variability according to phonetic context relative to the two back vowels. This statistical result can be interpreted in a similar light – although [i] has a more anterior lingual constriction location than [u], the tongue dorsum constriction in [i] constrains the movement of the tongue tip to such a degree that the constriction location of [θ] is retracted.

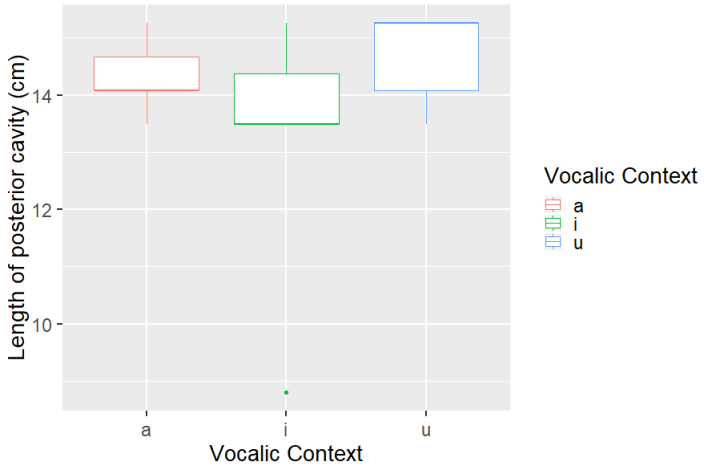

Figure 2.13: Posterior cavity length of [θ] by vocalic context

2.4.2 Burst results for [p] and [t]

The results for the [p] and [t] comparison would seem to resemble those for the dental fricatives in this experiment. Like [f] and [θ], [p] and [t] involve a labial and tongue tip constriction, respectively. Similarly, differences in anterior cavity volume for this consonant pair were smaller before [u] than [i]. An analysis similar to what was undertaken for the dental fricatives may be useful to clarify whether this result might reflect coarticulatory retraction on the part of [t]. Unlike [θ], the anterior cavity lengths of [t] do not differ significantly across any vocalic context, as seen in Figure 2.14, nor do those of [p]. Although [t] and [θ] are both produced with a tongue tip constriction, the two may differ their relative degree of articulatory constraint. These results would suggest that the constriction location of [t] varies less according to vocalic context than [θ], but there is as of yet little research into how these two consonants compare (see e.g., Fowler & Brancazio, 2000 for a discussion of the resistance of [d] and [ð] to vowel-to-vowel coarticulation).

An analysis of the posterior cavity length of [t] productions does not support the conclusion that [t] shows any difference in retraction according to vocalic context. Consequently, the apparent

increase in the similarity of the anterior cavities of [p] and [t] may be due to the complicating effects of the labial constriction on the calculation of the anterior edge of the consonantal constriction. For both consonants, the vocalic constriction lies behind both consonantal constrictions. Accordingly, absent an additional type of coarticulatory influence of the vowel on the consonant, the spectral characteristics of the burst would not be sensitive to any spatial feature that would vary by vocalic context.

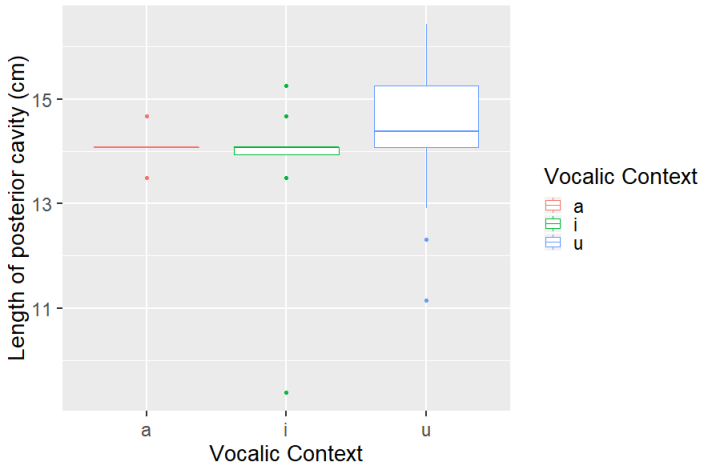

Figure 2.14: Posterior cavity length of [t] by vocalic context

2.4.3 The phonetic primitive

Perceptual asymmetry can inform gestural phonetic theories by providing an opportunity to better understand perceptual ambiguity. Taking a direct realist stance, Fowler (1996) describes articulatory events in speech as structuring the acoustic signal in an ordered manner such that the properties of the causative event can be recovered by listeners. In cases where listeners tend to confuse consonants that involve distinct active articulators and constriction locations, the ambiguity may in fact still exist in a space relevant to articulation – different types of productions may produce similar vocal tract configurations in certain phonetic environments, which then structure the acoustic signal similarly. Such an analysis may be possible for English rhotic approximants, as described in Chapter 1. Zhou et al. (2007) found that 'bunched' and retroflex rhotic approximants have similar vocal tract area functions despite differences in the articulators used to produce each rhotic. A gestural description that unifies these productions, as described in van Lieshout et al. (2008), may also involve a characterization that is ambiguous for the exact articulator used to generate the constriction target. While only a restricted set of articulators can produce two of the constrictions characteristic of rhotic approximants – a labial constriction (i.e., the lips) and pharyngeal constriction (i.e., the tongue root) – there are multiple ways that an anterior lingual constriction

can be produced (i.e., curling the tongue tip toward the palate or bunching the tongue body, among others).

The results reported in this chapter suggest that general vocal tract shape may be a relevant characteristic to consider when investigating confusions between productions that cross boundaries in terms of active articulator. The consonant pairs [k]-[t] and [k]-[p] both show greater similarity in their vocal tract geometries in the contexts that condition perceptual asymmetry. This articulatory description aligns closely with the contextual confusability of those consonant pairs.

CHAPTER 3

An Analysis of Spectral Features Contributing to Perceptual Asymmetry

As noted in Chapter 2, perceptual asymmetry in consonants is limited to specific consonant pairs and vocalic contexts. Some research (e.g., Plauché et al., 1997) has been undertaken to describe the acoustic factors that may explain distribution of this phenomenon across vocalic contexts. This chapter continues this work by identifying discrete spectral components in the acoustic signal that listeners find perceptually relevant when discriminating between consonant pairs that show perceptual asymmetry. The chapter begins with a review of the acoustic features used to distinguish place of articulation in stop and fricative consonants, followed by a short discussion of phonetic misperception and of methodologies that have been used to help identify features informative to perception. The main research questions of the chapter (§3.2) are addressed in two experiments: an exploratory analysis of the spectral properties of consonants that show perceptual asymmetry (§3.3), and a perceptual study that investigates listener perception of band-filtered speech (§3.4). The results of the two studies are discussed in §3.5.

3.1 Background

3.1.1 Perception of place in stops

A stop production is associated with several acoustic events including closure, burst release, aspiration, and the transition into and out of adjacent speech sounds. An articulatory description of each of these events can be found in §2.1.1. Different researchers have taken different, though not incompatible, approaches to the analysis of how these events might inform the listener about place of articulation.

One common approach (undertaken e.g., in Liberman et al., 1952, Dorman et al., 1977) has

been to treat the components of a stop production as possessing distinct cues to place of articulation localized in the time domain (and potentially the frequency domain). In a pioneering study on stop perception, Cooper et al. (1952) presented listeners with a variety of voiceless stop-vowel sequences where the peak frequency of the burst and the formant structure of the vowel in each syllable were independently varied. They found that listeners were sensitive to both cues when judging place of articulation. Similar results were observed in the perception of naturalistic speech as well. Schatz (1954), for example, presented listeners with /sCVC/ spliced syllables composed of a [s]-stop onset (isolated from an /sCV/ syllable) and a vowel-consonant rhyme. The place of articulation in the onset stop (as identified by the listener) depended on the place of articulation of the stop, the vowel it was spliced onto, and the vowel of the syllable it was extracted from.

Another approach centers around the analysis of stop events again as distinct cues, but with less granularity in the time and frequency domains. In such an approach, gross spectral differences may provide information about differences in place of articulation. Blumstein and Stevens (1979) proposed a set of spectral templates, intended to summarize a stop spectrum according to its compactness and the degree to which it increases in amplitude with frequency. Algorithms using this derived feature to classify stop place of articulation achieve an accuracy of well above chance (Blumstein & Stevens, 1979) and there is some evidence that listener perception of place also shows sensitivity to this measure (Plauché, 2001). Burst amplitude has also been identified as a cue that may play a role in perception, with tokens of [p] being more often perceived as [t] when the amplitude is artificially increased (Ohde & Stevens, 1983).

Researchers have also approached the perception of stop place by investigating time-varying features (see Forrest et al., 1988 for a description of a time-varying moment measure). Kewley-Port and Luce (1984), for example, identified three informative features, one of which (spectral tilt of burst onset) could be understood as temporally local just like Blumstein and Steven's spectral templates or burst peak frequency. The other two measures, however, took spectral changes across time into account: late onset of low frequency, corresponding to the eventual presence or absence of an F1 peak, and mid-frequency peaks extending over time, corresponding to the presence or absence of a peak between 1-3.5 kHz which lasts for longer than a certain threshold. Classification of stops using these criteria also achieved an accuracy of well above chance.

The analysis undertaken in this chapter most closely resembles the first approach, where information relevant to stop place of articulation is localized in the frequency domain. Such cues could, however, be easily incorporated into a framework that looks at frequency-general or time-varying features. A difference in peak frequency over a certain band, for example, could correspond to a change in the overall shape of the spectrum (and therefore its corresponding spectral template), or could be one of several spectral frames considered in time-varying measure.

Cooper et al. (1952), among many others, noted that the relation between articulation and

acoustics can be complex with respect to stop place of articulation. Describing the perception of synthesized syllables, the authors wrote that, "...bursts at 1440 cps are heard as identical sounds in acoustic terms, but they are heard as different speech sounds when paired with different vowels, e.g., pi, ka, pu." In this circumstance the same burst seemed to provide different information about place depending on the adjacent vowel. Analyses of stop perception have differed in the degree to which cues to stop place are assumed to vary with respect to phonetic context. Approaches range from a consideration of stop cues as invariant with respect to place of articulation (e.g., Stevens & Blumstein, 1978; Cole & Scott, 1974) to approaches that question the possibility of acoustic invariance for certain types of cues (e.g., locus equations in Fowler, 1994).

This chapter takes an intermediate approach – in identifying salient cues to stop place of articulation, cues are specified with respect to the contrast (e.g., cues that help to distinguish [k] from [t]). There is no assumption that the cues used to distinguish [k] from [p] are related to the cues used to distinguish [k] from [t]. However, this analysis does make the implicit assumption that a consonant pair relies on similar cues to a contrast across vocalic contexts. A parallel analysis that looks at consonantal cues specific to vocalic context would provide insight on the degree to which this assumption is reasonable.

3.1.2 Perception of place in fricatives

An articulatory description of dental fricatives is given in §2.1.2. Like stops, fricatives have distinct acoustic events associated with their production – specifically, frication and the transition into and out of adjacent segments. Spectral and non-spectral measures have both been used to characterize place of articulation for these consonants.

Measures related to the shape of the frication spectrum have been used to identify the place of articulation of a fricative. For example moment measures (i.e., spectral mean, variance, skew, and kurtosis) have also been taken for frication spectra; Jongman et al. (2000) found that English fricatives differ across several of these measures over the duration of the production; labiodental and interdental fricatives were distinguished by skewness and kurtosis (in addition to peak frequency).

Researchers have utilized additional measures to characterize fricative place including locus equations (e.g., Jongman et al., 2000), formant transitions (Harris, 1958), frication duration (Jongman, 1989), and frication amplitude. This chapter primarily examines time-localized spectral differences in frication between labiodental and dental fricatives. This choice is one of a variety that could have been made to uncover the differences between these fricative places. A time-varying analysis of frication spectra, or an analysis enriched with non-spectral information would likely generate additional insights into the difference between the two fricatives.

3.1.3 Phonetic misperception

Misperception can be conditioned by the communication channel as well as the acoustic structure of the segment. A noisy channel can impede the accurate identification of a consonant, whether the noise is white (e.g., Miller & Nicely, 1955; Cutler et al., 2004; Phatak et al., 2008) or speech-shaped (Phatak & Allen, 2007). Different phonetic contrasts also show varying degrees of robustness to noise. Miller and Nicely (1955), for example, showed that, with increasing noise, listeners tended to be more likely to misidentify a consonant's place of articulation than its manner of articulation. This variability in the degree to which certain contrasts can be maintained depends in part on the cues that inform that contrast. An aperiodic signal, for example, is more easily masked than a periodic signal by either periodic or aperiodic noise (Wright, 2004), and so a contrast (such as voicing in fricatives) that relies primarily on a periodic signal would preferentially be perceived by a listener in a noisy signal.

This chapter is focused on spectral differences in stop and fricative place of articulation that are localized in time and frequency. While there are certainly cases where a salient spectral feature shows strong robustness to noise (e.g., the characteristic peak frequency of [s]), the differences of interest in this chapter are not expected to show a comparable degree of robustness. In fact, the cues that aid contrast between these segments may not actually be able to survive the introduction of noise. Recalling Miller and Nicely (1955), listeners still readily confused [θ] and [f] for one another in high SNR conditions, with each identified for the other 26% and 9% of the time, respectively, at +12 dB SNR. Likewise, the confusion patterns characteristic of perceptual asymmetry that were identified in Winitz et al. (1972) were observed in a non-noisy environment. The cues that might help to distinguish between the consonants that show perceptual asymmetry might have low perceptual salience in the vocalic contexts that condition this asymmetry.

3.1.4 Identifying informative acoustic features

Synthesized and naturalistic speech have both been used to identify perceptually relevant acoustic cues. In cases like, for example, Cooper et al. (1952), where researchers used synthesized speech to evaluate the perceptual relevance of certain acoustic information, the design of the stimuli was constrained by explicit choices about which features were manipulated; the peak frequency of the stop burst and the formant structure of the vowel were each independently varied. This choice seemed to afford the researchers a clearer ability to infer the acoustic source of the change in listener perception.

The systematic manipulation of naturalistic stimuli can be used in an exploratory fashion as well as to evaluate hypotheses about the location and structure of certain cues. An analysis of this sort takes place in F. Li et al. (2010), where the researchers use a combination of noise mask-

ing, high and low-pass filtering, and temporal truncation to examine changes in the (simulated) perception of stop-vowel syllables according to each dimension of manipulation. Their analysis confirmed the relevance of a similar set of informative acoustic cues to those identified using other methodologies – burst acoustics and formant transitions – in addition to a novel set of distinguishing features that would warrant testing with human listeners. The high and low-pass banding used in Miller and Nicely (1955) indicate differences in perceptual outcome according to the frequency bands available to the listener. If there were existing hypotheses about which frequency-specific regions cue consonant place, then such a method might be useful to confirm that the absence of that region corresponds to a decrease in classification accuracy.

The experiments in this chapter use naturalistic speech explore the spectral cues to place of articulation as well as to confirm their perceptual relevance. In the future, the use of synthesized speech could be beneficial to more tightly constrain the phonetic information presented to the listener.

3.2 Research Questions

This chapter has two primary goals. It first intends to address:

RQ1

What spectral features distinguish consonant pairs that participate in perceptual asymmetry?

Researchers have already provided some indication of what these differences are. Differences in peak frequency, variance, skewness, kurtosis, or differences in energy over specific frequency bands could all play a role. For many of these consonant pairs, the knowledge is more concrete; for example, Plauché et al. (1997) highlighted the difference in spectral energy around 3 kHz as an informative difference between [k] and [t] in the context of a high front vowel.

This chapter also addresses a perceptual counterpart to RQ1:

RQ2

To what extent are distinctive spectral components in these consonant pairs perceptually relevant?

Variability in the spectral components distinguishing members of a consonant pair is predicted to have perceptual relevance to the listener. If, for example, /k/ and /t/ are distinguished by energy centered at 3-4kHz (as was found in Plauché et al., 1997 before high front vowels), then misidentifications should increase for the consonant pair if energy in this region is changed.

This chapter serves a necessary bridge between the previous and following chapters. Chapter 2 established that the phonetic context that conditioned perceptual asymmetry was also associated

with greater similarity in the spatial features relevant to spectral acoustics. Chapter 4 intends to explain how the spectral acoustics of the consonants might result in asymmetrical rates of confusion, and Chapter 5 addresses what role the confusability associated with these consonants might have on sound change. While a not insubstantial body of research exists that describes the spectral characteristics of these consonants (as described in §3.1.1 and §3.1.2), it is not necessarily clear which of features can be used to distinguish consonants from one another and which of these distinctive features can aid human listeners in the task of classification. This chapter uses machine learning and behavioral experimental data to identify those specific spectral regions along which differences in energy are relevant to the discriminability of the consonant pairs.

3.3 Experiment 3.1

Experiment 3.1 is an exploratory analysis that uses naturalistic speech tokens to identify spectral regions that distinguish the targeted consonant pairs (/k/-/t/, /k/-/p/, /p/-/t/, and /f/-/θ/) participating in perceptual asymmetry.

3.3.1 Design

3.3.1.1 Dataset

The data for the exploratory analysis come from the Buckeye Corpus (Pitt et al., 2005). Forty talkers (20 male, 20 female) from Columbus, Ohio participated in a conversational interview conducted by one of two interviewers. Each participant's recording ran for between 30 and 60 minutes. These conversations were originally recorded at a sampling rate of 48 kHz but were resampled to 16 kHz for the corpus.

For the present data set, 43,245 prevocalic tokens of [p], [t], [k], [θ], and [f] produced by 40 speakers were extracted from these conversations. Productions shorter than 20 ms were excluded from analysis. For productions 20 ms or longer, a 20 ms interval centered at the midpoint of the production was automatically isolated. This interval targets silence, burst, and aspiration for the plosive tokens and frication for the fricative tokens. The corpus was pre-transcribed. For the stop productions, the transcribed interval included closure as well as stop release and aspiration. One hundred tokens were checked for each consonant to verify that the section was capturing a portion of the stop release. For all fricative tokens checked, this region corresponded to frication. For the stop tokens, the majority captured aspiration or aspiration and the burst. Only a handful of cases per 100 isolated only silence.

The frequency of token occurrence for each prevocalic consonant is given in Appendix A. The relative frequencies of consonants showed a wide range of variability across the entirety of the data

set, from over 4000 tokens per category to as few as 43. In Chapters 4 and 5, differences in the frequencies will be seen to influence the outcome of certain classification techniques. Across this data set, productions of these consonants are less common before back vowels than in other vocalic contexts. The sampling used for acoustic analysis is representative of the distribution of prevocalic tokens as they appear in the data set, but some contexts are consequently sampled more often due to an imbalance by vocalic context. (An alternative analysis would also be possible where each vocalic context is sampled equally often.)

For each 20ms interval, a 26-filter log Mel-filter bank was extracted. Like a Fast Fourier Transform (FFT), a Mel-filter bank extracts acoustic energy over the frequency domain. The filter bank differs from FFT in two regards, however. The bins are evenly spaced on the Mel scale, which more accurately represents perceptual distances in frequency. Each filter in the Mel filter bank is the result of the application of a triangular window on the frequency domain, and so each filter captures a small range of frequency information. This has the effect of reducing the dimensionality of an otherwise highly autocorrelated signal. The center and cutoff frequencies for each filter can be found in Appendix B.

3.3.2 Analyses

After processing, a 26-filter Mel-frequency filter bank was extracted for each consonantal token. These filter banks served as an input to a random forest (RF), where the respective spectral energies at each of the 26 filters serve as input and the goal is binary classification (as either segment of a consonant pair).

3.3.2.1 Random forests

A random forest is "a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$, where Θ_k are identically distributed random vectors and each tree casts a vote for the most popular class at input x." (Breiman, 2001, p. 2) Each tree in the forest attempts to complete the same (classification) task but operates only on a random subset of the input features. As a result, each tree achieves a varying degree of success in the classification task according to the subset of features available.

In the parlor game Twenty Questions, a questioner is tasked with identifying the object the answerer is thinking of. The questioner might first ask 'Is the object larger than a bread box?' and then based on that answer, 'Is it an animal?' With each question, the questioner eventually separates out all other objects from the one the answerer is thinking of. The 'growing' of decision trees works in much the same way, by performing binary splits along the features available to the tree. While the questioner is ideally trying to separate out one object from all others in Twenty

Questions, a perfect decision tree would be able to assign the correct label to all objects after traversing its binary splits. One way of characterizing the state of the tree at a given node is Gini index. If the decision tree up to that node were to assign the majority label to all objects in the node, Gini index is the product of the proportion of correctly classified objects and the proportion of misclassified objects, summed across each class (James et al., 2013, p. 312). This value is non-negative and decreases to 0 as all objects in a node are correctly classified.

Again, in Twenty Questions, it might be useful to ask 'Is it custard-filled?' once one has learned that the answerer's object is a dessert item, but not necessarily at the start of the game, when nothing is known about the item. Similarly, a feature may vary in importance to classification depending on what other binary splits could be made according to the features available to the decision tree. Because each individual tree in a random forest is grown with a different sampling of features, it provides an opportunity to gain insight into how important a single feature might be to the classification task in general. Featural importance for the random forest models in the section are reported in mean decrease in Gini– the decrease in Gini indices for each node that was split using this feature, averaged across all trees where it appeared. Features that rank high in this measure of importance tend to more cleanly separate its input into the two classes of interest.

Random forests have been used before in phonetic research to identify the relative importance of acoustic features in helping to classify between different phonetic categories. Styler (2015), for example, had success using random forests to identify informative acoustic features for classifying nasality. Their use in this chapter is to identify spectral regions that might be important to the task of discriminating one member of a consonant pair from another[1].

Gini index is a common criterion used to determine how a data set should be split into child nodes. This criterion is sensitive to class imbalance. As the data set becomes more skewed toward a single category, the tree tends not to produce splits that isolate the minority category (Flach, 2003). Consequently, decision trees (and RF models by extension) are sensitive to class imbalance. To avoid results that are biased toward one consonant category, the two consonant categories are down-sampled (a subset of the majority consonant is sampled for a given tree) so that each consonant category occurs equally often in the model.

Five hundred individual trees were grown in each RF model. Growing many trees does not lead to overfitting (Breiman, 2001, p. 4), but RF built from too few trees seems to decrease the stability of mean decrease Gini across iterations of these models.

[1]The R package **`RandomForest`** was used in this chapter to build RFs.

3.3.2.2 /k/-/t/

In this section, a RF is used to identify spectral components important to the classification of /k/ and /t/. The output of the model[2] is plotted in Figure 3.1. In this plot, the x-axis corresponds to each of the 26 features (i.e., the Mel-frequency filter bank, arranged in order of increasing frequency), and the y-axis corresponds to mean decrease Gini. A higher value corresponds to increased importance for this feature. Two peaks over filters 12-18 (peak frequencies ranging from 1.25 - 3 kHz), and filters 23-26 (5.2-7.2 kHz) were identified.[3] These intervals correspond to the concave down regions centered at the highest points in the importance plot, as can be seen in Figure 3.1.[4] Both peaks are consistent with the spectral peaks of [k] and [t] predicted from the tube model of the vocal tract (see 2.1 for more information). With a tube length of 17.6 cm, an alveolar constriction is predicted to produce a singular peak at around 4 kHz, while a velar constriction is predicted to produce two, at around 1.5 kHz and 5 kHz (Stevens, 2000). These energy differences fall in the low range of each peak. Furthermore, Cooper et al. (1952) found that changes in peak frequency of the stop burst from about 1300 to 3000 Hz (depending on the vocalic context) elicited a change in perception between [k] and [t]. These results highlight a similar frequency range that may inform the perceived difference between the two consonants.

[2]One statistic commonly reported for random forests is Out-of-Bag (OOB) Error. When a random forest is grown, each individual decision tree is trained on a different subset of the data. OOB Error is calculated from the error achieved by each tree when classifying the subset of data unseen during training. Like test error, this measure can be used to characterize the degree to which the model is able to generalize to unseen data, and Breiman (1996) demonstrates the OOB Error (at least for the data sets used) achieves a comparable estimate of generalization error to test set error. OOB Error for this RF was 20.4%. This error rate is far above human listener error rates when discriminating between /k/ and /t/, but this task is somewhat difficult for the RF model, who only has the filter energies of 20 ms of audio as input for classification.

[3]To see whether the features identified by the random forest were better than noise, a separate RF was run identical to the one described, but with three additional columns composed of random samples from a uniform distribution. These variables had an average mean decrease Gini of 215, which suggests that the peaks identified in this model are more informative to the task than noise.

[4]This same criterion is used to identify regions of interest for the other consonant pairs. For an alternative method for characterizing feature importance, see §3.3.3 and Appendix C.

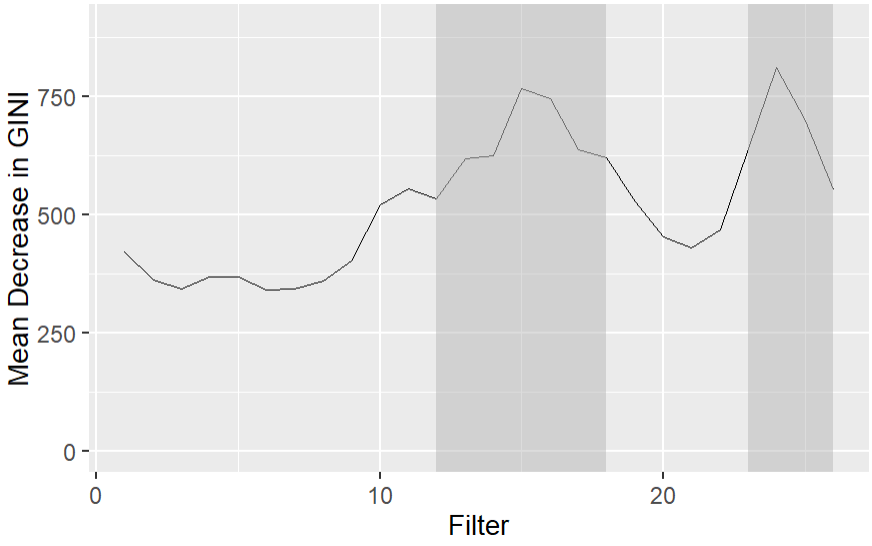

Figure 3.1: Mean decrease in Gini by filter for /k/-/t/ RF classifier. Filter regions targeted for further analysis are highlighted in gray.

As is evident in Figure 3.2, which plots filter energies in the two consonants by vocalic context, the lower frequency peak for the velar stop is clearly present before back and central vowels (at around filter 12). In a front vowel context on the other hand, the lower frequency peak appears at much higher frequency. Some of these peaks vary in strength by vocalic context. Across all vowel contexts, [t] has consistently higher energy than [k] over Filters 23 to 26, as opposed to the region from Filters 12 to 18, which vary according to vocalic context.

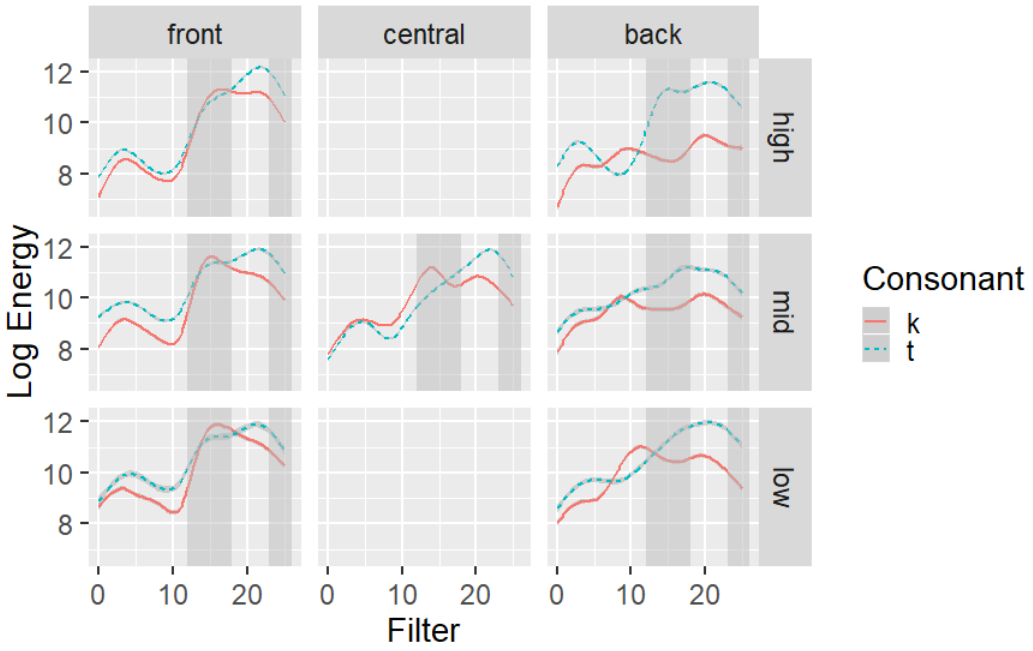

Figure 3.2: Filter energies for [k] and [t] by vocalic context (95% confidence intervals are present). Filter regions identified for further analysis are highlighted in gray.

3.3.2.3 /k/-/p/

A RF model was trained to classify [k] and [p] tokens . This model revealed two main peaks in importance at filters 12 (1.4 kHz) and 21 (4.3 kHz), as seen in Figure 3.3. As noted in §3.1.1, a velar burst is predicted to have spectral peaks at around 1.5 kHz and 5 kHz. Burst and aspiration noise for a bilabial stop has maximal energy at around 1 kHz (Stevens, 2000, p. 349). The filters highlighted by the RF classifier are roughly consistent with modeled predictions.

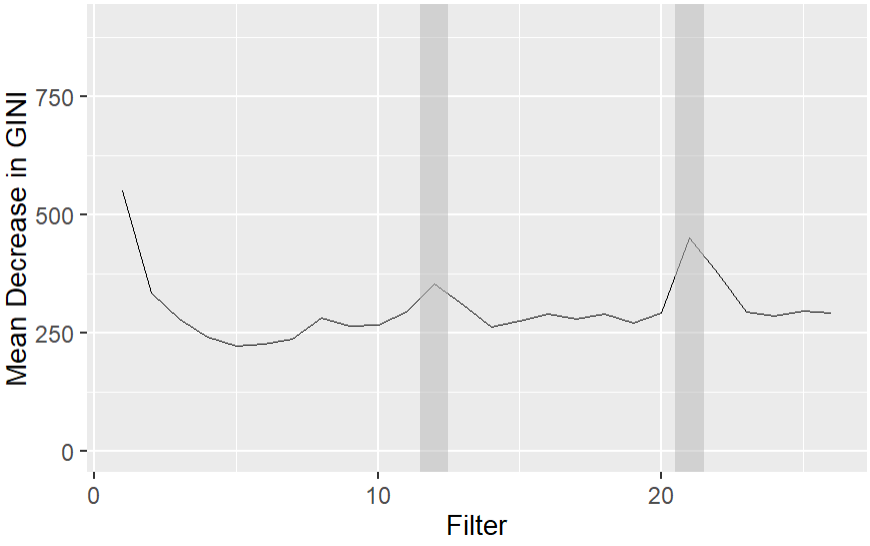

Figure 3.3: Mean decrease in Gini by filter for /k/-/p/ RF classifier.

Figure 3.4 presents a plot of filter energies by vocalic context. When [k] and [p] are both before high vowels, the characteristic peaks of [k] are of similar spectral energy to [p].

Figure 3.4: Filter energies for [k] and [p] by vocalic context (95% confidence intervals are present).

3.3.2.4 /p/-/t/

A RF was trained to classify /p/ and /t/ and identified strong peaks in importance in regions centered around filters 12 (1.4 kHz), 18 (3.0 kHz), and 22-25 (4.8-5.9 kHz), as seen in Figure 3.5. Recalling §3.3.2.2 and §3.3.2.3, the differences in the respective transfer functions of [p] and [t] are consistent with finding frequencies around 1 kHz informative to this contrast. Although the peak for [t] was predicted to occur at around 4 kHz, [t] energy (as seen in Figure 3.6) remains consistently high in the high frequency range - much higher than [p] energy in any context.

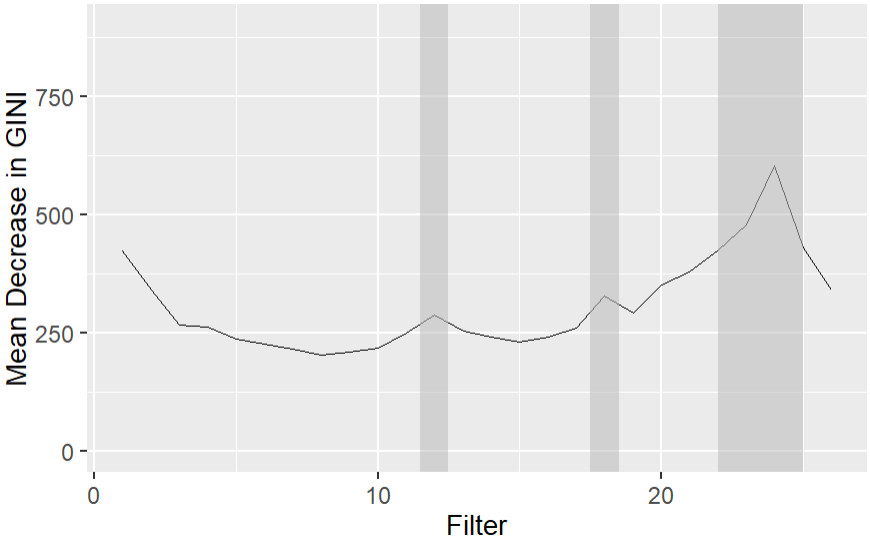

Figure 3.5: Mean decrease in Gini by filter for /p/-/t/ RF classifier.

Figure 3.6: Filter energies for [p] and [t] by vocalic context (95% CI present).

3.3.2.5 /θ/-/f/

A random forest classifier identified several regions of relative importance to the classification of the dental fricatives. Regions centered around filters 4-5 (0.3-0.4 kHz), filters 10-11 (1.1-1.3 kHz) and Filter 24 (5.9 kHz) were identified, as seen in Figure 3.7. There are comparatively few prior results that these can be directly compared against. Stevens (2000, p. 389) predicted that a labiodental fricative will have a front cavity resonance of about 10 kHz due to the small size of the cavity. An interdental fricative is likely characterized by a similar front cavity resonance due to its articulatory similarity. Stevens also described [f] as possessing a monopole noise source at the lower lip, which could contribute to the acoustics of [f] over a frequency range of less than 2.5 kHz. The articulation of an interdental fricative with the tongue may involve the inclusion of an additional noise source distinct from that of the lips, but it is unclear how this difference would be reflected in the spectral acoustics of [θ].

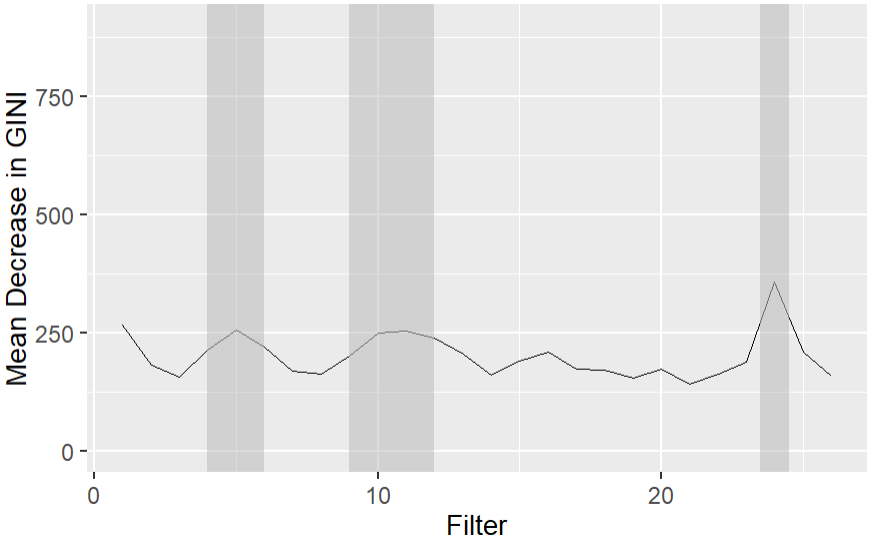

Figure 3.7: Mean decrease in Gini by filter for /θ/-/f/ RF classifier

A plot of the respective filter energies of the two dental fricatives (seen in Figure 3.8) suggests all three regions show some degree of variability according to vocalic context.

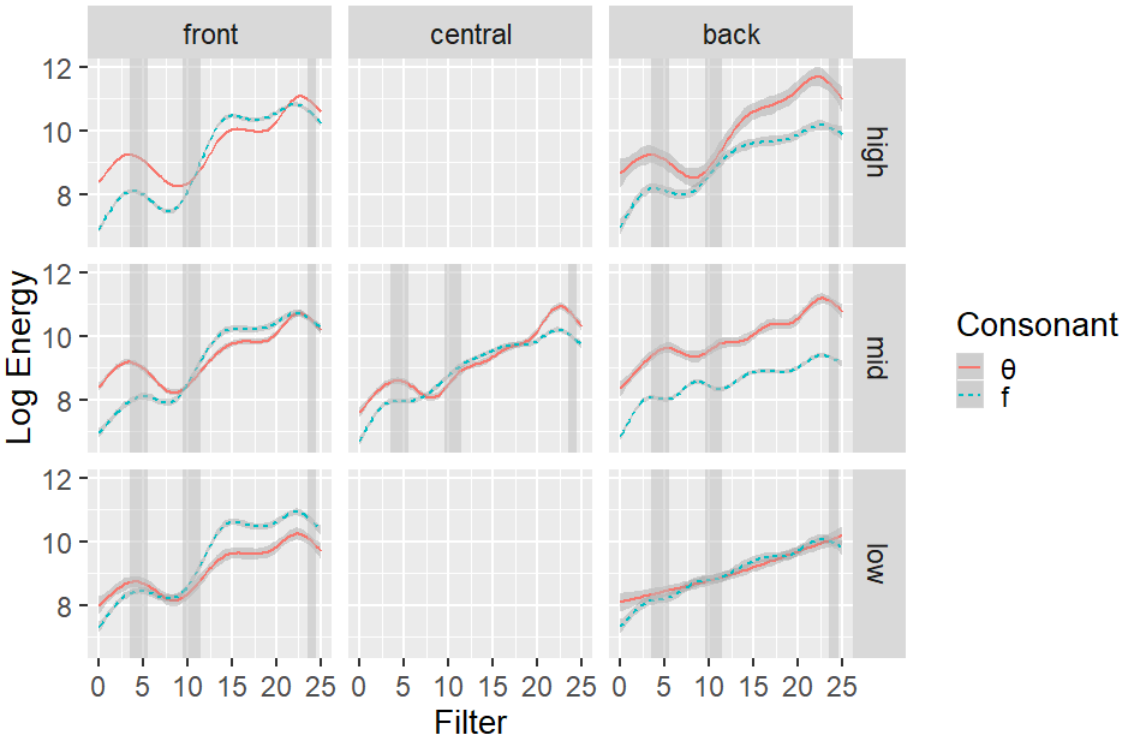

Figure 3.8: Filter energies for [θ] and [f] by vocalic context (95% CI is present).

3.3.3 Interim discussion

The goal of Experiment 3.1 was to identify spectral regions for each targeted consonant pair that provided contrastive information about place of articulation. The regions identified for each pair are summarized in Table 3.1:

	FILTERS
/k/-/t/	12-18, 23-26
/k/-/p/	12, 21
/p/-/t/	12, 18, 22-26
/f/-/θ/	4, 5 , 10, 11, 24

Table 3.1: Regions identified as 'important' for each consonantal contrast

As noted earlier, the features identified for the voiceless stops by the RFs do not depart strongly from what might be predicted from an acoustic model of stop production. The location and amplitudes of the spectral peaks for each stop place differ, and therefore the RF seemed to pick up on the frequency regions where a peak appears for one consonant but not the other.

The most extensive research about the spectral features distinguishing the targeted consonant pairs has been conducted for /k/and /t/. Plauché et al. (1997) found a high energy region which distinguished [ki] from [ti] at around 3-4 kHz. Guion (1998) found a similar result: [ki] tokens produced in citation form tended to have peak frequencies of around 3.5 kHz. Recalling that /k/ possesses two characteristic peaks, the frequencies described in these the areas outlined by the preceding three papers is likely reflective of a shift in the location of the low frequency peak – indeed in Guion (1998), the peak frequency of [k] before [ɑ] and [u] is about 2 kHz.

As noted in §3.3.2, the spectral filters targeted for further analysis were selected based on their associated mean decrease in Gini. One limitation with this method is that it is not possible to infer how sensitive the differences in mean decrease in Gini are to variability. One alternative method to calculate the effect of each feature on the model's performance is described in depth in Appendix C. This algorithm, adapted from Cafri and Bailey (2016), repeatedly estimates effect sizes for each feature by generating partial dependency plots from bootstrapped samples of the data. The largest effect sizes generated using this method tend to span the same regions identified using the method described in §3.3.2, but there are some places where the two methods disagree. Potential recommendations for the adjustment of boundaries based on the alternative method are listed in Table 3.6. Because the estimation of effect size is less predictable when the input features are correlated with one another, the method described in §3.3.2 (which avoids this issue) is preferred here:

	FILTERS TO INCLUDE
/k/-/t/	9, 10
/k/-/p/	22, 26
/p/-/t/	20, 21
/f/-/θ/	12

Table 3.2: Filter steps that might also be included based on the results of the analysis adapted from Cafri & Bailey (2016)

3.4 Experiment 3.2

The goal of this experiment is to test the degree to which the spectral regions identified in Experiment 3.1 hold perceptual relevance for listeners.

3.4.1 Methodology

3.4.1.1 Stimuli

The stimuli for this experiment were originally produced by a 30-year-old male speaker of American English who grew up in southern California. This speaker produced a variety of CVC syllables in the carrier phrase "Say X again". As seen in Table 3.3, all syllables were actual words in English except for the nonword 'thoo' (as there are no words in English with word-initial /θu/). The word-final consonant of each word was voiceless except for 'food' and the nonword 'thoo'. The vocalic contexts selected correspond to phonetic environments associated with perceptual asymmetry for some consonant pairs (i.e., [i] and [u]) as well as a vocalic context for which increased confusion is not predicted for any pair ([ɑ]).

	ɑ	i	u
p	POT	PEACH	POOP
t	TOT	TEACH	TOOTH
k	CAUGHT	KEEP	COOP
θ	THOUGHT	THIEF	*THOO
f	FOUGHT	FEET	FOOD

Table 3.3: Words elicited for Experiment 3.2

A CV sequence was isolated from each word, where the V portion was the first 60 ms of the total vowel (so as to minimize effects of V-C formant transitions). The intensity of the CV sequence was adjusted to be equal across all tokens. Six stimuli were created from each CV

syllable by isolating the consonant from the vowel, applying one of six band-reject filters to the entire consonant (i.e., burst and aspiration for stops and the entire duration of frication for the fricative), and splicing each consonant back onto the vowel. Each band-reject filter was spaced out in intervals of equal range in Mel scale. The cutoff frequencies are listed below in Table 3.4. For ease of reference in subsequent sections, Step refers to which of the six filters was applied to the CV syllable.

STEP	MIN FREQUENCY (HZ)	MAX FREQUENCY (HZ)
1	0	365
2	365	921
3	921	1768
4	1768	3056
5	3056	5016
6	5016	8000

Table 3.4: Experiment 3.2 filter boundaries

Impressionistically, the manipulated auditory stimuli sound like a brief CV syllable with some variability in the identity of the consonant. For example, the syllable [pi] sounds to me like [pi] when filter step 1 is applied but like [ki] when the filter step 3 is applied.

3.4.1.2 Participants

Participants were 19 undergraduate students at the University of Michigan. All were native speakers of American English and reported having no hearing problems.

3.4.1.3 Procedure

This experiment was conducted in a sound-attenuated booth in the UM Phonetics Lab on a Mac Mini over AKG K271 MK II headphones. The experiment was implemented in SR Research Experiment Builder. The primary linguistic task is blocked into two parts.

During Block 1, participants completed a three-alternative forced choice task. In each trial, the participant first sees three letters (‘p’, ‘t’, and ‘k’) spaced along the center of the screen. After 500 ms, the participant hears an auditory stimulus – one of the [p], [t], or [k]-initial filtered CV tokens as described in §3.4.1.1. The participant responds on a keyboard by pressing the key corresponding to the initial consonant they heard. The color and ordering of the letters on the screen match those of the stickers on the keyboard– for example, the leftmost choice on the screen and the sticker on the leftmost key of the keyboard are both red. The trial ends 500 ms after their response or 10 seconds after the audio stimulus was played. (The long trial time out time is intended to reduce

the likelihood that a late response is interpreted as a response for the next trial.) The ordering of the choices on the screen is consistent for a single participant but is counterbalanced across participants. The participant responds to each token four times, for a total of 216 trials in Block 1.

Block 2 is exactly like Block 1, except that the participant has two choices ('f' and 'th'), and the auditory stimuli they hear are one of the [f] or [θ]-initial filtered CV tokens. The participant responds to each token four times in this block as well, for a total of 144 trials.

At the end of the experiment, participants completed a questionnaire about their linguistic background.

