STRUCTURED VARIATION IN OBSTRUENT PRODUCTION AND PERCEPTION

by

Eleanor Chodroff

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

May, 2017

ABSTRACT

The acoustic-phonetic properties of speech sounds vary substantially across languages, and across talkers within a single language. The central thesis of this dissertation is that different aspects of phonetic realization are not independent but rather highly *structured* among phonetic categories across languages and talkers. Structured variation (defined as phonetic covariation among speech sounds) has important implications for the theory of phonetic realization as it applies to individual speakers and languages, and may also account for instances of generalized perceptual adaptation.

This dissertation makes several theoretical and empirical contributions to the field. First, I propose a principle of uniformity to account for phonetic covariation across talkers. This is formalized as three uniformity constraints that operate at the phonetics-phonology interface: pattern, target, and contrast uniformity. Pattern uniformity serves as a general constraint on phonetic implementation, requiring a similar structure of phonetic targets across talkers. Target and contrast uniformity directly influence the mapping from distinctive features to phonetic targets. Target uniformity requires similar (or identical) phonetic realization of a distinctive feature value, whereas contrast uniformity requires a comparable phonetic difference in sounds that contrast in a feature across talkers.

Second, I present several case studies of structured variation in stop consonant voice onset time (VOT) and sibilant fricative spectral shape to evaluate the predictions of uniformity. Variation and covariation in VOT were examined across American English talkers, children, and cross-linguistically. Structured variation was also investigated among sibilant fricatives across American English and Czech talkers. The structure that

emerged from our analyses provided strong evidence for target uniformity (and thus also for the more general notion of pattern uniformity).

Finally, listeners may exploit phonetic covariation to generalize talker-specific phonetic properties from one sound to another. Several experiments investigating generalized adaptation to talker VOT and fricative spectral shape were conducted. For fricatives, the phonetic covariation hypothesis was compared to a general auditory hypothesis based on spectral contrast and a cue-based normalization hypothesis. Generalization was observed for both VOT and fricative spectral shape in a manner consistent with phonetic covariation, but strong evidence was also found for spectral contrast, particularly in local contexts.

Thesis Committee

Jonathan Flombaum, Psychological and Brain Sciences

Jason Fischer, Psychological and Brain Sciences

Barbara Landau, Cognitive Science

Geraldine Legendre, Cognitive Science

Colin Wilson (primary advisor), Cognitive Science

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance and support from many people.

First, I would like to extend my heartfelt thanks to Colin Wilson for his generosity in knowledge, time, and patience. 'Generous' still feels like an understatement, and I can only consider myself very fortunate that I was the beneficiary. His mentorship and friendship were essential in guiding and sustaining me through this process. I cannot thank him enough for the great conversations, laughter, and of course, the accompanying food and drink.

I would also like to thank the members of my dissertation committee, Jason Fischer, Jon Flombaum, Barbara Landau, and Geraldine Legendre, as well as the members of my proposal committee, Paul Smolensky and Doug Whalen, for their insightful comments, questions, and suggestions. I am very humbled to have had such brilliant committees, and am grateful for all their feedback and support.

I also thank my mentors and collaborators in the Center for Language and Speech Processing, Jack Godfrey, Sanjeev Khudanpur, and Yenda Trmal. It has been a joy working with them, and I am grateful for the opportunity to learn from some of the best engineers in the field.

Thanks also goes to my friends in grad school for all the good times. I will look fondly on the many coffee / milkshake / lunch breaks and happy hours we've enjoyed over the years. I would especially like to thank my fellow CogSci grads for their substantial contributions to Formal Methods proofs. A special shout-out goes to Emily Atkinson who has not only been an amazing friend but also a great personal dissertation

mentor. I was very lucky to have had her help in everything from dissertation formatting to dissertation sympathy!

I would like to thank my research assistants from throughout the years. They were instrumental to data collection and processing, and I am very grateful for their help. I would especially like to thank Alessandra Golden for her commitment to the lab, her diligence and competence, and for being an amazing sounding board.

I also acknowledge my funding sources that allowed me to carry out this research: the DHS-USSS Forensic Services Division, Dolores Zohrab Liebmann Fellowship, and last but not least, the Distinguished Science of Learning Fellowship.