3.4.2 Predictions

Within a consonant pair, listeners are predicted to choose the competing consonant more often when a filter has modified energy in the regions identified as relevant to classification of the pair, as described in §3.3. Figure 3.9 is a schematic describing the process of generating predictions for how /k/-/t/ misidentifications will be affected by filter step. The filters identified as important by the RF for the consonant pair were 12-18 and 23-26. Because the filters are triangular, their energy is a weighted sum of energies at multiple frequencies. The frequencies (and their relative weights) selected by filters 12-18 and 23-26 are indicated on the blue curve. The boundaries of each notch step are marked with a dashed vertical line. Steps 1 and 2 do not overlap with any frequency components identified as important, so changing energy in this region is not expected to affect categorization outcomes. In contrast, steps 3-6 do, so they are predicted to affect classification outcomes of /k/-/t/.

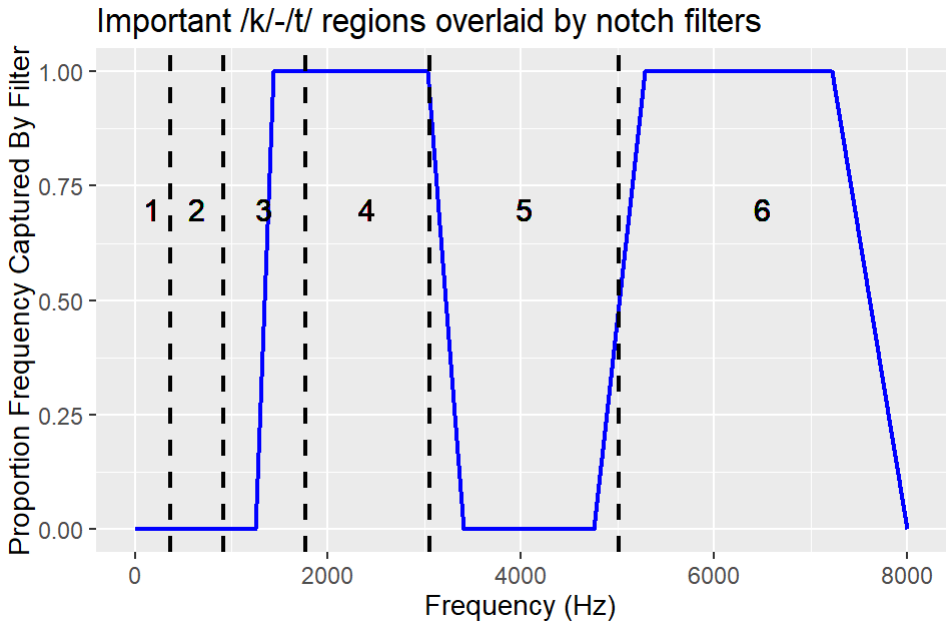

Figure 3.9: Notch boundaries and frequency regions 'important' to classification of /k/ and /t/

The same plot as Figure §3.3 for the other consonant pairs can be found in Appendix D, but the filter steps associated with increased misidentification rates are summarized in Table 3.5.

	FILTER STEPS ASSOCIATED WITH INCREASED ERROR RATES
/k/-/t/	3, 4, 5, 6
/k/-/p/	3, 4, 5
/p/-/t/	3, 4, 5, 6
/f/-/θ/	1, 2, 3, 6

Table 3.5: Steps predicted to correspond with higher classification error

3.4.3 Results

During each trial, one consonant was presented to the participant. Given the context-dependent nature of confusion for most of these consonant pairs, there is no expectation that filter step would have a uniform effect on listeners' choices across all vocalic contexts. Consequently, for each consonant pair, a separate logistic regression was run according to the consonant presented to the participant and its vocalic context. In each model, the dependent variable was a binary response (e.g., whether they chose /t/), and the independent variable was Step (as a categorical variable). For models where the effect of Step is significant, a post-hoc pairwise contrast is run between each step with a significant main effect, and all other steps. Because of the large number of comparisons

made, the p-value is adjusted using the Holm method, and significant contrasts are reported with $\alpha = 0.05$.

The risk of making a Type II Error (i.e. incorrectly failing to reject the null hypothesis) varies according to experimental effect size. Figure 3.10 plots the power of a binomial experiment with 19 participants at varying effect sizes.[5]With increasing effect size, the corresponding power of the study also increases. This figure suggests that this study has sufficient power to detect large differences in proportions, but not smaller ones. (See §3.4.4 for discussion.)

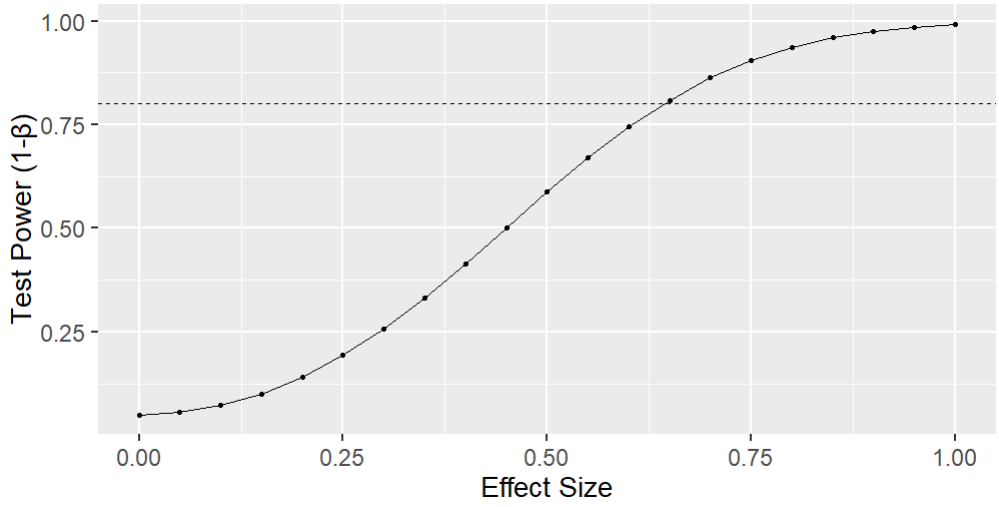

Figure 3.10: Experimental power plotted against effect size (n=19) The dotted line corresponds to a power of 0.8

3.4.3.1 /k/-/t/

Members of the consonant pair /k/-/t/ are predicted to show increased rates of misidentifying one for the another at filter steps 3-6.

Figure 3.11 plots the probability, by vocalic context, of /t/ responses to [k] stimuli. A peak in /t/ responses is visually apparent at Step 3. When a model was run on /t/ responses in the context of [ɑ], effects of Step 3 ($\beta = 2.96, z = 5.59, p < 0.0001$) and Step 5 ($\beta = 2.20, z = 2.04, p = 0.04$) were observed. Post-hoc pairwise comparisons for Step 3 (and the other steps) indicated that listeners were significantly more likely to choose /t/ in Step 3 than Step 1. No significant contrast was observed between Step 5 and the other steps, and no main effect was observed in other vocalic contexts.

[5]Effect sizes and corresponding power values were calculated using the **`pwr`** package in R.

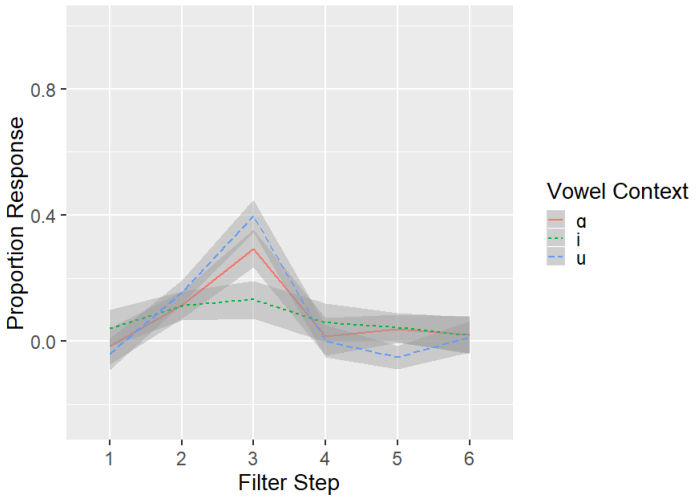

Figure 3.11: Misidentification rates of [k] as /t/

Figure 3.12 plots the raw response proportions of listeners' /k/ responses when presented with [t] stimuli. There is a significant effect of Step 4 on participants' likelihood to choose /k/ ($\beta = 2.76, z = 2.62, p = 0.009$) in the vowel context /i/, but none of the post-hoc pairwise comparisons achieved significance.

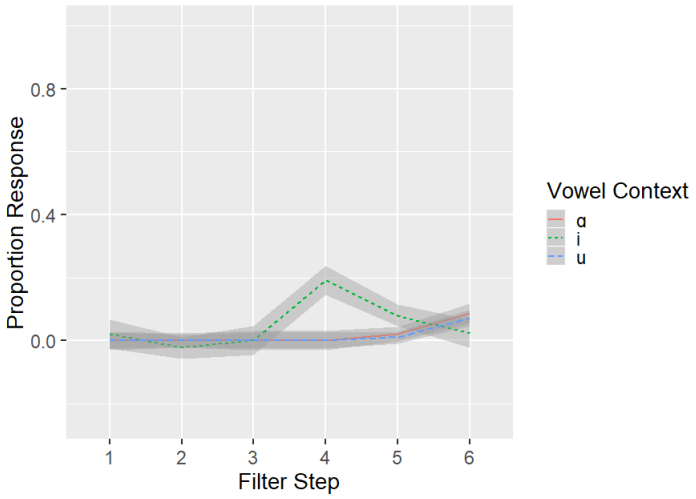

Figure 3.12: Misidentification rates of [t] as /k/

3.4.3.2 /k/-/p/

/k/ and /p/ are predicted to show increased confusion rates for steps 3 and 5.

For a model run on /k/ identifications when responding to [p] in the context of [i], a significant effect of Step 4 on participants' responses was found ($\beta = 2.89, z = 4.495, p < 0.0001$). Post-hoc

pairwise comparisons revealed that participants were more likely to respond /k/ for Step 4 than Steps 1, 2, 5, and 6. No effects were observed in other vocalic contexts.

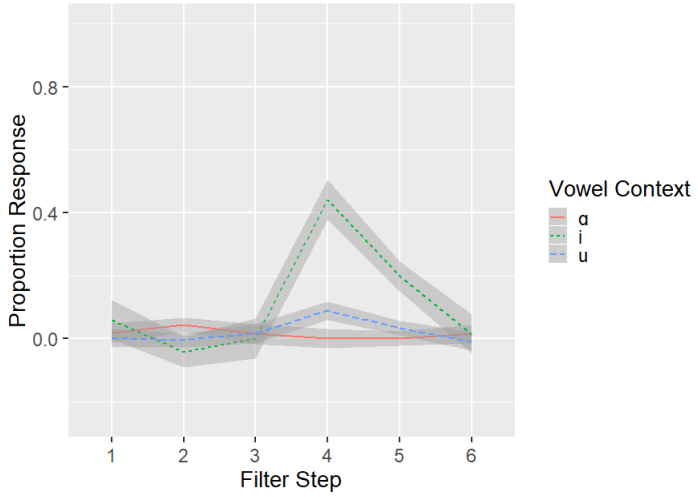

Figure 3.13: Misidentification rates of [p] as /k/

When the same model is run on participants' /p/ identifications when responding to [k], filter step was found to affect the participant's response in each vocalic context. Before [ɑ], there was a significant effect of step 3 ($\beta = 2.19, z = 2.04, p = 0.04$) on the participant's response, although none of the post-hoc pairwise comparisons with Tukey correction reached significance. Before [i], there were significant effects of steps 2 ($\beta = 2.66, z = 2.52, p = 0.01$) and 3 ($\beta = 3.03, z = 2.89, p = 0.004$). Post-hoc pairwise comparisons with Tukey correction revealed that, before [i], participants were more likely to respond /k/ for step 3 than steps 1, 4, and 6. Before [u], there was a significant effect of steps 3 ($\beta = -1.58, z = -3.65, p = 0.0002$) and 5 ($\beta = 2.32, z = 2.17, p = 0.03$). However, post-hoc pairwise comparisons indicate that, in this context, listeners were significantly less likely to choose /p/ in step 3 than steps 1, 4, 5, and 6. It is worth noting that listeners also tended to predominantly choose /p/ when they were presented with the syllable [ku], regardless of step, as is apparent in Figure 3.14. Such a result may offer support for the finding in Winitz et al. (1972) (though not Plauché, 2001) that listeners may misidentify [k] as /p/ more often when it appears before [u].

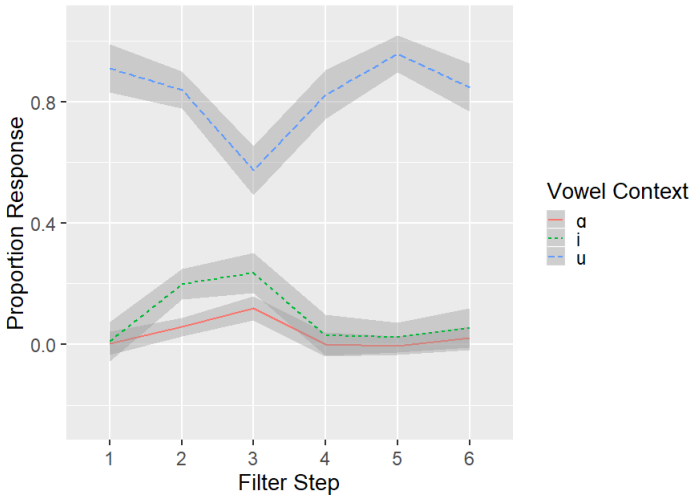

Figure 3.14: Misidentification rates of [k] as /p/

3.4.3.3 /p/-/t/

Members of the consonant pair /p/-/t/ are predicted to show increased rates of identification for one another when energies in filter steps 3, 4, and 5 are modified.

As is visually apparent in Figures 3.15 and 3.16, listeners appear not to vary in their identification rates of /p/ or /t/ according to step or vocalic context. Indeed, models find no significant effect of either predictor.

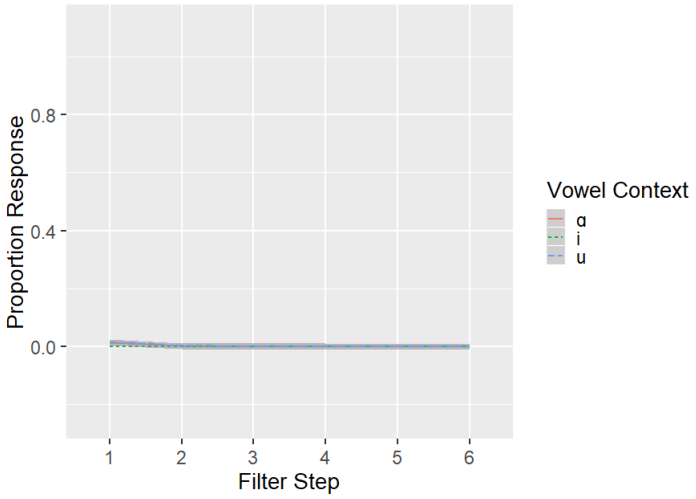

Figure 3.15: Misidentification rates of [t] as /p/

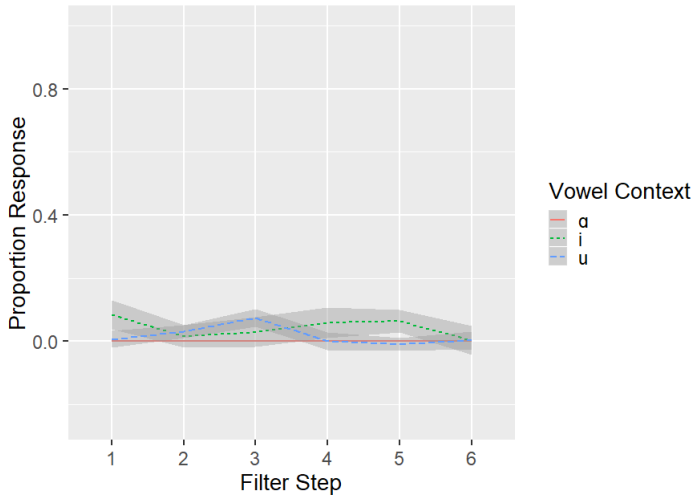

Figure 3.16: Misidentification rates of [p] as /t/

3.4.3.4 /θ/-/f/

Based on the results of Experiment 3.1, listeners are predicted to classify /θ/ and /f/ as one another more often in at steps 1,2, 3, and 6.

When a logistic regression was run on /θ/ identifications for participants' responses to [f], step was found to affect the participant's response before [ɑ] and [i]. Before [ɑ], there was a significant effect of Step 3 ($\beta = 1.30, z = 2.57, p = 0.01$) on participants' responses, and a post-hoc comparison with Tukey correction indicated that listeners were more likely to respond /θ/ in Step 3 than Step 2. Before [i], there was a significant effect of Steps 4 ($\beta = 1.32, z = 2.93, p = 0.003$) and 5 ($\beta = 1.69, z = 3.79, p = 0.0002$). A post-hoc comparison indicated that listeners were more likely to respond /θ/ in Step 5 than Steps 1, 2, 3, and 6, and were more likely to respond /θ/ in Step 4 than Step 1.

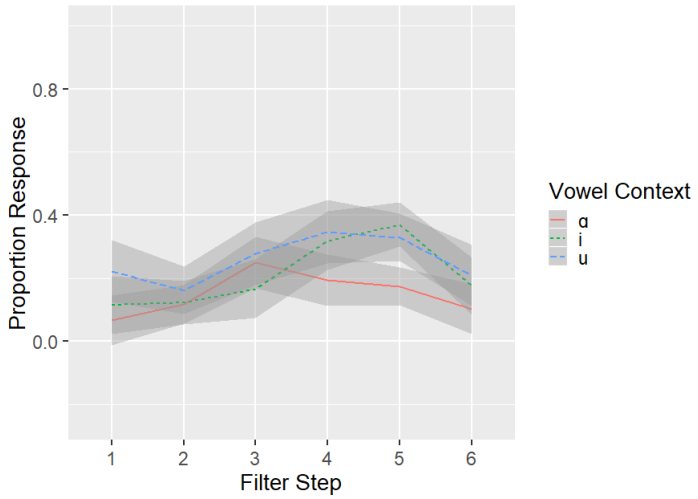

Figure 3.17: Misidentification rates of [f] as /θ/

When a logistic regression was run on /f/ identifications for responses to [θ], step was found to affect participants' responses before [i]. In this context, there was a significant effect of Steps 4 ($\beta = -0.85, z = -2.107, p = 0.04$) and 6 ($\beta = 1.09, z = 3.14, p = 0.002$) on listener response. A post-hoc comparison with Tukey correction indicated that listeners were more likely to respond /f/ in Step 2 than Step 4 and were more likely to respond /f/ in Step 6 than Steps 3 and 4.

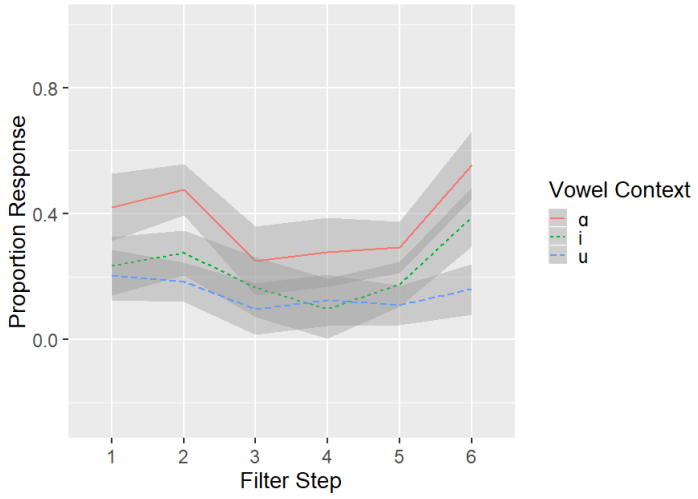

Figure 3.18: Misidentification rates of [θ] as /f/

3.4.4 Results summary

The results of Experiment 3.2 (summarized below in Table 3.6) tell a more complicated story than had been predicted by Experiment 3.1.

PAIR	PREDICTED	MISIDENTIFIED AS	/ɑ/	/i/	/u/	RESULTS SUMMARY
/k/-/t/	3,4,5,6	/k/	-	4	-	3,4,5
		/t/	3,5	-	-	
/k/-/p/	3,5	/k/	-	4	-	2,3,4,5
		/p/	3	2,3	3,5	
/p/-/t/	3,4,5,6	/p/	-	-	-	*no steps*
		/t/	-	-	-	
/θ/-/f/	1,2,3,6	/θ/	3	4,5	-	3,4,5,6
		/f/	-	4,6	-	

Table 3.6: Summary of results from Experiment 3.2. Results indicate steps with a significant main effect (relative to Step 1).

While filter step was a significant predictor of listener response, the results did not closely align with what had been predicted. For /k/ and /t/, the region corresponding to increased rates of listener misidentification was narrower than expected: steps 3, 4, and 5 affected listener response, but step 6 did not. For /k/-/p/ and /θ/-/f/, a wider range of filters than predicted affected listener response: most filters expected to influence response did so, but other filters also had an influence. For /p/ and /t/ no difference in response was observed according to filter step.

The discrepancy in the results of Experiments 3.1 and 3.2 may be due in part to the fact that the RFs and human listeners completed somewhat different tasks. Each random forest algorithm had to learn to classify consonants from filter energies extracted from a 20 ms sample of a stop or fricative, while human participants could listen to the entire consonant as well as a portion of the vowel when making their choice. Consequently, human listeners had access to spectral information about consonant place in the vowel or elsewhere in the consonant beyond the sample extracted for the RF. They could also make use of temporally varying information (like multiple stop bursts or formant transitions). The choice for the stimuli to include a portion of the vowel was a compromise between preserving naturalness for the listener and aligning the task performed by the listener and that performed by the RF. Early during the process of stimulus design for this experiment, it was found that listeners tended not to judge the isolated filtered consonants as sounding natural. An attempt to mitigate this issue was made by appending the consonant to its following vowel, which would help the stimulus to sound more like a (portion of a) naturally produced utterance. However, this approach created a considerable difference in the information that the RF (Experiment 3.1) and the human listeners (Experiment 3.2) had available during classification. Although this choice was made to better ensure that human listeners would respond to the stimuli as speech-like, this may have come at the cost of reducing the degree to which Experiment 3.1 could predict the results of Experiment 3.2.

Despite their statistical regularity, the informational cues identified by a RF are also not guaranteed to be the same features that listeners make use of in perception. The success of machine learning methods like RFs depends in part on the existence of structural regularities in the data. For the models used in this chapter, consonants of different places of articulation also have differing energies at different frequency regions, and this information can be exploited to create models that discriminate between places of articulation at an accuracy above chance. However, listeners are not necessarily sensitive to all regularities in the spectrum. Some differences may be impossible for listeners to exploit due to limitations of the human perceptual system. For example, while energy differences at a certain frequency region may robustly distinguish between consonants, the difference may be too small for differences in loudness to be perceived by a human often enough to reliably learn how to contrast the two with this information. It may also be the case that listeners have access to an acoustic property, but it is weighted low in perception. Under cue-based approaches to perception (discussed in further detail in §5.1.1), the act of perceiving involves the integration of multiple sources of information. Individual differences in the weighting of certain cues could lead some listeners to show comparatively low sensitivity to differences that an RF (or another listener) might otherwise pick up on.

As noted in §2.1.1, coarticulation between consonants and vowels can affect the articulatory events involved during a production as well as the acoustics of the production. The results of Experiment 3.1 reinforced such a finding. For example, [p], [k], and [t] all have low energy at around filter 12 (peak frequency: 1.4 kHz) before [i], while [p] and [k] have comparatively high energy in in this region in other vocalic contexts. Experiment 3.2 goes further to indicate that the effect of manipulating the spectral energy of consonant is also largely specific to vocalic context. For example, for [f] stimuli, step 3 influenced rates of /f/ identification in the context of [ɑ], but rates of /f/ identification did not change significantly by step in the context of [u]. Recalling Figure fig. 3.8, [f] and [θ] productions differ in energy in most regions of the spectrum before [u]; in this circumstance any individual filter could not block out the entire region over which the two consonants differ. In contrast, before [ɑ], the two consonants show comparatively little difference in the higher frequencies, which would mean that a filter applied at step 3 could have a more significant impact on the confusability of the pair.

The relation between the results of Experiments 3.1 and 3.2 may also be complicated by the number of participants in Experiment 3.2. The interpretation of null results for /p/ and /t/ may consequently be more difficult. In the condition where /k/ is the target and /t/ the competitor, for example, listeners are expected to have a high likelihood of categorizing the consonant as /k/ in Steps 1 and 2 (e.g., p=99%), and a low likelihood of categorizing the consonant as /k/ in the other steps. A 9% decrease in likelihood of /k/ response would correspond to an effect size of 0.44, and 19% decrease corresponds to an effect size of 0.73. If in fact the baseline rate of misidentification

is consistent with the assumptions above, then this experiment would likely have sufficient power to detect changes in response likelihood of 20% or more. In the case of /p/ and /t/, however, no significant change in misidentification likelihood was observed as a function of the filter applied. This lack of effect could be due to an absence of observable difference between filters, or it could indicate that the difference in response likelihood according to filter has a smaller effect size than had been observed for the other voiceless stop pairs (suggesting a Type II Error). Additional participants would help to clarify which of the two possibilities might explain this result.

The dental fricative results differed from those of the voiceless stops. Because the baseline misidentification rate was relatively high, a larger difference in proportion would be necessary in order for the experiment to have sufficient power to detect an effect size of comparable magnitude to that of the voiceless stops. A significant effect of filter was observed for this pair, but there may be still more obscured by a smaller effect size.

3.5 General discussion

The goal of this chapter was to identify distinctive spectral features for members of a consonant pair that are important to their classification. An exploratory analysis was performed in Experiment 3.1; RF models separated out frequency components of each consonant according to mean decrease Gini, a measure of feature importance. The perceptual significance of these spectral regions was further explored in Experiment 3.2, where listeners categorized band-stop filtered CV syllables. Taken together, the results of these experiments offer some insight into which spectral features may be important to the pairwise contrast of consonant pairs participating in perceptual asymmetry.

For /k/ and /t/, filters 12-18 and 23-26 were identified by a RF as features important to the classification to the two consonants, but in perception, listeners were only sensitive to changes in the region that included filters 12-18. This result is consistent with previous studies that modified the spectral properties of /k/ to affect listeners' categorization of the consonant. Plauché et al. (1997) applied a wide band-stop filter from 2.5-5 kHz (a region roughly spanning filters 16-23); listeners always categorized this modified stop as a /t/. In Chang et al. (2001), a 2 kHz band-stop applied over the peak frequency of the burst increased the frequency of /t/ categorizations.

For /k/ and /p/, listeners were sensitive to the manipulation of a wider frequency range than had been predicted from the RF model. The wider perceptual findings for /k/ and /p/ resemble the findings of Cooper et al. (1952) to some extent. In that study, consonants with burst peak frequencies ranging from 0.7-2.9 kHz could either be categorized as /k/ or /p/ depending on vocalic context. This range is consistent with the frequency range over which filters applied to either consonant can induce misidentifications for the other.

For /p/ and /t/, although filters 12, 18, and 22-26 were predicted to be important to the classi-

fication of the two consonants, listeners were no more or less likely to misidentify the consonants as one another according to step in any vocalic context. As discussed in the results summary for Experiment 3.2, the task and task taker for Experiments 3.1 and 3.2 were different, and so the types of differences captured by the RF may either be imperceptible or otherwise weighted low by the listener, or there are additional reliable cues to the production (like amplitude) that listeners can make use of to properly identify the consonant even if the spectrum has been filtered. A variety of other cues not addressed in this model may play a role here. Plauché (2001) noted that features like relative amplitude and linear fit (not dissimilar to the spectral templates mentioned in §3.1.1) were relevant to listener categorizations. It may well be possible that the confusability observed in /p/ and /t/ is not due to anything localized in the frequency domain of the consonant.

The reported results for /θ/ and /f/ appear to be novel. I am unaware of other research on the frequency components that distinguish the two dental fricatives or on the degree to which misidentification errors for these consonants show sensitivity to frequency filtering. Modifications along much of the spectrum seem to shift listeners' identification patterns in one direction or the other. Furthermore, Figure 3.18 suggests that /f/ identifications may show sensitivity to vocalic context. Indeed, a logistic regression model run with the dependent variable as /θ/ response and independent variables as step, vocalic context, and their interaction reveal a significant effect of [i] ($\beta = -0.17, z = -2.42, p = 0.01$) and [u] ($\beta = -0.19, z = -2.82, p = 0.005$) on listeners' responses. Identifications of /f/ occur before [ɑ] more often than before [u] or [i].

If [ɑ] conditions increased confusability between /θ/ and /f/, this conditioning context would set it apart from the other obstruent pairs looked at in this study, whose asymmetries are conditioned by [i] or [u]. [ɑ] may also condition perceptual asymmetry between [ɹ] and [l] (Müller 2010), but in this case, a pharyngeal constriction (characteristic of [ɑ]) is also present in either or both of those consonants. For [ɹ] and [l], the phonetic environment may create spatial similarity in the vocal tract between otherwise dissimilar consonantal productions. For [θ] and [f], however, none of the active articulators involved in the production of [ɑ] are shared by either consonant. Perhaps, rather than the two productions being made more similar in the context conditioning confusion, [θ] and [f] are made less similar by competing articulations from [i] and [u]. The tongue dorsum constrictions for [i] and [u] may compete with the tongue tip constriction for [θ], but not the labial constriction for [f]. This finding would be consistent with the observation in Chapter 2 that [θ] and [f] show the least articulatory similarity before [i].

CHAPTER 4

Evaluating a Probabilistic Account of Asymmetry in Perception

The preceding chapters have offered a description of the articulatory characteristics of four consonant pairs that show perceptual asymmetry, and an analysis of the frequency components that may best inform the contrast in place of articulation for each consonant pair. This chapter focuses specifically on the asymmetrical aspect of these confusions: that listeners tend to strongly favor one consonant over the other when hearing the consonant in isolation. A probabilistic account for why these confusions might show asymmetry is explored.

4.1 Background

4.1.1 Similarity and perceptual asymmetry

The traits of perceptual asymmetry would seem to show some overlap with similarity, the psychological relatedness of entities. As discussed in §1.1, letter pairs that show perceptual asymmetry (e.g., 'Q' and 'O') visually resemble one another. Likewise, in phonetic perceptual asymmetry, the consonant pairs share articulatory parameters, and consequently are characterized by similar acoustic events. Perceptual asymmetry is also characterized by perceptual confusion – listeners treat these entities as related enough to assign an item of one category the label of a different category. The goal of this chapter is to better understand the asymmetrical aspect of perceptual asymmetry, that is, why confusion rates are different for the two members of the pair. Similarity provides one route to understanding this phenomenon.

4.1.1.1 Similarity as a metric

One common treatment of similarity has been as a metric, a distance-like measure ranging from 0 (the items are identical) to positive infinity (the items are maximally distinct) with four specific properties:

I. Non-negativity: The similarity between two objects cannot be less than zero.

II. Indiscernibility: If two entities have a similarity of 0, then they are identical.

III. Symmetry: The similarity of A to B is the same as the similarity of B to A.

IV. Subadditivity: The similarity of A to B and the similarity of B to C are greater than the similarity of A to C.

This approach captures many properties intuitively associated with perception, including, for example, that nothing can be more similar to something else than to itself (according to Property I) and that as A-B and B-C become more and more similar to one another, we might expect A-C to also become more similar (according to Property IV). The properties associated with metrics also make it possible to use analytical techniques like Multidimensional Scaling (Shepard, 1962), which explicitly model similarity as a distance between entities. A metric-like approach to similarity appears in some theories of phonetic perception. The perceptual magnet effect (Kuhl et al., 1992), for example, relies on the warping of similarity distances near a category prototype.

Under a purely metric understanding of similarity, members of a pair would be equally similar to one another because of Property III. In a case like [f] and [θ] identification, where listeners show preference for one segment in a forced choice identification, it is difficult to see how such a measure alone could predict a difference in confusion rates.

One strategy to accommodate asymmetries in identification has been to incorporate a bias measure, which can shift the likelihood of response toward one member of a pair. The Similarity Choice Model (Shepard, 1957; Luce, 1963) describes the conditional probability of a response given a stimulus as a function of the similarity of the two stimuli and the individual's bias toward one category. In this model, given in 4.1, the probability of a listener choosing j given a stimulus i is the similarity of the two stimuli (η_{ij}) multiplied by the bias toward a j response (b_j), normalized by the sum of the respective similarities of i and all other entities multiplied by their respective biases. In this model the similarity measure η is a metric; the degree to which there is preference for one segment over another is represented independently of their similarity to one another. The bias is also a model parameter – it is defined independently of the model. While it might be able to accurately describe the pattern of consonant confusions observed in perceptual asymmetry, this model does not provide insight into why bias would appear.

$$p(R_j|S_i) = \frac{b_j \eta_{ij}}{\sum_k b_k \eta_{ik}}$$

Equation 4.1: The Similarity Choice Model (SCM) (Shepard 1957)

4.1.1.2 Similarity as a metric

If entities are understood as possessing features, the similarity of two entities can be approached by considering their shared and distinct attributes. This feature matching approach, taken in, for example, Tversky (1977) (also generalized to fuzzy set theory in Shiina, 1988), imposes fewer assumptions on the properties associated with similarity than the metric approach, like symmetry. In such models, if the features of B not shared by A are perceptually weighted more strongly than the features of A not shared by B then, in a comparison between the two, B will have greater similarity to A than vice versa. This seemingly unintuitive outcome captures the observation that non-prototypical examples of a certain category tend to be judged as more similar to the prototypical member of that category than vice versa.

Equation 4.2 is one implementation of an approach to perception that accommodates a featural understanding of similarity. A retooling of Equation 4.1, it defines similarity and bias with respect to the features that distinguish the two categories. The feature m is an element of the set of attributes that are not shared between categories I and J, and a_m and d_m are the respective likelihoods that each of these features do or do not appear in an event.

$$\eta_{ij} = \pi \frac{a_m d_m}{(1 - a_m)(1 - d_m)}$$
$$m \in [(I \cup J) - (I \cap J)]$$

$$\frac{b_j}{b_o} = \pi \frac{a_m (1 - d_m)}{d_m (1 - a_m)}$$
$$m \in J$$

Equation 4.2: Extensions to SCM to accommodate features (Nosofsky 1991)

The revised similarity measure here depends on the likelihood that each feature distinguishing categories I and J appear in a stimulus. For each feature, the addition and deletion probabilities are multiplied and divided by the product of one minus the addition and deletion probabilities. Each of these values is then multiplied together. The more likely these features are to appear, the less similar these categories are.

To calculate bias, for each feature present in J specifically, the addition probability and one minus the deletion probability are multiplied together and divided by the deletion probability and one minus the addition probability multiplied together. Each of these values are multiplied. The more likely a feature of J is to appear in the stimulus and the less likely it is to be deleted, the stronger the bias is toward J.

For this chapter, phonetic similarity between consonant tokens is treated more like a metric than featural. Consonantal tokens are not described by the presence or absence of a cue but by their values in a two-dimensional phonetic space. The following section lays out the reasoning for why the assumption of symmetry in metric measures is not as problematic for perceptual asymmetry as it might seem.

4.1.2 Perception and categorial structure

As described above, featural and metric-based treatments of similarity have been used to perform pairwise comparisons between entities. The metric-based theories give the impression of difficulty accommodating the unequal confusion patterns characteristic of perceptual asymmetry. In fact, it seems possible that perceptual asymmetry could arise from category structure, independent of the approach taken to similarity.

Throughout this section, I will refer to two hypothetical situations that might help to better illustrate the role that category level considerations might play in eliciting asymmetrical confusion rates. 'Similar' in this section could describe a feature-matching or metric approach.

Imagine a listener hears a voiceless plosive on its own. Circumstances have conspired in such a way that the listener knows for a fact that it is a /k/ or /t/, but they are not immediately sure which. Consistent with the results of Winitz et al. (1972), /k/ and /t/ in isolation are not very confusable for one another if extracted from an /ɑ/ or /u/ context but are more confusable if extracted from an /i/ context.

SITUATION 1*

Before high front vowels, [k] and [t] sound dissimilar to [k] and [t] in other phonetic contexts.

It could be possible that the acoustics of [k] and [t] both change with respect to vocalic context such that they are specifically similar to each other, but not similar to either category in other phonetic contexts.

SITUATION 2*

1. *[k] and [t] before high vowels both sound similar to [t] in other phonetic contexts*
2. *[k] and [t] before high vowels both sound dissimilar to [k] in other phonetic contexts*

It could also be possible that the acoustics of [k] change before a high front vowel, but those of [t] do not. In this case, [k] (before high vowels) sounds similar to [t] but not to [k] in other contexts, while [t] still sounds like other members of its category.

Equation 4.3 is an example of a perceptual model that would allow for the possibility of perceptual asymmetry, despite using metric similarity. This equation is another modification of the Similarity Choice Model which accommodates category structure; the probability of category response depends on a category bias parameter (B_J), the similarity between the entity and all other members of the category (η_{ij}), and bias values for each entity in that category (b_j). Ignoring the effect of category or entity bias, if a token is dissimilar to members of a category A but similar to members of category B, then the token will be more likely to be identified as category B. Thus, if the conditions in Situation 2* are assumed, then there should be perceptual asymmetry between /k/ and /t/ before high front vowels – [k] before high front vowels will be misidentified more often than [t] before high front vowels. In contrast, if Situation 1* is assumed, there should not be asymmetry.

$$p(R_J|S_i) = \frac{B_J \sum_{j \in C_j} b_j \eta_{ij}}{\sum_K B_K \sum_{k \in C_K} b_k \eta_{ik}}$$

Equation 4.3: Extensions to SCM to accommodate categories (Nosofsky 1987)

Equation 4.3, from Nosofsky (1987), resembles exemplar theoretic approaches to perception. In the model outlined in Pierrehumbert (2001), perception involved a comparison of the targeted entity with exemplar traces within a threshold level of similarity. Within this window, the algorithm assigned a category label score as the sum of the exemplar traces of that category (weighted against the recency of the traces). Once again, under Situation 2*, such a model could implement perceptual asymmetry – a production of [k] before high front vowels would be near exemplar traces of /t/, as would a production of [t] before high front vowels.

A recent exemplar-theoretic approach to perception subdivides category comparisons into typicality, which asks 'how good is the token as a realization of its identified category' and discriminability, which asks 'how likely is the token to be a realization of its identified category' (Todd et al., 2019). Equation 4.3 most clearly captures discriminability by comparing the relative similarities of an entity to entities of each category. Typicality, the absolute similarity of a token to other entities within a category is not clearly captured, however – an atypical token with a marginal similarity to entities of any category will still be identified in the category to which it is most similar in Equation 4.3, whereas it may not in a model that considers typicality.

A model that looks at category discriminability and typicality may still show perceptual asymmetry in Situation 2*, but not Situation 1*. Under Situation 2*, the similarity of [k] before high

vowels to [t] in other contexts means that it (and [t] before high vowels) would be a likely realization of /t/ and potentially a good production of it. In this case, both could reasonably be classified as /t/. In contrast, in Situation 1*, both /k/ and /t/ would likely be rated atypical for either category and would fare comparably well in either category with respect to discriminability concerns.

Tversky (1977) makes mention of category-level comparisons within his discussion of prototypicality, the degree to which an entity "exemplifies the category to which it belongs." He formulated the prototypicality of an entity as a function of the features that are shared or distinct between it and all other items in that category. A perceptual model that considers featural category prototypicality would likely also find perceptual asymmetry in Situation 2* but not Situation1*. A production of [k] before high vowels would share more features with productions of [t] and fewer with productions of [k] (if it were truly similar to /t/ in a featural sense), while in Situation 1*, neither production is prototypical of either category.

4.1.3 Categorial structure in probabilistic models

Probabilistic models can also account for perceptual asymmetry in a comparable way to what had been described in §4.1.2. For these models, phonetic categories are defined as probability distributions. Because there are not discrete tokens to compare as in the previous models, the calculations used to determine which category is most likely given a production's acoustics are different.

Equation 4.4 is a rendering of Bayes' Rule for phonetic categorization; the conditional likelihood $p(C_i|x)$ of a category choice (given the acoustics of a token) is a function of the category's prior probability $p(C_i)$ and the likelihood that category would produce a token with those acoustics $p(x|C_i)$. The considerations that go into the generation of the posterior probability can be thought of as a probabilistic counterpart to Equation 4.3. The category bias present in Equation 4.3 is analogous to the category prior in Equation 4.4.

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{\sum_{k=1}^{K} p(x|C_k)p(C_k)}$$

Equation 4.4: Bayes' Rule as applied to phonetic categorization

Unlike Equation 4.3, no explicit similarity calculations are made between the entity and other potential tokens of the category. Likelihood probability appears to make a similar contribution to similarity in the model, however. In Equation 4.3, the calculation in the numerator performed for each category involved the summation of the similarities (multiplied by biases) between the stimulus and all other tokens in that category. Taking an exemplar-based approach also, Nosofsky (1990) argues that a similarity measure can be reframed as a likelihood – how likely the stimulus

was to be generated by the same distribution as each token within the category, summed over all tokens in that category. If a stimulus has a high likelihood of being generated by a certain category, it is more likely that other tokens in that same category will be similar to it.

A classification algorithm that chooses the category that maximizes $p(C_k|x)$ is a Bayes Optimal Classifier, and this technique gives the highest classification accuracy on average given the data. This technique has been used as a model of phonetic perception (e.g., Norris & McQueen, 2008; Clayards et al., 2008; Kirby, 2010; Sonderegger & Yu, 2010; Kronrod et al., 2016). Figuring out which category maximizes the posterior probability depends on the exact priors and likelihood probabilities of each category. If there are two categories and they have equal prior probabilities, the two likelihood probabilities alone could be compared to identify the category. Situations 1* and 2* are reframed in terms appropriate to Equation 4.4.

SITUATION 1

1. $p([k]|/k/)$ *and* $p([t]|/k/)$ *are comparable and small before high front vowels*
2. $p([k]|/t/)$ *and* $p([t]|/t/)$ *are comparable and small before high front vowels.*

SITUATION 2

1. $p([k]|/t/) > p([k]|/k/)$ *before high front vowels*
2. $p([t]|/t/) > p([t]|/k/)$ *before high front vowels*

Situation 1 suggests that [k] and [t] taken from a high vowel context are unlikely productions of either category, while Situation 2 suggests that they are likely productions of /t/ but not of /k/. If the prior probabilities for the two categories were equal, then in the case of Situation 2, $p(C|[k])$ and $p(C|[t])$ are both maximized by /t/. A Bayes Optimal Classifier would categorize both tokens as /t/. In contrast, in Situation 1, it is not clear which category would maximize the posterior probability, and so the outcome of the classification does not suggest perceptual asymmetry as clearly.

This chapter takes the approach described in this section, but there is no expectation that this choice of model would impact whether perceptual asymmetry could be observed, relative to any of the approaches described in §4.1.2. Independent of the model involved or the understanding of similarity taken, it seems this phenomenon can emerge if a situation like Situation 2/2* is possible. There may also be other ways that perceptual asymmetry can emerge from the models described in this section beyond those described by Situation 2/2*. If bias or category prior favors one category over another, for example, then individuals might tend to identify one category more often even under Siutation 1/1*.

4.2 Research Questions

The goal of this chapter is to answer the following question:

RESEARCH QUESTION I

Why do the consonant pairs /k/-/t/, /k/-/p/, /p/-/t/, and /θ/-/f/ show confusion patterns that are specifically asymmetrical?

Asymmetry could potentially appear due to bias (described in §4.1.1.2), or differences in the prior likelihood of a category (described in §4.1.3). While differences in bias or category prior alone are not explored as the sole mechanism driving perceptual asymmetry, they do make an appearance in the models explored in Chapter 5. The following experiments explicitly explore the role of likelihood probability in confusion asymmetries – that consonants may be less similar to their own category than to another in specific phonetic environments. Specifically, the following hypotheses are explored, taking /k/ and /t/ as an illustrative example:

HYPOTHESIS 1A

$\frac{p([k]|/t/)}{p([k]|/k/)}$ *is largest in the context that conditions asymmetry*[1]

The favored category (e.g., /t/) in a perceptual asymmetry pair is predicted to show the greatest likelihood of generating both tokens (e.g., [k] and [t]) relative to the disfavored category, in the phonetic context that conditions the asymmetry. In contrast:

HYPOTHESIS 1B

$\frac{p([t]|/t/)}{p([t]|/k/)}$ *will not be smaller in the context that conditions asymmetry than in other contexts.*[2]

For tokens within the favored category, multiple situations are possible that could favor asymmetry. Tokens of [t] before high front vowels could be even more likely to be generated by /t/ (relative to /k/), in which case the increased [k] confusion rate is compounded by the decrease in [t] confusion rates. It may also simply be that tokens of [t] are comparably likely to be generated by /t/ as in any other context, in which case the increased [k] confusion rates could still drive asymmetry.

The ratio in likelihood probabilities is also predicted to have perceptual relevance – identifications of /t/ are expected to increase as the $p(x|/t/)$ increases (where x is an aocoustic), relative to $p(x|/k/)$, as described in the hypothesis below.