Finally, I thank my family for their unconditional love and support, for trusting me in my decisions to explore, and for instilling in me the importance of hard work, creativity, and curiosity. I would never have gotten here without them.

Portions of this dissertation appear in the following publication:

Chodroff, E. & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30-47. https://doi.org/10.1016/j.wocn.2017.01.001

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

The phonetic category [k^h], as found at the beginning of the word 'cat', involves several articulatory gestures and timing relationships: a constriction at the soft palate with the tongue body, vocal fold abduction (glottal spreading), and vocal fold frication following the release of the oral constriction (aspiration). The complex articulation gives rise to several defining acoustic properties in the temporal and spectral domains, such as voice onset time (VOT), or the duration from the stop release to the onset of voicing, and the spectral shape of the release. These respectively approximate the laryngeal and oral gestures.

The articulatory and acoustic instantiations of a speech sound derive from a chain of phonological and phonetic processes. The theory in this thesis assumes at least two levels of representation: a phonological representation and a phonetic representation, identified in Figure 1.1 as the phonological surface form and the phonetic targets (see also Keating, 1990; Cohn, 1993; and Fruehwald, 2017 for comparable characterizations of the phonetics-phonology interface). The phonological surface form is assumed to contain a sequence of segments that form a word or phrase and can be represented by distinctive features, metrical and prosodic structure, and other phonological specifications. The surface segment is then linked to a set of phonetic targets, defined in an articulatory and/or auditory space, which forms the abstract planning code for the physical instantiation of the phonetic category. The mapping from the phonological

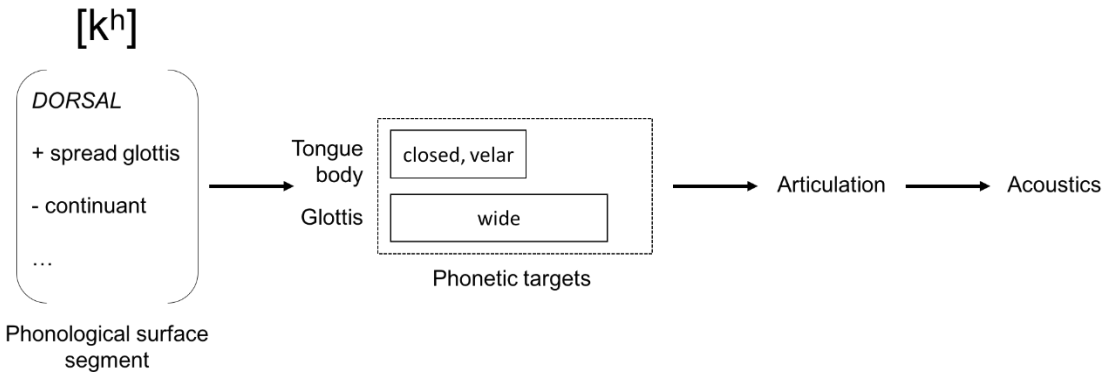
representation to the phonetic representation is referred to as the *phonetic implementation* or *realization*.[1]

Figure 1.1. Phonetics-phonology interface.

A specific example is shown for [k^h] in Figure 1.2 in which the phonological surface segment, composed of a set of distinctive feature values is realized as a set of phonetic targets. For [k^h], these may include laryngeal targets, such as the duration and magnitude of the glottal spreading gesture, as well as its relative timing to the constriction. In addition, there may also be specifications for targets related to the manner and place of articulation, such as the duration and location of the constriction, as well as specifications for the amount and rate of airflow. Note that while the targets are discussed in an articulatory space, auditory targets may also exist. The phonetic targets are then physically instantiated, giving rise to a physical articulation and resulting acoustic signal.

[1] Throughout the dissertation, the term '(phonological) surface segment' is used interchangeably with the terms 'allophone' and to a certain extent 'phonetic category.' It is, however, acknowledged that 'phonetic category' has weaker associations to any phonological structure than terms such as 'surface form' or 'allophone' and may also refer to the perceptual representation of a speech sound (e.g., Miller, 1994).

Figure 1.2. Simplified representation of the phonetics-phonology interface for [k^h].



Evidence from language- and talker-specific phonetics has revealed extensive variation in the articulatory and acoustic instantiations of phonological surface segments. For example, the average [k^h] VOT mean in American English is 80 ms (Lisker & Abramson, 1964), whereas in Navajo, the mean [k^h] VOT is 154 ms (Cho & Ladefoged, 1999). Analogous variation is found across talkers: within American English, the talker mean VOTs for [k^h] range over more than 40 ms (Theodore et al., 2009). As reviewed in the following sections, the variation in the phonetic realization of surface segments is considerable. The central thesis of this dissertation is that different aspects of phonetic realization are not independent but rather *covary* among segments across languages and talkers. This kind of structured or patterned variation has previously been investigated for vowels and, to a more limited extent, other classes of sound. Structured variation has important implications for (i) the theory of phonetic implementation and (ii) perceptual adaptation for human and automatic speech recognition (ASR) systems.

The fact that a single surface segment can be realized with a wide variety of phonetic targets indicates complexity and some degree of idiosyncrasy in phonetic implementation. The extent of variation in the phonetic implementation of individual speech sounds across languages is reviewed in section 1.1. Variation in phonetic

realization across talkers is briefly summarized in section 1.2. As the central empirical base of the dissertation is cross-talker variation, this will be further discussed in the relevant chapters. Section 1.3 provides an overview of *structured variation*, or dependencies in phonetic realization within and among speech sounds, and in section 1.4, I propose a principle of uniformity, formalized in two specific constraints that restrict variation in phonetic implementation, giving rise to patterns of covariation among phonological segments. The quantitative predictions of these constraints are outlined in section 1.5. Finally, in section 1.6, I consider the implications of structured variation for perceptual adaptation.

1.1 Cross-linguistic variation

The following section presents a review of some aspects of cross-linguistic variation in phonetic implementation. This review is by no means extensive, but rather serves to highlight a sampling of differences in the mapping of a common phonological segment to the corresponding phonetic targets and resulting physical instantiation. Cross-linguistic phonetic variation is not limited to any phonetic category or manner class, but instead appears as an inherent property of phonetic realization more generally.

1.1.1 Vowels

Numerous studies have identified cross-linguistic differences in the realization of vowel categories in their defining resonant frequencies or formants (e.g., F1, F2). Lindau & Wood (1977) found significant acoustic differences in several vowel categories common to both Yoruba and Edo. While both languages have the same seven vowel categories, the high, high-mid, and low-mid vowels are spaced differently within the two languages in the F1-F2 acoustic space. In Yoruba, the high and high-mid vowels are

closer together than those of Edo, and the high-mid and low-mid vowels are more dispersed than those of Edo. Similar cross-linguistic studies have also identified significant differences in the phonetic realization of vowel categories even when produced by a bilingual talker (Disner, 1983), and after correcting for talker differences (Chung et al., 2012).

1.1.2 Stops

An important phonetic property of a stop consonant is its voice onset time (VOT), the duration measured from the release of the stop to the onset of vocal fold vibration. VOT serves as a primary correlate of voice contrasts (e.g., [k^h] vs. [g]; Lisker & Abramson, 1964) and a secondary correlate of place contrasts (e.g., [k^h] vs. [p^h]; Peterson & Lehiste, 1960; Klatt, 1975). As mentioned at the beginning of the Introduction, the VOT of the same stop category varies considerably by language: Cho & Ladefoged (1999) reported a mean VOT of 154 ms for Navajo [k^h], a voiceless aspirated stop (on the basis of recordings described in McDonough & Ladefoged, 1993), but other languages in their survey have far lower means for the same phonetic category (e.g., 84 ms in Hupa) and lower values have been reported in several studies of American English (e.g., Lisker & Abramson, 1964: 80 ms; Klatt, 1975: 70 ms; Byrd, 1993: 52 ms). The cross-linguistic variation in VOT for [k^h] is paralleled in other aspirated stops such as [t^h] and [p^h], as well as in other stop voicing categories (Cho & Ladefoged, 1999).