[1] An exemplar-theoretic similarity-based account might frame H1a as: "For a consonant pair that shows perceptual asymmetry, the category less often identified in its conditioning phonetic context (e.g., /k/) will show the greatest similarity to the favored category (/t/), relative to its own category."

[2] And for an exemplar theoretic approach: "Productions of consonants that are more similar to the other category will be more likely to be identified by listeners as the other category."

HYPOTHESIS 2

Listeners will identify [k] and [t] as /t/ more often as $\frac{p(x|/t/)}{p(x|/k/)}$ *becomes larger.*

As noted earlier, perceptual asymmetry is observed in an isolated consonantal context, as well as in a CV context. The hypotheses tested in this chapter rely on the assumption that categorization choice is sensitive to the context-general distributions of the two competitor consonants (e.g., $p(x|/t/)$ vs. $p(x|/t/, /i/)$). This assumption can remain theory-agnostic when considering the consonant-only case of perceptual asymmetry. In this condition, contextual information is absent from the signal, and so it would be more plausible that the token would be compared against other tokens within the categories /k/ and /t/.

It is more difficult for this assumption to remain theory-agnostic if the results in this chapter are expected to speak to perceptual asymmetry occurring in a CV context. The vowel provides information about phonetic context, and listeners show sensitivity to this type of information in perception. If the acoustic information about a consonant varies accordingly by phonetic context, then there may be no reason to assume that all the different phonetic contexts of a token are organized as a single distribution in perception. Perhaps the two distributions under comparison are also conditioned on the phonetic context. However, some theories of phonetic perception do posit the presence of an invariant acoustic cue to a speech sound (e.g., Blumstein & Stevens, 1979), in which case it would seem to not make a difference whether the consonant distributions were also conditioned on phonetic context. The least theory-laden interpretation of the relevance of the following experiments would be that they provide insight into why listeners show perceptual asymmetry when listening to isolated consonants.

4.3 Experiment 4.1

4.3.1 Methodology

Experiment 4.1 serves as a theoretical test of Hypotheses 1a and 1b – that consonant pairs might be distributed in such a way that perceptual asymmetry may emerge simply from the fact that, under certain phonetic conditions, the acoustic realization of one segment is more similar to another category than to its own. As described in §1.4, Plauché (2001) pursued a similar direction of analysis in her dissertation. This chapter also takes up the mantle of a Bayesian framework to probe the source of asymmetries, but from a different perspective. Most straightforwardly, this chapter looks at the localized spectral characteristics of burst, aspiration, and frication, while Plauché looked at less local frequency characteristics, as well as those specified in the time domain and in the vowel. The two approaches also differ in how categories are structured. For Plauché (2001), featural overlap between consonants were analyzed along a single dimension at a time, while in this

chapter the category distributions (and therefore their overlap) are defined in two dimensions, and this methodology could easily be extended to an even higher dimensional setting to accommodate additional acoustic cues.

Finally, whereas for Plauché (2001) consonants were matched for vocalic context during comparisons, this chapter adopted an alternative approach, looking at how a consonant in a specific vocalic context relates to the competitor consonant category. Such an approach may be appropriate in a context where the listener in fact only hears stop and burst information. In fact, the voiceless stop perceptual asymmetries tested in this dissertation appear robustly in situations where the listener does not have vowel information to clarify phonetic context, and so the results of this chapter may help to explain this difference. A parallel analysis that compares each consonant within vocalic context is also possible.

4.3.1.1 Dataset

The dataset for this experiment is the same as for Experiment 3.1 (in §3.3).

The results of the experiments of the previous chapter offer intuition about spectral regions within which differences in energy can be used by algorithms to discriminate between consonants (Experiment 3.1) and are relevant to listeners in perception (Experiment 3.2). A simplified phonetic measure was chosen based on these experiments. Except in the case of /p/ and /t/, where listeners did not show any difference in identification rate, the features adopted for the study are the consensus of those identified from the two experiments. These measures are intended to serve a similar function to F1 and F2 for vowels. Though vowels can be defined along a variety of phonetic dimensions, one common method of description is in terms of these acoustic values. This representation excludes some perceptually relevant information, but also describes coherent categories while still preserving some acoustic structure that listeners find useful in perception.

For /k/ and /t/, energy differences over the range of filters 12-18 were relevant to listeners, but not differences over the range of filters 23-26. Like adjacent points in a time signal, the energies of adjacent frequency bands can be correlated with one another. In designing a phonetic space for /k/ and /t/, a principal component analysis (PCA) was run on values within this filter range to identify uncorrelated linear combinations of these filter energies. The loadings for the first two principal components (PCs)[3]are plotted in Figure 4.1: the first component is roughly proportional to the average energy across this region, while the second component corresponds to a weighted difference in the energy from filters 12-14 and that of filters 15-18. The first two PC scores for the /k/ and /t/ tokens are reported here and elsewhere as $PC1_{kt}$ and $PC2_{kt}$.

[3]The first two principal components accounted for 90% of the variance.

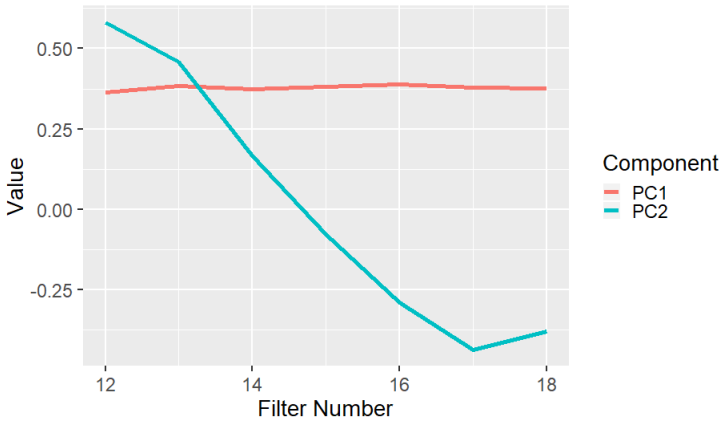

Figure 4.1: Loadings for PCA run for /k/ and /t/ energies over filters 12-18

For /k/ and /p/, filters 12 and 21 were identified as important to classification, but in perception, listener responses were sensitive to a wider range of productions, including not just filter steps 3 and 5, but also 2 and 4. The intersection of these two results, which works out to the energies at Filters 15 and 21, respectively, are selected as the phonetic dimensions for /k/ and /p/ (described here and elsewhere as E_{F15} and E_{F21}).

For /θ/ and /f/, Filters 4, 5[4], 10, 11, and 24 were identified as important to classification by the RF, but listeners seemed generally sensitive to the changes in the spectrum across much of its range. For this reason, a PCA was run that includes filters 4, 5, 10, 11, and 24[5]. As seen in Figure 4.2, the loadings are structured similarly to those for /k/ and /t/ in Figure 4.2 – the first component is like an average of the filter energies, and the second component corresponds to a weighted difference between the energies at low frequencies and those at high frequencies. The principal components contrasting the two consonants are described here and elsewhere as $PC1_{\theta f}$ and $PC2_{\theta f}$.

[4]There was no significant main effect of Step 2 on participant response for either /f/ or /θ/ (relative to Step 1), but this step was associated with a significantly higher rate of confusion than Step 4.

[5]The first two PCs account for 89% of the variance in the data.

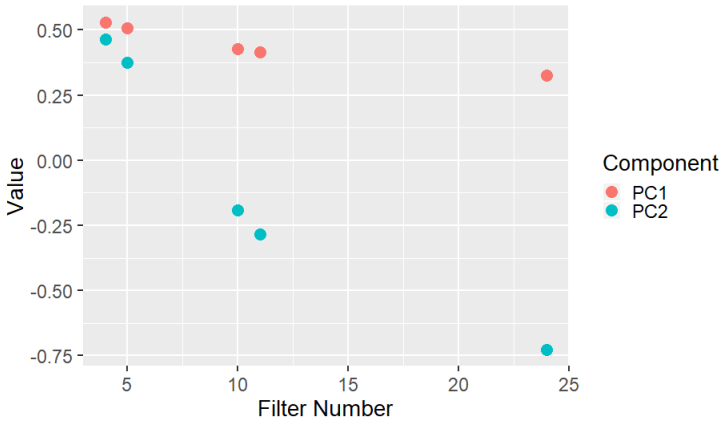

Figure 4.2: PC loadings for /θ/-/f/ contrast over filter steps 4, 5, 10, 11, and 24. Unlike Figure 4.1, the features input to the PCA are not all consecutive

For /p/ and /t/, Filters 12, 18, and 22-25 were identified as important to classification by the RF, but in perception, listeners did not change their response according to filter step. In contrast to the other consonant pairs, the perceptual results do not give a clear indication of an adequate phonetic representation for this consonant pair based on its spectrum. With these caveats in mind, a subset of the features identified from the RF are used to define a phonetic space for /p/ and /t/. Recalling Figure 3.6, filters 22-25 appear to capture a similar difference between /p/ and /t/ to that captured by filters 23-26 for /k/ and /t/. In both cases /t/ has a much higher energy over this region than does the other consonant (as can be seen in Figures 3.2 and 3.6, respectively). Filters 22-25 are not included in the description of /p/ and /t/ because this region shows higher energy for /t/ (relative to /k/ and /p/) across all vocalic environments, and listeners seem not to show sensitivity to differences in energy in this region. The two features for /p/ and /t/ are their respective energies at filters 12 (E_{F12}) and 18 (E_{F18}). The phonetic dimensions for each consonant pair are summarized in Table 4.1.

CONSONANT PAIR	DIMENSION 1	DIMENSION 2
/k/-/t/	$PC1_{\text{kt}}$ (Filters 12-18)	$PC2_{\text{kt}}$ (Filters 12-18)
/k/-/p/	Energy at Filter 12	Energy at Filter 21
/p/-/t/	Energy at Filter 12	Energy at Filter 18
/θ/-/f/	$PC1_{\theta\text{f}}$ (Filters 4,5,10,11,24)	$PC2_{\theta\text{f}}$ (Filters 4,5,10,11,24)

Table 4.1: Acoustic dimensions for each consonant pair

4.3.2 Analysis

To evaluate H1a and H1b, the likelihood that each category in a consonant pair would produce a consonant is compared for each individual consonant in the pair. In the phonetic context con-

ditioning confusion, the likelihood of the favored category generating a consonant with acoustics consistent with the token is predicted to be the highest relative to the favored category. To maintain continuity with other chapters, analyses are reported with respect to three phonetic contexts: before high front vowels ([i,ɪ]), before high back vowels ([u,ʊ]), and before the low back vowel ([ɑ]).

To quantify overlap between consonant categories, this section uses a statistic adapted from multivariate analysis of variance (MANOVA). In this procedure (a multivariate extension to ANOVA), one is interested in the simultaneous comparison of multiple response variables across categories as, for example, when comparing two vowel categories defined in F1 and F2. One of the test statistics for MANOVA is the Pillai-Bartlett trace (also called Pillai score), a value ranging from 0 to 1, where higher values suggest a greater contribution of a factor to the model. This measure has also been used in phonetic analyses before (e.g., Hay et al., 2006, Hall-Lew, 2010) as a measure of pair-wise category overlap, where higher values correspond to two categories with a greater degree of separation from one another. Unlike model parameter estimates, there is no significance assigned to this score, and no ready way to compare whether one overlap is significantly greater than another with this tool. However, the Pillai score can help make concrete some of the visually apparent differences among the phonetic categories. While this score has been used to quantify degree of merger among vowel categories, there is no assumption in this chapter that greater or lesser overlap in these categories implies partial or complete merger of these consonant categories in production. However, higher overlap, associated with increased acoustic similarity, is expected in the phonetic contexts that condition confusability.

4.3.2.1 /k/-/t/

The probability densities for the two categories in the context of high front, high back, and low back vowels are plotted in Figure 4.3 along the phonetic dimensions identified in fig. 4.1. Distributions of /k/ (but not /t/) appear to show variability by vocalic context.

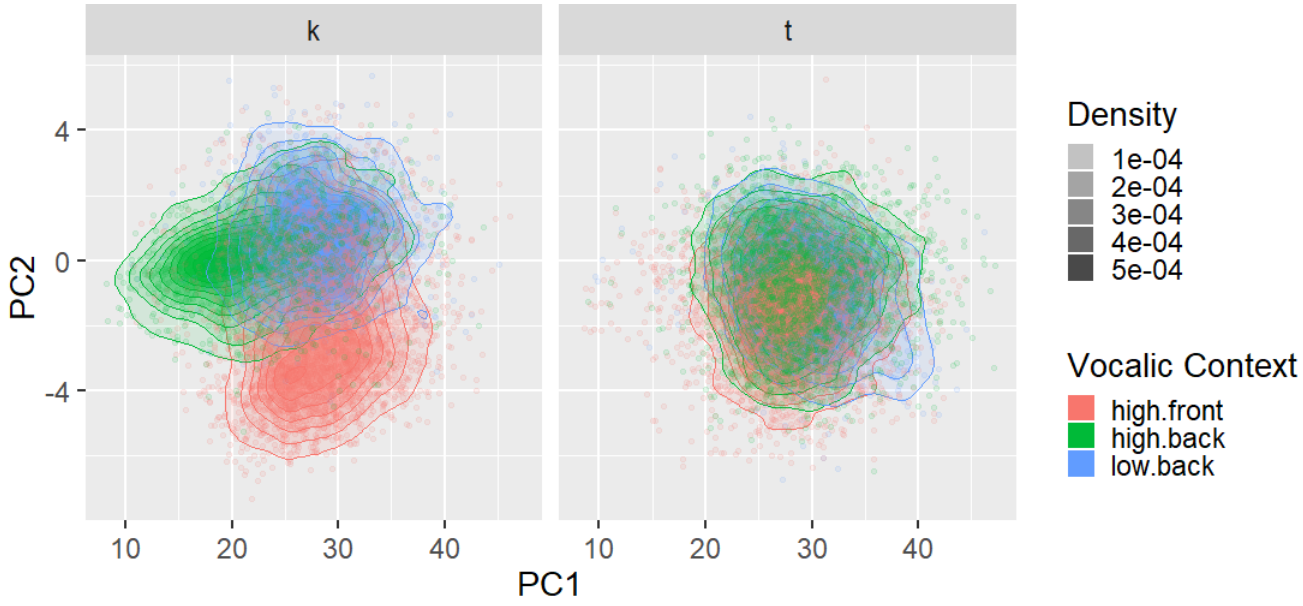

Figure 4.3: Densities for /k/ and /t/ by vocalic context

Indeed, differences between /k/ and /t/ in contextual variability cause them to have differing degrees of within-category overlap. Table 4.2 lists the pairwise Pillai scores for /k/ in each vocalic context plotted in Figure 4.3; small values are interpreted as being due to greater overlap between categories. There is comparatively little overlap between [k] tokens before [u] and before [i].

	LOW BACK (LB)	HIGH FRONT (HF)	HIGH BACK (HB)
LB	-	0.18	0.18
HF		-	0.27
HB			-

Table 4.2: Pairwise within-category Pillai scores for /k/ by vocalic context (/k/-/t/)

For productions of /k/ before high front vowels, PC1 (the overall intensity of the spectrum from 1.3-3.3 kHz) tends to be higher and PC2 (the weighted difference in energies in the 1.3-2.1 kHz and 1.7-3.4 kHz bands) tends to have a wider range than in other vocalic contexts. A low PC2 value in the high front vowel context is expected, consistent with one of /k/'s characteristic peaks (normally located at around 1.5 kHz) being raised due to coarticulation with [i]. The high energy at around 3 kHz has been identified as the approximate location of the peak frequency for the [k] in [ki] in Plauché et al. (1997) and Chang et al. (2001). A high PC2 value in the back vowels reflects the presence of a high-energy peak at around 1.5 kHz. A low PC1 for /k/ before high back vowels may reflect a decrease in the high frequency energy due to the presence of a labial constriction ahead of the consonant release. Such a result would be consistent with the effect the labial constriction has on the amplitude of higher formants (Stevens, 2000, p. 344).

In contrast, [t] tokens appear to show strong overlap with one another in just about every category, as seen in Table 4.3 reflecting the stability of [t] productions across vocalic contexts. Its characteristic high intensity peak at around 4.5 kHz tends to appear in each vocalic context, as can be seen in Figure 3.2.

LB	HF	HB	
LB	-	0.03	0.00
HF		-	0.05
HB			-

Table 4.3: Pairwise within-category Pillai scores for /t/ by vocalic context (/k/-/t/)

The differences in the consonant distributions by vocalic context also affect the degree of overlap between the two consonant categories. Table 4.4 lists the Pillai score between [k] tokens (in a specific vocalic context) and all [t] tokens. There is greater overlap between the /t/ distribution and [k] tokens in high front and low back vowel contexts than in high back contexts. In contrast, /t/ overlaps considerably with the /k/ category in just about every context, as indicated by the low Pillai scores in Table 4.5. /k/ before [i] shows high overlap with /t/, and less overlap with /k/ in other contexts, while /t/ shows high overlap with /k/ in every context tested. These differences in overlap may translate into differing likelihood probability differences by vocalic context. Before [i], a production with acoustics like a /k/ before a high front vowel is predicted to have an especially high likelihood of having been generated by /t/ and a somewhat low chance of having been generated by /k/. This prediction is tested below.

	LB	HF	HB
/k/	0.06	0.04	0.11

Table 4.4: Pillai scores for /k/ (by vocalic context) and /t/

	LB	HF	HB
/t/	0.01	0.03	0.00

Table 4.5: Pillai scores for /t/ (by vocalic context) and /k/

To estimate likelihood probability – how likely it is for a category to generate a consonant that has acoustics of a certain sort, kernel density estimation is used. Just like the process by which plots were generated in Figure 4.3, no strong assumptions are made about the underlying structure of the data, which could be advantageous for irregularly shaped categories, like /k/. A smoothed distribution is estimated that approximates the shape of the data. Density estimation for

each consonant category was performed using the **kde** function in the **ks** R package . Tokens in all vocalic contexts served as input to this model. As shown in Figure 4.4, the density estimates for each category closely resemble the shape of the data plotted in Figure 4.3.

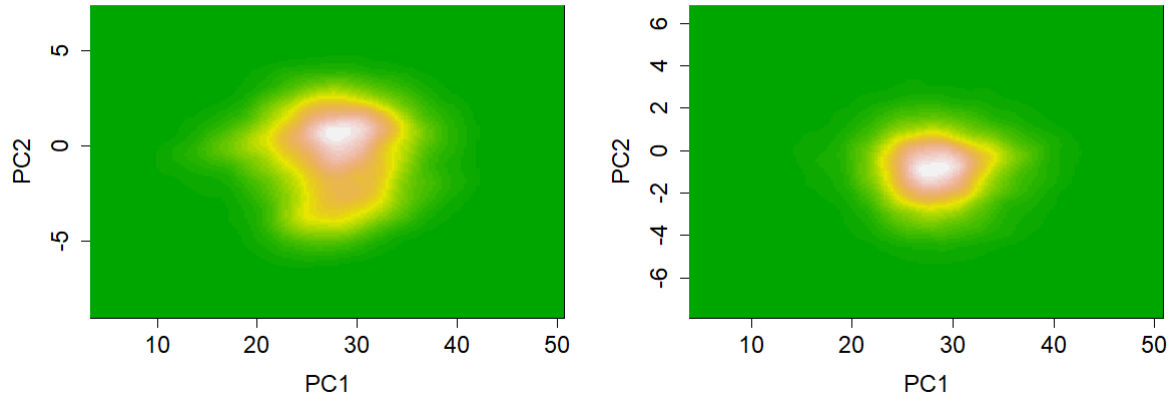

Figure 4.4: Kernel density estimates (KDEs) of /k/ (left) and /t/ (right); reds/whites indicate higher probability density

To test H1a – that $\frac{p([k]|/t/)}{p([k]|/k/)}$ is largest before high front vowels – the two likelihood probabilities are compared for each token. The KDEs for each category were used to generate these values. For $p([k]|/k/)$, for example, the distribution defined in the left panel of Figure 4.4 was used to estimate the value; a token closer to the white region in the distribution would have a higher likelihood probability for /k/ because more [k] tokens tend to have similar acoustics. An omnibus ANOVA was run for [k] tokens, with the independent variable as vocalic context, and the dependent variable as the log difference in likelihood probabilities. The results were significant $[F(2, 5898) = 47.0, p < 0.0001]$. Post-hoc pairwise comparisons for each vocalic context with Holm correction showed that the difference in likelihood probabilities in the high front vowel context was significantly greater than in the high back vowel context ($p<0.0001$) and the low back vowel context ($p<0.0001$), as can be seen in the left panel of Figure 4.5, a plot of individual token difference values by vocalic context. This difference in the low back vowel context, however, did not differ significantly from the high-back vowel context ($p=0.13$). As expected, the vocalic context corresponding to consonant confusions is the same context where the difference in likelihood probabilities is smallest.

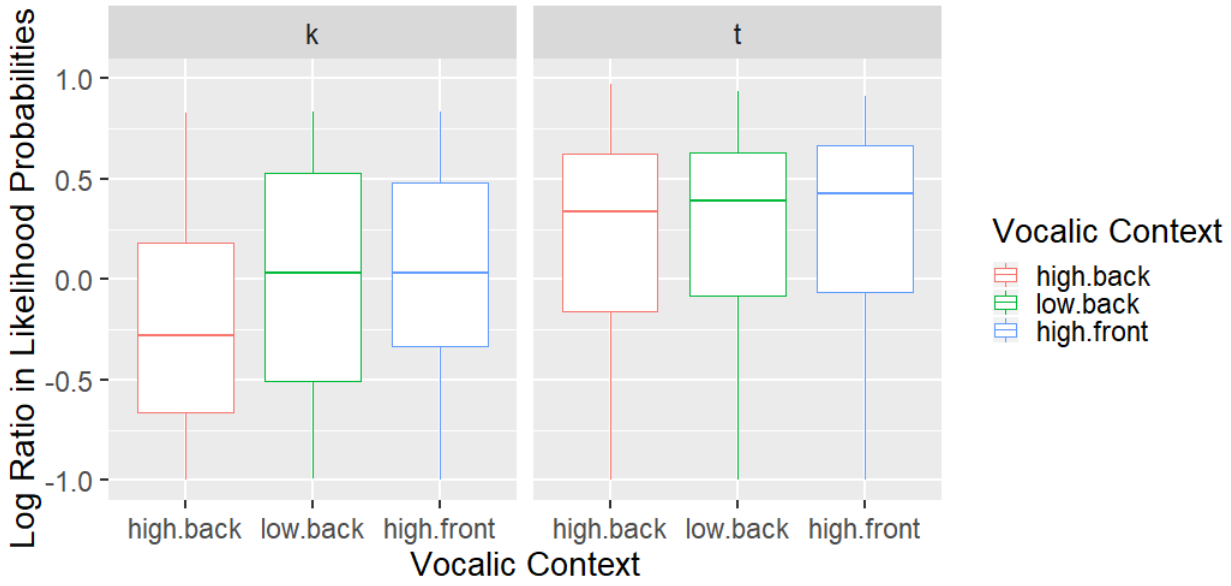

Figure 4.5: Token-wise log ratio likelihood probabilities of /k/ and /t/ – positive values mean $p(x|/t/) > p(x|/k/)$

Tokens of [t] (as seen in the right panel of 4.5) seem to behave differently. Unlike /k/, the median log ratio in likelihood probabilities is greater than zero in every vocalic context tested, suggesting that most of [t] tokens are more likely to be generated by /t/ than /k/. An ANOVA run with identical structure to the model in the previous paragraph also indicates that there is a significant difference in log ratio in likelihood difference according to vocalic context for [t] tokens $[F(2, 7279) = 21.37, p < 0.0001]$, and post-hoc pairwise comparisons indicate that the log ratio in likelihood probabilities is significantly greater before high front vowels than before high back vowels ($p<0.0001$). Productions of [k] and [t] are both more likely to be generated by /t/ before high front vowels, as predicted.

4.3.2.2 /k/-/p/

The probability densities of /k/ and /p/ are plotted in Figure 4.6. While the [k] tokens are the same as those described in §4.3.2.1, they are represented along different phonetic dimensions. Both consonants are plotted with respect to their energies in the 12th and 21st filters (see §4.3.1.1 for more information).

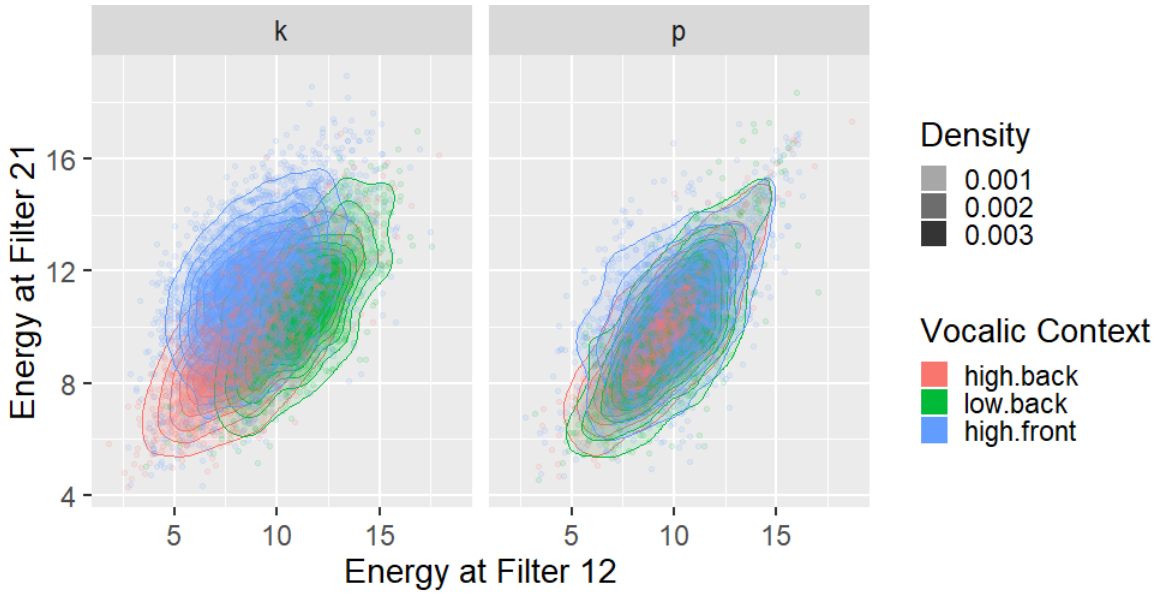

Figure 4.6: Densities for /k/ and /p/ by vocalic context

Just as in §4.3.2.1, /k/ shows phonetic variability with respect to vocalic context. For /k/ there is greater overlap between the high front and high back vowels than other vowel pairings, as can be seen in Table 4.6. Tokens of [k] before high vowels appear to have lower energy at around 1.4 kHz (at the 12th filter), which is consistent with the characteristically low frequency peak of /k/ being shifted to a higher frequency due to coarticulation with [i]. In the high front and high back vowel contexts, /k/ appear to have differing energies at around 4.3 kHz (at the 21st filter). The presence of a labial constriction may contribute to low energy of [k] productions before /u/ at both filters.

	LB	HF	HB
LB	-	0.24	0.21
HF		-	0.15
HB			-

Table 4.6: Within-category Pillai scores for /k/ (/k/-/p/)

In contrast to /k/, /p/ shows near identical degrees of overlap in every context, as can be seen in Table 4.7.

	LB	HF	HB
LB	-	0.02	0.02
HF		-	0.02
HB			-

Table 4.7: Within-category Pillai scores for /p/ (/k/-/p/)

/k/ also shows greater overlap with /p/ before high back vowels and low back vowels (as seen in Table 4.8), while /p/ shows comparable overlap with /k/ in every vocalic context tested (as seen in Table 4.9). This difference in overlap pattern may in fact make it more likely for a production with acoustics like /k/ before [u] to be more likely to have been generated by /p/ and /k/. On the flipside, the mismatch in degree of overlap between /k/ and /p/ before [i] may make it possible that a production with acoustic like /p/ before [i] is more likely to be generated by /k/ than /p/. Both hypotheses are tested.

	LB	HF	HB
/k/	0.04	0.24	0.03

Table 4.8: Pillai scores for /k/ (by vocalic context) and /p/

	LB	HF	HB
/p/	0.02	0.01	0.01

Table 4.9: Pillai scores for /p/ (by vocalic context) and /k/

Kernel density estimation is used again to generate likelihood probabilities for each category. The probability distributions created for each category, pictured in Figure 4.7, resemble those seen in Figure 4.6.

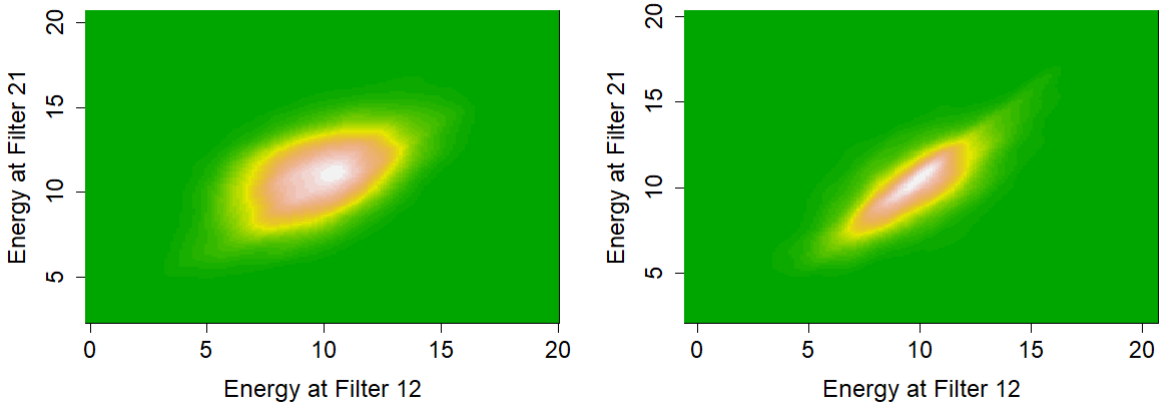

Figure 4.7: KDEs of /k/ (left) and /p/ (right); reds/whites indicate higher probability density

Because perceptual asymmetry between /k/ and /p/ appears to favor /p/ before high back vowels (Winitz et al., 1972 and §3.4.3.2; though see Plauché, 2001), the likelihood probabilities are compared for [k] tokens. The log ratios between the two are plotted in Figure 4.8, with more

positive values corresponding to a higher likelihood probability of being generated by /p/ relative to /k/. In the left panel, token-wise differences appear to be highest in the high back vowel context. An omnibus ANOVA was run for /k/, with vocalic context as the independent variable and the log-ratio of $p(x|/p/)$ and $p(x|/k/)$ as the dependent variable. The model output $[F(2, 5898) = 472, p < 0.0001]$ was significant. Post-hoc pairwise comparisons showed that the log ratio in likelihood probabilities was significantly higher in the high back vowel context than in the high front (p<0.0001) but not the low back (p=0.25) vowel context, contrary to expectation. This difference was also significantly higher in the low back than the high front vowel context (p<0.0001).

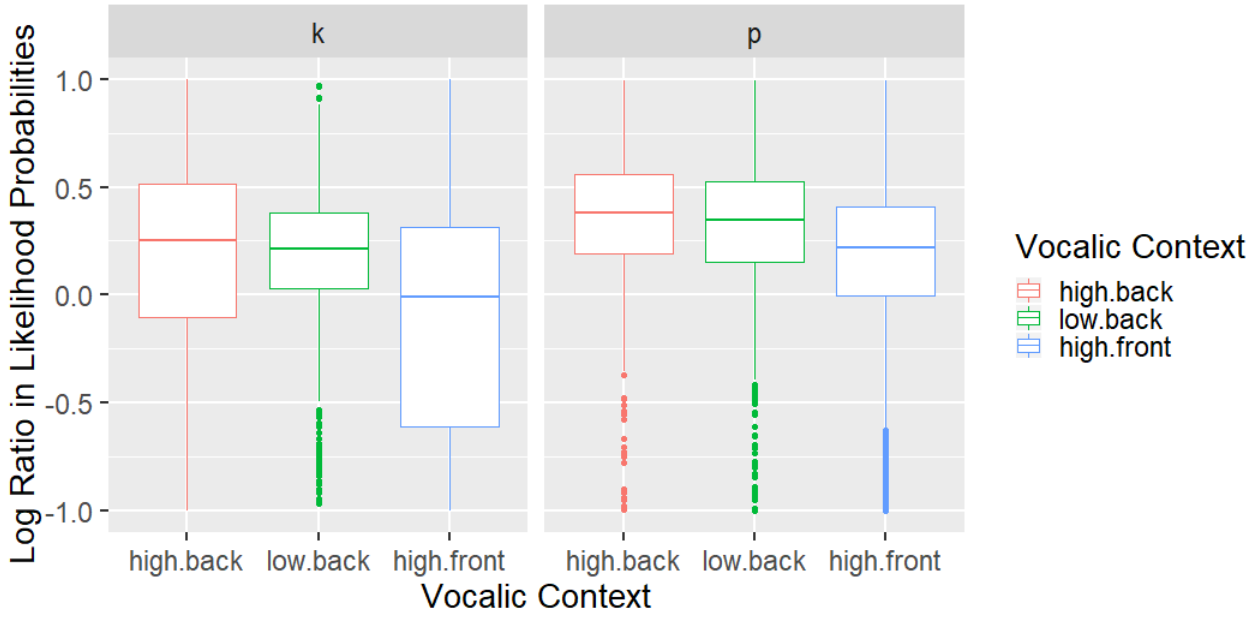

Figure 4.8: Log ratios in likelihood probabilities of /k/ and /p/ – positive values mean $p(x|/p/) > p(x|/k/)$

Perceptual asymmetry between /k/ and /p/ may also favor /k/ before high front vowels (Winitz et al., 1972; but see Plauché, 2001). In the right panel of Figure 4.8, it seems that the likelihood ratios (as in the left panel, higher values correspond to a greater likelihood of being generated by /p/) in high-front vowels are lower than in high back and low back vowel contexts, suggesting a greater likelihood of being generated by /k/. An omnibus ANOVA was run for /p/ tokens, with the same structure as the previous model. This model $[F(2, 2642) = 79.63, p < 0.0001]$ was significant, so post-hoc pairwise comparisons were made for each vocalic context with Holm correction. Again, the dependent variable is the same as in the previous model with [k] tokens, so more positive values correspond to a higher likelihood that /p/ generated the token (and a lower probability that /k/ generated it). As expected, the log ratio in likelihood probabilities before high front vowels was significantly lower than before high back vowels (p<0.0001) or low back vowels (p<0.0001). This

difference in the low back vowel context was not significant before high back vowels (p=0.29).

The same results described above can be reframed to evaluate H1b for each direction of confusion for this consonant pair. For confusions of /k/ and /p/ (before [u]), which favor /p/ before high back vowels, the prediction is that the log-likelihood ratio for [p] productions should not be lower before [u] than any other vocalic context. The results of the previous paragraph show this to be the case – the log ratio in likelihood probabilities before /u/ was significantly higher than before /i/ and did not significantly differ with [ɑ]. Similarly, for confusions of /k/ and /p/ (before [i]), which favor /k/ before high front vowels, the log likelihood ratio for [k] productions is lower before [i] than before all other vocalic contexts, as would be predicted.

4.3.2.3 /p/-/t/

The densities of /p/ and /t/are plotted in Figure 4.9 with respect to vocalic context and energies in the 12th and 18th filters (see §4.3.1.1).

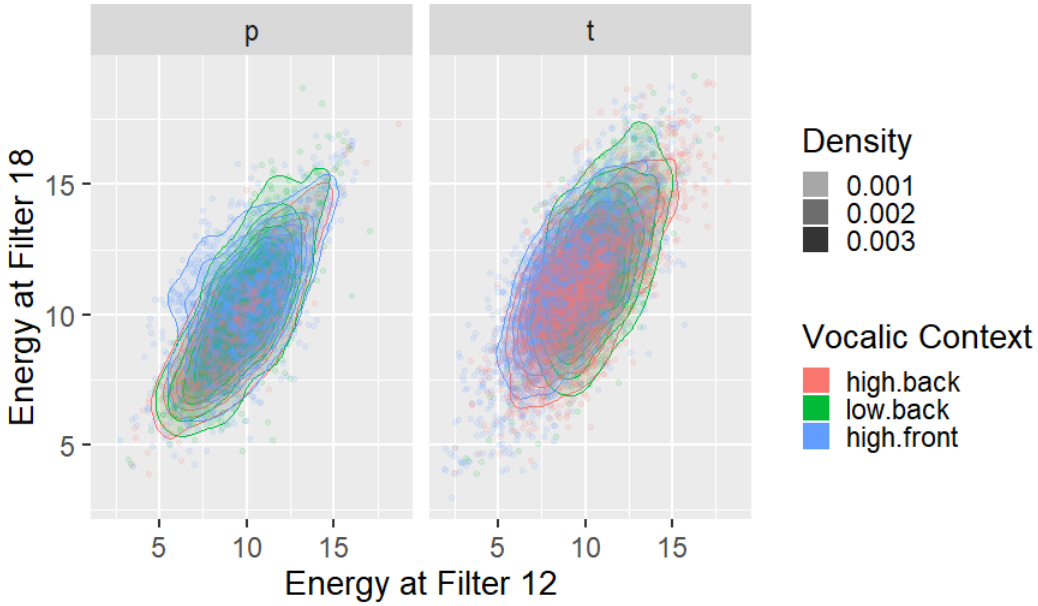

Figure 4.9: Densities for /p/ and /t/ by vocalic context

In the previous two models, /k/ was the category that tended to show more context-dependent variability than its consonant counterpart. For this consonant pair, neither member appears to show variability by context. Indeed, both consonants show strong within-category overlap of Pillai scores for all context pairings, as can be seen in Tables 4.10 and 4.11.

	LB	HF	HB
LB	-	0.00	0.02
HF		-	0.03
HB			-

Table 4.10: Within-category Pillai scores for /p/ (/p/-/t/)

	LB	HF	HB
LB	-	0.04	0.01
HF		-	0.04
HB			-

Table 4.11: Within-category Pillai scores for /t/ (/p/-/t/)

In contrast, there does appear to be some difference in overlap between /p/ and /t/. /p/ shows comparable overlap with /t/ in every vocalic context tested (Table 4.12). However, /t/ shows less overlap with /p/ before high front vowels (Table 4.13). This observation would be consistent with the vocalic context associated with the perceptual asymmetry for /p/ and /t/, which happens before high front vowels. Perhaps there are fewer [t] tokens that sound like /p/ than there are [p] tokens that sound like /t/.

	LB	HF	HB
/p/	0.01	0.02	0.02

Table 4.12: Pillai scores for /p/ (by vocalic context) and /t/

	LB	HF	HB
/t/	0.05	0.19	0.07

Table 4.13: Pillai scores for /t/ (by vocalic context) and /p/

Kernel density estimates were taken for the two categories (shown in Figure 4.10), and resemble the distributions observed in Figure 4.9.

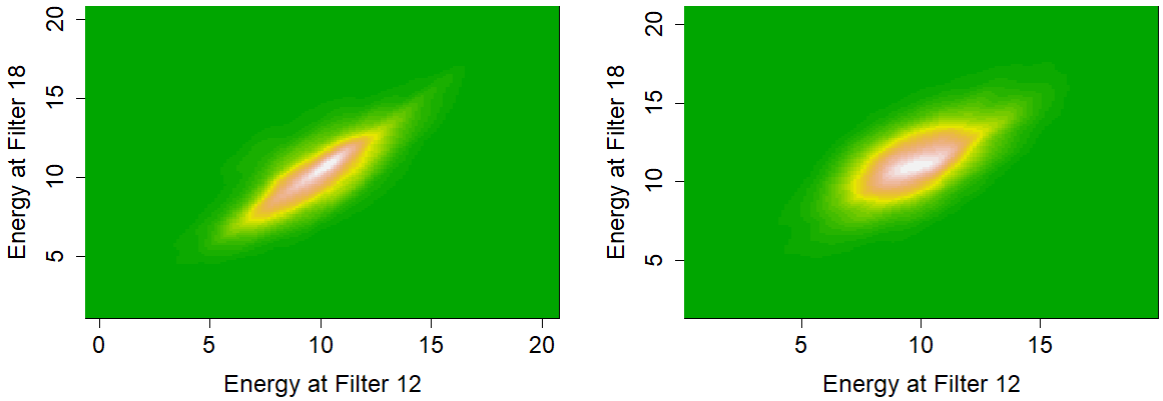

Figure 4.10: KDEs of /p/ (left) and /t/ (right); reds/whites indicate higher probability density

Confusions between /p/ and /t/ favor /t/ before high front vowels. A comparison of probabilities was performed just like in previous sections and those differences are plotted in the left panel of Figure 4.11. In this figure the log likelihood ratios in the high front vowel context seem higher than in the low back context, but the relationship between the other pairings is less visually apparent. The output of an omnibus ANOVA run for /p/, with vocalic context as the independent variable and the log ratio of $p(x|/t/)$ and $p(x|/p/)$ as the dependent variable, was significant $[F(2, 2642) = 45.8, p < 0.0001]$. Post-hoc pairwise comparisons for vocalic contexts showed that the difference measure was significantly higher in the high front than in the high back ($p<0.0001$) and the low back ($p<0.0001$) vowel contexts, as expected. The difference was also significantly higher before low back vowels than before high back vowels ($p<0.0001$). This result does not go against predictions but was not expected.

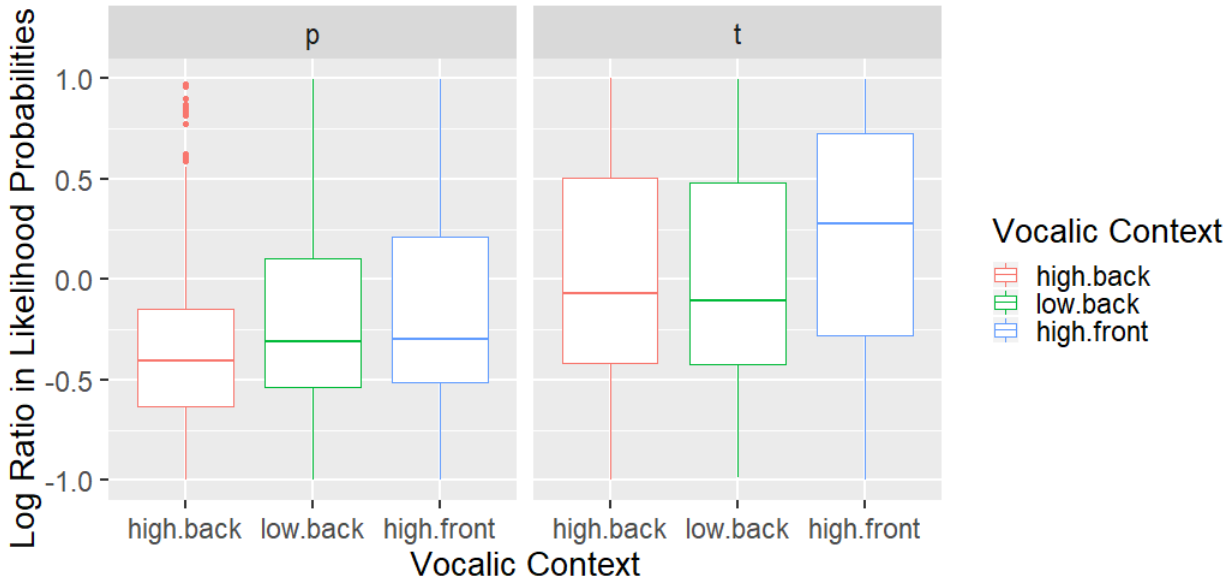

Figure 4.11: Log ratio in likelihood probabilities for /p/ and /t/ – positive values mean $p(x|/t/) > p(x|/p/)$

As in §4.3.2.1, it is worth considering /t/ in parallel to /p/. An ANOVA run with parallel structure to that of the previous paragraph reveals a significant effect of vocalic context on the log ratio of likelihood probabilities of /t/ (Figure 4.11, right panel). Post-hoc pairwise comparisons reveal that the log likelihood ratio is significantly higher in the context of high front vowels than high back ($p<0.001$) and low back ($p<0.001$) vowels. The log-ratio of likelihood probabilities for /t/ before low back vowels is also higher than before high back vowels ($p<0.001$). Like §4.3.2.1, the context that favors confusion is the context where both [p] and [t] are the most likely to have been generated by /t/ (relative to /p/).

4.3.2.4 /θ/-/f/

Figure 4.12 shows the probability densities of /f/ and /θ/, plotted along two PCs as defined in §4.3.2.1. Like /p/ and /t/, each fricative shows relatively little variability with respect to vocalic context.

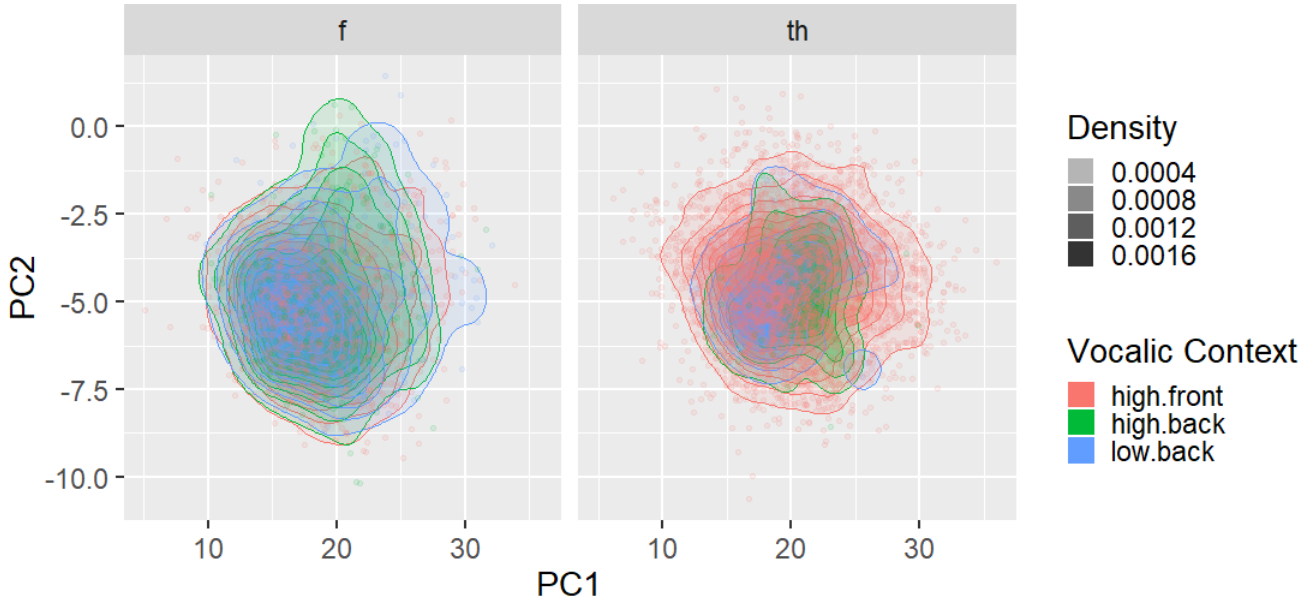

Figure 4.12: Densities for /f/ and /θ/ plotted by vocalic context

In Tables 4.14 and 4.15, the pairwise within-category Pillai scores for these tokens indicate strong overlap across just about every vocalic context tested.

	LB	HF	HB
LB	-	0.00	0.03
HF		-	0.00
HB			-

Table 4.14: Within-category Pillai scores for /θ/ (/θ/-/f/)

	LB	HF	HB
LB	-	0.01	0.00
HF		-	0.00
HB			-

Table 4.15: Within-category Pillai scores for /f/ (/θ/-/f/)

Similarly, /f/ and /θ/ show strong overlap with one another in every vocalic context tested. Before high front vowels there is a comparatively high inter-category Pillai score, but these values are still small. There is neither a strong prediction from the data nor from the prior literature about what context would favor asymmetry between /θ/ and /f/.

	LB	HF	HB
/θ/	0.00	0.03	0.00

Table 4.16: Pillai scores for /θ/ (by vocalic context) and /f/

	LB	HF	HB
/f/	0.00	0.03	0.00

Table 4.17: Pillai scores for /f/ (by vocalic context) and /θ/

Kernel density estimation is used to generate likelihood probabilities for each category, the results of which are visualized in Figure 4.13 (compare with Figure 4.12).

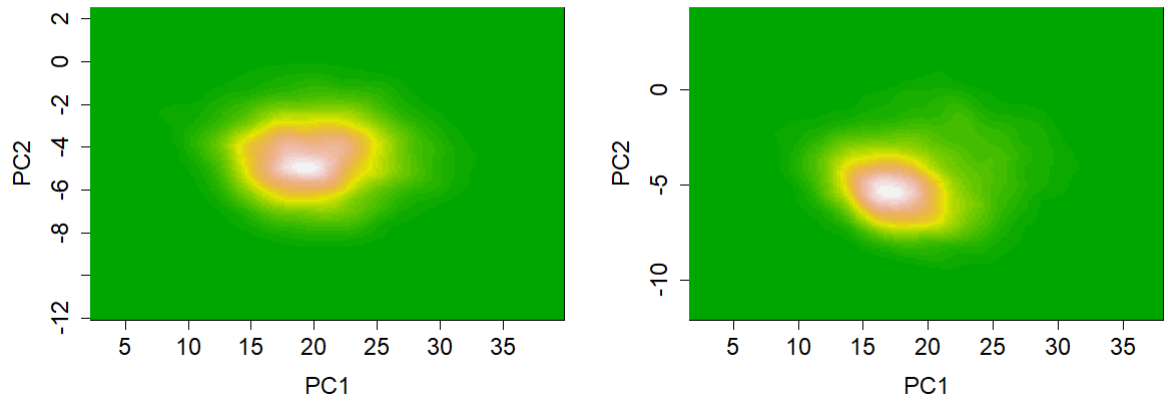

Figure 4.13: KDEs of /θ/ (left) and /f/ (right); reds/whites indicate higher probability density

Perceptual asymmetry between /f/ and /θ/ favors /f/, with no context identified in the literature (but perhaps /ɑ/ from §3.5). The differences between the two likelihood probabilities are plotted in the left panel of Figure 4.14. An omnibus ANOVA was run for /θ/, with vocalic context as the independent variable and the log ratio in the likelihood probabilities as the dependent variable. This model was not significant $[F(2, 4152) = 0.94, p = 0.38]$, so post-hoc pairwise comparisons were not run. The results of §3.4 suggest that these two consonants are relatively confusable in all vocalic contexts, perhaps more so before [ɑ], but the output of this model does not suggest a vocalic context that would condition asymmetry.

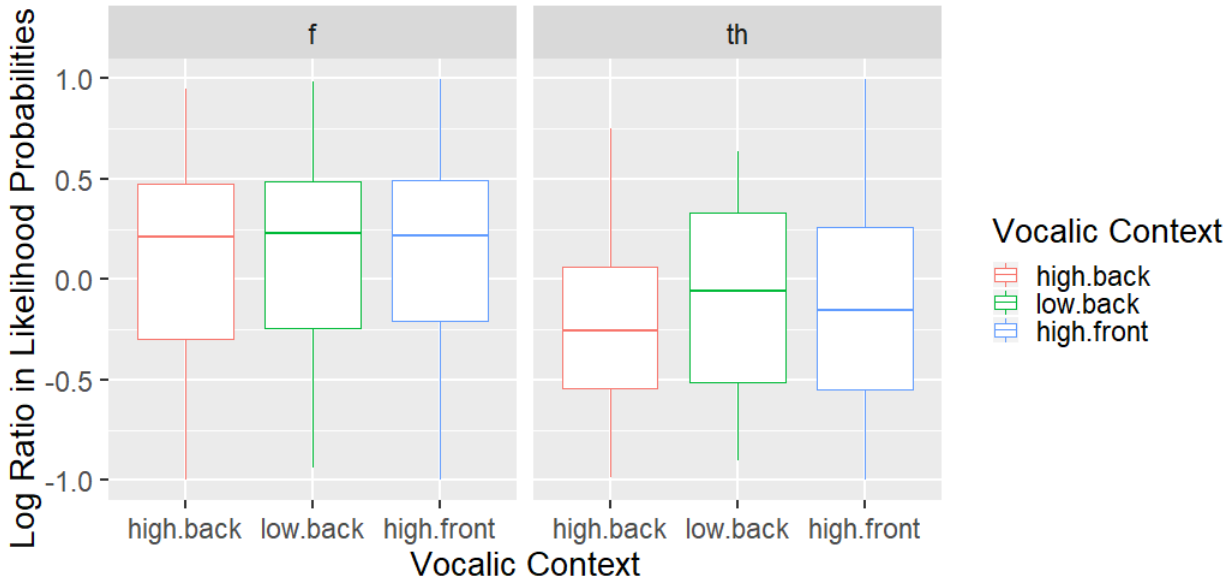

Figure 4.14: Log ratio in likelihood probabilities for /θ/ and /f/ – positive values mean $p(x|/f/) > p(x|/\theta/)$

For /f/, there is little variability also – an ANOVA run on the log-likelihood ratio of /f/ reveals no significant effect of vocalic context $[F(2, 1644) = 0.09, p = 0.92]$. Vocalic context appears not to make [f] more or less likely to be generated by /f/ relative to /θ/.

4.3.3 Interim discussion

Experiment 4.1 took a probabilistic approach to attempt an explanation for why certain consonant pairs show might show asymmetric confusion patterns in restricted contexts. The results suggest that the acoustics of the disfavored consonant in an asymmetry pair become more like those of the favored category (relative to its own category) in the phonetic context conditioning asymmetry. For this experiment, the difference in likelihood probabilities between the two categories was compared for each vocalic context. It was predicted that this difference in likelihood probabilities would be smallest in the vowel context conditioning asymmetry. The results of the ANOVAs and pairwise comparisons based on the four consonant pairs are summarized in Table 4.18.

CONSONANT PAIR	FAVORS	H1A PREDICTIONS	H1A RESULTS	H1B PREDICTIONS	H1B RESULTS
/k/-/t/	/t/	LB, HB $<$ HF	LB, HB $<$ HF	LB, HB $\leq$ HF	HB $<$ HF, LB
/k/-/p/	/p/	LB, HF $<$ HB	HF $<$ HB, LB	LB, HF $\leq$ HB	HF $<$ HB, LB
/k/-/p/	/k/	LB, HB $<$ HF	LB, HB $<$ HF	HB, LB $\leq$ HF	HB, LB $<$ HF [6]
/p/-/t/	/t/	LB, HB $<$ HF	LB $<$ HB $<$ HF[7]	LB, HB $\geq$ HF	HB $<$ LB $<$ HF
/θ/-/f/	/f/	???	*no diff*	???	*no diff*

Table 4.18: Experiment 4.1 results summary (HF=High Front, HB=High Back, LB=Low Back). Green highlight indicates results consistent with predictions; yellow highlight indicates results partially inconsistent with predictions.

H1b was fully supported by the data – in the context conditioning asymmetry, the log-ratio of likelihood probabilities for tokens of the favored consonant was always higher than or comparably high to all other contexts. As for H1a, the results largely follow the pattern that the vocalic context conditioning asymmetry is the context with the likelihood difference measure. For /k/-/t/, /k/-/p/ (favoring /k/), and /p/-/t/, the differences in the context of high front vowels were smaller than the differences in all other vocalic categories, as predicted. In the case of /k/-/p/ (favoring /p/), however, the difference in the predicted vocalic context was not significantly smaller than the differences in all other tested contexts.

There is no indication in the literature of a vocalic context that would condition perceptual asymmetry for dental fricatives. The results of Experiment 4.1 suggest that there may not be a single vocalic context that conditions confusability for this consonant pair. This trait would set the fricative pair apart from the voiceless stops, whose misidentifications tend to be much rarer outside the context associated with confusion. It would, however, be consistent with the acoustic descriptions of the pair, which offer few spectral traits that distinguish between the two. These results are especially interesting in light of Experiment 3.2, where listeners tended to misidentify [θ] more often in the environment of [ɑ] (see Experiment 3.2). Perhaps this difference in misidentification rate observed in that experiment depends on cues present in the vowel.

[6] Because /k/-/p/ are the same consonant pair with two predicted confusions, these results are the same as are listed for the H1a Results

[7] H1 predicts a specific relationship between the HF condition and the LB and HB conditions. There are no predictions about how the HB and LB conditions should relate to one another.

4.4 Experiment 4.2

4.4.1 Methodology

The results of Experiment 4.1 suggest that the acoustics of consonant categories may be distributed in a way that could lead to increased identification errors in the vocalic contexts that show perceptual asymmetry. However, the perceptual consequences of a difference in likelihood probabilities is not clear. The goal of this second experiment is to see whether differences in likelihood probability can predict listener classification errors.

4.4.1.1 Stimuli

The stimuli for this perceptual experiment are tokens consisting of the isolated burst and aspiration of a stop ([p],[t],[k]) or the frication portion of a fricative ([f],[θ]), drawn from the vocalic context associated with increased confusion rates. These tokens are a subset of those that were used in Experiment 4.21. Based on the results of Experiment 4.1, each individual consonant in a pair is associated with its own likelihood probabilities for each category. Some are rated as being more likely to be generated by one category and others are rated as being less likely for this to have happened. For a consonant category in a pair (e.g., /k/ of /k/-/t/), all the consonants were ordered by this measure. The 15 consonants that were most likely to be generated by their category were selected – from these tokens, five stimuli were chosen for this experiment that were correctly labeled, free of noise, and not excessively short. The stimuli were bounded on both sides by 50 ms of silence. Fifteen more tokens were selected that were the least likely to be generated by their category, and from these five stimuli were selected for the study according to the same criteria. There were 20 stimuli in each pair (2 consonants per pair x 2 categories (more likely category A/more likely category B) x 5 tokens), for 100 unique stimuli. The consonant pairs addressed in this experiment are summarized in Table 4.19.

CONSONANT PAIR	FAVORS	VOCALIC CONTEXT
/k/-/t/	/t/	[i]
/k/-/p/	/p/	[u]
/k/-/p/	/k/	[i]
/p/-/t/	/t/	[i]
/θ/-/f/	/f/	[ɑ][8]

Table 4.19: Consonant pairs and contexts investigated in Experiment 4.2

[8]This vocalic context was chosen based on the results of Experiment 3.2.

Figure 4.15 plots PC values for tokens selected for the /k/-/t/ contrast. In the left panel, the red tokens are the five of the 15 [k] (before high front vowel) tokens with the highest probability of being generated by /k/ (relative to /t/). The blue tokens are five of the 15 tokens with the lowest probability of being generated by /k/ (relative to /t/). In the right panel, the blue points are five of the 15 /t/-productions with the highest probability of being generated by /t/ (relative to /k/), and the red points are five of the 15 productions with the lowest probability. Notice that the five 'unlikely /k/' /k/ tokens are grouped closely together with the 'unlikely /k/' /t/ tokens. The same is not true for the 'likely /k/' /k/ tokens, which are relatively distant from the 'likely /k/' /t/ tokens. Recalling Figure 4.3, [k] and [t] have differently shaped distributions, and so there may not be /t/ tokens in the same region as the 'likely /k/' /k/ tokens. Figures E.1-E.4 in Appendix E include plots of the stimuli selected for the other four consonant pairs.

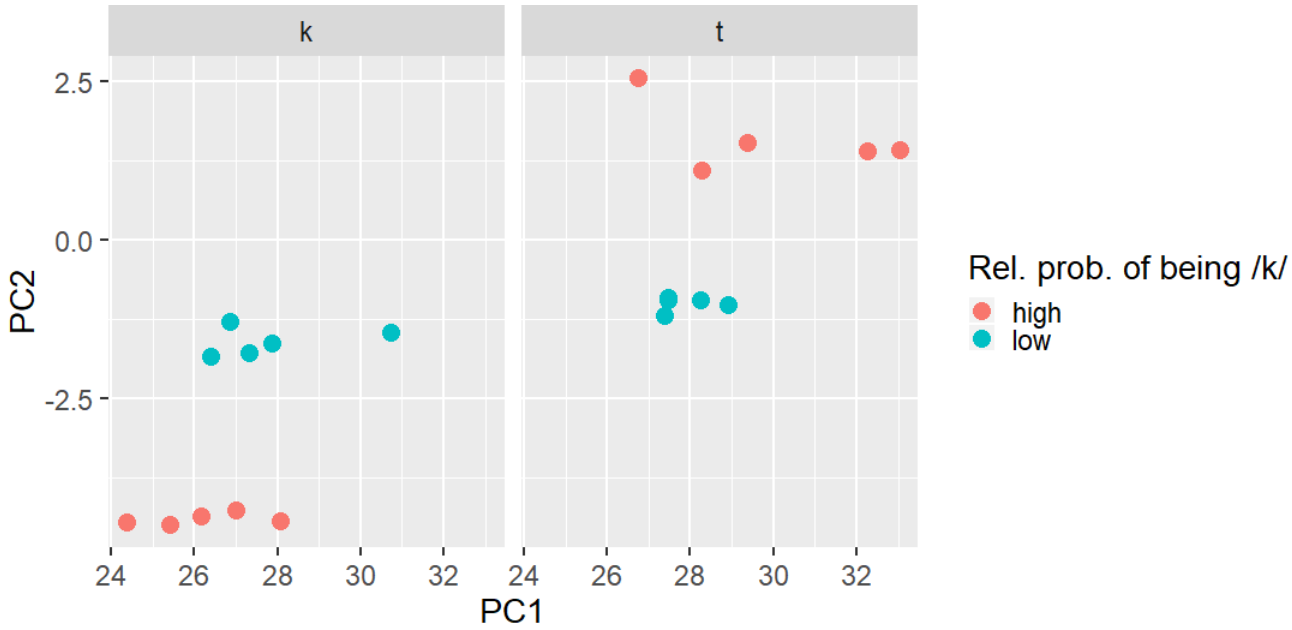

Figure 4.15: /k/ and /t/ stimuli selected for Experiment 4.2

The calculations used to determine the relative likelihood of each token being either member of a consonant pair do not compare perfectly to the task the participants must perform. Out of concerns for perceptibility, listeners are not just presented with a 20 ms clip of speech for each token. However, as a result, listeners have access to information about the duration of the consonant, for example, while the model does not. Not every type of difference can be tested, but it is possible to figure out whether duration differences reliably predict grouping. Figure 4.16 presents box plots of token duration by consonant pair. The [f] and [θ] tokens rated more likely to be generated by /f/ are shown in one boxplot for that consonant pair, and those rated more likely to be generated by /θ/ appear in the other. The blue plots correspond to the lower duration category in the group.

A logistic regression was run with category (i.e., higher/lower duration group) as the dependent variable and duration, consonant pair, and the interaction of the two as the independent variables, and this model does not reveal a significant effect of any of the factors. While Figure 4.16 suggests slight differences in duration, these differences do not seem reliable enough to significantly predict grouping.

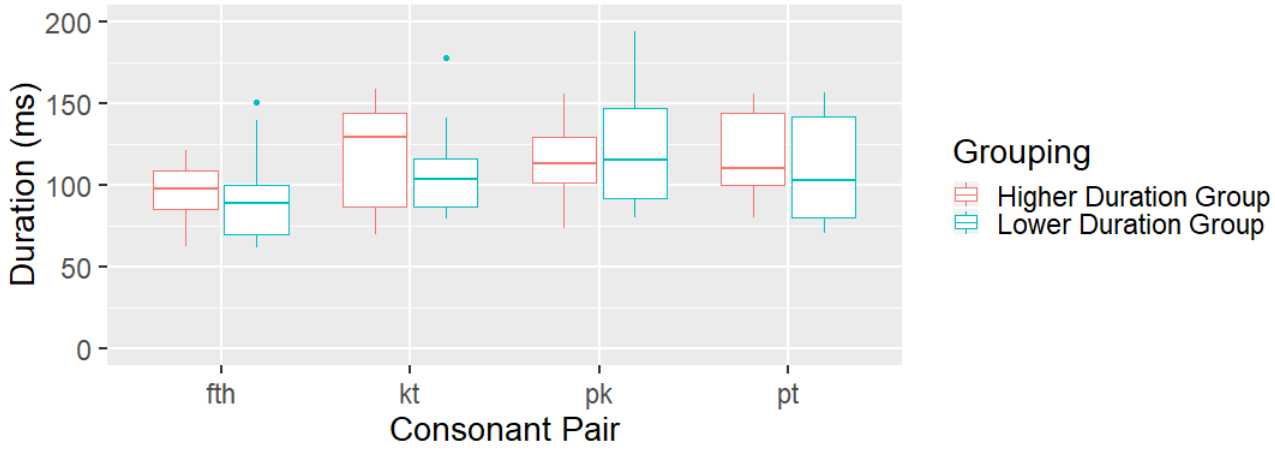

Figure 4.16: Token durations by consonant pair and grouping

4.4.1.2 Participants

Participants in this study are the same as those who participated in Experiment 3.2. The order of the perceptual tasks in Chapters 3 and 4 is counterbalanced across participants.

4.4.1.3 Procedure

The experimental setting and equipment were the same as for Experiment 3.2. Participants completed a two-choice forced choice task. In each trial, the participant first sees two letters (e.g., 'p' and 't') spaced along the center of the screen. These consonants (letters) are members of a pair that show confusion asymmetry. After 500 ms, the participant hears an isolated consonant (e.g., [p] or [t]) and responds on a keyboard by pressing the key corresponding to the consonant they heard. The color and ordering of the letters on the screen match those on the keyboard – for example, the leftmost choice on the screen and the key on the keyboard corresponding to the left choice are both red. 500 ms after their response (or 10 seconds after the audio stimulus played) the trial ends. The ordering of the letters is held constant for each participant but counterbalanced across participants. The participant responds to each stimulus four times, for a total of 400 trials.

At the end of both experiments (this experiment and Experiment 3.2), participants complete a questionnaire about their linguistic background.

4.4.2 Predictions

In §4.3, vocalic context was a predictor of the difference in likelihood probabilities between the categories of a consonant pair. In this experiment, consonant group is predicted to affect listeners' categorization strategy. For example, a listener is predicted to be more likely to categorize a [k] (before a high front vowel) as /t/ if it came from the 'unlikely /k/' group than if it came from the 'likely /k/' group.

4.4.3 Results

For each consonant pair, a binomial regression is run, with the dependent variable as their binary response (e.g., /k/ or /t/) and the fixed predictors as grouping (e.g., 'likely /k/'/'unlikely /k/'), the consonant presented to the listener, and their interaction, with participant as a random factor.

4.4.3.1 /k/-/t/

For [k] and [t] (extracted from an [i] context), there was a significant effect of consonant ($\beta = 5.53, z = 12.27, p < 0.0001$) and grouping ($\beta = -1.32, z = -3.46, p = 0.0005$) on participant likelihood to respond /k/ in the two-choice categorization task. The log-odds of a listener categorizing a token as /k/ decreases by 1.32 if the token is in the Low ('unlikely /k/') group. There were no other interaction effects observed. These results are reflected in Figure 4.17, which plots probabilities of a /k/ response by grouping and consonant (averaged by participant). There is greater inter-listener variability in response when hearing [k] than when hearing [t], but in both cases, participants are more likely to choose /k/ when the token is the High ('likely /k/') group.

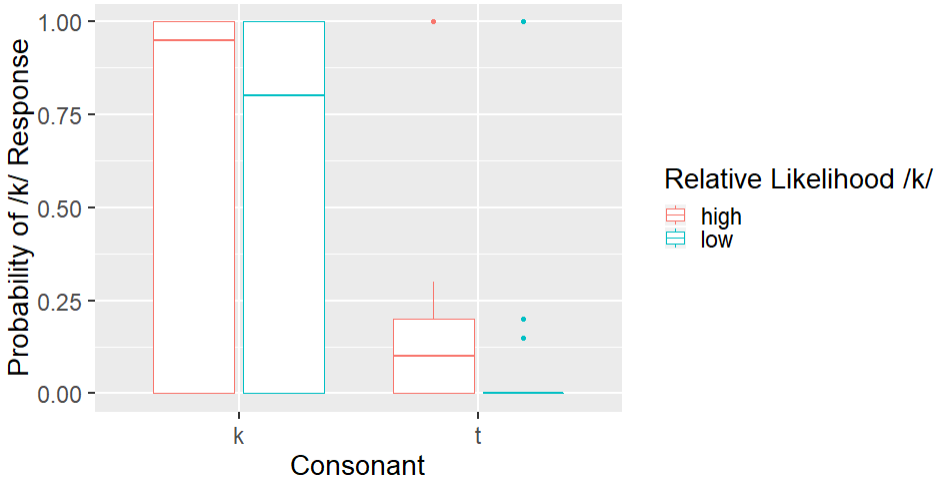

Figure 4.17: Plot of average listener likelihood to respond /k/ by consonant and grouping (in the context of [i])

4.4.3.2 /k/-/p/

In contrast to [k]-[t], the results of [k] and [p] do not follow predictions for either the original high front or high back vowel context. Before high front vowels, there are increased rates of confusion between /k/ and /p/, which appears to favor /k/ (Winitz et al., 1972) or /p/ (Plauché, 2001). Figure 4.18 plots probabilities of /p/ response by grouping and consonant for the high front vowel context. While the /p/ response is less likely for the Low ('unlikely /p/') group when hearing [k(i)], as expected, it appears to be greater for the Low group when hearing [p(i)], which is contrary to expectations. A logistic regression confirms these results: there was a significant effect of consonant ($\beta = 1.92, z = 10.35, p < 0.0001$) and grouping ($\beta = -0.46, z = -2.90, p = 0.004$) on listeners' likelihood to respond /k/ as well as an interaction effect ($\beta = 0.60z = 2.23, p = 0.03$), contrary to predictions. This result suggests that the grouping effect is not affecting response in a consistent manner across consonants.

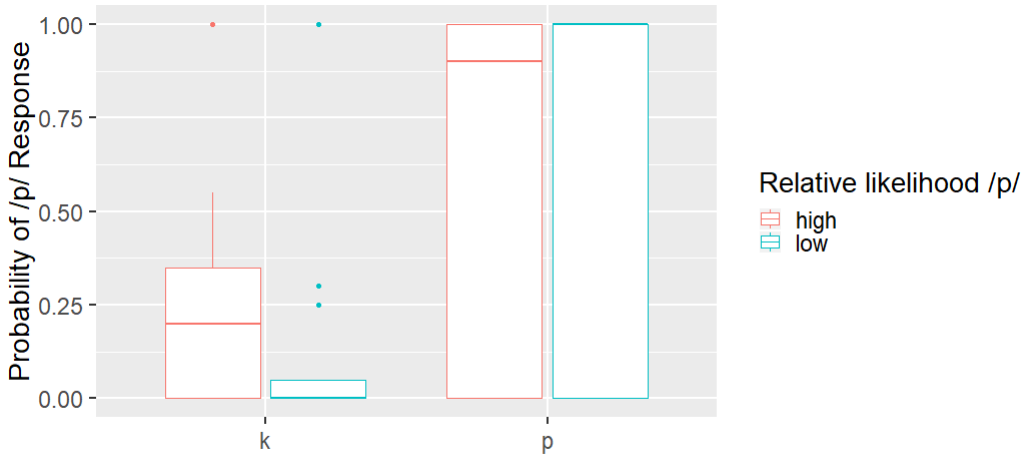

Figure 4.18: Plot of average listener likelihood to respond /p/ by consonant and grouping (in the context of [i])

There are also increased rates of confusion between /k/ and /p/ before high back vowels. Figure 4.19 plots participant probabilities of choosing /k/ when responding to [k] or [p] extracted from the context of [u]. While /k/ responses are less likely in the Low ('unlikely /k/') group when responding to [k(u)], there does not appear to be much difference by grouping when responding to [p(u)]. A logistic regression confirms these observations: there was a significant effect of consonant ($\beta = 2.54, z = 12.97, p < 0.0001$) and grouping ($\beta = -0.73, z = -6.05, p < 0.001$) on participant likelihood to respond /k/ as well as an interaction effect ($\beta = 0.87, z = 3.50, p = 0.0004$), which was contrary to predictions. As with the original [i] context, the effect of grouping on listener response for the original [u] context is not consistent across consonants.

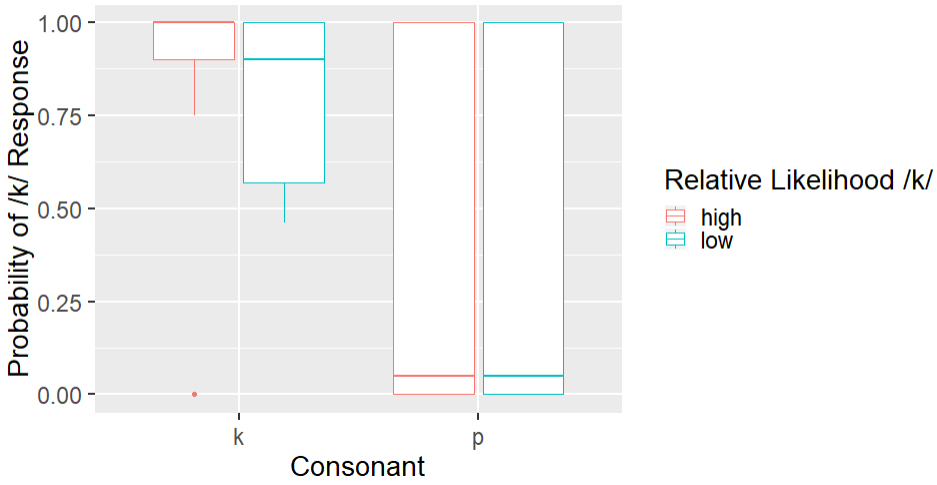

Figure 4.19: Plot of average listener likelihood to respond /k/ by consonant and grouping (in the context of [u])

4.4.3.3 /p/-/t/

Before high front vowels, listeners show increased rates of confusion between /p/ and /t/. As is apparent in Figure 4.20, there appears to have been greater consistency across speakers in their response to [t] than to [p] (extracted from an [i] context). Nonetheless, as predicted, there was a significant effect of consonant on response ($\beta = 8.27, z = 12.90, p < 0.0001$) indicating that listeners were more likely to choose /p/ when they in fact heard [p]. There was also a weak effect of grouping observed ($\beta = -1.12, z = -2.15, p = 0.03$); when listeners heard a token from the Low (unlikely 'p') group, their log odds of choosing /p/ would decrease by 1.12. No significant effect of interaction was observed.

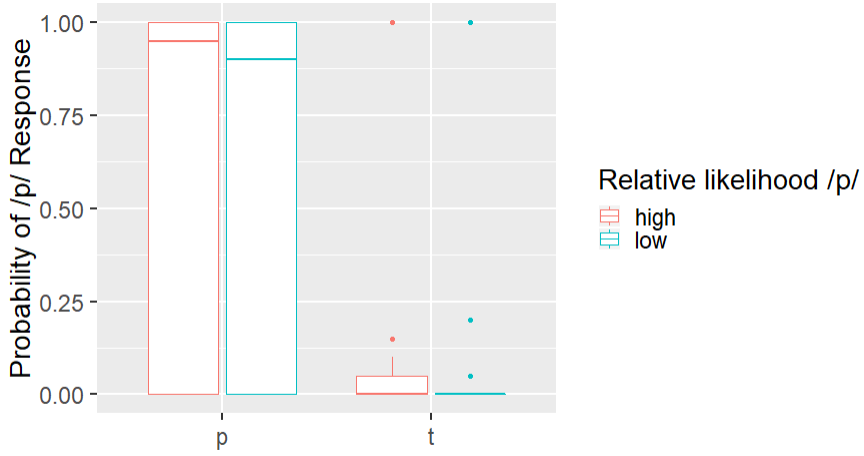

Figure 4.20: Plot of average listener likelihood to respond /t/ by consonant and grouping (in the context of [i])

4.4.3.4 /θ/-/f/

This consonant pair is associated with increased confusability, but the precise environment in which it occurs has not been specified in the literature. The results of Experiment 3.2 suggest that there are higher rates of confusion for /θ/ extracted from an [ɑ] context, relative to other contexts. As can be seen in Figure 4.21, listeners appear more likely to respond /θ/ when listening to the High ('likely /θ/') group tokens for both consonant categories. As predicted, there is a significant effect of consonant $(\beta = 2.04, z = 8.19, p < 0.0001)$ and grouping $(\beta = -0.90, z = 3.85, p = 0.0001)$ on participant likelihood to respond /θ/, but no significant interaction effect. The log-odds of a listener categorizing a token as /θ/ decreases by 0.90 if the token is in the Low ('unlikely /θ/') group.

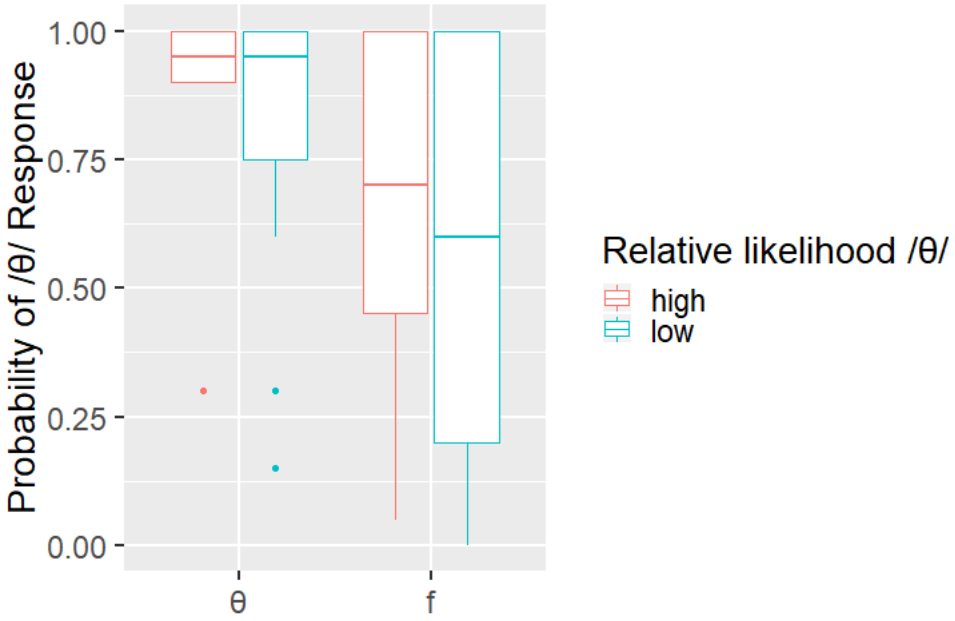

Figure 4.21: Plot of average listener likelihood to respond /f/ by consonant and grouping (in the context of [ɑ])

4.4.4 Results Summary

The purpose of this experiment was to determine the extent to which modeled differences in likelihood probability hold perceptual relevance for the listener when categorizing consonant pairs in the context that conditions asymmetry. The results of this experiment are summarized in Table 4.20. The results' partial support of the initial hypothesis suggests that a more nuanced reformulation of H2 might be needed to better understand why grouping had a consistent effect on listener response on only three of the four consonant pairs explored.

PAIR	EFFECT OF GROUPING
/k/-/t/	Consistent b/w consonants
/k/-/p/	Inconsistent b/w consonants
/p/-/t/	Consistent b/w consonants
/θ/-/f/	Consistent b/w consonants

Table 4.20: Summary of results for Experiment 4.2; Green highlighting indicates results that matched predictions, and red highlighting indicates results that disagreed with predictions

Listeners performed as expected on the consonant pair /θ/ and /f/ - they were more likely to categorize [f] and [θ] tokens as /θ/ if the tokens came from the 'likely /θ/' group. This result suggests a consistent relationship between the information used by listeners to classify the dental fricatives and the information used in the model used to rank their relative probabilities of having been generated by /θ/. It is possible that the spectral features identified in §3.3 specifically drive this difference in identification rate. It is also possible that differences in the phonetic distributions of the two consonant pairs as represented here are simply correlated with other perceptually relevant features in the acoustic signal which the model did not have access to. This experiment is not able to tease apart which of the two accounts led to a consistent effect for this consonant pair.

Listeners also performed as expected for the consonant pair /k/-/t/. They were more likely to categorize a [t] or [k] as /k/ when they came from the 'likely /k/' group. If the results of Guion (1998) are considered, which showed that listeners were more likely to misidentify a [k] production as /t/ when produced during rapid speech, it would seem probable that [k] tokens would vary in the degree to which they could be perceived as /k/. In contrast, that /t/ participates in perceptual asymmetry with both /p/ and /k/ (with /t/ being the favored outcome) suggests that listeners would tend to correctly identify [t] as /t/ when it is produced. However, listeners did show variability in their likelihood to categorize [k] as /k/ according to grouping.

Listener responses to the stop contrast /p/-/t/ are also as expected. Recalling the perceptual experiment in Chapter 3 (Experiment 3.1), filter step did not predict the misidentification rates of the two consonants for one another. In Experiment 4.2, however, listeners were more likely to respond /p/ when the token comes from the High ('likely p') grouping. As mentioned earlier, it may be possible that the spectral differences captured by the RF model (Experiment 3.1) do not directly drive differences in perception (consistent with the results for /p/-/t/ in §3.4.3.3) but are correlated with other types of perceptually relevant acoustic variability. Differences in energy at the frequency components identified by the RF may co-occur with other perceptually relevant acoustic features. It may also be the case that the type of manipulation performed in §3.4.1.3 was not extensive enough to cause changes in listener perception. Because only one filter was ever applied at a time, listeners still had access to other spectral information in the signal to identify

place of articulation.

For /k/-/p/ (in both vocalic contexts), the effect of grouping was inconsistent across consonants. Listeners were less likely to categorize [k] as /k/ when it came from the Low ('unlikely /k/') group but were not less likely to categorize [p] as /k/ when it came from the same group. As mentioned in §4.3.2.2, /p/ showed extensive category overlap across all vocalic contexts with /k/. In contrast, /k/ showed varying degrees of overlap with /p/ according to vocalic context. If /p/ productions generally overlap the space of possible /k/ productions, then it may be the case that 'unlikely /k/' tokens do not differ enough acoustically from a 'likely /k/' token to be perceptually relevant to the listener. In contrast, there might be a wider acoustic distance between 'likely /p/' [k] tokens and 'unlikely /p/' [k] tokens.

4.5 General Discussion

The goal of this chapter was to explore a probabilistic account for why confusions within consonant pairs show perceptual asymmetry. In Experiment 4.1 (§4.3), a comparison of likelihood probabilities across consonant categories in an asymmetric pair revealed that the modeled relative likelihood of the less favored (i.e., more confusable) consonant being generated by the category of the favored consonant was predicted by the vocalic context. That is, in the context that conditioned asymmetry, the less favored consonant tended to have the highest likelihood of being generated by the other consonant (relative to its own category; see Table 4.18). For the results of Experiment 4.1 to provide insight on the asymmetrical confusion rates of human listeners, these differences in likelihood probability would need to be shown to be perceptually relevant. Experiment 4.2 (§4.4) set out to investigate whether such relevance exists. Tokens of consonants within each of the four targeted pairs (from Experiment 4.1) were ranked according to their relative probability of having been generated by one consonant category or the other, and subsets of the highest and lowest ranked consonants along this measure were selected. Listeners then performed a binary forced-choice categorization of these consonants. Their responses offer some reason to consider these difference measures as perceptually relevant. For /k/-/t/, /p/-/t/, and /θ/-/f/ (though not /k/-/p/ in either vocalic context), being drawn from the High ('likely A') category makes the listener more likely to categorize the consonant as /A/. A [θ] and [f] drawn from the High ('likely /θ/') category are both more likely to be categorized as /θ/ than tokens drawn from the Low ('unlikely /θ/') category. Although the grouping choice appears to show perceptual relevance to the listener, it is not necessarily the case that the exact features identified by the RFs in §3.3 are the same that directly drive differences in perception for listeners. For these consonant pairs, the features captured in the phonetic dimensions explored here may co-vary with the other types of information listeners use in perception.

In her 2001 dissertation, Plauché had some success relating distinctive acoustic features identified in her study to observed patterns of perceptual asymmetry. However, for some pairs, she was unable to identify an acoustic feature that behaved in the expected way. For example, the pair [ki] and [ti] show significant symmetric overlap with one another in VOT, as demonstrated in her research, but she noted that this symmetry failed to suitably explain why /t/ is favored in that confusion. This chapter takes a different approach by looking at overlap and likelihood probabilities in a multidimensional space – for this pair [k] shows variable degrees of overlap with /t/ according to the vocalic context of [k] along phonetic dimensions defined based on their spectra, while [t] shows a consistent degree of overlap with /k/ in all contexts tested. This finding offers an alternative perspective as to why /t/ is favored in the confusion, which seems consistent across the consonant pairs tested in Experiment 4.1.

As reported in the literature so far, perceptual asymmetry is a laboratory phenomenon. Listeners show unequal rates of confusion when listening to isolated syllables or (in the case of voiceless stops) when listening to the consonant alone. The results reported for Experiments 4.1 and 4.2 speak most closely to a situation where listeners hear isolated consonants, in contrast to Plauché (2001), which moves in the direction of a context-dependent analysis. Both offer accounts that speak to the phenomenon in the contexts in which it has been observed. To the extent that this confusion does occur in more naturalistic communication settings, both analyses would need to be expanded much further to accommodate additional layers of information (e.g., lexical status, prosodic environment). At the very least, a more comprehensive account of perceptual asymmetry may blend the two approaches used here and in Plauché (2001), by incorporating a multidimensional analysis of likelihood probability along spectral and non-spectral cues and context-specific and context-general information about a category.

The results of Experiment 4.2 suggest that likelihood probabilities can concretely predict listener rates of confusion. Significant work on the modelling of phonetic processes under a Bayesian framework has been undertaken (e.g., Clayards et al., 2008; Kirby, 2010). These models may also be used to identify tokens whose phonetic properties could affect the outcome of listener perception in a predicted way, as seems to have happened in Experiment 4.2.

CHAPTER 5

Modeling the Role of Perceptual Asymmetry in Sound Change

This chapter situates perceptual asymmetry within the context of diachronic change across a speaker community. Computational modeling is used to explore the potential effect of perceptual asymmetry on the transmission of phonetic categories across speaker generations. In §5.1, a brief overview is given of the literature on sound change and approaches taken to simulate language communities over time. Following this introduction, the design, hypotheses, and results of the modeling experiment are reported.

5.1 Background

5.1.1 Perceptual Asymmetry & Sound Change

The literature identifies several sound changes that resemble the perceptual asymmetries whose articulatory, acoustic, and perceptual characteristics have been analyzed in the earlier chapters. For changes of this sort, depicted in Table 5.1, the consonant undergoing the change and the output of the change tend to correspond to the consonant pair participating in laboratory-observed perceptual asymmetries. Specifically, the consonant undergoing the change corresponds to the less often identified member of the pair in the laboratory. In addition, these sound changes tend to be unidirectional. Changes that work in the reverse direction are either rare or unattested altogether.[1]For example, the diachronic change of /θ/ to /f/, observed in some English varieties (Stuart-Smith et al., 2007; Schleef & Ramsammy, 2013) and perhaps in Rotuman (Blevins, 2004), tends to mirror the synchronic identification asymmetry between [θ] and [f]. Furthermore, only the change from

interdental to labiodental place of articulation has been observed. The reverse is unattested.

The resemblance of these diachronic processes to established perceptual asymmetries in terms of place of articulation (in Table 5.1, compare consonants in columns 1 and 3 to those in column 6) and conditioning context (compare column 3 to column 7) invites the possibility that identification asymmetry may facilitate these sound changes. Theories of sound change that root the listener at the center of the process may offer insight into how perceptual asymmetry might help to condition change. Each of these accounts tends to be applied within an approach that assumes that the information required for perception is fully specified by the acoustic signal – an assumption broadly compatible with the framework of direct perception. Under these accounts, the information required for perception is fully specified by the acoustic signal. Consequently, the outcome of perception is, as described by Fowler 'only as successful as the specifying information provided by the acoustic signal' (Fowler, 1996). If a production lacks robust, perceptible cues to help the listener discriminate between possible phonetic candidates, listeners would not be able to reliably avoid misperception.

[1] Change of /k/ to /p/ in a labializing context may be an exception here. In pre-Seneca, for example, /p/ also changed to /kw/ (Chafe, 1964), which works in the opposite direction of the /k/ to /p/ sound change.

Original cons.		Reflex	Sound Change Environment	Varieties undergoing this change	Associated Consonant Pair	Context Conditioning Asymmetry
/k,g,x/	>	/tʃ,dʒ,ʃ,ʒ/	/j/ front vowels, when palatalized	e.g., Slavic (Gardiner, 2008)	/k/-/t/	/i/[2]
/p,b,m/	>	/t,tʃ,d,dʒ,n,ɲ/	/j/, when palatalized	e.g., Czech Dialects (Andersen, 1973); Northern Tai Dialects (F. K. Li, 1977)	/p/-/t/	/i/
/p,b/	>	/c,ɟ/	/j/	Albanian Dialects[3] (Desnitskaya, 1968)	/k/-/p/	/i/
/k,g/	>	/p,b/	/w,v/, when labialized	e.g., Ancient Greek, Celtic dialects (Ohala & Lorentz, 1977)	/k/-/p/	/u/
/θ,ð/	>	/f,v/	consonant-adjacent, coda position[4]	English dialects (Stuart-Smith et al., 2007); Rotuman (Blevins, 2004)	/θ/-/f/	*No context identified*[5]

Table 5.1: Sound changes (columns 1-4) that resemble perceptual asymmetry (columns 5-6)

.

Under the 'innocent misapprehension' family of accounts (e.g., Ohala, 1981; Yu, 2010), successful perception involves the listener figuring out which phonetic features of a production are inherent to the production and which arise due to phonetic or prosodic context. A failure to accurately determine whether a feature is intrinsic or extrinsic (e.g., context-induced) to that production may create some of the preconditions for a sound change. A listener who fails to assign contextual variation to its environment (hypocorrection) may instead recover a different phonetic

[2]An assumption made here in connecting perceptual asymmetry to sound change is that the vowel conditioning asymmetry has a coarticulatory effect on the consonant. A [k] produced before a high front vowel shows coarticulatory palatalization and therefore before a palatal approximant.