1.1.3 Fricatives

The acoustic-phonetic characterization of fricative place of articulation is largely carried by its spectral shape, or distribution of energy across the frequency spectrum. Standard measures of the energy distribution include the energy-weighted mean

frequency, or spectral center of gravity (COG), the spectral peak of the distribution, among others. Cross-linguistically, the spectral realization of sibilant fricatives varies considerably. Heffernan (2004) identified significant differences in the mean COG of [s] between Canadian English and Japanese. Despite controlling for gender differences, Canadian English had a lower COG [s] than Japanese. A comparable finding was observed between American English and Japanese in Li et al. (2007): the COG above the F2 region of American English [s] was numerically lower than that of Japanese [s]. Qualitative assessments of variation have also been reported for COG and overall spectral shape across fricatives collected in a standardized manner from a wide range of languages (Nartey, 1982; Gordon et al., 2002).

Cross-linguistic variation has also been identified in the dynamics of fricative spectra, as well as the articulation. For example, the trajectory of spectral peak in English [s] was relatively flat compared to the trajectory of spectral peak in Japanese [s], which peaked sharply around the middle of the sibilant (Reidy, 2016). In articulation, Fuchs & Toda (2010) observed that German [s] was produced with a wider constriction width, and thus lower acoustic spectral peak, compared to English [s].

1.1.4 Nasals

Cross-linguistic variation also exists among nasal consonants, as demonstrated by both acoustic and perceptual studies. Harnsberger (2000) examined how listeners of a variety of languages (including Malayalam, Marathi, Punjabi, Tamil, Oriya, Bengali, and American English) identify nasal stops produced by Malayalam, Marathi, or Oriya speakers. While many of these languages have a common inventory of nasal phonetic categories, the instances of a given nasal as produced by speakers of one language are not

always reliably identified as that nasal category by listeners of another language. For example, Tamil speakers failed to categorize many of the Malayalam, Marathi, or Oriya [n]s as [n], despite also having a native phonetic category of [n]. This would suggest that the phonetic realization of [n] may differ substantially across these languages.

1.1.5 Liquids

In the articulation of a liquid consonant such as /l/ or /r/, there is typically an anterior and posterior gesture of the tongue; the presence and relative timing of these gestures can differ extensively across languages (Gick et al., 2006). On the basis of articulatory data from Western Canadian English /l/, Quebec French /l/, Serbo-Croatian (dark) /l/, Squamish Salish /l/, Beijing Mandarin /r/, Gick et al. (2006) found substantial variation in the form of the anterior and posterior tongue gestures, as well as their timing relationship. Among languages specifically with the alveolar lateral approximant [l] (Western Canadian English, Quebec French, and Squamish Salish), the anterior and posterior gestures were found to take several forms in prevocalic position: a tongue tip raising and fronting gesture, as well as a tongue dorsum backing gesture were present in both English and Salish [l], whereas a tongue tip raising and fronting gesture, but no tongue dorsum backing gesture were found in French [l].

1.1.6 Additional examples

The above sections provide an overview of documented instances of cross-linguistic variation in segmental realization that is certainly non-exhaustive. In addition to vowels and consonants, there have also been studies of cross-linguistic differences in the phonetic realization of lexical tone (e.g., Francis et al., 2008), speaking rate (Barik, 1977), and prosodic structure (e.g., Cho & Keating, 2001). These cases shed light on the

breadth of permissible variation not only in the realization of vowels and consonants, but in phonetic features more generally.

1.2 Cross-talker variation

The extent to which the phonetic implementation of a single category can differ ranges not only across languages, but also across talkers within a single language. As the content of the dissertation largely focuses on cross-talker variation, this section will provide a brief survey of previous findings for each manner class, with extended review and additional case studies for stop consonants and sibilant fricatives in Chapters 2 and 3.