[3]Rankin (1981) also notes that some Rumanian dialects underwent a change whereby labial segments changed to place to velar before a palatal approximant. However, there is considerable variability across dialects in what the labial has become, ranging from a single velar stop to a labial-dorsal consonant sequence (i.e., [pkj], [pç], [ptʃ]). For dialects with a single velar stop, the original labial may have changed by means of an intermediate stage resembling the consonant cluster, as is suggested by Thomason (1986). If so, it is not clear whether the perceptual asymmetry of [k] and [p] alone can appropriately explain the emergence or disappearance of these consonant clusters.

[4]TH-Fronting in some English-speaking communities appears to be a gradient phenomenon. The two conditions listed here are environments where TH-Fronting occurs more frequently in English as spoken in Edinburgh (Schleef & Ramsammy, 2013)

[5]But see §3.5

category from what the speaker intended, as would a listener who misattributes a phonetic detail of a segment to its environment (hypercorrection). For example, coarticulatory factors can cause a production of [u] to have slightly lower pitch after a [b] than [p]. (House & Fairbanks, 1953) If a listener hearing [bu] takes phonetic environment into account, they may attribute low pitch in the vowel to the consonant. However, if the listener fails to account for the phonetic environment where the vowel appears, then might identify low pitch on [u] as intrinsic to the production. In such a case of hypocorrection, coarticulatory variation in pitch is instead treated as inherent to the vowel.

An innocent misapprehension approach offers a clear prediction for how some perceptual asymmetries could condition sound change. For all consonant pairs investigated here, except for /θ/ and /f/ (though see §3.5), perceptual asymmetry is only observed in a certain context. Hypocorrection could cause a listener to misidentify, for example, a [k] before a high front vowel as an intended production of a /t/.

Confusions of /θ/ and /f/ are not described in the literature as being conditioned by vocalic context, and the results of Experiment 4.1 more concretely suggest that asymmetry may not be conditioned by a vocalic context when the listener hears the isolated fricative (though see Experiment 3.2 for evidence of a conditioning context in a CV context). A general tendency for listeners to recover /f/ from [θ] productions could lead to sound change just as in the previous paragraph, and there are a few possibilities for why confusions could be context independent under this approach. First, it may be the case that coarticulation of /θ/ with vowels generally introduces phonetic features consistent with an [f] production that listeners have difficulty correctly attributing to their environment. It could also be the case that there are few robust features inherent to a /θ/ production that distinguish it from an /f/. In absence of information robustly specifying consonant place, confusions are more likely.

Phonetic perception has also been understood as a process of categorization informed by the integration of several phonetic cues (e.g., Toscano & McMurray, 2010). This framework is consistent with the finding that speech sounds tend to differ along many phonetic features (Wright, 2004, among many others). Unlike the innocent misapprehension approach, cue-based approaches to sound change (as taken e.g., in Beddor, 2009; Kirby, 2013; Coetzee et al., 2018; Kuang & Cui, 2018) do not assume that the listener misperceives the production of a speaker. Instead, individual differences in the features attended to in perception and produced in one's speech lead to variability in how listeners categorize speech sounds. Cue-based approaches to sound change can accommodate differing assumptions about whether a shift in primary cue is understood as occurring in perception first, in production first, or if both move in lockstep. Recalling the example of [bu], two cues are available to the listener to indicate that the initial consonant is voiced: low voice onset time (VOT) and low F0 at the start of the vowel. A listener could make use of either cue or

both in perception, and another listener could choose a different strategy. In a perception-driven model, a community of speakers might shift from using both features to using F0 information alone to identify the voicing status of the consonant, and this change becomes reflected in the production of speakers in the community. In contrast, a production-driven model might suggest that the reduction or loss of VOT (due to independent circumstances) as a useful feature could lead community members to further enhance F0 differences to maintain a voicing contrast.

Cue-based approaches also offer a straightforward prediction of how perceptual asymmetry could condition sound change, at least when the sound change is not a merger, a loss of contrast between two categories. Productions of [k] before a high front vowel are perceptually similar to [t] and [tʃ] productions, which suggests overlap in the features of pre-[i] [k], [t] and [tʃ]. If a community of listeners were to heavily weight these shared cues, these listeners-turned-speakers would systematically realize /ki/ as [tʃi] (or [ti]) and may (re-)interpret /ki/ as /tʃi/ (or /ti/). In the case where the sound change is a merger, as appears to be the case for TH-Fronting in Glaswegian English, for example, the story might become more complicated. Cue-weighting accounts appear to effectively describe sound changes which involve a change in the phonetic features used to maintain a contrast. If there were insufficient cues to maintain the contrast, then just as in the innocent misapprehension approach, the cues present would not uniquely specify either segment in a reliable manner.

Hypo- and Hyper-articulation Theory (Lindblom, 1990; Lindblom et al., 1995) provides another account for how perceptual asymmetry might condition sound change that also considers speaker-specific factors. Also under this model, phonetic variability arises in part from the speaker's need to balance two competing constraints - intelligibility for the listener and their own articulatory effort. The phonetic targets achieved at any given moment are theorized to optimally satisfy these constraints. Under this model, an utterance that is more strongly coarticulated likely reflects a phonetic target more strongly constrained by speaker-centered needs. Under some interactional settings, listeners may be afforded a greater ability to attend to the fine phonetic detail of such a production. Under these circumstances, this phonetic variant can then be reproduced in their own speech. A tendency for this phonetic variant to be perceived and then reproduced can then lead to greater change across the community.

Guion (1998)'s study of /k/ and /t/ perception found that [ki] syllables tend to lead to more confusions in rapid speech than in slow speech. This arguably more strongly coarticulated speech may reflect a hypo-articulated variant due to production constraints during rapid speech. If so, then listeners have preferential access to the phonetic detail of these [t]-sounding /k/s, and may then produce [t]-sounding /k/s in their own speech. For consonant pairs showing perceptual asymmetry, if the disfavored category (e.g., /θ/) generally sounds more like the favored one (/f/) in hypo-articulated speech, then a similar mechanism to that described for /k/and /t/ could potentially apply

to any of the pairs.

The following section describes strategies used to simulate the evolution of linguistic categories over time. This discussion remains largely agnostic to the theories described above. However, a discussion of sound change theory in light of the simulation results of this chapter can be found in §5.4.

5.1.2 Computational modeling of sound change

Weinreich et al. (1968) outlined several of the fundamental questions that surround the study of sound change. The question addressed by this chapter is largely one of constraints – why phonological systems tend to change in a restricted set of ways. Historical materials offer suggestive evidence for which types of sound change are possible or common, and experimental behavioral methodologies provide insight into the structure of human speech production and perception. Computational modeling is an additional means to investigate how listener-speaker interactions shape the long-term evolution of phonetic categories. Under these frameworks, a subset of the factors inherent to communication, processing, and storage are explicitly modeled. The interaction of factors and phonetic structure can condition change in the tokens over time. Because of the small number of factors explored in each, computational models offer a well-controlled environment in which to build predictions of how certain synchronic factors might interact with one another.

It can also be useful to explicitly manipulate the parameters of computational models to get a better idea of how different factors interact with each other. Such analyses can offer some insight into the degree to which some results are sensitive to certain factors. Kirby and Sonderegger (2015), for example, undertook some exploration of sound change outcomes as a function of parameter values (e.g., categoricity bias strength) and differing perceiver models (e.g., Maximum Likelihood vs. Bayesian models).

Computational models of sound change have differed over a variety of design choices. Several common areas of difference are described below. These choices described in the literature help to inform the model design for this chapter.

5.1.2.1 Malleability of adult phonetic categories

Sound change models can differ in the degree to which adult speakers' phonetic categories are fixed, a choice with a foot in theories of language acquisition and language change. Some models (e.g., Fulop & Scott, 2019) adopt an intergenerational transmission approach – adults have stable phonetic categories, while those of children are malleable. By receiving linguistic input from adults, children develop a mature phonological system, at which point their categories become fixed and serve as input to a successor generation. Such a model implies that linguistic structure

not present in the parent's speech can still emerge in the child's, an assumption that receives support from other bodies of literature where children have been observed to acquire linguistic structure absent from their input (e.g., Goldin-Meadow, 2005 for the development of linguistic structure among home-signers).

For other models, the phonetic categories of adults and children are equally susceptible to change. These models assume that the phonetic category of an individual can change in structure over one's life through interaction with other individuals. This assumption receives support from longitudinal studies of individual speakers over the course of their adult lives, where individuals' phonetic targets are observed to drift over time (e.g., Sancier & Fowler, 1997; Harrington et al., 2000).

5.1.2.2 Phonetic category structre

Sound change models also differ in how phonetic categories are structured. Several implementations (e.g., Pierrehumbert, 2001) have adopted an exemplar-theoretic approach, where each category is composed of a set of tokens. In these models, production, perception, and category update depend on the manipulation of phonetic tokens. A speaker might produce a phonetic token by randomly drawing from a set of tokens of a certain category (as in e.g., Pierrehumbert, 2001, Todd et al., 2019) and applying noise and/or bias to the selected token. A listener would then categorize the token by comparing it against stored tokens from each potential category (a similar process is described in Section §4.1.2 of Chapter 4). In contrast, other implementations (e.g., Kirby, 2014) have taken a distributional approach, where each category is modeled as a probabilistic distribution. Although communication still involves the transmission of discrete tokens, under these models, processes like production and category update involve, for example, sampling and parameter estimation, respectively, which preserve the distributional structure of the category. A speaker in a model might produce a token by sampling from a category distribution. A perceiver would then categorize this token by comparing the respective likelihoods that each category would have generated a token of that sort.

These approaches differ representationally and in their underlying theoretical frameworks but may in fact produce similar results under certain conditions. Modeling the evolution of phonetic categories under a variety of phonetic biases, Sóskuthy (2013) found that the respective outcomes of exemplar-theoretic and distributional models showed little difference after many iterations.

5.1.2.3 Perception and production

The problem of phonetic perception is generally framed as finding the most appropriate category for a token given its acoustic/phonetic structure. For exemplar-theoretic approaches, this can in-

volve explicit comparison of the token to other phonetically similar exemplars. This family of models varies further over which measures serve as input to the comparison – ranging from phonetic similarity (Pierrehumbert, 2001) to category typicality and discriminability. (Todd et al., 2019) For distributional approaches, phonetic categories are not structured appropriately for comparisons between tokens. Instead, categories are compared according to the likelihood that the category would generate such a token. These calculations can take the form of a Maximum Likelihood estimation (Kirby & Sonderegger, 2015) or a Bayes Optimal Classification (Kirby, 2014), with the latter model allowing the prior probability of phonetic categories to be specified.

The process by which a phonetic token is produced also differs between models. In distributional approaches, tokens are generated by random sampling from category distributions. In contrast, under exemplar-theoretic approaches, phonetic tokens are generated by first selecting a specific exemplar and then applying any of several adjustments, including noise or a bias factor.

5.1.2.4 Perception and production

Because of differing structures, exemplar-theoretic and distributional approaches to sound change approach the storage of phonetic tokens differently. In exemplar-theoretic models, the token is explicitly added to the category when appropriate. In the approach taken by Pierrehumbert (2001), individual units are weighted in such a way that tokens encountered earlier affect the structure of the category less significantly. This can be implemented by assigning a weight to tokens which exponentially decays with time or by forgetting tokens altogether (e.g., Harrington & Schiel, 2017; Harrington et al., 2018). In both cases, a token is increasingly likely to have low weight with increasing time. This decay weighting may also depend on factors like category typicality.

For distributional approaches to modeling, the parameters that define the category distribution must be re-estimated for the category to change. When category distributions are assumed to be a Gaussian mixture, parameter estimation can take the form of Expectation-Maximation (EM), an iterative algorithm to find the parameters with the maximal likelihood given the data. The parameters of each category distribution could also be estimated separately.

Many models of sound change have offered a similar treatment toward misclassifications (e.g., Harrington & Schiel, 2017; Harrington et al., 2018; Todd et al., 2019) – instances where the category produced by the speaker differs from the category identified by the listener. In these cases, the token is typically not stored at all. The current study treats the outcome of misclassified tokens as a set of parameters that vary across simulations.

5.2 Research question

This study is motivated by the following question:

RESEARCH QUESTION

What effect can perceptual asymmetry have on the long-term stability of phonetic categories?

While sound changes described in Table 5.1 resemble certain cases of perceptual asymmetry, the exact mechanism by which it can condition change is not clear. The goal of this study is to better understand what effect perceptual asymmetry might have on the acquisition and evolution of consonants over several generations.

5.3 Experiment 5.1

5.3.1 Model architecture and methodology

The simulations employed in this study follow an agent-based model (ABM) design in which members of a speaker community are explicitly modeled as entities with the ability to generate phonetic tokens, categorize tokens, and acquire a phonological system. Each agent is defined by a set of stored tokens and a phonological system.

Recalling §5.1, the simulations undertaken this chapter seem most compatible with a vastly simplified understanding of the cue-based account of sound change. Listeners may or may not identify a token as the same category as what the speaker intended, but these differences between listeners and speakers are not random – a token with the same acoustics will always be categorized in the same way by the listener, reflecting the consistent (but listener-specific) categorization strategies of a cue-based perceptual approach. This model also assumes strict parity between the information used to categorize consonants in perception and the features speakers can use in production. Category movement is also possible in phonetic space to maintain or enhance a contrast between consonants.

5.3.1.1 Malleability of adult phonetic categories

Agents are modeled as having fixed phonetic categories after having reached maturity, one of two standard options for a model (as described in §5.1.2.1). While an adult may experience changes to their phonetic category over their lifetime that are in fact consistent with the sound changes investigated here, that mechanism of change is not explored in these models.

5.3.1.2 Phonetic token structure

In Experiments 3.1 and 3.2, several spectral features were identified that were relevant to place contrasts within consonant pairs. These simulations use the same features used in Experiments 4.1 and 4.2 to model the change of phonetic categories over time. Each phonetic token is then defined as a two-dimensional vector. Additional information about how these features were obtained, as well as their perceptual significance, can be found in §4.3.1.

A phonetic token also has two category labels associated with it – the category from which the speaker generated the token and the category assigned by the listener. Whether a listener stores a token depends in part on whether a token is classified correctly.

5.3.1.3 Phonetic category structure

An agent's phonological system is modeled as a multivariate Gaussian mixture model, given by the following formula below. This model choice captures several intuitions about how a phonetic system might be organized.

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

Equation 5.1: Multivariate Gaussian mixture model

Phonetic categories differ in overall frequency, and phonotactic probability may have a substantive effect on linguistic processing, as is suggested in Vitevitch and Luce (1999). Correspondingly, different distributions in this equation have a different prior likelihood of occurrence. This trait is captured by the mixing parameter π_k, which ranges from 0 to 1 inclusive. A speaker agent will be less likely to produce a token from a distributional category with a lower mixing parameter.

There are two features that define each token in a phonetic space. The relative likelihoods that each category produces a token with certain acoustic properties is represented by a multivariate normal distribution. Such a distribution is defined by a mean vector (μ_k,), where the distribution is centered, and a covariance matrix (Σ_k), the shape and dispersion of the category.

Each phonetic category C_k is then defined by a mean vector, covariance matrix, and mixing parameter. Equation 5.1, which describes the marginal probability of a mystery token having acoustics of a certain sort, is the sum of the likelihoods that each category would produce a token with those acoustics, multiplied by the mixing parameter for that category.

5.3.1.4 Production and perception

A speaking agent generates a token by random sampling from the multivariate Gaussian mixture. No additional adjustment is made to the token.

When presented with a token, the listening agent attempts identification of the category by applying Bayes' Rule (visualized in Equation 4.4). The category with the highest posterior probability given the acoustics of the token x is selected as the listener-assigned category of the token.

5.3.1.5 Storage and category generation

Once a token has been categorized by a learner agent, any of several outcomes are possible. If the label assigned by the speaker agrees with that assigned by the listener, the token is stored by the agent. If the two labels disagree, then the token may still be stored with some chance s. This variable reflects the possibility inherent to all three sound change accounts (see §5.1.1) that a listener may recover a novel category different from what the speaker intended. This parameter s is varied between the models.

If the token is misclassified but stored, it can be stored as the category intended by the speaker with probability p or stored as the category intended by the speaker with chance $(1 - p)$. The parameter p is also varied between models. A high p represents a communicative setting rich in contextual information beyond the phonetic features that the listener can use to identify the category intended by the speaker. For example, if a listener hears "I used the [ti] to open the door", even though the production sounded like a /t/, knowledge of the lexical items of English and which would be likely to appear in such a statement would likely lead the listener to expect that the speaker intended to produce /k/. In contrast, a low p represents a communicative setting where the listener cannot use additional contextual information to decide on the category intended by the speaker. For example, if a listener instead heard the singular syllable [ti], they would lack additional information to determine the intended category of the listener. This situation could also arise for a learner when there are few minimal pairs to distinguish a contrast. While /θ/ and /f/ are contrastive in English, there are fewer minimal pairs to contrast the two than there are for /k/ and /t/ (before high front vowels). If a listener hears a statement like "I took a [bæf] last night", the listener relies more heavily on the acoustics of the consonant to identify the category intended by the speaker, as there is no competing word *baff that would contrast in meaning with this production.

Once a learner agent has categorized a threshold value of tokens, this token set is used to generate an adult phonological system. Category parameters are estimated using an Expectation-Maximization algorithm.

5.3.1.6 Model algorithm

This following algorithm is run in all simulations in this chapter modeling the evolution of phonetic categories over time. The structure of this simulation is partially modeled off that of Kirby (2014), but with several adjustments.

INITIAL GENERATION

1. N adult agents are generated with initial phonological system parameters estimated from a naturalistic corpus of tokens.
2. N child agents are generated with initial phonological system parameters identical to those of the adult generation.
3. Each adult agent generates T tokens.
4. Each child agent is randomly paired with a single adult agent and categorizes the T tokens generated by the adult. This choice differs from Kirby (2014), where each learner agent samples randomly from a matrix of productions by all speaker agents, but is a not uncommon model of population structure. As described in Kirby (2013), this mode of transmission involves parallel diffusion chains. Each agent receives linguistic input from a unique adult agent and transmits language to a single child agent. The partial restriction of learner agents' input captures the intuition that listeners do not actually receive input equally from all adults in a community.
5. The T tokens are (potentially) stored by the listener according to the settings of the simulation. The conditions under which an agent stores a token is of interest to this experiment.
6. Each child agent re-estimates their parameters according to the tokens that were successfully stored during the learning phase. These agents become adults, and the previous generation of agents no longer serve as input to the simulation.

SUBSEQUENT GENERATIONS

7. N child agents are generated with initial phonological system parameters estimated from the entire pool of successfully stored tokens from last generation.
8. Repeat steps 3-7.

5.3.1.7 Dataset

The data for these simulations come from the Buckeye Corpus (Pitt et al. 2005) and is described in detail in §3.3.1.1.

5.3.1.8 Dependent Variables

In each generation, listener agents each develop a phonological system according to the tokens they categorized and stored. Two parameters from the phonological system of each are extracted at each iteration for analysis for each agent: the mixing parameter and mean feature vector and for each category.

5.3.2 Predictions

Each simulation varies in the acoustics of the phonetic categories and in several perceptual factors. These parameters and their predicted effect on the stability of phonetic categories are discussed below.

5.3.2.1 Storage Parameter (s)

In the ABM simulations, s describes how likely it is for listener agents to store a token when the category identified by the listener differs from what the speaker had intended, as described in . This parameter can range from 0 (tokens are only stored when the listener and speaker agree) to 1 (all tokens are always stored; see §5.3.2.2).

$s \approx 0$: If s is small, then tokens that more typical of the competing category will tend not to be stored. This tendency is predicted to cause the categories to drift apart. Over generations categories will become increasingly populated just with tokens that are typical of their respective category. In cases where the two categories strongly overlap, one category may be consistently miscategorized by the listener, in which case this category will become infrequent for the listener, leading to a lower mixing parameter. Under this assumption, the disfavored category in a perceptual asymmetry pair would tend to drift away from the favored category but would also become infrequent. While the divergence in category mean (on its own) is not consistent with the associated sound changes, a change in frequency would be consistent with a partial merger of the two categories.

5.3.2.2 Storage Parameter (p)

In this model, a token may still be stored if the listener and speaker do not agree on the category assigned to it. If so, the token is either stored as the category intended by the speaker (with p probability) or as the category identified by the listener (with $(1 - p)$ probability).

$p \approx 1$ (and $s \approx 1$): When p is close to 1, the listener can be understood as having an effective ability to recover the speaker's intended category, independent of the acoustics of the production. This circumstance is comparable to instances where a listener can make use of non-phonetic linguistic information (e.g., word/non-word status) to distinguish between the possible segmental

choices. The distributions of phonetic categories acquired by the listener should closely resemble those generated by the speaker, so there would be little change in category parameters between generations.

$p \approx 0$ (and $s \approx 1$): When p is small, the listener is unable to draw on additional information to figure out which category the speaker intended. As noted in §5.3.2.1, such a situation might occur if, for example, two consonants tend to have few or infrequently occurring minimal pairs, in which case incorrect category assignment will oftentimes not lead to misperception at other levels of linguistic structure (e.g., identifying the wrong word). The dental fricative pair, for example, has few minimal pairs; if [θ] is misidentified as /f/ in the word 'bath' it is unlikely that the listener will recover any other word.

Tokens in one category that have low conditional probability relative to the other category are stored in the second category. As a result, the only stored tokens in each category should be those with a higher probability of occurring in their respective categories. Different regions of overlap will prefer one category over the other. Over generations, this may have an effect of moving the categories apart into distinct regions of the acoustic space.

Under this assumption, most productions from the disfavored category in a consonant pair showing perceptual asymmetry might tend to be categorized as the favored category, while the infrequent remaining minority would show divergence from the favored category. In that most of the tokens from the disfavored category are categorized as the favored category, this case would seem to resemble a partial merger of the two categories. However, a change in category mean without an associated change in frequency would not be consistent with the sound changes associated with perceptual asymmetry.

5.3.2.3 Mixing parameter (π)

The initial mixing parameter associated with each category is predicted to affect the degree of movement of each category as well as how the two categories change in frequency over subsequent generations. If the two categories are overlapping, a category with a smaller mixing parameter is predicted to show more movement in phonetic space than the more frequent category.

In Experiments 4.1 and 4.2, the prior probabilities for each model were set at 0.5. Equal prior probabilities for each category means that each is treated as equally likely before the listener receives any acoustic information about the token. In this chapter, a value must also be chosen for the initial mixing parameter for each consonant category. This value will then act as the initial prior probability for the category in perception and production.

There appears not to be a set standard in phonetic models of perception with a prior-like measure about what the initial parameter should be set as. When this choice is mentioned explicitly, researchers have often chosen the same initial prior (Sóskuthy, 2015; Kronrod et al., 2016) for

each category. This assumption would imply that listeners are equally likely to categorize a token as one category or the other without any acoustic information about the token beforehand. It further implies that the phonetic structure of the two categories is the sole factor driving perceptual outcomes. Under such a model, confusions between /k/ and /t/ in the environment of high front vowels would be explicitly driven by the shapes of the distributions for /k/ and /t/. The results of Experiment 4.1 suggest that likelihood ratio measures consistent with the expected asymmetry appear even when the two categories are assumed to have equal prior.

Listeners also make use of non-phonetic factors during perception; they show sensitivity to phonotactic legality (Massaro & Cohen, 1983; Pitt, 1998) as well as phonological frequency (Pitt & McQueen, 1998), among a variety of other factors. Differences in the relative frequency of occurrence of a token (relative to the environment in which it appears) may affect the listener's prior expectation of which token they will hear. In such a case, it might be possible to choose initial mixing parameters for categories according to their empirically measured relative frequencies in the language. For the results reported in §5.3.3, the probability of the consonant's occurrence by context is selected as the initial mixing parameter for each category in the simulation.

Because the mixing parameter at each generation also informs the listener's prior, if one category has a much higher parameter than others, then the listener agent will show a bias toward identifying tokens as this category. Over successive generations, this can have a compounding effect on the relative frequencies of the categories.

5.3.2.4 Vocalic context

The consonant pairs analyzed in this study are predicted to show instability in the vocalic contexts associated with perceptual asymmetry. The consonants /k/ and /t/, for example, are predicted to show the most movement before high front vowels, and less in other vocalic contexts. Although confusions are comparatively rare outside of the vocalic context conditioning asymmetry, the two categories do still show overlap outside of these phonetic contexts (see the Pillai scores in Experiment 4.1), and some degree of instability is still possible.

5.3.2.5 Contextual considerations during perception

As noted in §5.3.1.4, token categorization depends in part on the likelihood that each category would generate a phonetic token with its acoustics. Simulations can differ in whether perception takes vocalic context into account. For one class, perception involves the comparison of context-general category distributions; such a model reflects a communicative situation where listeners lack access to additional information about phonetic context beyond what is present in the speech sound itself. Within the scope of perceptual asymmetry, this analysis (the approach taken in Experiment

4.1 and 4.2) would provide insight on the emergence of asymmetry when listening to isolated consonants, as has been observed in Winitz et al. (1972). This simulation style is also consistent with approaches to perception that assume the presence of invariant phonetic cues, like Stevens and Blumstein (1978). For another class of models, perception may involve the comparison of context-specific category distributions, which would reflect a situation where listeners do have access to information about the environment in which a token appears. This method of analysis would provide insight on the finding that perceptual asymmetry appears when listeners are presented with full speech syllables (Plauché et al., 1997; Plauché, 2001; Chang et al., 2001). Such an analysis is also consistent with an assumption that listeners make use of phonetic context during perception (e.g., Mann, 1980; Mann & Repp, 1980).

5.3.3 Results

The simulated results for Experiment 5.1 are organized by consonant pair and model settings. The pairs /k/-/t/ and /p/-/t/ showed similar results under identical parameter settings. For the sake of clarity and by way of illustration, the results of the ABM simulations involving /k/-/t/ are described in close detail, and the simulated results for /p/ and /t/ are described in comparison to the /k/-/t/ models.

The results of the other two consonant pairs differed from predictions and, as a result, are also described in detail in this section.

Per §5.3.2.5, models were run where perception was either vowel-dependent or vowel-independent. For the sake of clarity in describing these results, the vowel-dependent simulations are reported in the following sections. Some of the context-independent simulations suffered from a tendency for categories with low initial mixing parameter to disappear (i.e., take on a mixing parameter value of 0; also discussed in §5.4). For simulations where disappearing categories were not an issue, the results looked analogous to the context-dependent simulations.

For all plots in this section, a 95% confidence interval is plotted in gray over the smooth curves. In many cases, these intervals make it visually apparent whether change has been observed over the course of the simulation. In cases where change is not clear from the plots, statistical tests are run to clarify the outcome of the model.

5.3.3.1 ABM simulations of /k/-/t/

In this section, the results of /k/-/t/ simulations are reported. In each model, the initial mixing parameter, covariance matrix, and mean vector are identical. Two vocalic contexts are compared, which differ in the degree to which stability is predicted across time. Because velar palatalization is conditioned by neighboring palatal segments, the segment pairs /k/ and /t/ before high front vowels

(i.e., [i] and [ɪ]) are predicted to show instability across generations under some circumstances. In contrast, /k/ and /t/ before low back vowels are predicted to remain comparatively stable over time. Each of these simulations was run over 40 generations, where each generation was composed of 20 agents, and each learner agent categorized 100 tokens total per generation. The number of tokens and agents are smaller than comparable simulations performed in Kirby (2014) by a factor of 5, but increasing the number of agents and tokens beyond the values used here does not appear to affect the directionality of results.

5.3.3.1.1 s=0

For simulations where $s = 0$, learner agents only store a phonetic token if the category intended by the speaker agent matches that identified by the listener. The two categories are predicted to show some degree of instability – the region of overlap between the two categories is expected to decrease, corresponding to increased distance between the two categories in featural space.

/k/-/t/ BEFORE HIGH FRONT VOWELS

As visualized in Figure 5.1, over successive generations, the prior probability of /k/ (visualized in orange) falls from 48% to approximately 10%. In contrast, the prior probability of /t/ (visualized in blue) rises from 52% to approximately 90%. /k/ becomes much less frequent by the 40th generation. Such an occurrence may be consistent with partial merger, where two phonetic categories come to be represented by a single category in a specific context. In this case, it would appear the two are merging in such a way that /t/ is the predominant category preceding high front vowels.

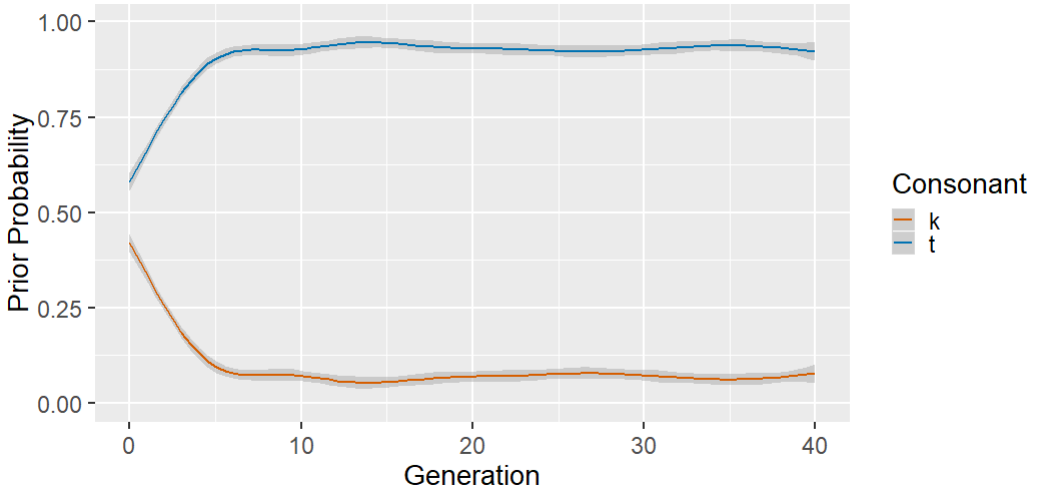

Figure 5.1: Simulated prior results of /k/-/t/ before high front vowels (s=0)

The featural values between the two categories are given in Figure 5.2. For this model, PC1 corresponds to overall energy centered around 1.4 to 3 kHz, while PC2 corresponds to a weighted

difference between the upper and lower halves of that frequency band. In contrast to the dramatic change in relative category frequency, these values showed comparative stability. A t-test revealed, however, that PC2 (right panel) achieved a significant difference between categories by the 40th generation $[t(19) = 5.40, p < 0.0001]$. This change might be viewed as dissimilation – while most of the /k/ distribution seems to have merged into /t/, the remaining members of the category have taken on a feature that more robustly contrasts with /t/. There also appears to be higher inter-agent variability in the category means for /k/ than /t/.

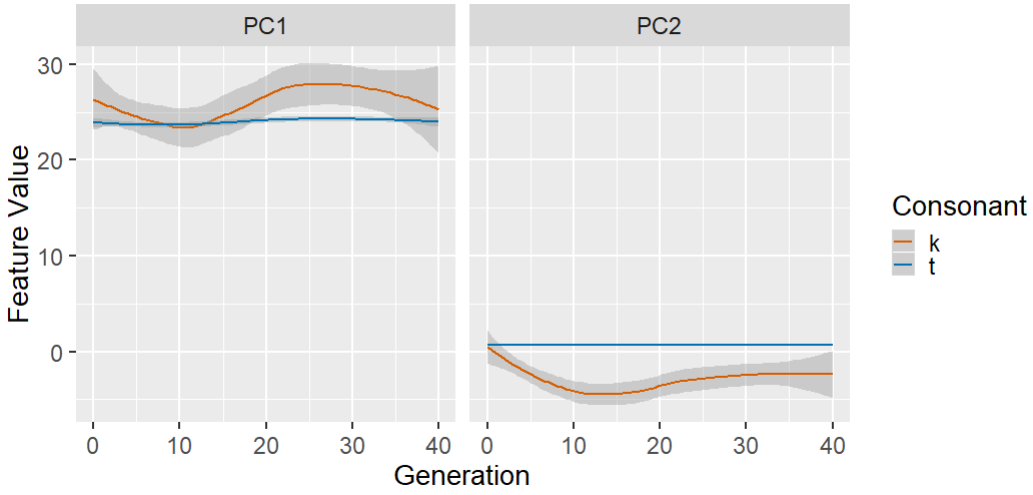

Figure 5.2: Simulated featural results of /k/-/t/ before high front vowels (s=0)

/k/-/t/ BEFORE LOW BACK VOWELS

As expected, the prior values for /k/ and /t/ before low back vowels are stable across generations. As visualized in Figure 5.3, neither category shows much movement over successive generations. A t-test comparing the agent-wise difference in category means at the 1st and 40th generations reveals neither divergence nor convergence in PC1 $[t(19) = 0.77, p = 0.45]$, but some divergence in PC2 $[t(19) = 5.96, p < 0.0001]$.

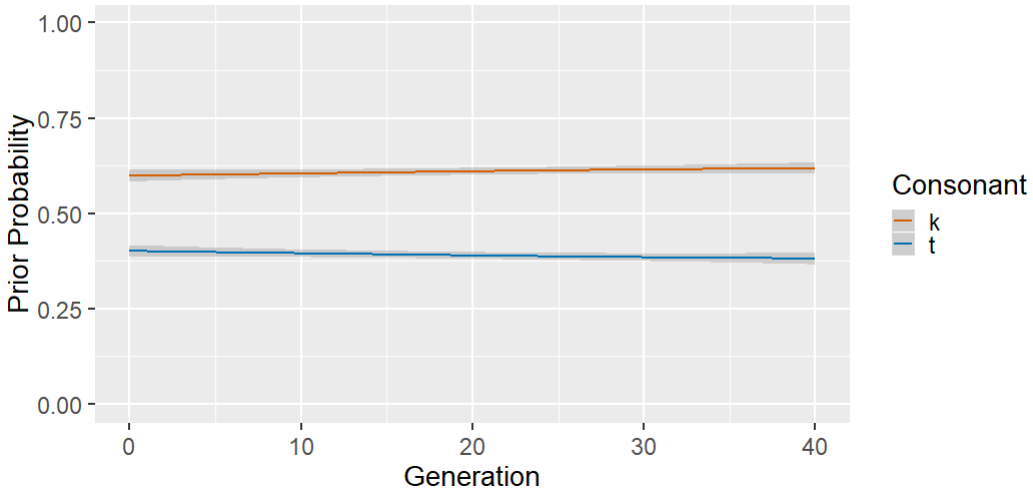

Figure 5.3: Simulated prior results of /k/-/t/ before low back vowels (s=0)

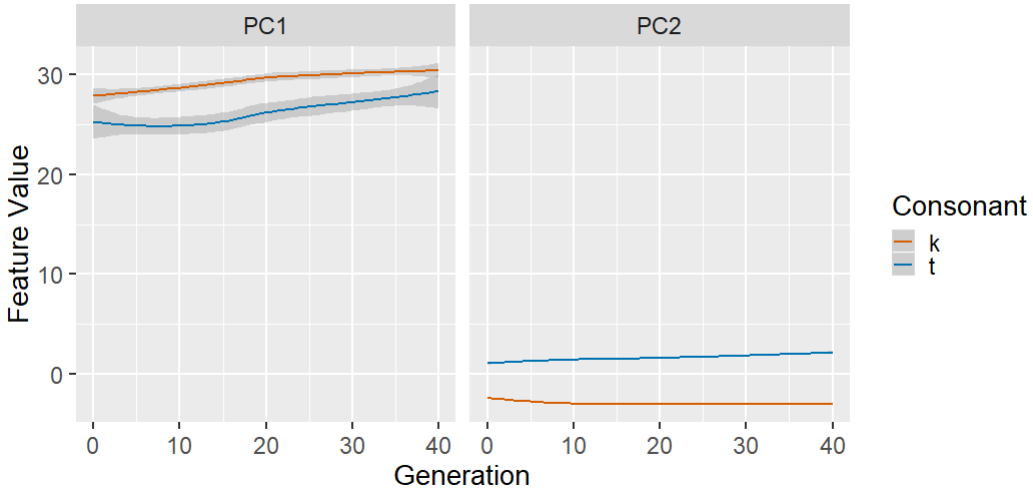

Figure 5.4: Simulated featural results of /k/-/t/ before low back vowels (s=0)

5.3.3.1.2 s=1, p=0

For simulations where $s = 1$ and $p = 0$, learner agents always store a token that they encounter. Regardless of what the speaker intended, each token is categorized as the label assigned by the listener. In such a circumstance, if the two categories overlap and one is more frequent, the less frequent category is predicted to decrease in probability over time. The less common overlapping category is also predicted to move in a direction that decreases overlap with the more frequent category.

The evolution of priors resembles the results observed when $s = 0$. As seen in Figure 5.5, the prior probability of /k/ drops to below 10%.

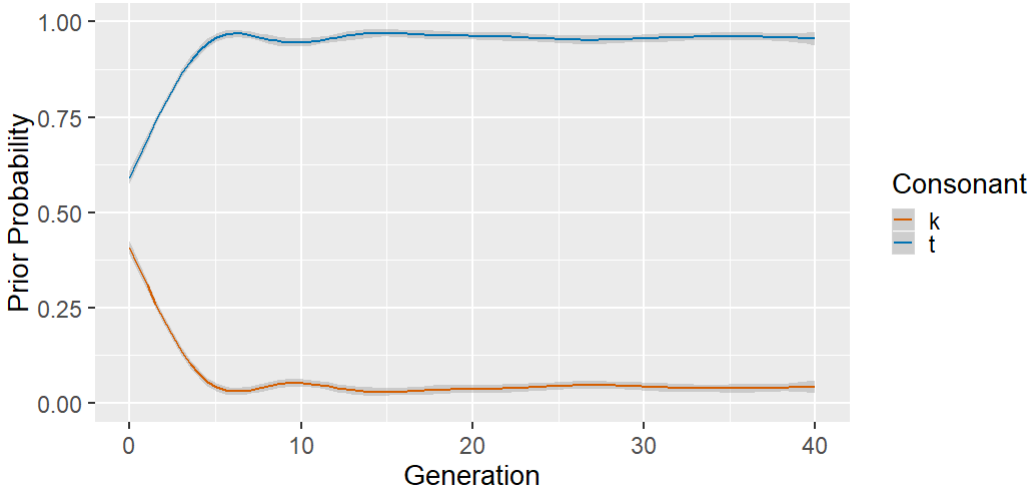

Figure 5.5: Simulated prior results of /k/-/t/ before high front vowels (s=1, p=0)

The evolution of featural parameters differs between when $s = 1$ and $s = 0$. When $s = 1$, /k/ and /t/ clearly diverge along both phonetic dimensions after 40 generations (as seen in Figure 5.6), and /k/ shows the greatest movement over the course of this simulation. As noted above, each token is stored as the category identified by the listener, which corresponds to the category most likely to generate token. When the two categories overlap and differ in mixing parameter, tokens with acoustic properties that fall in this region of overlap tend to be identified as the category with the higher mixing parameter. Tokens identified as the infrequent category will be less frequent (causing a decrease in prior probability) and will tend to occur outside of the region of overlap, leading to movement away from the more frequent category.

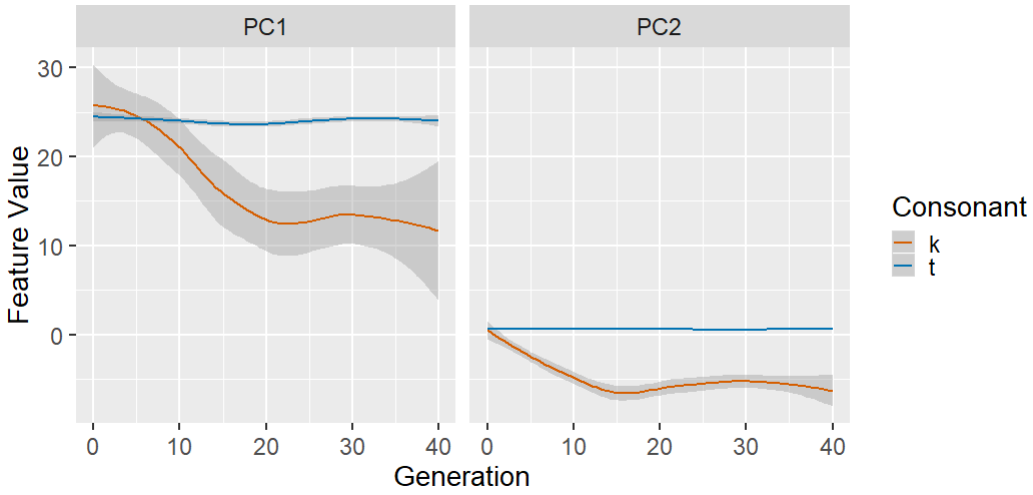

Figure 5.6: Simulated featural results of /k/-/t/ before high front vowels (s=1, p=0)

/k/-/t/ BEFORE LOW BACK VOWELS

The respective prior values for /k/ and /t/ before low back vowels show linear movement across generations. The two categories show a slow reversal in prior probability, but no other changes are apparent.

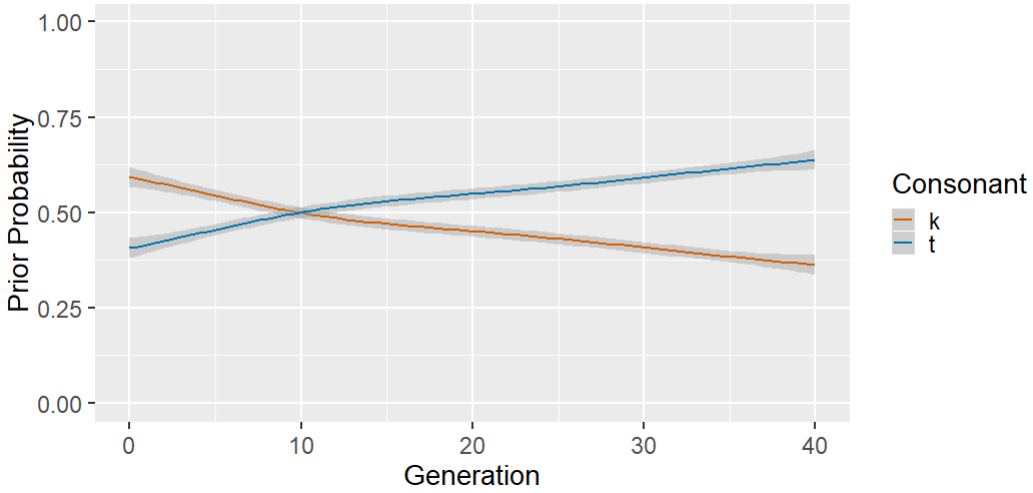

Figure 5.7: Simulated prior results of /k/-/t/ before low back vowels (s=1, p=0)

The featural values for /k/ and /t/ once again show stability across generations, as seen in Figure 5.8. Neither category shows consistent movement across either featural dimension.

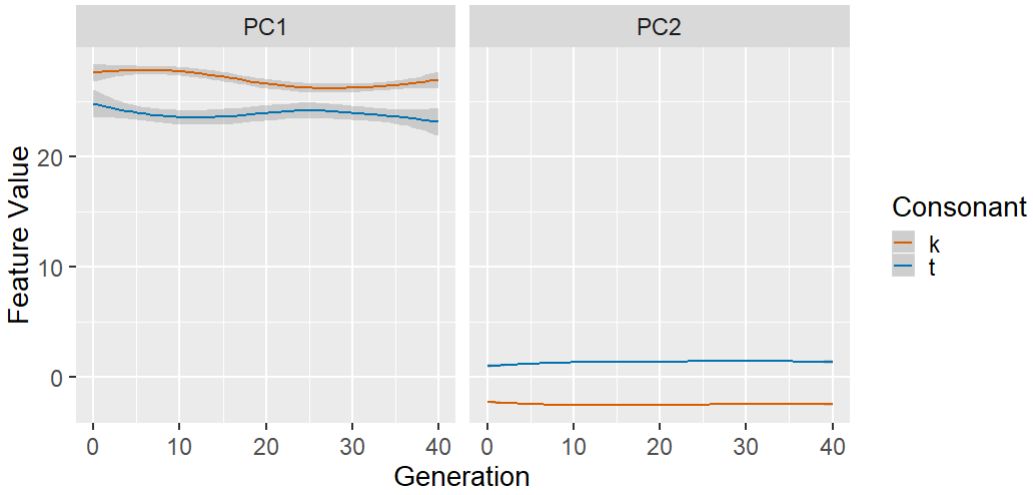

Figure 5.8: Simulated featural results of /k/-/t/ before low back vowels (s=1, p=0)

5.3.3.1.3 s=1, p=1

For simulations where $s = 1$ and $p = 1$, learner agents store every token they encounter as the category intended by the speaker. Little category movement is predicted to be observed.

/k/-/t/ BEFORE HIGH FRONT VOWELS

Unlike the previous models, the prior values between /k/ and /t/ show general stability, as in Figure 5.9. The prior probability of /k/, for example, does not appear to differ significantly between the initial parameters and the 40th generation $[t(19) = 0.42, p = 0.68]$.

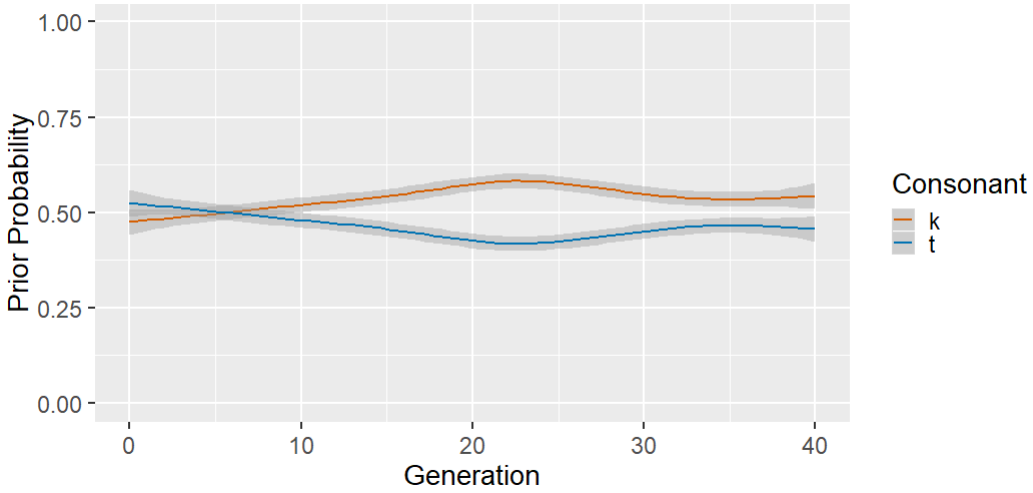

Figure 5.9: Simulated prior results of /k/-/t/ before high front vowels (s=1, p=1)

Likewise, the featural values for both categories, given in Figure 5.10, do not diverge or converge over time. Unlike prior simulated results for /k/ and /t/ before high front vowels, the featural

means for the two categories overlap closely.

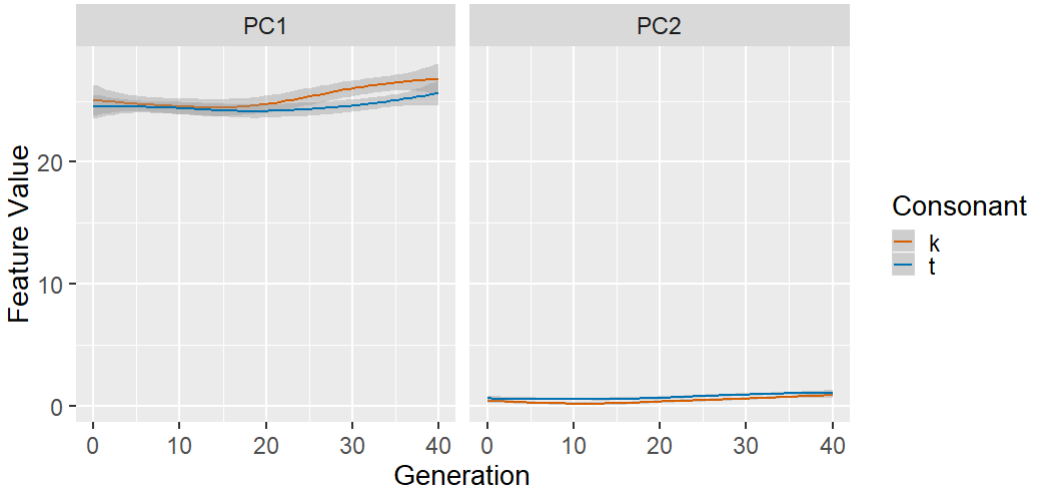

Figure 5.10: Simulated featural results of /k/-/t/ before high front vowels (s=1, p=1)

/k/-/t/ BEFORE LOW BACK VOWELS

The results of /k/ and /t/ before low back vowels resemble those before high front vowels. There is very little movement in the respective prior probabilities of the two categories (Figure 5.11).

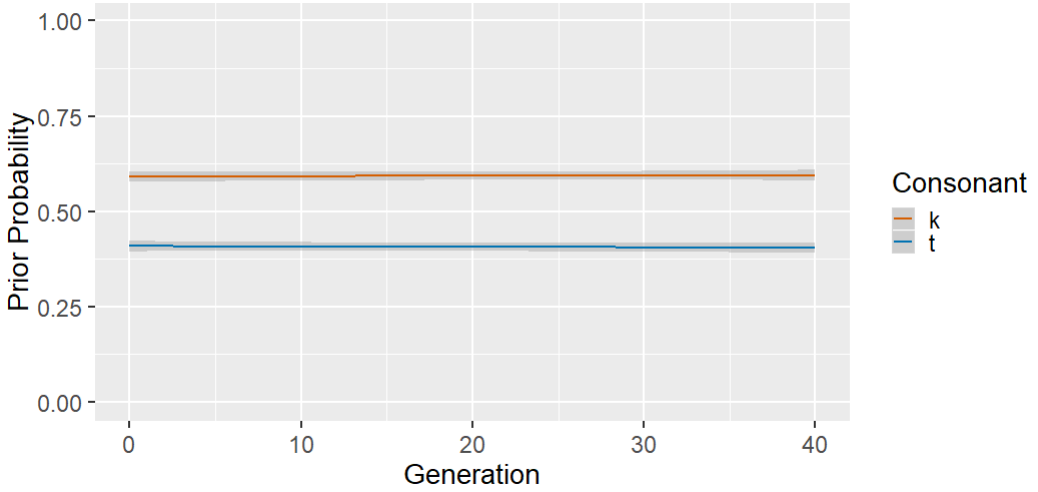

Figure 5.11: Simulated prior results of /k/-/t/ before low back vowels (s=1, p=1)

Likewise, there is little movement between the two categories in featural space (Figure 5.12).

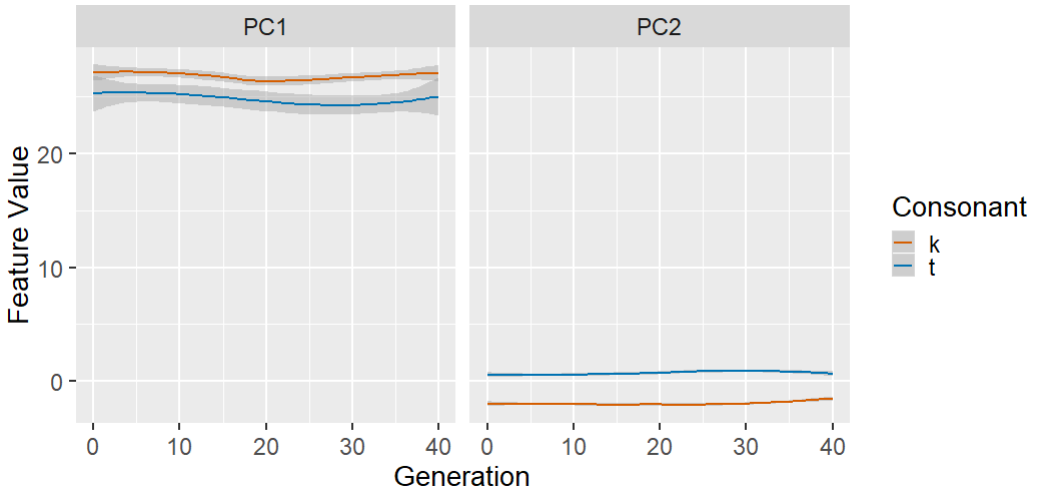

Figure 5.12: Simulated featural results of /k/-/t/ before low back vowels (s=1, p=1)

5.3.3.2 ABM simulations of /p/-/t/

In this section, the results of /p/-/t/ simulations are reported. As for /k/-/t/, the initial category means and mixing parameters were identical for all simulations. The confusability of /p/ and /t/ in the context of high front vowels bears a resemblance to cases of labial palatalization, wherein phonologically or contextually palatalized labial segments take on a dental, alveolar, or postalveolar place of articulation. Consequently, category instability between /p/ and /t/ is predicted to occur before high front vowels, but not in other vocalic contexts.

Because the simulated results for these models are similar to those that have been described in detail for /k/-/t/, the /p/-/t/ results are only summarized here. The full results of the simulations are given in Appendix F.

For /p/-/t/, as for /k/-/t/, when tokens were only categorized when the category of the listener and speaker agreed (s=0) and when tokens were always categorized as the category identified by the listener (s=1, p=0), the prior probabilities of both categories diverged before high front vowels. For this stop pair, the change was so complete that /p/ disappeared completely as a category for listeners in both models. Prior to the disappearance of the /p/ category, featural divergence was also readily apparent under these parameters.

Before low back vowels, /t/ is somewhat more frequent than /p/. There was instability in category mean and mixing parameter, but to a much less extensive degree than in the simulations of /p/-/t/ in the context of high front vowels. In these models /t/ showed a modest increase in mixing parameter relative to /p/, and the two categories diverged in category mean, with /p/ showing higher inter-agent variability in category mean than /t/. When phonetic tokens were always stored as the category intended by the speaker (s=1, p=1), then the simulations showed relative stability in prior,

and inconsistent movement in category mean.

5.3.3.3 ABM simulations of /k/-/p/

The results of the /k/-/p/ simulations are reported in this section. The increased confusability of /k/ and /p/ in the context of high back vowels (see, for example, Experiment 3.2) bears a resemblance to instances of sound change whereby labialized velar segments (e.g., /k^{w}/) take on a labial place of articulation (e.g., /p/; see Table 5.1). Because phonologically labialized velar stops are not identical to velar stops that are coarticulatorily labialized due to an adjacent /u/ vowel, a few assumptions must be made for observations of perceptual asymmetry to speak to such a sound change. In these simulations, tokens of [k] extracted from a pre-[u] context (which show perceptual asymmetry) are assumed to approximate the acoustic properties of a labialized velar stop (or one that appears before [w]). If this assumption is invalid, then the existence of perceptual asymmetry for this consonant pair would not clearly relate to the sound change it has been associated with. These simulations proceed with the intuition that the contextual labialization of a [k] production before /u/ might be relevant to the confusability and category stability of a [k^{w}] production, but this assumption warrants testing.

Category instability between /k/ and /p/ is predicted to occur before high back vowels, but not low back vowels.[6]

5.3.3.3.1 s=0

Recalling §5.3.2.1, when the learner only stores tokens where the listener and speaker agreed on the category, divergence in category mean and changes in category frequency were predicted in the vocalic context conditioning perceptual asymmetry.

/k/-/p/ BEFORE HIGH BACK VOWELS

The prediction is largely confirmed for this segment pair. Because /p/ is lower in frequency than /k/ in this vocalic context, it follows that /p/ would show greater movement over time and would decrease in frequency (Figure 5.13), as happened in §5.3.3.1.1 for /k/ in /k/-/t/. This result, however, follows the pattern opposite the sound change this consonant pair is associated with. Whereas labialized velars tend to become labial segments, the outcome of this simulation is an increase in the frequency of the velar segment. The featural values of /p/ also diverge strongly, as seen in Figure 5.14.

[6]There is also a sound change between /k/ and /p/ that favors /k/ before high front vowels, and this confusion may be associated with cases where a labial segment becomes a palatal in a palatalizing context. This model was not run for the sake of a tractable results section.

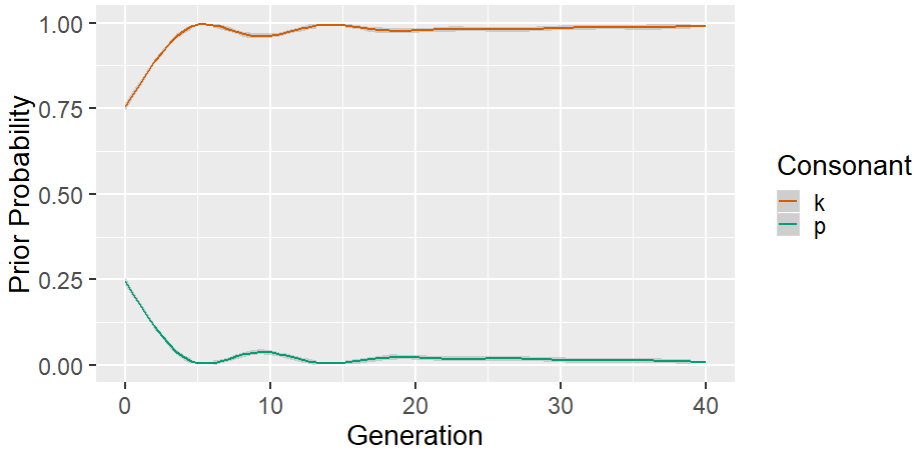

Figure 5.13: Simulated prior results of /k/-/p/ before high back vowels (s=0)

In this vocalic context, there is a large difference in the relative frequencies of the two categories – /k/ occurs 70% of the time while /p/ occurs only 30% of the time. The results reported in this section reflect simulations where the mixing parameters for the categories were defined according to their relative frequencies. An alternative analysis for the high back vowel condition is offered in §5.4.2, where the initial mixing parameter of each category is 0.5. Under these conditions, the results look consistent with the direction of sound change expected for this consonant pair.

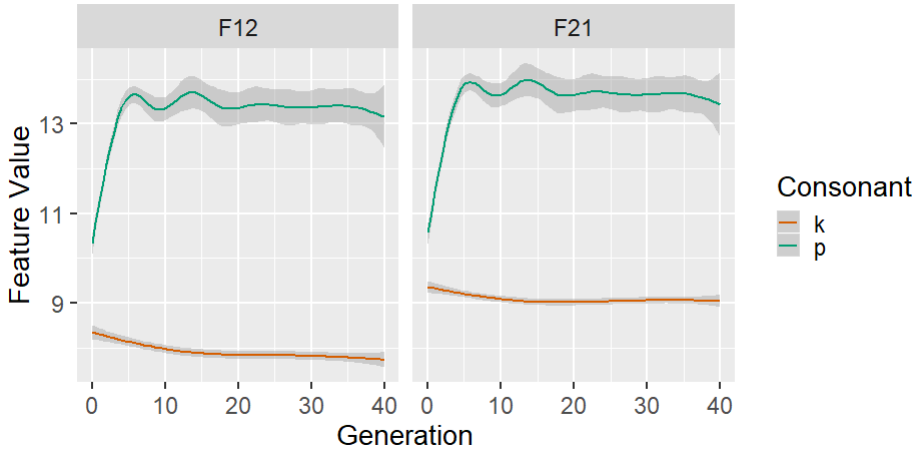

Figure 5.14: Simulated featural results of /k/-/p/ before high back vowels (s=0)

/k/-/p/ BEFORE LOW BACK VOWELS

As in the /k/-/t/ simulations before low back vowels, a condition where stability is predicted, little change is expected, but /p/ does become less common over the course of the simulation, as shown in Figure 5.15. Likewise, both categories diverge in their category means, as is apparent in Figure 5.16. Such a change would seem to reflect slight dissimilation between the two categories.

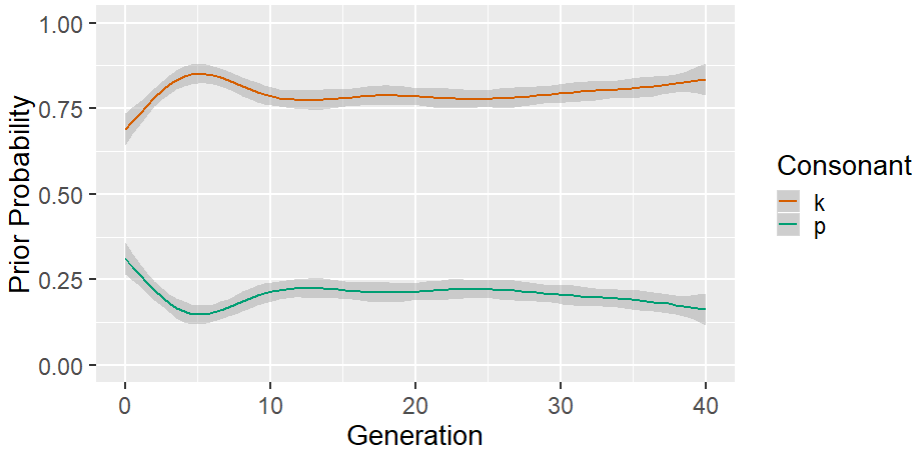

Figure 5.15: Simulated prior results of /k/-/p/ before low back vowels (s=0)

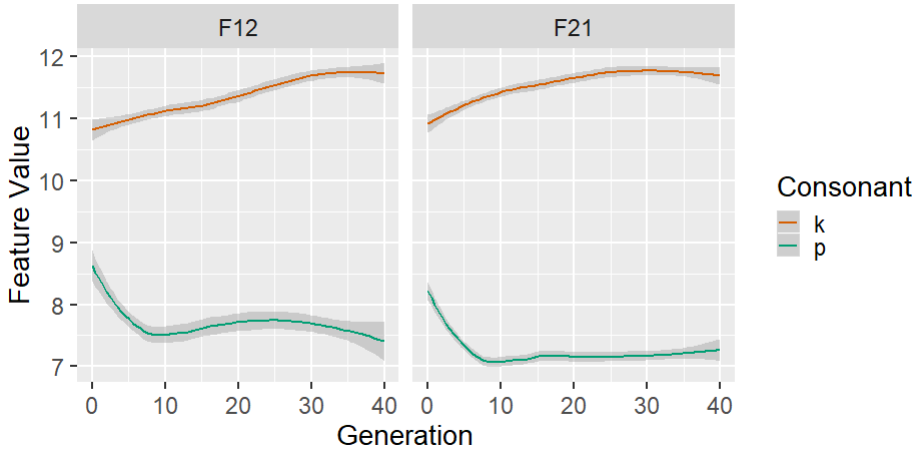

Figure 5.16: Simulated featural results of /k/-/p/ before low back vowels (s=0)

5.3.3.3.2 s=1, p=0

When $s = 1$ and $p = 0$, as in §5.3.3.1.2, the tendency for a listener to store tokens in the category assigned by the listener is predicted to lead to divergence in prior and category mean in the vocalic context that facilitates confusion.

/k/-/p/ BEFORE HIGH BACK VOWELS

The predicted pattern is observed here but, just as in §5.3.3.3.1, the pattern of movement is opposite the direction of the corresponding sound change, as seen below in Figures 5.17 and 5.18.

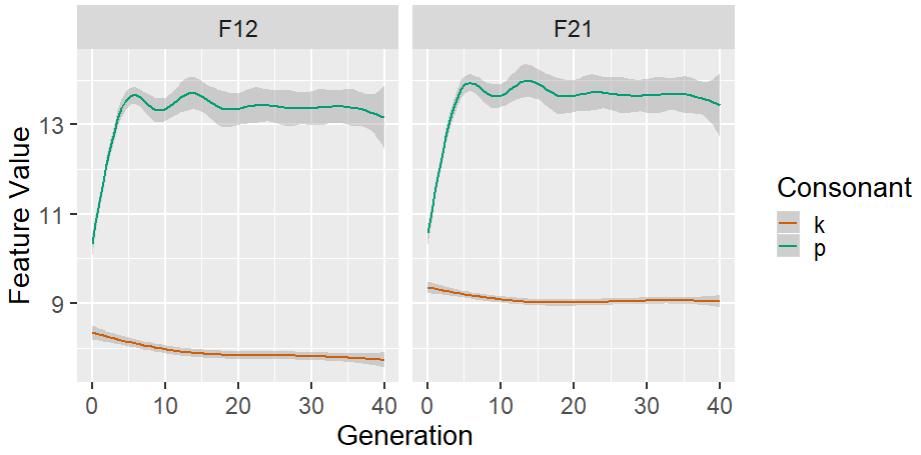

Figure 5.17: Simulated prior results of /k/-/p/ before high back vowels (s=1, p=0)

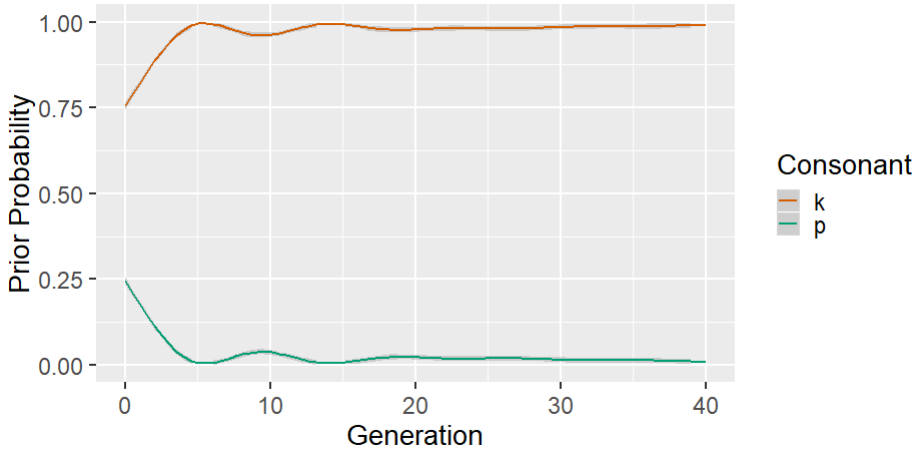

Figure 5.18: Simulated featural results of /k/-/p/ before high back vowels (s=1, p=0)

/k/-/p/ BEFORE LOW BACK VOWELS

As in §5.3.3.1.2, the two categories are stable in prior probability (as seen in Figure 5.19) but show divergence in featural values (Figure 5.20).

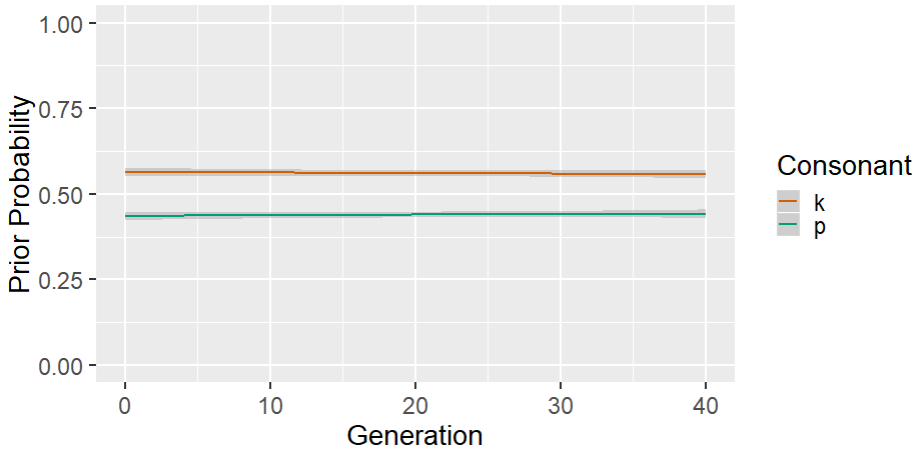

Figure 5.19: Simulated prior results of /k/-/p/ before low back vowels (s=1, p=0)

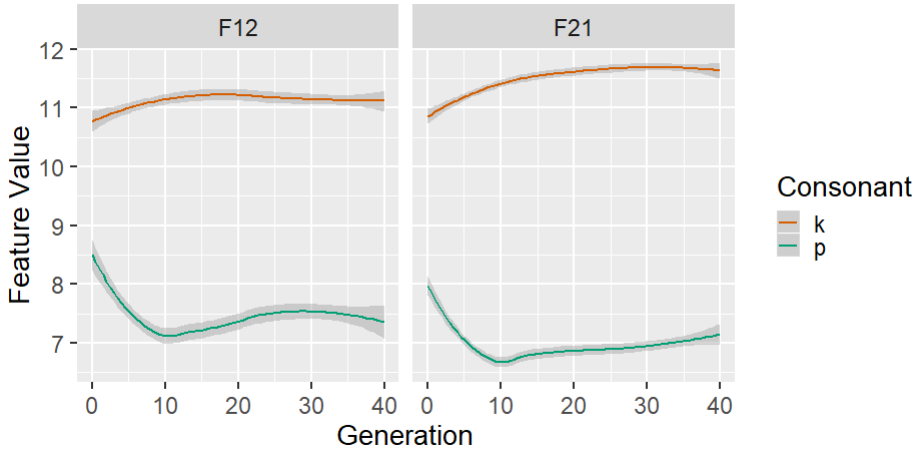

Figure 5.20: Simulated featural results of /k/-/p/ before low back vowels (s=1, p=0)

5.3.3.3.3 s=1, p=1

When $s = 1$ and $p = 1$, learners always store tokens as the category intended by the speaker. Regardless of vocalic context, this condition is uniformly predicted to correspond to category stability.

/k/-/p/ BEFORE HIGH BACK VOWELS

Before high back vowels, there is neither convergence nor divergence in E_{F12} $[t(19) = 1.53, p = 0.14]$ or E_{F21} $[t(19) = 1.46, p = 0.16]$ and no overall change in prior probability (seen in Figures 5.21 and 5.22), as predicted.

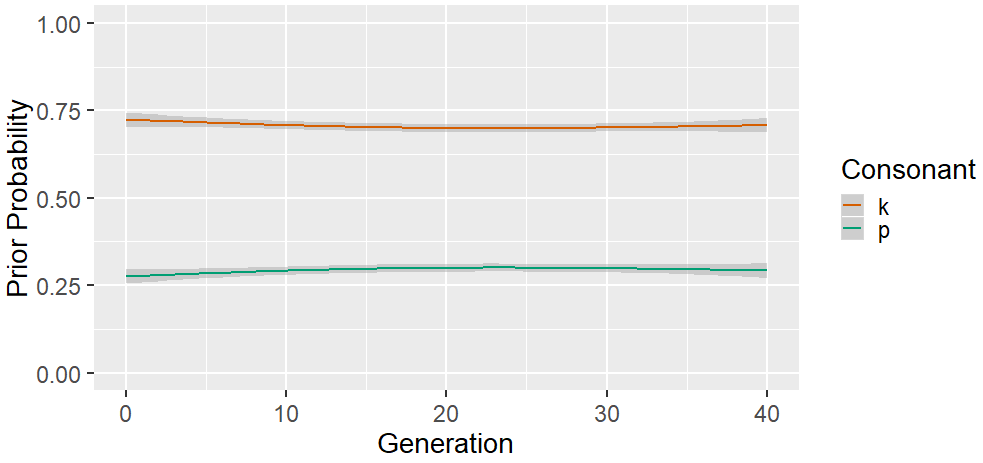

Figure 5.21: Simulated prior results of /k/-/p/ before high back vowels (s=1, p=1)

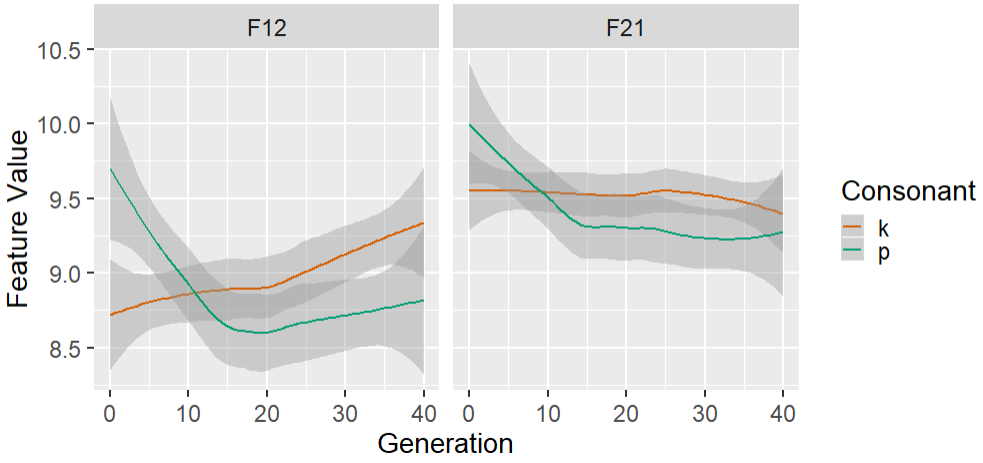

Figure 5.22: Simulated featural results of /k/-/p/ before high back vowels (s=1, p=1)

/k/-/p/ BEFORE LOW BACK VOWELS

Likewise, before low back vowels, there is neither divergence nor convergence in the mean of E_{F12} $[t(19) = 1.79, p = 0.09]$ or E_{F21} $[t(19) = 1.69, p = 0.11]$ or prior probability, seen in Figures 5.23 and 5.24.

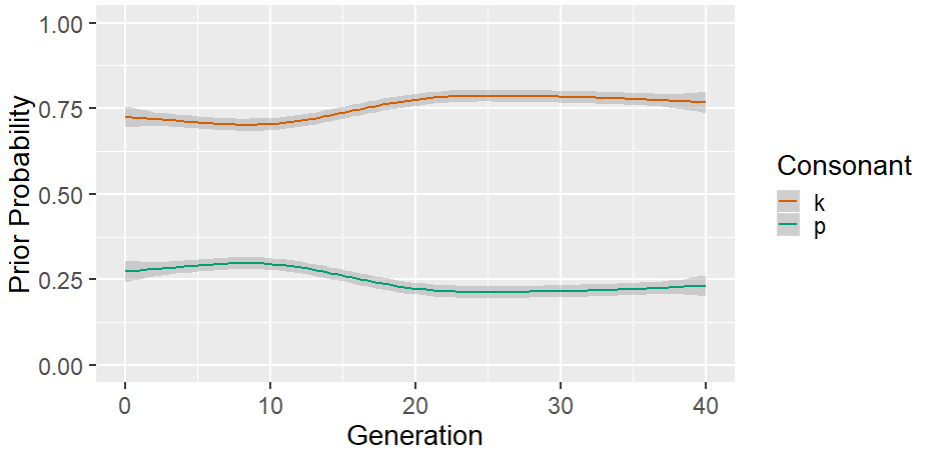

Figure 5.23: Simulated prior results of /k/-/p/ before low back vowels (s=1, p=1)

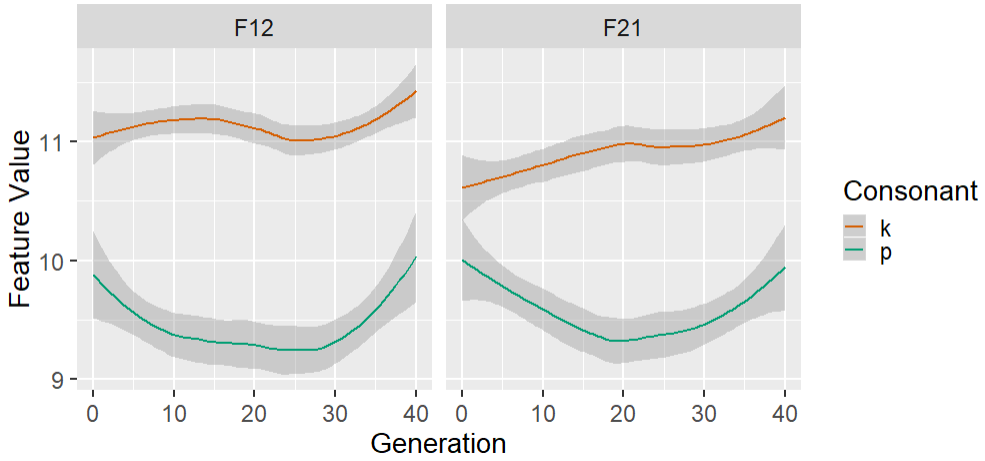

Figure 5.24: Simulated featural results of /k/-/p/ before low back vowels (s=1, p=1)

5.3.3.4 ABM simulations of /θ/-/f/

In this section, the results of /θ/-/f/ simulations are reported. The confusability of /θ/ and /f/ resembles instances of sound change where interdental fricatives take on a labiodental place of articulation. Category instability between /θ/ and /f/ is predicted, but not necessarily in a certain vocalic context. There is tentative evidence from §3.5 that [ɑ] may be a context that conditions confusability more than [i], but Experiment 4.1 suggests that this effect may only appear when the listener has consonantal and vocalic information.

5.3.3.4.1 s=0

In this condition, learners only store correctly identified tokens. As predicted and observed in

§5.3.3.1.1, §5.3.3.2, and §5.3.3.3.1, divergence in prior probability and (sometimes) category mean happen under these settings. This outcome is not observed for /θ/ and /f/, however. The prior probability remains stable (Figure 5.25) and there is divergence in the mean of $PC1_{\theta f}$ $[t(19) = 4.14, p = 0.0006]$ but not $PC2_{\theta f}$ $[t(19) = 0.41, p = 0.69]$ of the two categories (Figure 5.26).

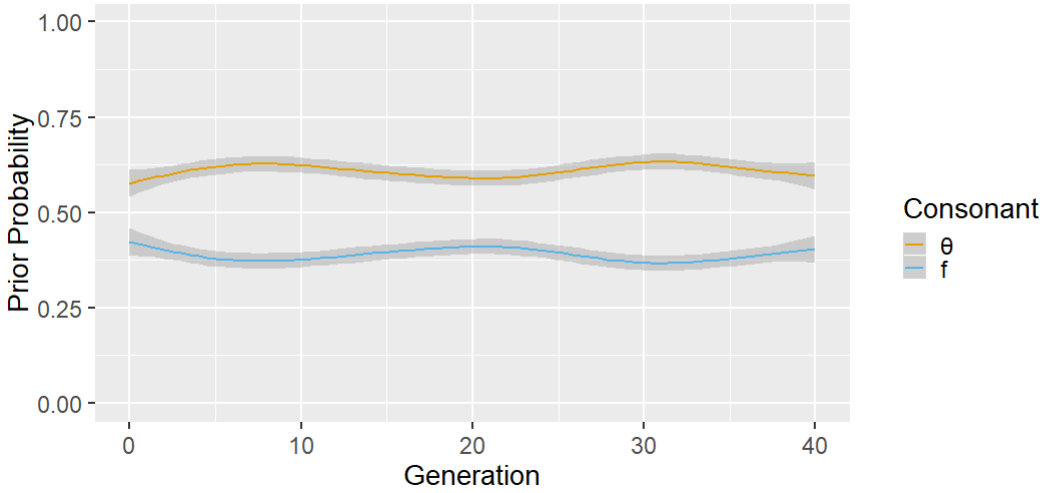

Figure 5.25: Simulated prior results of /θ/-/f/ (s=0)

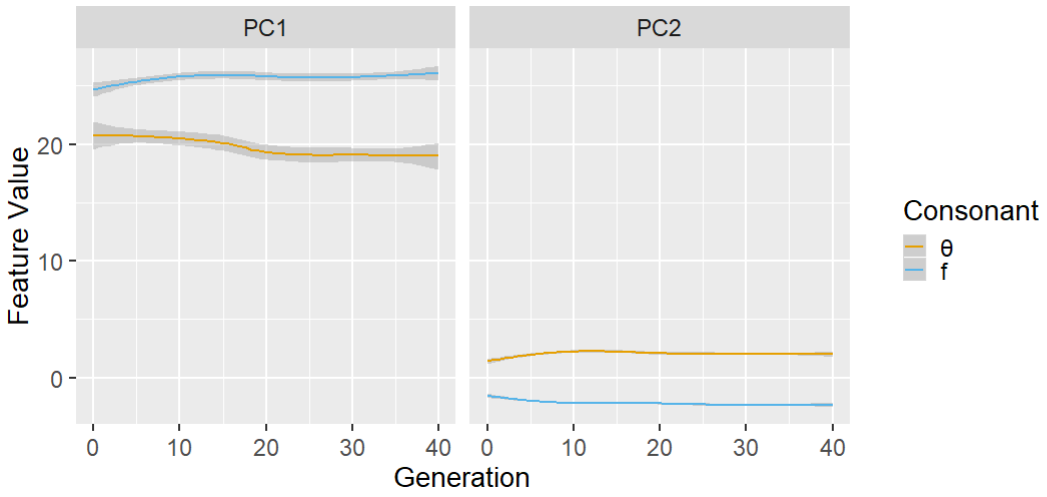

Figure 5.26: Simulated featural results of /θ/-/f/ (s=0)

5.3.3.4.2 s=1, p=0

As seen in §5.3.3.1.2, §5.3.3.2, and §5.3.3.3.2, divergence in prior probability and category mean tends to occur when listeners always categorize each token as the label they assigned to it. Once again, however, this outcome does not hold for /θ/ and /f/, where the two categories show reversal in their prior value and show divergence in $PC1_{\theta f}$ $[t(19) = 6.21, p < 0.0001]$ but not $PC2_{\theta f}$

$[t(19) = 0.31, p = 0.76]$.

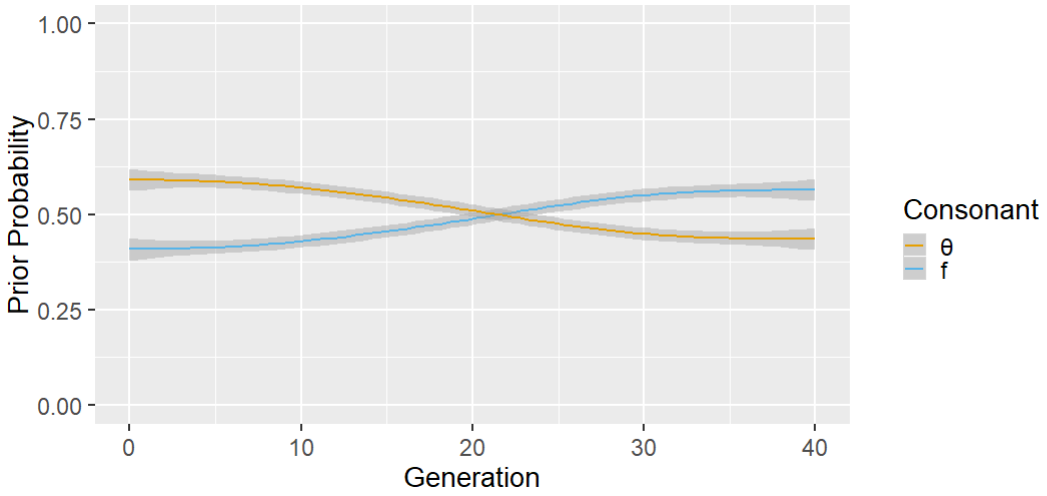

Figure 5.27: Simulated prior results of /θ/-/f/ (s=1, p=0)

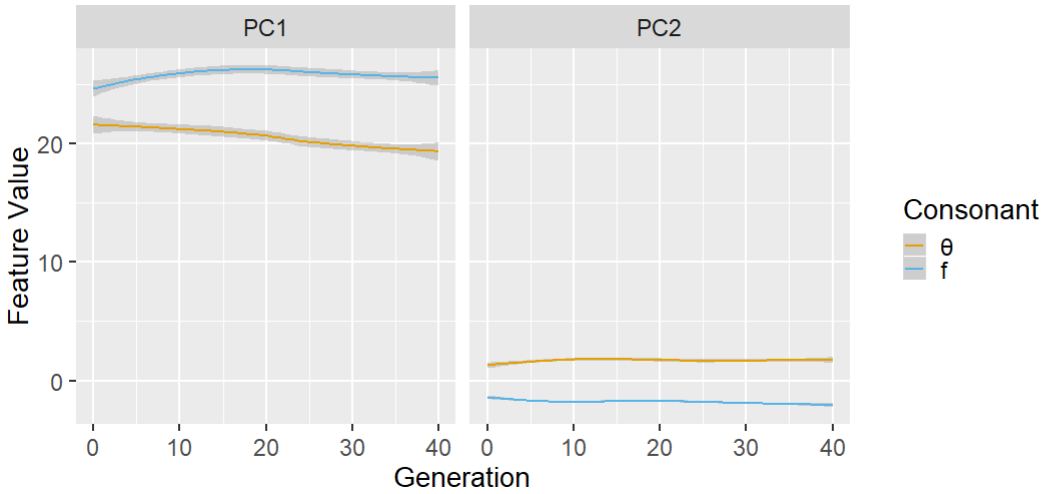

Figure 5.28: Simulated featural results of /θ/-/f/ (s=1, p=0)

5.3.3.4.3 s=1, p=1

When $s = 1$ and $p = 1$, the categories are predicted to show stability. The pair /θ/-/f/ is no exception; there is little movement in frequency or category featural values, as seen in Figures 5.29 and 5.30.

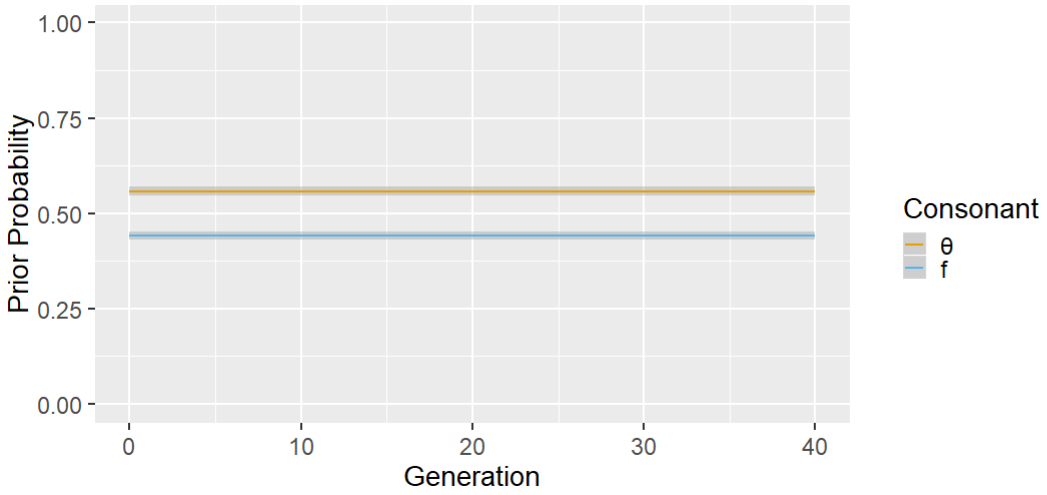

Figure 5.29: Simulated prior results of /θ/-/f/ (s=1, p=1)

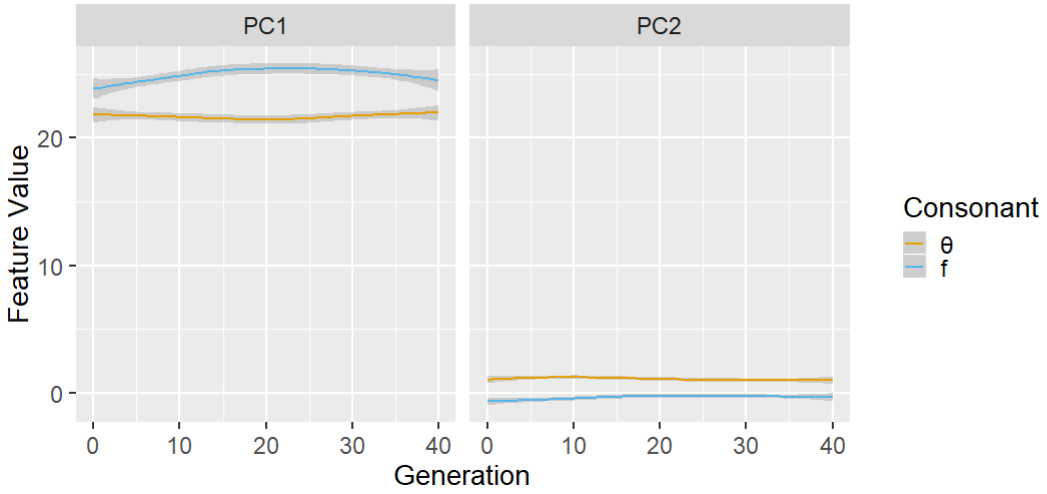

Figure 5.30: Simulated featural results of /θ/-/f/ (s=1, p=1)

5.4 Discussion

Researchers have suggested a conditioning relationship between perceptual asymmetry and sound change, but the mechanism by which one can lead the other is unclear. This chapter used computational modeling to test several hypotheses about how assumptions related to phonetic perception, acquisition and category structure could inform the potential role played by this process. The results of the models are summarized in Tables 5.2 and 5.3.

CONTEXTS CONDITIONING ASYMMETRY							
		s=0		s=1, p=0		s=1, p=1	
	CONTEXT	PRIOR	MEAN	PRIOR	MEAN	PRIOR	MEAN
/k/-/t/	[i,ɪ]	/k/ ↓	*divergence*	/k/ ↓	*divergence*	*no change*	*no change*
/k/-/p/	[u,ʊ]	/p/ ↓	*divergence*	/p/ ↓	*divergence*	*no change*	*no change*
/p/-/t/	[i,ɪ]	/p/ ↓	*divergence*	/p/ ↓	*divergence*	*no change*	*no change*
/θ/-/f/	*none tested*	*no change*	*divergence*	/θ/↓	*divergence*	*no change*	*no change*

Table 5.2: Summary of simulation results under phonetic contexts that condition asymmetry. Green: results conform to prediction. Red: results inconsistent with predictions.

CONTEXTS NOT CONDITIONING ASYMMETRY							
		s=0		s=1, p=0		s=1, p=1	
	CONTEXT	PRIOR	MEAN	PRIOR	MEAN	PRIOR	MEAN
/k/-/t/	[ɑ]	*no change*	*divergence*	/t/ ↓	*no change*	*no change*	*no change*
/k/-/p/	[ɑ]	/p/ ↓	*divergence*	*no change*	*divergence*	*no change*	*no change*
/p/-/t/	[ɑ]	/t/ ↓	*divergence*	/t/ ↓	*no change*	*no change*	*no change*

Table 5.3: Summary of simulation results under phonetic contexts that were not understood to condition asymmetry. Green: results conform to prediction. Red: results inconsistent with predictions

5.4.1 The role of model parameters on simulation outcomes

When perceptual asymmetry is posited to condition sound change, the misidentification of the speaker's intended category is portrayed as playing a concrete role in the process. Such a possibility is captured by models where $s = 1$ and $p = 0$, which could reflect a situation where there is insufficient information to clarify the intended production of the speaker beyond the phonetic information available to the listener. Under these circumstances, one of the two categories tends to become less frequent, and the enlarged (or shrunken) categories also diverge from one another, potentially approximating a partial merger. The underlying assumption for such a model is not inconsistent with any of the approaches to sound change described in §5.1.1. The listener recovers a novel phonetic target unlike the one intended by the speaker, either due to a failure to account for phonetic context, or an idiosyncratic cue-weighting strategy, or superior attention to the phonetic signal.