Individual differences have been noted in several vowel and consonant categories. Within the F1-F2 vowel space, distinct vowel categories as produced by different talkers can overlap entirely within the same acoustic-phonetic range (e.g., Peterson & Barney, 1952; Hillenbrand et al., 1995). For example, instances of talker-specific means for the vowels [ɪ], [ɛ], and [æ] all coexist within the F1 range of 500 to 600 Hz and the F2 range of 2200 to 2600 Hz (Hillenbrand et al., 1995). Correspondingly, the same vowel category can take on a range of formant values, depending on the speaker. Among stop consonants, many studies have identified significant talker differences in the realization of stop VOT in American English, even after controlling for differences in speaking rate (e.g., Allen et al., 2003; Theodore et al., 2009). Among sibilant fricatives, Newman et al. (2001) found substantial differences in the means, variances, and distributional overlap of [s] and [ʃ] COGs. As in vowel production, some talkers' [s] COGs overlapped almost entirely with other talkers' [ʃ] COGs. Finally, talkers differ considerably in the articulation of [ɹ] in American English. Delattre & Freeman (1968) identified minimally six unique tongue shapes ranging from a highly retroflexed articulation to a 'bunched'

tongue articulation in which the tongue body is raised and retracted while the tongue tip is lowered (see also Hagiwara, 1995, Westbury et al., 1998).

1.3 Structure in phonetic variation

The previous sections review a sampling of observed variation in the phonetic realization of a single sound category across languages and talkers. This variation is further compounded by additional factors including but not limited to coarticulation with neighboring sounds, lexical factors, and speaking rate. Phonetic variation may be structured in meaningful ways such that variation may be represented by either a talker or a listener in a lower-dimensional space than is measured in the articulatory or acoustic space. The term 'structured variation' can, however, refer to several types of dependency. To understand the different types of structure, I will define the surface phonetic variation with a random phonetic variable μ_{ijk}^{c}, measured on the phonetic dimension *c*, whose realization minimally depends on the talker *i,* language or dialect *j,* and phonetic category (phonological surface form) *k* (Figure 1.3).

Figure 1.3. Characterization of a phonetic variable

$$\mu_{ijk}^{c}$$

c = phonetic dimension
i = language / dialect
j = talker
k = category

Phonetic variables could, for example, represent means of individual formants in e.g. Hertz or Mel, stop consonant VOT in milliseconds, or dynamic properties such as a change in spectral peak over the course of a fricative. Note that μ_{ijk}^{c} is intended as a mean value on a measured dimension (e.g., mean VOT for [k^h] as produced by speaker *j* of language *i*).

One way of defining structured variation has been to identify extra-linguistic factors, such as social and temporal variables, that may determine the realization of the phonetic variable (e.g., Labov, 1966; Foulkes et al., 2010; see also temporal dependencies: sound change, Sonderegger et al., 2017). The relevant social and temporal variables may further refine expectations about the realization of a single phonetic variable for a given talker or language. Socially and temporally structured variation plays an important role in determining the most likely realization of a set of phonetic variables, but this does not necessarily assume any inherent dependencies among the phonetic variables that comprise a talker-specific grammar.[2,3]

As with any set of random variables, the realization of one phonetic variable may be related to, or share a dependency with the realization of a second phonetic variable. Non-independence between variables can take on many mathematical forms; however, the present dissertation explores the extent to which phonetic variables are *linearly* related through linear correlations and types of linear regression. Other dependencies between phonetic variables could include co-occurrence, ordinal rankings, other non-linear relationships, among others.

Dependencies among phonetic variables within a grammar have previously been explored in a variety of ways. These include dependencies among multiple phonetic

[2] As an example, varieties of New York City English have both th-stopping and a low degree of rhoticity in postvocalic [r], in which the words 'that' may be produced as 'dat' and 'car' as 'cah' (Labov, 1966; Newlin-Lukowicz, 2013). This correlation within New York City, however, does not necessarily entail an inherent dependency between th-stopping and rhoticity: varieties of Irish English are also characterized by th-stopping, but a high degree of rhoticity in postvocalic [r] (Hickey, 1989; Hickey 2007a; Corrigan, 2010). This is also related to *bricolage*, in which speakers construct their social identity by 'picking and choosing' realizations of linguistic variables (e.g., Eckert, 2008).

[3] The notion of social coherence addresses dependencies between linguistic variables and sociolects, and may also account for covariation of linguistic variables within a sociolect. The relation between this idea and the ideas in the present dissertation will be considered in Chapter 5.

correlates for a single phonetic category (e.g., Shultz et al., 2012; Dmitrieva et al., 2015; Kirby & Ladd, 2015), dependencies among multiple category-internal time points along a single phonetic dimension (e.g., Sussman et al., 1991; Reidy, 2016), and dependencies among multiple phonetic categories along a single phonetic dimension (e.g., Joos, 1948; Nearey, 1978; Cho & Ladefoged, 1999; Theodore et al., 2009). In the following sections, I review the relevant literature for each of these types of dependencies, as they form an important foundation for understanding the ways in which variation can be structured among phonetic variables.

1.3.1 Dependencies among multiple phonetic correlates

Dependencies among multiple phonetic correlates have been examined in vowels and stop consonants. Within vowels, mean F2 and F3 have been shown to be positively correlated across talkers (Rose, 2010). Nearey (1989) also reported high pairwise correlations of talker-specific means among f0, F1, F2, and F3 in log frequency, ranging from $r = 0.82$ to $r = 0.87$. Relatedly, Assmann et al. (2008) reported a high correlation between the geometric mean F0 and geometric mean of F1, F2, and F3 across young talkers, ages 5 to 18, at $r = 0.80$. This relationship was somewhat weaker across female talkers ($r = 0.46$) than across male talkers ($r = 0.87$). These analyses, however, were performed over all vowels together, so it remains unclear which categories, if any, exhibit these dependencies most strongly. Vowel height (as measured by F1) and vowel duration are also known to covary across vowels in many languages (e.g., Lindblom, 1967; Maddieson, 1997); however, Toivonen et al. (2015) reported no such correlation within individual vowel categories across tokens.

Within stop consonants, the relationship between VOT and f0, both cues to the voicing contrast, has been extensively examined to determine whether the relationship enhances the phonological contrast (positive correlation) or preserves the contrast through compensation (negative correlation; e.g., Shultz et al., 2012; Dmitrieva et al., 2015; Kirby & Ladd, 2015, 2016; Clayards, 2018). While a few trends have been observed in this relationship, the correlations tend to vary by language and study.

Negative correlations have been observed between f0 and VOT for word-initial voiceless labials in American English (e.g., Dmitrieva et al., 2015) and Italian (Kirby & Ladd, 2015). However, this correlation was not significant in French (Kirby & Ladd, 2015) and a second study of American English (Clayards, 2018). For word-initial voiced labials, f0 and VOT were positively correlated in American English (Dmitrieva et al., 2015), Italian (Kirby & Ladd, 2015), and French (Kirby & Ladd, 2015), but this correlation did not reach significance in Spanish (Dmitrieva et al., 2015) and a second study of American English (Clayards, 2018). For intervocalic [p] and [b], no covariation was observed between f0 and VOT for either Italian or French, and the talker-specific correlations between f0 and VOT varied substantially (Kirby & Ladd, 2016). Clayards (2017) also examined token-by-token correlations between following vowel duration and f0, as well as following vowel duration and VOT, but found only weak relationships between these correlates.

Linear discriminant analyses have also been used to determine the relative weighting of f0 and VOT in the labial voice contrast. Shultz et al. (2012) found a moderate negative correlation between f0 and VOT coefficients across talkers of American English with $r = -0.42$, indicating that talkers may 'trade-off' in their use of

cues to create a contrast in the voice feature. However, using the same analysis, Clayards (2017) found the exact opposite pattern between f0 and VOT coefficients with $r = 0.41$. The explanation for this discrepancy remains unclear.

1.3.2 Dependencies among multiple category-internal time points

For a given phonetic category, there may also be correlations of a single phonetic correlate between two separate points in time. The trajectory of a spectral property has been shown to depend on the phonetic category, as well as the language. Sussman et al. (1991) introduced the notion of locus equations for stop place of articulation, in which the F2 trajectory following a stop consonant strongly depends on the place of articulation. In particular, the linear regression between the F2 measured directly following the stop consonant and the F2 measured at the vowel midpoint yields high accuracy in classifying stop place of articulation. Therefore, for the phonetic realization of a [b], for example, the F2 at the vowel midpoint will in part depend on the F2 directly following the stop.