For many computational models of sound change (e.g., Harrington et al., 2018), the storage of phonetic tokens is at least partially contingent on whether the label identified by the listener matches the category intended by the speaker. This standard assumption appears to concretely influence the long-term evolution of phonetic categories. For simulations where $s = 0$, phonetic

categories tended to show divergence and changes to mixing parameter over the time-course of the model. Such models performed similarly to simulations where perceived tokens were always stored in the category assigned by the listener.

When phonetic tokens were uniformly stored as the category intended by the speaker ($s = 1$, $p = 1$), the priors and means of the category tended to remain stable for the duration of the simulation. These simulations could potentially represent communicative contexts where multiple channels of information are available to the listener to help identify the speaker's intended production, which would appear to help facilitate stability. The listener may be able to use knowledge about other levels of linguistic structure to correctly identify the category of the consonant.

5.4.2 Unexpected results for /k/-/p/: the role of prior in simulation outcomes

The outcomes for /k/-/p/ were unexpected. Confusions of /k/ and /p/ before high back vowels are associated with change in place of articulation of the velar consonants, but this consonant pair evolved in the opposite direction - /p/ (rather than /k/) tended to become infrequent over a few generations. Prior probability may play a role in this outcome. Because /k/ is extremely frequent, listener agents will tend to categorize all tokens as /k/, which prevents most /p/ tokens from being stored, and which then likely further reduces the mixing parameter for /p/.

Figures 5.31 and 5.32 show a simulation of /k/ and /p/ before high back vowels (when s=0) with the initial prior manually adjusted, such that both categories have a frequency of 0.5 (their actual prior probabilities are 73% and 27%, respectively, in the original models). Despite having equal probability, /p/ becomes the more frequent category over 40 generations, consistent with what was observed in /k/-/t/ and /p/-/t/.

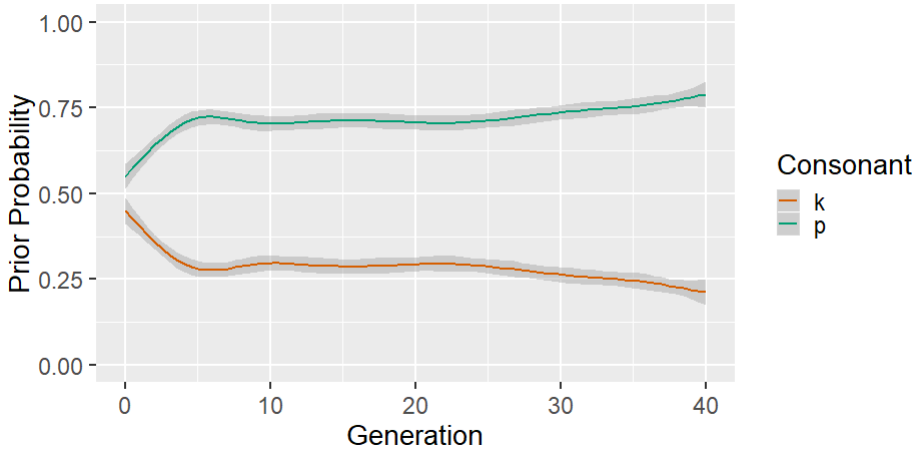

Figure 5.31: Simulated change in prior for /k/-/p/ before high back vowels (s=0). Initial priors have been artificially set at 0.5 for each category

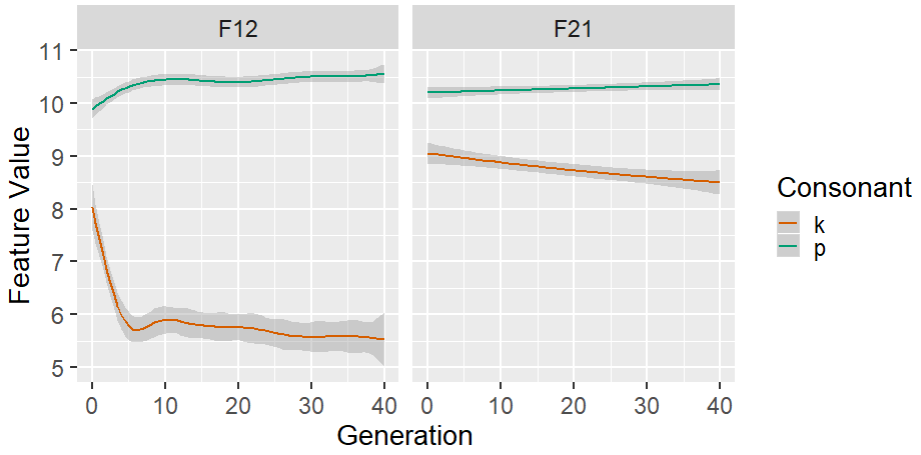

Figure 5.32: Simulated change in category means for /k/-/p/ before high back vowels (s=0). Initial priors have been artificially set at 0.5 for each category

The results of these simulations underscore the importance of choosing a suitable mixing parameter for models like these. Only when the mixing parameter for the two were equal did the category instability work in the direction predicted based on the sound change. Under the assumption that the asymmetry captured by [k] and [p] in the context of [u] is indeed an appropriate choice to simulate stability or phonetic change in labialized velars (i.e., the unexpected result is not a result of an inappropriate choice of likelihood probability distribution for the /k/ category), it is worth asking how a suitable mixing parameter could have been identified a priori. As noted earlier, several models of phonetic perception have simply assigned categories equal prior probabilities (e.g., Norris & McQueen, 2008), but this choice does not seem to acknowledge the sizeable body of literature showing that perception is not simply sensitive to phonetic distributions alone.

While listeners may be sensitive to the frequency of a category, raw probabilities may not be properly representative of the effect of frequency on category bias. A listener may not necessarily assume that a category is twice as likely to appear simply because it appears twice as often as another category. A transformation of probabilities may end up being more appropriate. In fact, if small differences in frequency map to even more negligible differences in category prior, the assumption of roughly equal priors (as has been done in other models) may not prove so inappropriate.

5.4.3 Unexpected results for the dental fricatives

Confusions of /θ/ and /f/ have been associated with the sound change of a dental fricative to a labiodental fricative. In contrast to all other consonant pairs, the two categories show relative stability in all parameter settings reported. Unlike /k/ and /p/, the relative frequencies of the two

categories are fairly similar (/f/ and /θ/ have frequencies 45% and 55%, respectively), though this was also true for /k/ and /t/.

As noted in §5.3.1.3, each category distribution is defined also by a covariance matrix, which describes the shape of the distribution. The determinant of this matrix gives the generalized variance, which can describe how scattered a distribution is around its mean. Figures 5.33 and 5.34 compared generalized variance of simulations of /f/ and /θ/ when s=0 and s=1, p=1. When s=0, (which is associated with some type of instability for all simulations in the vocalic context conditioning asymmetry), the generalized variance for /f/ (the less frequent category) falls to far below that corresponding value for /θ/. In contrast, when s=1, p=1, the parameter uniformly associated with category stability, the two categories show a comparable change in variance. These results suggest that the dental fricatives may be showing variance, though not by change in featural category mean. Both consonants are becoming more compact, with /f/ showing a steeper rate of contraction. This result seems to demonstrate divergence (as seen in the other consonant pairs) but without category movement.

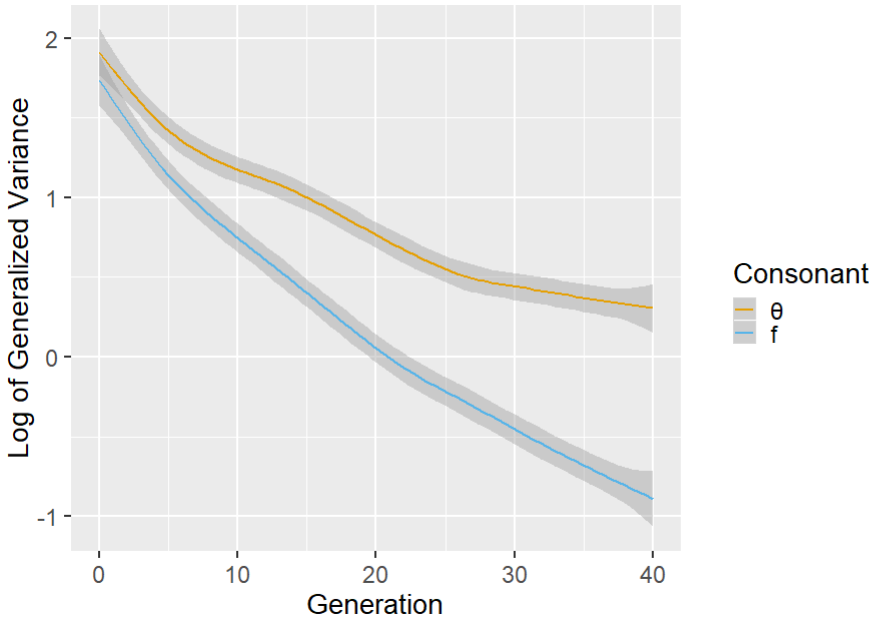

Figure 5.33: Simulated change in generalized variance for /θ/-/f/ (s=0)

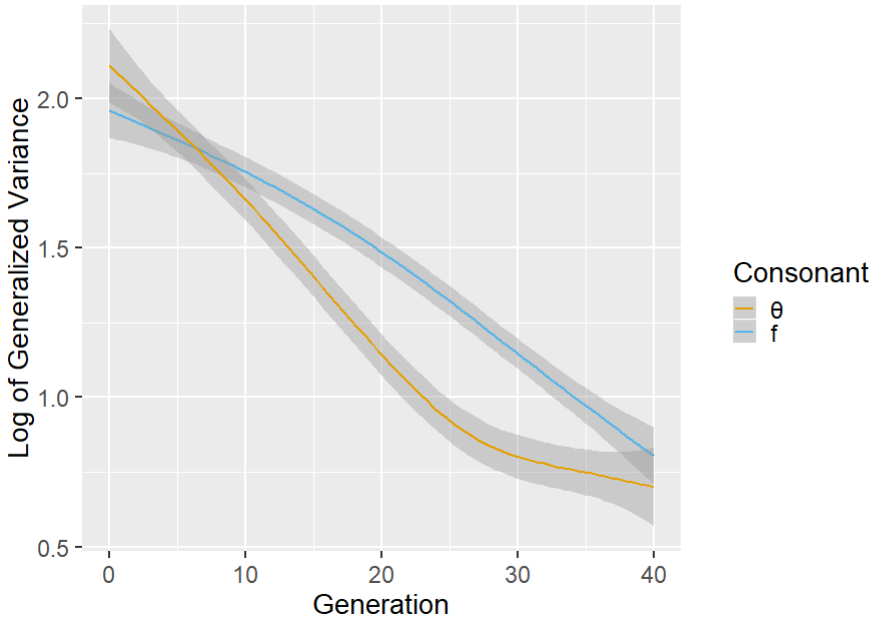

Figure 5.34: Simulated change in generalized variance for /θ/-/f/ (s=1, p=1)

This unexpected outcome for the dental fricatives may have something to do with the nature of overlap for the pair. Category movement in these models is constrained to some extent by the original extent of the consonant category. If tokens in a certain region of the distribution of this distribution tend to be correctly classified more often than tokens generated in another region, then the category mean may shift toward that region of the original distribution over successive generations. As noted in Experiment 4.1, however, /θ/ and /f/ showed extensive degrees of overlap with one another in every vocalic context tested. There may not have been a region for either consonant where productions in the region would be more reliably classified as the category. For multivariate gaussians, the highest likelihood probability occurs near the category mean, so perhaps tokens generated here were most likely to be perceived correctly, in which case there could be a decrease in variability over generations without movement in category mean.

5.4.4 Model limitations

For some models (particularly those that implemented vowel-independent perception), overlapping categories with extremely low mixing probability ($< 10\%$ initially) tended to decrease in frequency to 0%, except for simulations where $s = 1$ and $p = 1$. Increasing the number of tokens encountered by each agent in a generation to 1000 (in simulations not reported in these results) did not appear to fix this problem, so it is unclear if this issue is inherent to the model, or if the number of tokens needs to be increased by several factors more.

A general tendency in all simulations was for the generalized variance (the determinant of the

covariance matrix) for each category to decrease over successive generation. This outcome appears to be consistent with the model used in Kirby (2014) and appears to be a result of population structure in the simulations. Each learner agent is initialized on a mean vector, and a covariance matrix whose parameters are estimated from all the teacher agents of that generation. For each generation, the parameters of the adult learners are ultimately a mixture of the parameters of the teacher agents from the previous generation. Such 'blended inheritance' can have the effect of reducing variance, as is shown by Boyd and Richerson (1985). This general decrease in variance is non-optimal for the simulation of a speaker community; there is no reason to expect that listeners would also behave in this manner. However, it is difficult to see how this outcome could be avoided in an iterated learning scenario, as is used here and in most intergenerational models of sound change.

5.4.5 Revisiting perceptual asymmetry and sound change

Taken together, these simulations offer several predictions about how perceptual asymmetry might help to condition sound change from the standpoint of intergenerational transmission. The results of these models suggest that some factors may allow perceptual asymmetry to contribute to category instability consistent with sound change.

1. PERCEPTUAL FACTORS

   (a) (s=1, p=0): A learner lacks reliable access to non-phonetic information about a token's category identity (e.g., there are few contrasts between the two consonants), or

   (b) (s=0): The learner tends not to learn about phonetic categories from tokens that are not highly discriminable.

2. PRIOR/MIXING PARAMETER CONSIDERATIONS

   (a) When two categories show overlap, a category with significantly larger prior is expected to further increase in frequency relative to the other category.

While it is not possible to decisively demonstrate that these sound changes in fact require these conditions, components of these predictions can be evaluated. For example, perceptual learning studies can offer insight on how parallel channels of information available to the listener or the distributional properties of tokens influence the learnability of phonetic categories.

The theories of sound change described in §5.1.1 describe how a listener might come to perceive a phonetic token that is unlike what the speaker had intended, which could in turn be reproduced and lead to wider language change. When learner agents had perfect access to the intended category of the speaker, the categories tended to remain stable over time. This result creates the

most friction with the Hypo- and Hyper- Articulation model, where access to the fine phonetic detail of a (presumably correctly categorized) speech token affords listeners the ability to reproduce this form, potentially leading to community wide change. These results suggest that change might only be possible if the listener-turned-speaker produces this novel phonetic form at a rate more frequent than it would normally appear in speech randomly. If the novel phonetic form is only produced with the same frequency as it occurs naturally, then it is unclear if this would have any effect on the stability of phonetic categories over generations. This hurdle is ultimately easy to surmount, however – if this phonetic variant were to acquire, for example, some type of socioindexical value, then speaker's use of this form might increase far beyond its natural rate of occurrence.

The results of this simulation offer similar degrees of support to innocent misapprehension and cue-based approaches to sound change. Under the assumption that the listener would always store a token in the category the listener identified it as, instability was observed among the phonetic categories, which often led to divergence, and in some cases led to a shift in mixing parameter for the two categories.

Finally, the condition where listeners do not learn from miscategorized tokens (i.e., when s=0) does not seem to fit easily into any of the models, even though this does appear to be a relatively standard assumption in designing sound change models. Under an innocent misapprehension approach, a listener that failed to account for the phonetic environment of [k] in [ki] might recover /t/ or /tʃ/ instead. Under a cue-based account, a listener may weight perceptual cues in such a way that [ki] is perceived in the same category as /tʃ/. In both cases, though, the listener would need to be able to then reproduce the tokens whose categories do not match with those of the speaker. If listeners do not store the misclassified tokens, only tokens that are less likely to be confused for the other category are stored – specifically the [k] tokens that sound less like /t/ or /tʃ/. This would give the listener-turned-speaker less access to /t/ or /tʃ/-sounding [k] tokens in production. While this model assumption does give results consistent with a communicative setting where listeners store all tokens in the categories they assigned them to, it also aligns somewhat poorly with some listener-based accounts of sound change. Perhaps listeners really only have access to the phonetic variability present in tokens that have been correctly classified, but it is as of yet unclear whether this is a realistic description of how listeners actually perceive speech.

5.4.6 Thinking ahead about sound change modeling

The simulations performed in this chapter provide a specific type of insight. They describe how sampling randomly from a Gaussian mixture can lead to changes in mixing parameter and category mean over successive iterations depending on the initial conditions of the mixture and differing assumptions about what sampled tokens can serve as input to the estimation of the next model's

mixture parameters. The changes in the model can be likened to changes in the acoustics of a consonant over time, and the storage conditions in the simulation can be likened to the way in which a listener might process stimuli, but there would need to be extensive additional theoretical work done for this model to speak more directly about what role a synchronic feature could play in a sound change.

It would first be necessary to clarify how the acquisition of phonetic categories should be modeled. In the simulations, a learner agent is initialized with the parameter distribution of the generation, implying that a first-language system is shaped largely by the speakers with whom they interact in a community. While that assumption is likely not unreasonable, individuals are not in fact born with a well-developed phonetic system comparable to that of their community. Similarly, children do not abruptly acquire an adult phonological system after having stored enough tokens. A more comprehensive simulation of phonetic category acquisition would likely reflect incremental learning of these categories.

This chapter explores the effect of differing storage criteria on the simulation outcome. To the extent that the feasibility of each condition depends on factors that naturally vary across language varieties (e.g., the presence of contrastive pairs in the language), these simulations can provide a rough prediction about when perceptual asymmetry is more likely to contribute to category instability. However, it may also be possible that one of these criteria is inherently inconsistent with how human phonetic processing works. In this case, it would be possible to narrow the space of storage criteria explored to those consistent with human perception.

Finally, a more comprehensive model of intergenerational transmission would not necessarily involve the exchange of isolated phonetic tokens. Listeners typically have access to multiple levels of linguistic structure, and so it is not necessarily safe to assume that the candidate phonetic categories are the only structures under considerations during phonetic perception, as happens in this simulation. A somewhat more comprehensive model of communication is employed in Fulop and Scott (2019), where individuals produce monosyllabic words instead of isolated speech sounds. Such a system could be further embellished by incorporating additional layers of linguistic structure, including prosodic and semantic context.

The simulations provide a concrete step toward understanding what role perceptual asymmetry can play in sound change in that they lay out predictions of the circumstances where phonetic instability over time might be more likely. For this experimental paradigm to speak more directly to what will happen in a speaker community over time, however, more research needs to be done to clarify how the process of perception and language acquisition should be modeled.

CHAPTER 6

Discussion

The primary focus of this dissertation was perceptual asymmetry in consonants. Because this perceptual phenomenon relates to several different phonetic sub-disciplines, asymmetry was studied from a variety of perspectives. Chapter 2 undertook an articulatory analysis of consonant pairs showing perceptual asymmetry, with the intent of investigating the degree to which perceptual similarity among members of each consonant pair was mirrored by spatial similarity in the vocal tract when these consonants are produced. Chapter 3 combined an exploratory acoustic analysis with a perceptual experiment to identify spectral regions relevant to the discrimination of the members of a consonant pair. Chapter 4 evaluated the extent to which a Bayesian framework helps explain why confusions show asymmetry, and whether differences in likelihood probability have a perceptual consequence for listeners. Finally, in response to suggestions in the literature that perceptual asymmetry may condition certain sound changes, Chapter 5 used ABM simulations to identify the conditions under which stability or long-term phonetic change would occur. Although the analyses in this dissertation focused on the spectral properties of these consonants, those are not the only cues that distinguish the members of the pair. Future analyses of these tokens would benefit from a more comprehensive analysis that also considers the role of non-spectral phonetic cues, as was done in Plauché (2001).

Section §6.1 discusses the results for each consonant pair, and §6.2 discusses the results general to all four.

6.1 Pair-specific findings

6.1.1 /k/-/t/

Researchers (e.g., Guion, 1998; Chang et al., 2001) have given by far the most attention to confusions between /k/ and /t/, which favor /t/ in the environment of a high front vowel. Its putative

'poster-child' status in perceptual asymmetry may be justified in that it provides the clearest depiction of the features characteristic of the phenomenon. The results of this dissertation are no exception; the findings for this consonant pair tended to align with predictions across all experiments.

In production, the vocalic constrictions of [i] interact with the consonantal constrictions of either [t] or [k] to produce vocal tracts with similar shape. Consequently, the spatial characteristics of the vocal tract relevant to stop burst and aspiration characteristics are most similar between /k/ and /t/ before [i] (Experiment 2.1).

Acoustically, differences in energy over 1.3-3 kHz and 5.2-7.2 kHz can be used to contrast /k/ and /t/ (Experiment 3.1), but listeners only showed sensitivity to variability in energy over the lower frequency range. Within this acoustic region, productions of [k] show variability. The distribution of /k/ shows more overlap with /t/ before high front vowels than before other vowels. In contrast, the distribution of [t] productions showed what appeared to be comparable overlap with /k/ in all vocalic contexts tested. In fact, [k] and [t] productions show the highest likelihood of being generated by /t/, relative to their likelihood of being generated by /k/ (Experiment 4.1). The modeled difference in likelihood of being generated by /k/ or /t/ also seemed to have perceptual relevance to listeners – they were more likely to categorize a [k] and [t] as /t/ if it came from the High 'likely /t/' grouping (Experiment 4.2).

6.1.2 /k/-/p/

Unlike /k/-/t/, the pair /k/-/p/ was potentially associated with two asymmetries that appear to work in opposite directions regarding whether /k/ is favored. Winitz et al. (1972) and Plauché (2001) found that listeners tended to favor one consonant before [u] and a different consonant before [i] (though their results were not consistent about which were favored).

The articulatory results were consistent with both researchers' findings of greater perceptual similarity in the environment of high vowels than in the environment of /ɑ/. Productions of these consonants involve a labial (for [p]) or velar (for [k]) constriction. In the context of [u], these constrictions can overlap with those of the vowel, which also involve a velar and labial constriction. In contrast, productions of /k/ and /p/ before [i] involve the presence of a palatal constriction (due to [i]), which reduces the size of the vocal tract in the regions responsible for aspiration characteristics (in the case of [p]) and burst and aspiration characteristics (in the case of [k] and [p]). The productions of these consonants tended to show significantly smaller differences before high vowels than before [ɑ] (Experiment 2.1).

RF models identified two distinct spectral regions at around 1.4 kHz and 4.3 kHz where differences in energy at this region seemed to inform the classification of the consonants (Experiment

3.1). The regions identified were roughly consistent with the peak frequency locations for /k/ and /p/ predicted by a tube model of the vocal tract. However, listeners' identifications of the consonants in a CV context showed sensitivity to changes in energy over a wider range of frequencies (Experiment 3.2).

/k/ tended to show more variability than /p/ within these spectral regions. The spectral characteristics of /k/ tended to show greater overlap with those of /p/ before high back and low back vowels, while /p/ tended to show high spectral overlap with /k/ in every vocalic context explored. Productions of /p/ showed the highest modeled probability of having been generated by /k/ before high front vowels (consistent with Winitz et al. 1972's finding), and /k/ showed a higher modeled probability of having been generated by /p/ before high back vowels than before high front vowels, though there was no significant difference between high back vowels and low back vowels (Experiment 4.1). However, the results of Experiment 4.2 did not offer clear support to the claim that these differences in likelihood probability had perceptual relevance for listeners.

6.1.3 /p/-/t/

Like /k/-/t/, /p/ and /t/ were associated with a single asymmetry that favored /t/ before high front vowels Winitz et al. (1972). However, unlike /k/-/t/, the conditioning vocalic context for /p/-/t/ does not involve a constriction that occurs between the two consonantal constrictions. For this pair, the palatal constriction (for [i]) occurs posterior to both consonantal constrictions and so did not appear, in the articulatory analysis, to increase the similarity of the vocal tract in a way relevant to burst characteristics. The results of Experiment 2.1 also suggest that the palatal constriction increases the similarity of the consonants in a manner relevant to aspiration characteristics, but not burst acoustics.

The spectral regions identified by RFs were consistent with acoustic models of labial and alveolar stop releases, but the loss of energy in any one of these regions did not affect listeners' likelihood to categorize one stop as the other. Unlike the previous two pairs, /p/-/t/ showed little phonetic variability according to vocalic context. The high front vowel context was the only vocalic context in which the phonetic distribution of [t] productions showed comparatively little overlap with the distribution of [p] productions. As predicted, however, /p/ had the highest modeled likelihood of being generated by /t/ before high front vowels, and this likelihood measure seemed relevant to listeners. Participants were significantly more likely to categorize a [p] and [t] as /t/ when the consonant appeared in the 'likely /t/' grouping (Experiment 4.2).

6.1.4 /θ/-/f/

The voiceless dental fricatives have received a sizeable amount of perceptual study, but it remains unclear what reliable acoustic differences exist between the two consonants, to what extent these cues are present in the frication noise itself, and which of these differences are perceptually relevant for the listener's discrimination of the two.

Narayanan et al. (1995) noted that speakers varied in the degree of tongue body raising and tongue tip protrusion in the production of [θ], but because schwa was the only context used in that study, the effect of vowel coarticulation on these productions was unclear. In the production experiment conducted for this study (Experiment 2.1), the cavity anterior to the dental constrictions showed a greater difference for /θ/ and /f/ before high front vowels (relative to high back vowels), a result which may be due to the effect of the palatal [i] constriction on the tongue tip constriction of /θ/. The results may be consistent with the finding in Experiment 4.1 that the targeted spectral properties of both fricatives showed a somewhat smaller degree of overlap before high front vowels than other vocalic contexts.

A RF model revealed several frequency components informative to the discrimination of /θ/ from /f/ (Experiment 3.1), suggesting that the cues present in the spectrum are diffuse. In perception, a wide range of manipulated spectral components was associated with increased identification errors (Experiment 3.2), suggesting that any information useful to the discrimination of the pair might be spread across the frequency range.

In the phonetic space identified in Experiment 3.1, productions of /θ/ tended to not differ in the likelihood of being generated by /f/ (relative to their likelihood of having been generated by /θ/) according to vocalic context (Experiment 4.1). This result contrasts with the finding of Experiment 2.1 that the two fricatives showed greater articulatory dissimilarity before high front vowels, and the finding of Experiment 3.2 that [ɑ] may condition increased [θ] confusions. If confusion asymmetry is conditioned by vocalic context, the acoustic cues informing this asymmetry may not be located in the frication spectrum alone.

6.2 General findings

6.2.1 ABM results

While consonants pairs that show asymmetry in the laboratory (e.g., /k/-/p/ in the environment of [u]) resemble comparable sound changes (e.g., /k/ becoming /p/ when labialized), the exact mechanism by which one contributes to the other has not been outlined. The simulations of Chapter 5 suggest that perceptual asymmetry can be consistent with either long-term phonetic stability or change, depending on the structure of the phonetic categories, the parallel channels of information

available to the listener during perception, and whether misidentified tokens are assumed to have an effect on phonetic categories.

When the two categories showed little overlap, then, regardless of the additional assumptions of the model, one phonetic category tended not to become more predominant, and the category means of the two categories tended not to diverge. In contrast, in cases where the two categories showed higher degrees of overlap, the ultimate outcome of the model depended on additional assumptions.

Category stability across time tended to be observed when the model assumed that listeners were able to perfectly recover the category intended by the speaker, and all perceived tokens were stored by the listener. The recovery assumption may be consistent with most natural communicative settings, where listeners have access to visual, lexical, semantic, and pragmatic information that can aid them in figuring out what production was intended by the speaker.

When either of the assumptions in the previous paragraphs did not hold, then category instability tended to be observed for overlapping categories. This outcome was observed when listeners stored only tokens identified as the same category as the speaker intended; perhaps category typicality or discriminability (as suggested by Todd et al., 2019) plays a role in phonetic category storage. This outcome also emerged when listeners stored all tokens as the category they perceived it as possibly because, when there are few parallel channels of information available to the listener to determine which of the two productions was intended, the listener must rely on the acoustics cues within the token itself to discriminate between the two possibilities. Under both settings, one phonetic category tended to become predominant over time, and the means of the two categories tended to diverge from one another. This result is consistent with the partial merger of the two categories.

As token classification took place using a Bayes Optimal Classifier, differences in prior probability played a large role in which category would become predominant over generations. In the consonant pair /k/-/p/ (in the environment of /u/), /p/ became predominant (as expected) when the prior for the two consonants was 0.5, but /k/ instead became favored (against expectations for that vocalic context) if the prior for /k/ was 0.7 (consistent with the relative frequency for /k/ vs. /p/ in that environment). There are few recommendations in the literature about how to go about deciding on a prior value for such a model, but it is worth noting that it is possible to affect the phonetic category outcome according to this measure.

Taken together, these results suggest that consonant pairs that show perceptual asymmetry can exist comfortably within a sound system in a language, if there are additional non-phonetic channels of information that allow the listener to recover the intended production of the speaker. In cases where this assumption becomes false – perhaps a separate change happens in the language that reduces the number of minimal pairs for members of a consonant pair – then the two categories are predicted to show instability. Which category becomes predominant depends in part on their

prior probability and in part on how the two categories overlap.

The ABM simulations used for Chapter 5 could be significantly enriched so as to draw closer parallels with how individuals acquire and transmit language. It would be worth introducing classification decisions that are not binary – listener agents might learn and categorize tokens that appear from a variety of categories. Token classification also does not have occur at the level of (context-dependent) segment. Agents of an enriched model could instead categorize biphones, syllables, or even full words. The size of the community could be larger per generation and each learner could acquire their speech from multiple teacher entities. Finally, as the phonetic categories of adults are known not to crystallize fully in real life, instability in teacher agent categories could be incorporated into the model. Instead of just talking to learner agents, teacher agents could also converse with one another.

6.2.2 Perceptual asymmetry and the phonetic primitive

Consonant pairs that show perceptual asymmetry tend to involve the confusion of productions that differ in active articulator, place of articulation, or both. While this mismatch may be immaterial to an auditory theorist, it may prove theoretically interesting to a gestural theorist. The ambiguity between the two productions may, for example, be purely acoustic, a coincidental spectral overlap between otherwise dissimilar productions. It could also be that this acoustic ambiguity is mirrored by ambiguity in the spatial configuration of the vocal tract – the acoustic signal fully specifies the shape of the vocal tract, but different articulatory events could produce such a configuration. Chapter 2 suggests that the latter may be possible. The articulatory settings analyzed in the study (relevant to either burst, aspiration, or frication acoustics) tended to show the greatest similarity in the vocalic context that conditioned perceptual asymmetry, suggesting that the vocal tract is in fact taking on similar configurations along spatial parameters relevant to the production's spectral acoustics. In a gestural approach, the productions would create a vocal tract shape ambiguous between two articulatory settings.

Recalling Miller and Nicely (1955), as the signal to noise ratio of the speech decreased, listeners tended to categorize consonants less accurately, but the confusion patterns were structured. For example, listeners would tend to confuse place of articulation before they would misidentify manner of articulation. The confusions explored in this dissertation seem to reflect a structure to the confusion as well, whereby the rough vocal tract shape is correctly identified while the specific articulators used to create the shape are not. This structure could reflect the granularity of information specified by the acoustic signal. It may be possible for the signal to not reliably specify articulator but reliably specify vocal tract shape.

There are other productions that show a mismatch in articulatory and acoustic variability. These

pairs would provide cases where the predicted similarity in vocal tract shape could be tested. English rhotic approximants have been discussed throughout this dissertation, but [ʃ] and [ʂ] (a voiceless retroflex fricative), for example, are another type of acoustically similar productions with different constriction locations and which recruit distinct articulators. It would be interesting to see to what extent these productions in fact create a similar vocal tract shape despite their articulatory difference. Similarly, while [l] and [ɹ] do not normally show high misidentification rates for one another, there may be increased confusability in favor of [l] in the environment of /ɑ/ (Müller, 2010). The additional pharyngeal constriction characteristic of both [ɑ] and [ɹ] may play a role in this asymmetry, but the extent to which confusability between these sonorants can be explained by vocal tract shape is as of yet unclear.

6.2.3 Building and testing predictions with humans

The results of Experiment 3.1 offer additional reason to consider RF models an effective technique to locate phonetic differences between speech sounds. For the voiceless stops, the regions identified tended to correspond straightforwardly to the differences in peak frequency location predicted by acoustic models of the vocal tract. Such an analysis could also be extended to non-spectral measures (e.g., duration, amplitude) and measures that vary across time (e.g., change in energy over certain time points of the production). Experiment 3.2, however, suggests that it is not possible to directly infer what listeners will be able to use in speech from these results - the perceptual results of Experiment 3.2 only showed ballpark similarity to what had been predicted from Experiment 3.1. In §3.5, there was a discussion of where these differences may have come from, in part due to how the tasks in Experiment 3.2 differ from those of Experiment 3.1.

An RF is not identical to the human perceptual system in form or function, so the categorization strategies of the two will still likely differ even if the two are completing the same task. Reliable differences between two categories that are imperceptible to listeners may still be captured by a RF, in which case the features important to a human would be a subset of those useful to the algorithm. Furthermore, if listeners assign variable weights to different phonetic cues, not even every perceptible phonetic difference will affect how a listener categorizes a speech sound. In an ideal case (i.e., one where both the model and listeners were relying on the same input to perform their classification), a RF would be able to capture the broadest set of phonetic differences between categories, and listeners would make use of a subset of these features to contrast the speech sounds.

6.2.4 Perceptual asymmetry and beyond

The results of Chapter 4 (along with those of Plauché, 2001) offer support for thinking about perceptual asymmetry from a Bayesian framework – in contexts favoring perceptual asymmetry,

productions of the disfavored category are more likely to have been generated by the favored category than by their own. This in turn leads to greater classification errors in the disfavored category.

Despite existing justification for this framework, more testing is needed to further clarify how perceptual asymmetry works. For example, it is unclear if these findings generalize to other obstruents, or to sonorant consonants, which also participate in perceptual asymmetry. Second, the choice of prior in Bayesian models could be refined. As this choice has been shown in Chapters 4 and 5 to affect the outcome of classification (and simulation), additional work is warranted to determine what an appropriate initial mixing parameter would be for listeners. Perhaps fitting a Bayesian model to perceptual results would allow for an appropriate prior to be identified.

The methodology used in Chapter 4 could also be used to identify new candidate speech sound pairs that might demonstrate perceptual asymmetry. The role of prior and the category distributions in determining classification outcomes offer clear predictions about how the acoustics of the speech sound category will relate to its confusability with another category.

APPENDIX A

Buckeye Corpus Token Frequencies

[p]	FRONT	CENTRAL[1]	BACK
HIGH	1512	-	453
MID	1352	852	517
LOW	232	-	680

Table A.1: [p] tokens extracted from the Buckeye Corpus

[t]	FRONT	CENTRAL	BACK
HIGH	4098	-	2539
MID	2297	3188	858
LOW	445	-	645

Table A.2: [t] tokens extracted from the Buckeye Corpus

[k]	FRONT	CENTRAL	BACK
HIGH	3790	-	1202
MID	1869	3458	909
LOW	875	-	909

Table A.3: [k] tokens extracted from the Buckeye Corpus

[θ]	FRONT	CENTRAL	BACK
HIGH	4065	-	47
MID	738	484	200
LOW	187	-	43

Table A.4: [θ] tokens extracted from the Buckeye Corpus

[1]The only monophthongal vowels classified as central are /ʌ/ and /ə/.

[f]	FRONT	CENTRAL	BACK
HIGH	1129	-	220
MID	547	834	1301
LOW	472	-	298

Table A.5: [f] tokens extracted from the Buckeye Corpus

APPENDIX B

Mel-frequency Filterbank Information

FILTER	MIN. FREQUENCY (HZ)	PEAK FREQ.	MAX. FREQ.
F1	31	63	94
F2	94	125	188
F3	156	219	281
F4	250	313	375
F5	344	406	469
F6	438	500	594
F7	531	625	719
F8	656	750	875
F9	781	906	1031
F10	938	1063	1219
F11	1094	1250	1406
F12	1281	1438	1625
F13	1469	1656	1844
F14	1688	1875	2094
F15	1906	2125	2375
F16	2156	2406	2688
F17	2438	2719	3000
F18	2750	3031	3375
F19	3063	3406	3781
F20	3438	3813	4219
F21	3844	4250	4719
F22	4281	4750	5250
F23	4781	5281	5844
F24	5313	5875	6500
F25	5906	6531	7188
F26	6563	7219	7969

Table B.1: Mel-frequency filterbank information

APPENDIX C

An Alternative Method of Characterizing Filter Importance

C.1 Algorithm

1) A random forest was first grown with 4000 tokens (=2000 tokens randomly sampled from each category with replacement). 2) The partial dependence of each variable was estimated. Partial dependence describes the marginal effect of a feature on the predicted outcome of the response variable.[1] 3) The proposed effect size for the feature was then calculated by fitting a least squares model with response based on the tree-averaged predicted values obtained in Step 2. The dependent variable in this model was the tree-averaged likelihood of a consonant classification response. 4) Steps 1-3 were repeated 19 times, for 20 bootstrap samples.

C.2 Limitations

Adjacent filter energies tend to be correlated with one another. Because partial dependence is calculated on a marginal distribution, this calculation relies on an assumption that data points populate the region under examination. If two (or several) features are correlated with one another, the model may need to extrapolate to regions of the parameter space poorly populated by data points in the training set. Depending on how the model accounts for these regions, a violation of this assumption can lead to unpredictable estimates of effect.[2] Despite the limitations of the method described in §3.3, features with high correlation tend to behave more predictably – they

[1] Partial dependence plots were generated using the **`iml`** package in R.

tend to show similar mean decreases in Gini.

C.3 Results

C.3.1 /k/-/t/

Figure C.1 plots the estimated effects of each filter on the response variable (classification as /k/ or /t/). The shaded regions are the filters that had been targeted for further analysis using the importance plots in §3.3.2. The filters in red have effect sizes that differ significantly from zero. The regions targeted using the method in §3.3.2 tend to correspond to filters whose effect size differs from zero (except for Filters 9 and 10). However, the magnitude of effect size for filters within this region shows considerable variability. For example, the magnitude of effect in Filters 15 and 16 is several orders greater than the other filters, even though all were identified as relatively important with the method in §3.3.2. Since the energy in adjacent filters is expected to be correlated, it is unclear whether the large differences in effect size observed for this consonant pair should be expected.

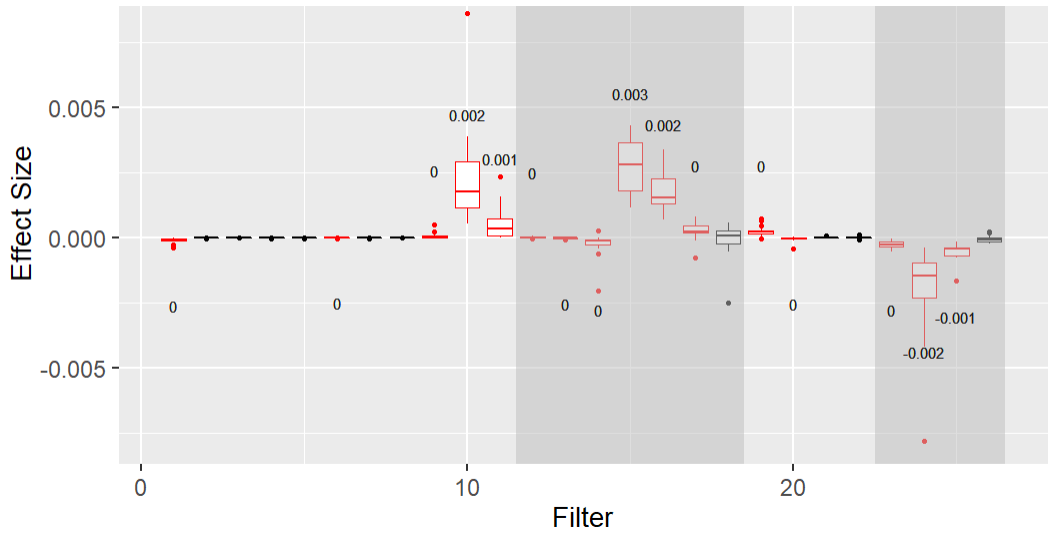

Figure C.1: Estimated effects of filters in /k/-/t/ classification. For filters whose means differ significantly from zero (shown in red), the sample mean of the effect is shown to three decimal places. Filters whose effects do not differ significantly from zero are shown in black. A positive mean effect indicates that the listeners are more likely to classify the token as /k/ with increasing energy at that filter

[2]Accumulated local effects (Apley & Zhu, 2016) are an alternative to partial dependence that calculate the effect of the feature on the response variable over a conditional (not a marginal) distribution. This methodology avoids the prediction of values outside the training set and so would be expected to treat correlated variables in a more reasonable manner.

C.3.2 /k/-/p/

Figure C.2 shows a plot of the estimated effect size for each filter in the task of /k/-/p/ classification. The two filters identified using the importance plot correspond to an effect size that differs from zero, but only the effect of Filter 21 showed an effect size with large magnitude. Filters 22-26 all have effects that differ significantly from zero, but this difference was not apparent in the importance plot.

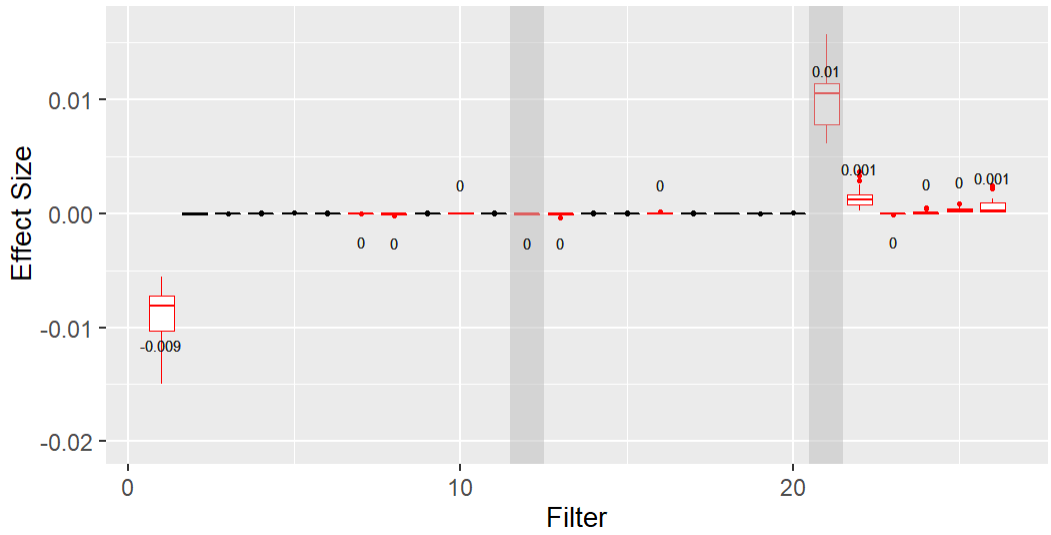

Figure C.2: Estimated effects of filters in /k/-/p/ classification. A positive mean effect indicates that the listeners are more likely to classify the token as /k/ with increasing energy at that filter.

C.3.3 /p/-/t/

The estimated effect of each filter on the classification of /p/ and /t/ is shown in Figure C.3. The regions targeted by the analysis in section §3.3 tend also to be the regions whose effect size differ significantly from zero and whose magnitude tend to be relatively large (except Filters 20 and 21).

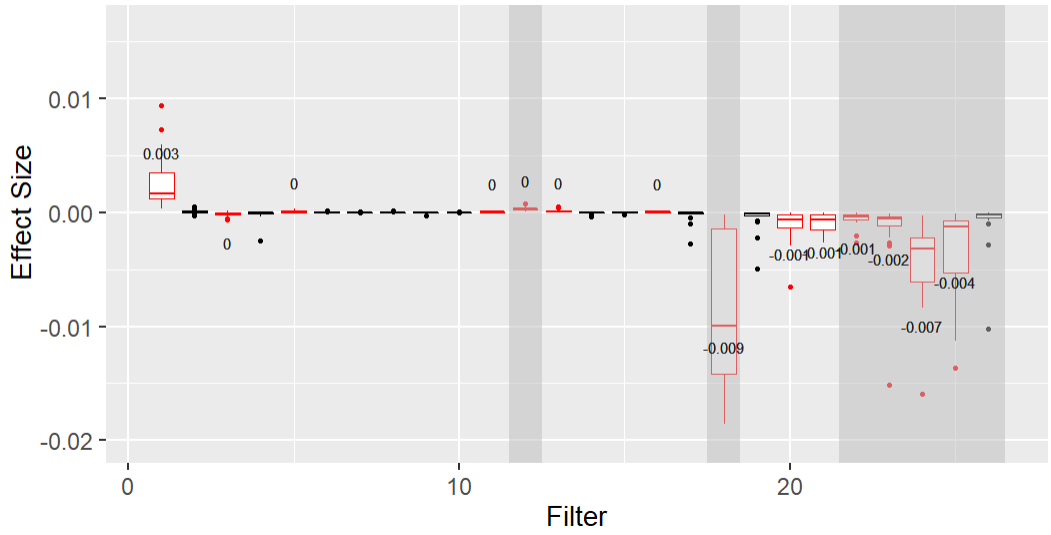

Figure C.3: Estimated effects of filters in /p/-/t/ classification. A positive mean effect indicates that the listeners are more likely to classify the token as /p/ with increasing energy at that filter

C.3.4 /θ/-/f/

Figure C.4 shows the estimated effect for each filter in the task of /θ/ and /f/ classification. Unlike the previous consonant pairs, most filters have an effect that differs significantly from zero. However, the regions with the largest effect size tend to agree with the filters identified using the importance plots.