Furthermore, dependencies among phonetic variables at different time points of a category have also been shown to vary across languages. For example, the peak frequencies over the course of a Japanese [s] follow a different trajectory from that of American English [s] (Reidy, 2016). Specifically, the slope between the peak frequency at the midpoint of the [s] and the peak frequency at the end of the [s] is steeper in Japanese than the corresponding slope in an American English [s].

1.3.3 Dependencies among multiple phonetic categories

Dependencies among multiple phonetic categories have long been observed for vowels: talkers have relatively congruent F1-F2 vowel spaces that can be mapped to one another by (log-)linear translations, suggesting a high degree of covariation among vowel

categories (e.g., Joos, 1948; Nearey, 1978; Nearey & Assmann, 2007). Consistent formant frequency ratios between vowel categories along spectral and temporal dimensions are also largely preserved across different speaking rates and styles (Smiljanic & Bradlow, 2008; DiCanio et al., 2015). Fruehwald (2017) also demonstrated that the phonetic properties of vowel categories tend to shift in parallel in diachronic sound change, indicating that the phonetic realization of one vowel category can be highly dependent on the phonetic realization of a second vowel category.

Dependencies among multiple phonetic categories along a single phonetic dimension have also been found for sibilant fricatives: while the distribution of one talker's [s] COG may overlap almost entirely with another talker's [ʃ] COG, each talker nonetheless maintains a systematically higher COG for [s] than for [ʃ] (Newman et al., 2001), and the differences among talker's fricative systems on the COG dimension have been modeled with a single linear offset (McMurray & Jongman, 2011).

Several studies have also reported dependencies among stop consonant VOT. Most notably, in many (if not all) languages that have both [p^h] and [k^h], the value for the aspirated labial stop is lower than that of the aspirated velar stop (e.g., Fischer-Jørgensen, 1954; Peterson & Lehiste, 1960; Lisker & Abramson, 1964; Cho & Ladefoged, 1999). This generalization has also been found to apply to the voiceless unaspirated stops across languages (e.g., Cho & Ladefoged, 1999).

Moreover, previous studies have identified a tight positive (linear) correlation between the mean VOTs of [p^h] and [k^h] across talkers of the same language (e.g., Zlatin, 1974; Koenig, 2000; Newman, 2003; Solé, 2007; Theodore et al., 2009). Previous research has observed that the VOT values of different stops covary across speakers in

laboratory speech (i.e., single words produced in isolation or in carrier phrases). In the earliest relevant study, Zlatin (1974) reported moderate correlations of talker-specific VOT means among voiceless stops (ranging from $r = 0.54$ to $r = 0.57$) and among voiced stops ($r = 0.46$ to $r = 0.54$). Correlations between stops of different voicing specifications and between stops differing in both place and voice were inconsistent in Zlatin's study, most failing to reach significance. Subsequent studies include Koenig (2000), who observed a significant correlation of median VOTs between word-initial [p^h] and [t^h] across adult and child talkers ($r = 0.78$), and Newman (2003), who found significant correlations among voiceless stops ($r = 0.88$ to $r = 0.96$) and among voiced stops ($r = 0.54$ to $r = 0.75$), but much weaker relations between stops differing in voice ($r = -0.06$ to $r = 0.37$) in CV syllable productions by adults.

In addition, Theodore et al. (2009) made the important observation that the difference in VOT means for [p^h] and [k^h] was relatively constant across talkers — a clear indicator of covariation between these two stops. Theodore et al. further established that the relationship between [p^h] and [k^h] remained even when the potentially confounding factor of utterance-level speaking rate was taken into account (using the method of Allen et al., 2003). Approximately constant differences in talker-specific mean VOT between [p^h], [t^h] and [k^h] were also observed in Scobbie (2006) for Shetland English, as well as for speakers of Southern British English and Catalan at varying speech rates (Solé & Estebas, 2000; see also Solé, 2007). Solé & Estebas (2000) found that the pattern in English holds most clearly for labial and velar stops, with the VOT of the coronal stop perhaps varying more idiosyncratically across talkers or rates. This is likely related to other findings that aspirated coronal stops do not consistently conform to the

generalization that VOT increases with more posterior place of articulation (e.g., Whalen et al., 2007).