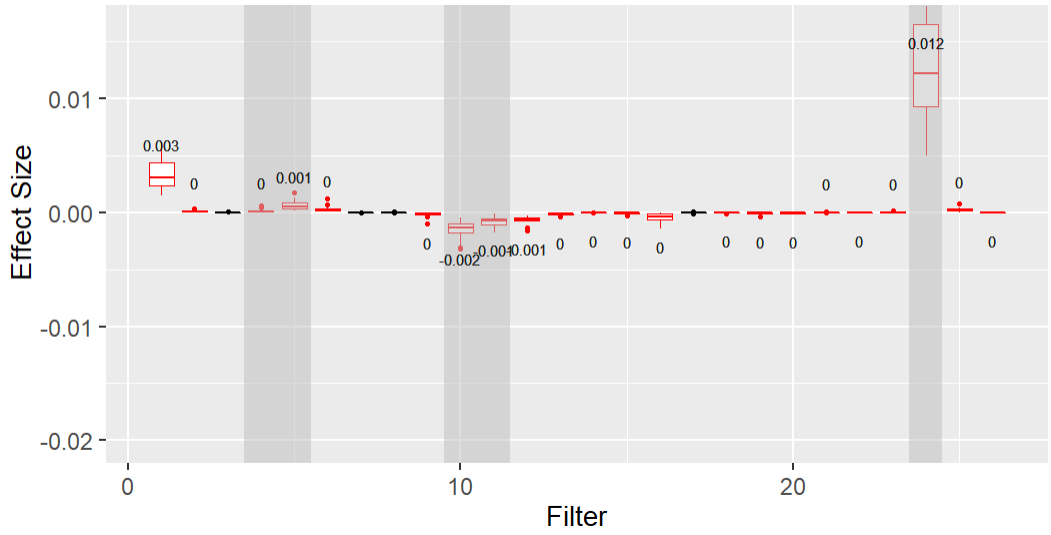

Figure C.4: Estimated effects of filters in /θ/-/f/ classification. A positive mean effect indicates that the listeners are more likely to classify the token as /θ/ with increasing energy at that filter

APPENDIX D

Spectral Bands Identified from Experiment 3.1 Overlaid by Filters

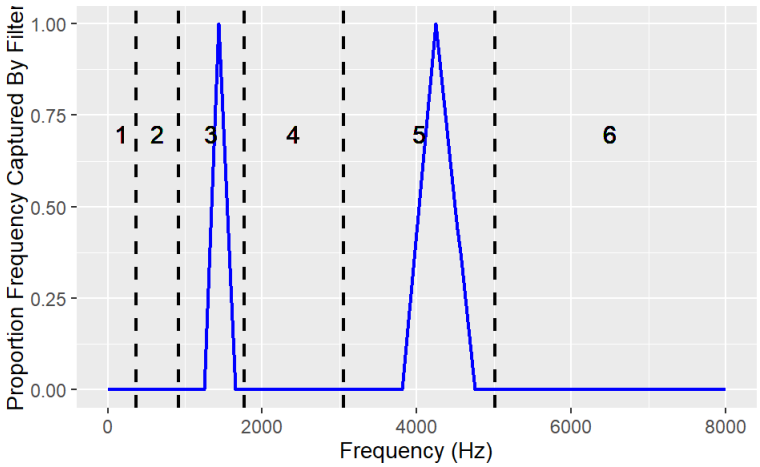

Figure D.1: Notch filter boundaries and frequency regions 'important' to classification of /k/ and/p/

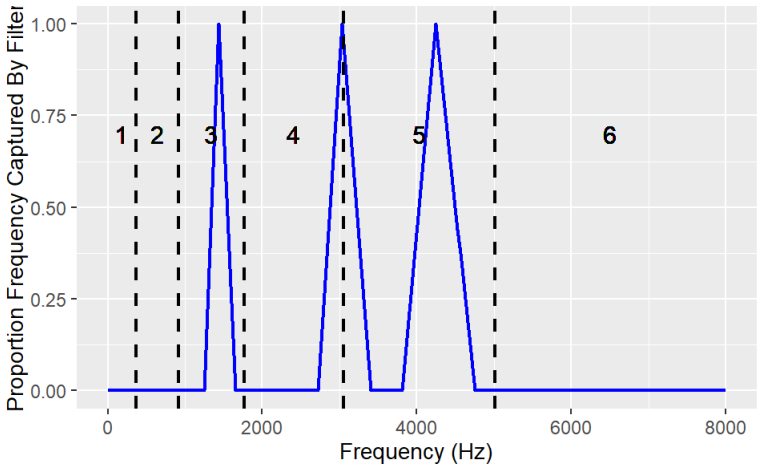

Figure D.2: Notch filter boundaries and frequency regions 'important' to classification of /p/ and/t/

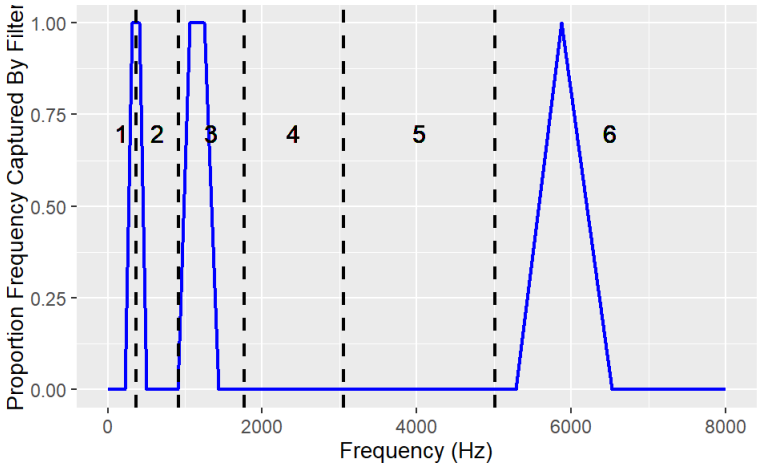

Figure D.3: Notch filter boundaries and frequency regions 'important' to classification of /θ/ and/f/

APPENDIX E

Plots for Experiment 4.2

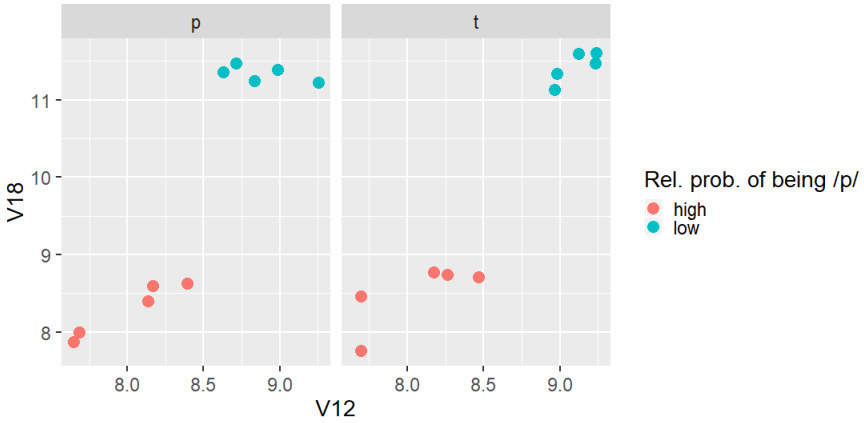

Figure E.1: /p/ and /t/ stimuli selected for Experiment 4.2

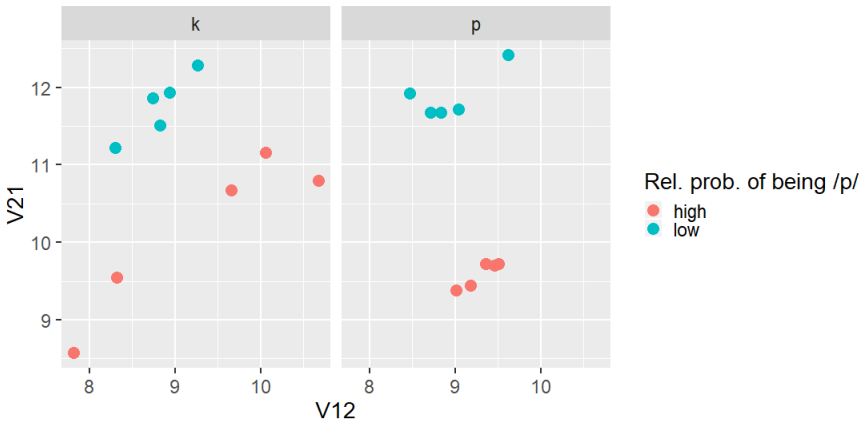

Figure E.2: /k/ and /p/ stimuli selected for Experiment 4.2 (before [i])

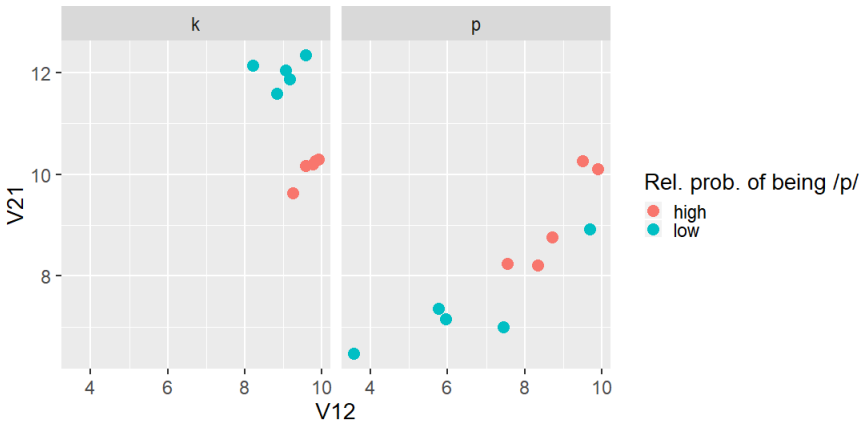

Figure E.3: /k/ and /p/ stimuli selected for Experiment 4.2 (before [u])

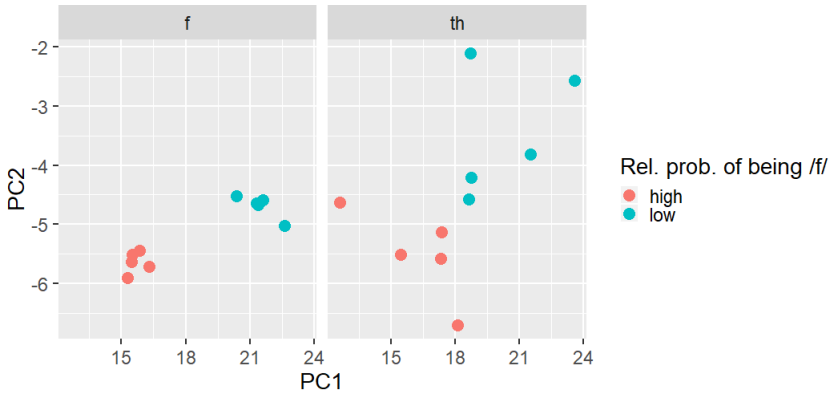

Figure E.4: /θ/ and /f/ stimuli selected for Experiment 4.2

APPENDIX F

ABM Simulations of /p/ and /t/

F.1 s=0

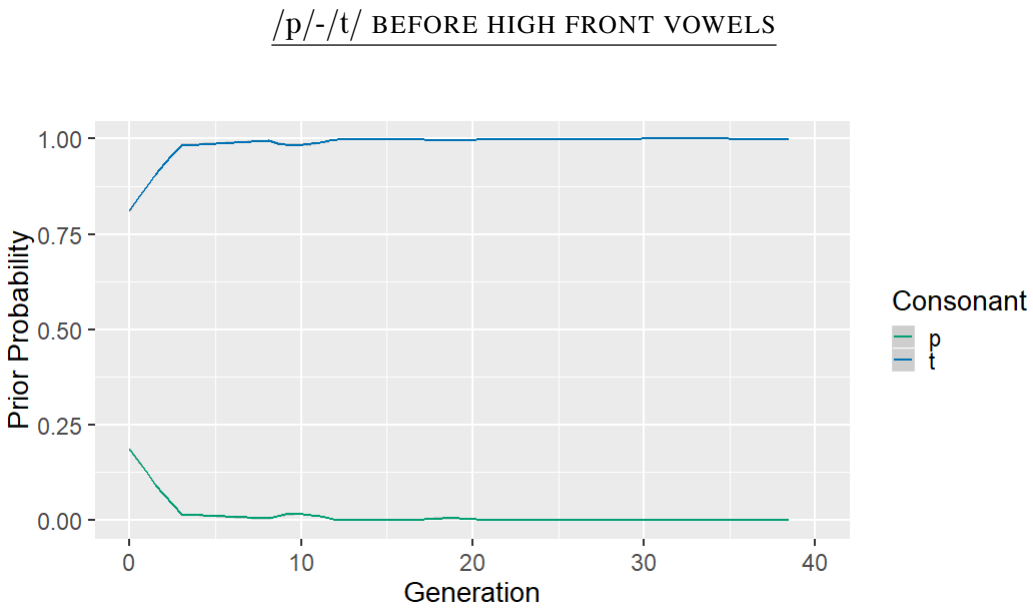

Figure F.1: Simulated prior results of /p/-/t/ before high front vowels (s=0)

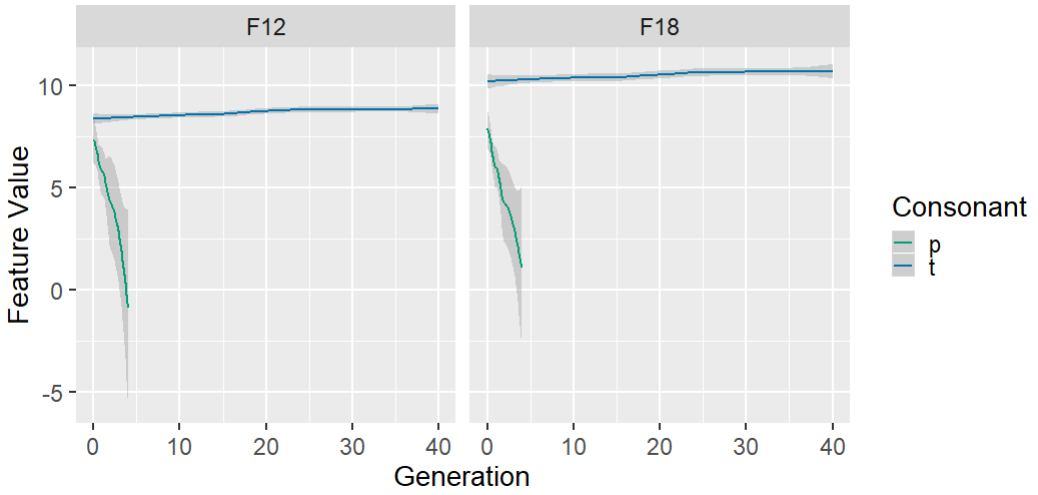

Figure F.2: Simulated featural results of /p/-/t/ before high front vowels (s=0)

/p/-/t/ BEFORE LOW BACK VOWELS

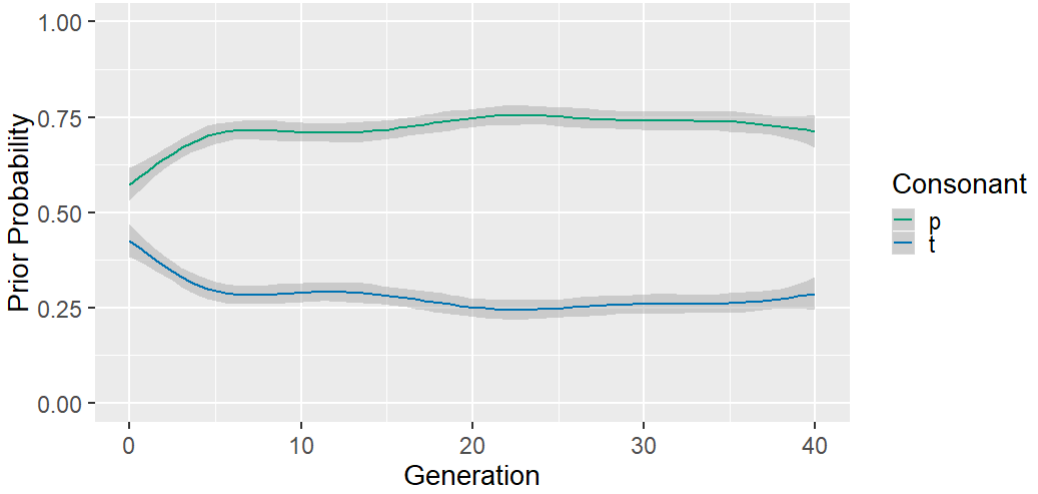

Figure F.3: Simulated prior results of /p/-/t/ before low back vowels (s=0)

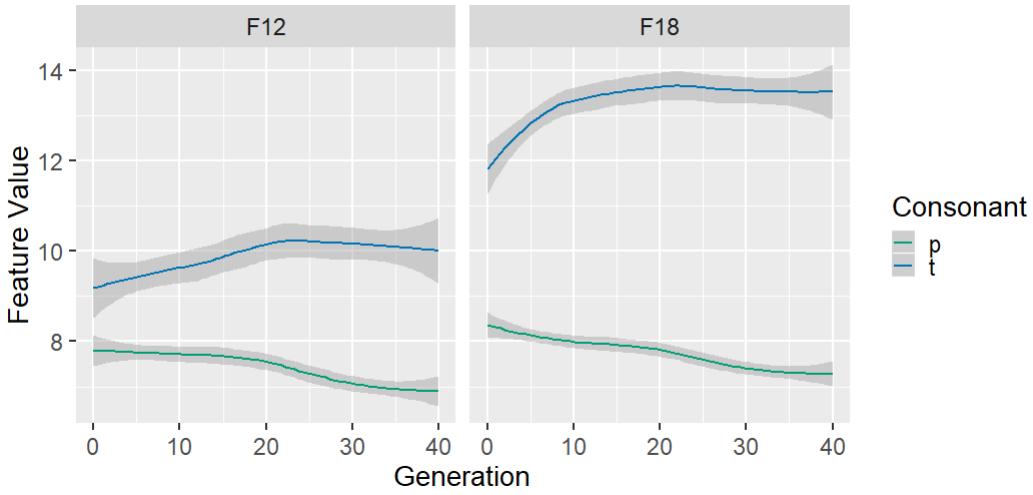

Figure F.4: Simulated featural results of /p/-/t/ before low back vowels (s=0)

F.2 s=1, p=0

/p/-/t/ BEFORE HIGH FRONT VOWELS

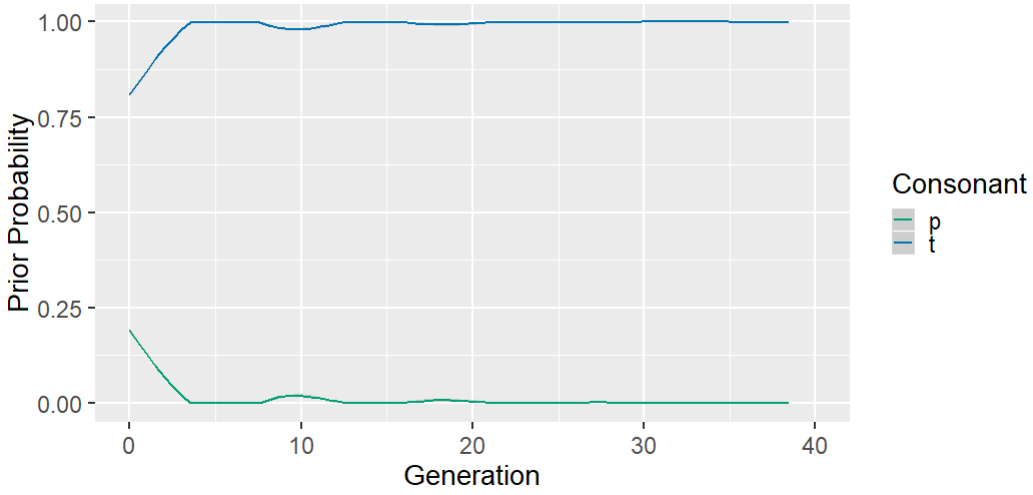

Figure F.5: Simulated prior results of /p/-/t/ before high front vowels (s=1, p=0)

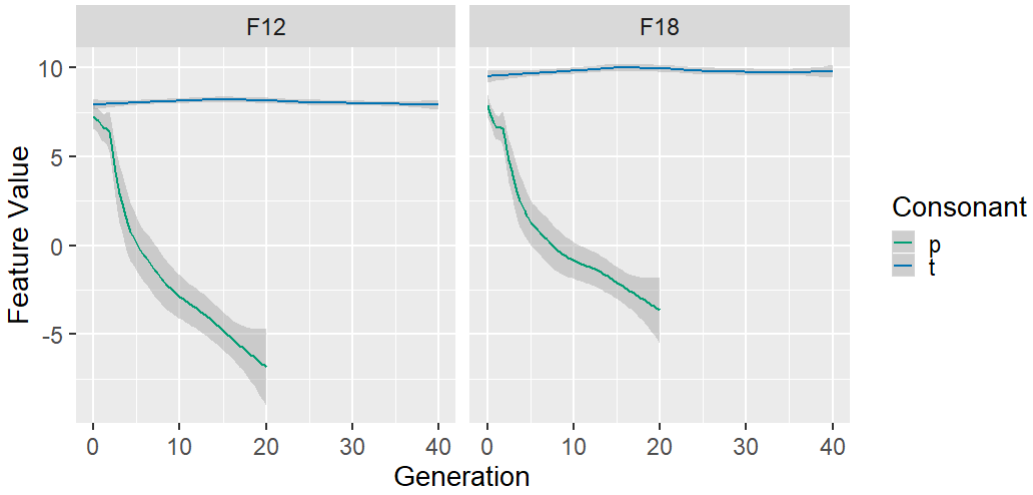

Figure F.6: Simulated featural results of /p/-/t/ before high front vowels (s=1, p=0)

/p/-/t/ BEFORE LOW BACK VOWELS

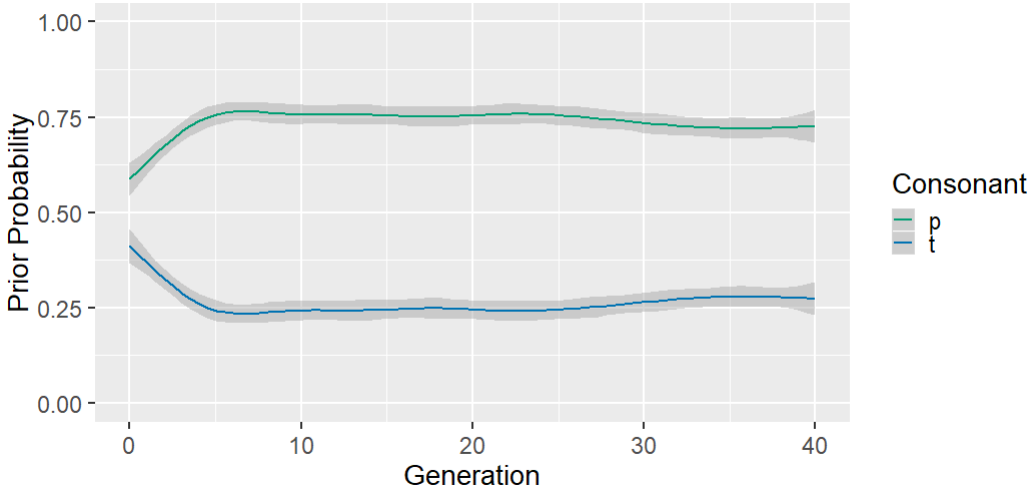

Figure F.7: Simulated prior results of /p/-/t/ before low back vowels (s=1, p=0)

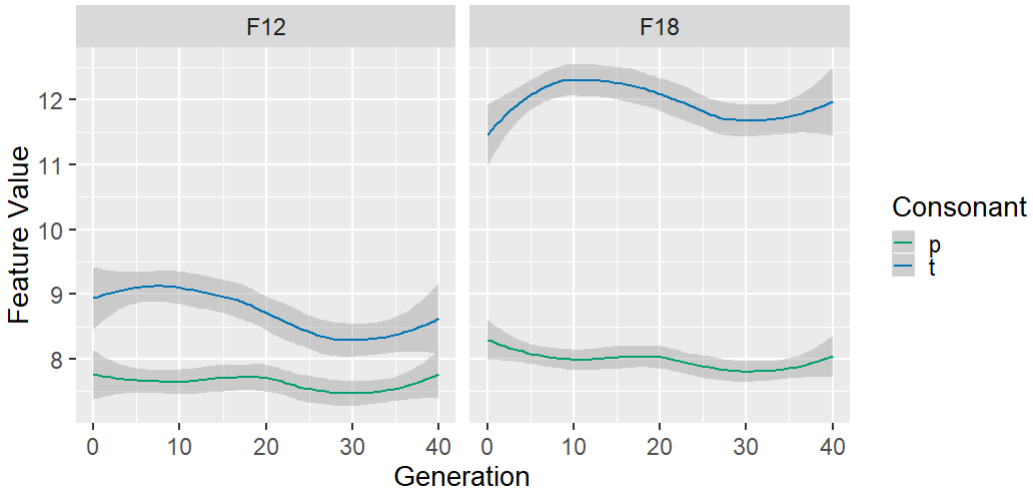

Figure F.8: Simulated featural results of /p/-/t/ before low back vowels (s=1, p=0)

F.3 s=1, p=1

/p/-/t/ BEFORE HIGH FRONT VOWELS

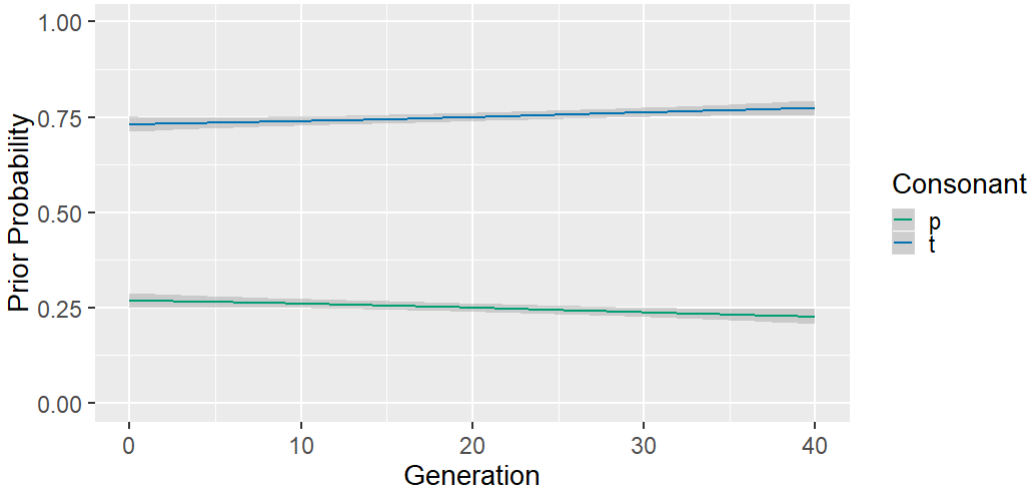

Figure F.9: Simulated prior results of /p/-/t/ before high front vowels (s=1, p=1)

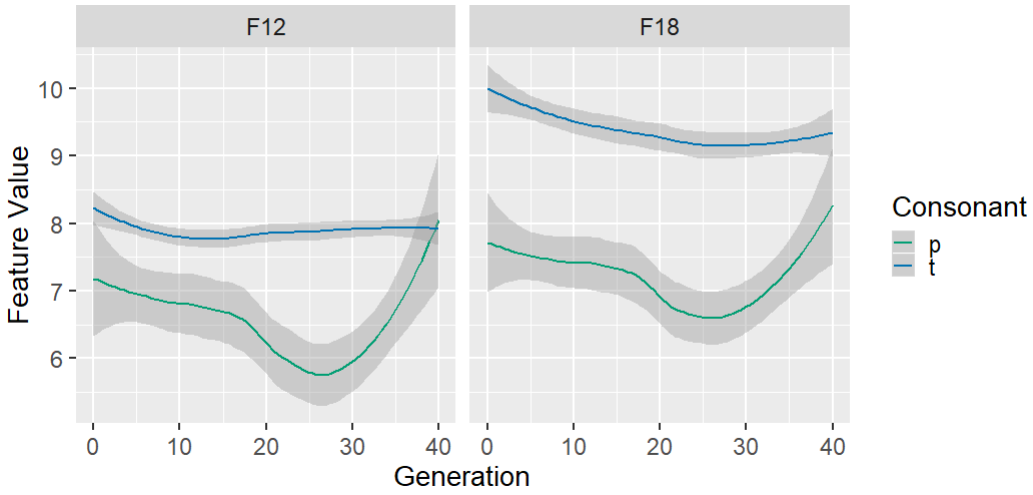

Figure F.10: Simulated featural results of /p/-/t/ before high front vowels (s=1, p=1)

/p/-/t/ BEFORE LOW BACK VOWELS

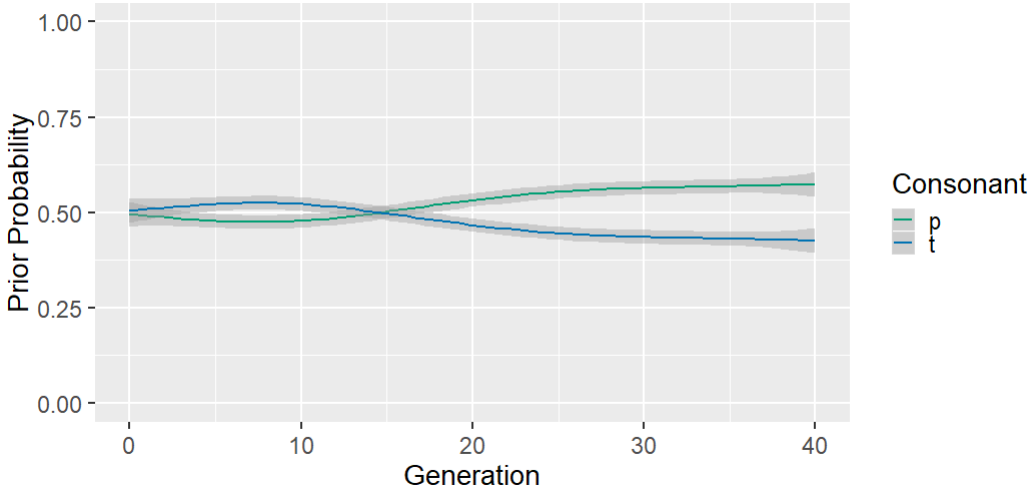

Figure F.11: Simulated prior results of /p/-/t/ before low back vowels (s=1, p=1)

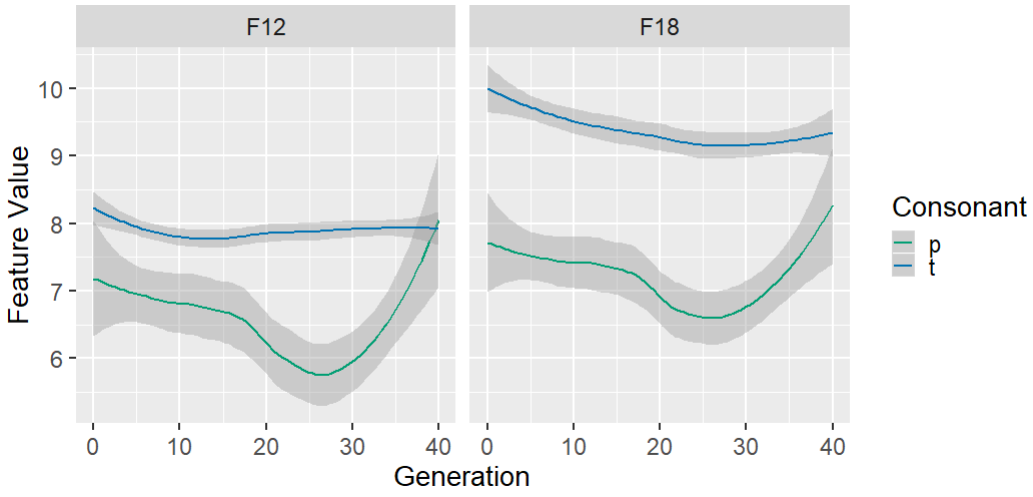

Figure F.12: Simulated featural results of /p/-/t/ before low back vowels (s=1, p=1)

REFERENCES

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *The Journal of the Acoustical Society of America*, *101*(2), 1078–1089.

Andersen, H. (1973). Abductive and deductive change. *Language*, *49*(4), 765–793.

Apley, D. W., & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, *85*(4), 785–821.

Best, C. T., Mathur, G., Miranda, K. A., & Lillo-Martin, D. (2010). Effects of sign language experience on categorical perception of dynamic ASL pseudosigns. *Attention, Perception, & Psychophysics*, *72*(3), 747–762.

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge, UK: Cambridge University Press.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S. (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, *120*(4), 1791–1794.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180.

Cafri, G., & Bailey, B. A. (2016). Understanding variable effects from black box prediction: Quantifying effects in tree ensembles using partial dependence. *Journal of Data Science*, *14*(1), 67–95.

Carney, P. J., & Moll, K. L. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica*, *23*(4), 193–202.

Chafe, W. L. (1964). Another look at Siouan and Iroquoian. *American Anthropologist*, *66*(4), 852–862.

Chang, S., Plauché, M., & Ohala, J. J. (2001). Markedness and consonant confusion asymmetries. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 79–101). Academic Press.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., & Wissing, D. (2018). Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics*, *66*, 185–216.

Cole, R. A., & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review*, *81*(4), 348.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, *24*(6), 597–606.

Cutler, A., Smits, R., & Cooper, N. (2005). Vowel perception: Effects of non-native language vs. non-native dialect. *Speech Communication*, *47*(1-2), 32–42.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678.

Delattre, P., & Freeman, D. C. (1968). A dialect study of American R's by X-ray motion picture. *Linguistics*, *6*(44), 29–68.

Desnitskaya, A. (1968). *The Albanian language and its dialects*. Leningrad: Nauka.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*(2), 109–122.

Flach, P. A. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In T. Fawcett & N. Mishra (Eds.), *Proceedings of the 20th international conference on machine learning* (pp. 194–201). Menlo Park, CA: AAAI.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115–123.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics*, *55*(6), 597–610.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1730–1741.

Fowler, C. A. (2007). Speech production. In M. G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 489–501). Oxford, UK: Oxford University Press.

Fowler, C. A., & Brancazio, L. (2000). Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation. *Language and Speech*, *43*(1), 1–41.

Fulop, S. A., & Scott, H. (2019). Vowel System Sandbox: Complex System Modelling of Language Change. *Journal of Open Research Software*, *7*(1).

Gardiner, S. C. (2008). *Old Church Slavonic: an elementary grammar*. Cambridge, UK: Cambridge University Press.

Garner, W., & Haun, F. (1978). Letter identification as a function of type of perceptual limitation and type of attribute. *Journal of Experimental Psychology: Human Perception and*

Performance, *4*(2), 199.

Garrett, A., & Johnson, K. (2013). Phonetic bias in sound change. In A. C. L. Yu (Ed.), *Origins of Sound Change: Approaches to Phonologization* (pp. 51–97). Oxford, UK: Oxford University Press.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*(5), 425–431.

Goldin-Meadow, S. (2005). *The Resilience of Language: What gesture creation in deaf children can tell us about how all children learn language*. New York/Hove: Psychology Press.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*(1), 43–53.

Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica*, *55*(1-2), 18–52.

Hall-Lew, L. (2010). Improved representation of variance in measures of vowel merger. In *Proceedings of Meetings on Acoustics 159* (Vol. 9, pp. 1–10).

Harrington, J., Kleber, F., Reubold, U., Schiel, F., & Stevens, M. (2018). Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in Cognitive Science*, *10*(4), 707–728.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature*, *408*(6815), 927–928.

Harrington, J., & Schiel, F. (2017). /u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change. *Language*, *93*(2), 414–445.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, *1*(1), 1–7.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484.

Hogg, R. M. (1979). Old English palatalization. *Transactions of the Philological Society*, *77*(1), 89–113.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-based methods". In *An Introduction to Statistical Learning: with Applications in R* (pp. 303–335). New York, NY: Springer New York.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, *85*(4), 1718–1725.

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

Kewley-Port, D., & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics*, *35*(4), 353–360.

Kinsler, L. E., Frey, A. R., Coppens, H., & Sanders, J. V. (1982). *Fundamentals of acoustics* (3rd ed.). New York, NY: John Wiley and Sons, Inc.

Kirby, J. (2010). *Cue selection and category restructuring in sound change* (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Kirby, J. (2013). The role of probabilistic enhancement in phonologization. In A. C. L. Yu (Ed.), *Origins of Sound Change: Approaches to Phonologization* (pp. 228–246). Oxford, UK:

Oxford University Press.
Kirby, J. (2014). Incipient tonogenesis in Phnom Penh Khmer: computational studies. *Laboratory Phonology*, *5*(1), 195–230.
Kirby, J., & Sonderegger, M. (2015). Bias and population structure in the actuation of sound change. *arXiv preprint arXiv:1507.04420*.
Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.
Kuang, J., & Cui, A. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics*, *71*, 194–214.
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.
Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, *127*(4), 2599–2610.
Li, F. K. (1977). A handbook of comparative Tai. *Oceanic Linguistics Special Publications*(15), i–389.
Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 497–516.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of perception of speech revisited. *Cognition*, *21*, 1–36.
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). Dordrecht, Netherlands: Kluwer Academic.
Lindblom, B., Guion, S., Hura, S., Moon, S.-J., & Willerman, R. (1995). Is sound change adaptive? *Rivista di linguistica*, *7*, 5–36.
Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), (pp. 103–189). New York, NY: John Wiley & Sons, Inc.
Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.
Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228.
Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics*, *34*(4), 338–348.
Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3.
Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.
Müller, D. (2010). Phonetic factors influencing /l/-rhoticisation in greek. In A. Botinis (Ed.), *Proceedings of the third ISCA Tutorial and Research Workshop on Experimental Linguistics.* Athens, Greece: University of Athens.
Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *98*(3), 1325–1347.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical psychology*, *34*(4), 393–418.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*(1), 94–140.

Ohala, J. J. (1978). Southern Bantu vs. the world: The case of palatalization of labials. In J. J. Jaeger et al. (Eds.), *Proceedings of the 4th annual meeting of the Berkeley Linguistics Society* (Vol. 4, pp. 370–386). Berkeley, CA: Berkeley Linguistics Society.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior - Chicago Linguistic Society* (pp. 178–203). Chicago, IL: Chicago Linguistics Society.

Ohala, J. J. (1997). Comparison of speech sounds: Distance vs. cost metrics. In S. Kiritani, H. Hirose, & H. Fujisaki (Eds.), *Speech production and language: In honor of Osamu Fujimura* (pp. 261–270). Berlin/New York: Mouton de Gruyter.

Ohala, J. J., & Lorentz, J. (1977). The story of [w]: an exercise in the phonetic explanation for sound patterns. In K. Whistler et al. (Eds.), *Proceedings of the 3rd annual meeting of the Berkeley Linguistics Society* (Vol. 3, pp. 577–599). Berkeley, CA: Berkeley Linguistics Society.

Ohde, R. N., & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *The Journal of the Acoustical Society of America*, *74*(3), 706–714.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, *121*(4), 2312–2326.

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, *124*(2), 1220–1233.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (Vol. 45, pp. 137–157). Amsterdam, Netherlands: John Benjamins Publishing.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & psychophysics*, *60*(6), 941–951.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*(3), 347–370.

Plauché, M. (2001). *Acoustic cues in the directionality of stop consonant confusions* (Unpublished doctoral dissertation). University of California, Berkeley, Berkeley, CA.

Plauché, M., Delogu, C., & Ohala, J. J. (1997). Asymmetries in consonant confusion. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *5th European Conference on Speech Communication and Technology.*

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, *22*(4), 463–475.

Rankin, R. L. (1981). *On Palatalization as a Phonetic Process* (Vol. 6; Working Paper). Lawrence, KS: University of Kansas.

Recasens, D., Pallarès, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, *102*(1), 544–561.

Repp, B. H., & Lin, H.-B. (1989). Acoustic properties and perception of stop consonant release transients. *The Journal of the Acoustical Society of America*, *85*(1), 379–396.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436.

Schatz, C. D. (1954). The role of context in the perception of stops. *Language*, 47–56.

Schleef, E., & Ramsammy, M. (2013). Labiodental fronting of /θ/ in London and Edinburgh: a cross-dialectal study. *English Language & Linguistics*, *17*(1), 25–54.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shiina, K. (1988). A fuzzy-set-theoretic feature model and its application to asymmetric similarity data analysis. *Japanese Psychological Research*, *30*(3), 95–104.

Sonderegger, M., & Yu, A. C. L. (2010). A rational account of perceptual compensation for coarticulation. In S. Ohlsson (Ed.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).

Sorensen, T., Skordilis, Z. I., Toutios, A., Kim, Y.-C., Zhu, Y., Kim, J., Lammert, A. C., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., & Narayanan, S. (2017). Database of Volumetric and Real-Time Vocal Tract MRI for Speech Science. In F. Lacerda, D. House, M. Heldner, & J. Gustafson (Eds.), *Proceedings of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017): Situated Interaction* (pp. 645–649). Baixas, France: ISCA - International Speech Communication Association.

Sóskuthy, M. (2013). *Phonetic biases and systemic effects in the actuation of sound change* (Unpublished doctoral dissertation). The University of Edinburgh, Edinburgh, UK.

Sóskuthy, M. (2015). Understanding change through stability: A computational study of sound change actuation. *Lingua*, *163*, 40–60.

Stevens, K. N. (1993). Models for the production and acoustics of stop consonants. *Speech Communication*, *13*(3-4), 367–375.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). Cambridge, MA: MIT press.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368.

Stuart-Smith, J., Timmins, C., & Tweedie, F. (2007). 'Talkin'Jockney'? Variation and change in Glaswegian accent. *Journal of Sociolinguistics*, *11*(2), 221–260.

Stungis, J. (1981). Identification and discrimination of handshape in American Sign Language. *Perception & Psychophysics*, *29*(3), 261–276.

Styler, W. (2015). *On the acoustical and perceptual features of vowel nasality* (Unpublished doctoral dissertation). University of Colorado, Boulder, Boulder, CO.

Takemoto, H., Honda, K., Masaki, S., Shimada, Y., & Fujimoto, I. (2006). Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *The Journal of the Acoustical Society of America*, *119*(2), 1037–1049.

Thomason, S. G. (1986). On changes from palatalized labials to apical affricates. *International Journal of American Linguistics*, *52*(2), 182–186.

Tiede, M., & Whalen, D. (2015). GetContours: An interactive tongue surface extraction tool. In *Proceedings of Ultrafest VII.* Hong Kong, China.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1–20.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

van Lieshout, P., Merrick, G., & Goldstein, L. (2008). An articulatory phonology perspective on rhotic articulation problems: A descriptive case study. *Asia Pacific Journal of Speech, Language and Hearing*, *11*(4), 283–303.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248–1266.

Weinreich, U., Labov, W., Herzog, M., Lehmann, W. P., & Malkiel, Y. (1968). Empirical Foundations for a Theory of Language Change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics.* Austin, TX: The University of Texas Printing Division.

Williams, D., & Julesz, B. (1992). Perceptual asymmetry in texture perception. *Proceedings of the National Academy of Sciences*, *89*(14), 6531–6534.

Winitz, H., Scheib, M. E., & Reeds, J. A. (1972). Identification of stops and vowels for the burst portion of/p, t, k/isolated from conversational speech. *The Journal of the Acoustical Society of America*, *51*(4B), 1309–1317.

Wood, S. A. J. (1982). *X-ray and model studies of vowel articulation* (Vol. 23; Working Paper). Lund Unversity Department of Linguistics.

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 34–57). Cambridge, UK: Cambridge University Press.

Yu, A. C. L. (2010). Perceptual compensation is correlated with individuals' "autistic" traits: implications for models of sound change. *PloS one*, *5*(8), e11950–e11950.

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English/r. *The Journal of the Acoustical Society of America*, *123*(6), 4466–4481.

Zhou, X., Espy-Wilson, C. Y., Tiede, M., & Boyce, S. (2007). An articulatory and acoustic study of" retroflex" and" bunched" American English rhotic sound based on MRI. In *Eighth*

Annual Conference of the International Speech Communication Association. Antwerp, Belgum: ISCA.