1.4 Principle of uniformity

Structured variation has important implications for the theory of phonetic realization as it applies to individual speakers and languages, and may also account for instances of generalized perceptual adaptation. The previous section summarized observed dependencies among phonetic variables along a measurable physical dimension of speech, e.g., μ_{ijk}^{c}. However, any observed structure in the physical output of speech must be conditioned on structure in the phonetic inputs or targets (e.g., Keating, 1985: 126-127). Recall that for any given phonological surface segment, there may be several associated phonetic targets, that may be defined along articulatory and/or auditory dimensions. As shown in Figure 1.4, the phonetic target is characterized as a value along a phonetic target dimension *t* and is dependent on the language or dialect *i*, speaker *j,* and phonetic category (or phonological surface segment) *k*.

Figure 1.4. Characterization of a phonetic target

$$\tau_{ijk}^{t}$$

t = phonetic target dimension
i = language / dialect
j = talker
k = category

The phonetic variable μ_{ijk}^{c} should reflect τ_{ijk}^{t}, and ideally, the phonetic correlate *c* should be chosen such that it closely reflects the underlying target dimension *t*. In the present thesis, I assume that phonetic targets most directly correlate with articulations (but note that this may not always be the case). Identifying measures of the acoustic signal that most clearly reflect articulation is part of a larger enterprise within phonetics

(e.g., Koenig et al., 2013; Shadle et al., 2014), but for most traditional phonetic variables, the relationship between the acoustics and articulation is not one-to-one. Biomechanical and aerodynamic consequences of an assumed articulatory-defined phonetic target will also need to be considered when measuring an acoustic-phonetic variable as a proxy for articulation and abstract phonetic target.

While there are several important types of dependency among phonetic variables, the current thesis focuses primarily on linear dependencies among multiple phonetic categories along a single phonetic dimension. This covariation is accounted for with a principle of *uniformity* in the phonetic realization of phonological surface forms that constrains the phonetic targets τ_{ijk}^{t} within and across talkers. Strong covariation among phonetic categories suggests a principle or constraint of uniformity, in the sense of "uniform or parallel behavior of members of a class" (Keating, 2003), and would impose a common relational structure or pattern on the phonetic systems of languages and talkers. In this section, I propose and discuss three constraints of uniformity that place limitations on the mapping from allophones to phonetic targets that could potentially give rise to patterns of covariation.

Very generally, phonetic covariation among speech sounds would be observed provided all talkers converged on a highly similar *pattern* of phonetic targets across phonetic categories. This could be accounted for by the most general constraint to be discussed — *pattern uniformity*, which is defined as follows:

> **Pattern uniformity**: within a language i, for every speaker j, the difference[4] between phonetic targets τ_{ijk}^{t} for phonological surface segments k_1 and k_2 is uniform across talkers

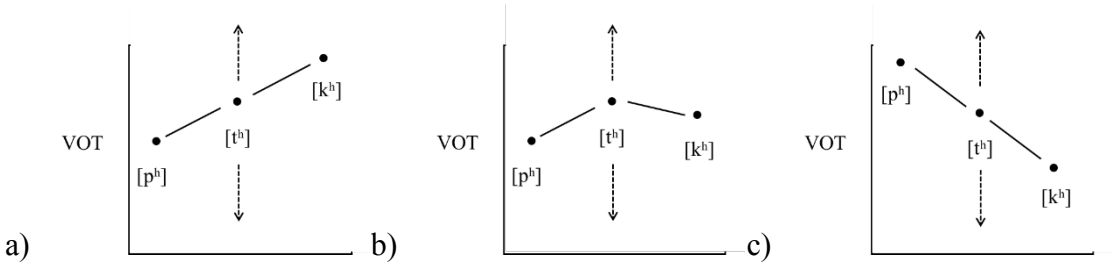
The notion of pattern uniformity resembles the emphasis in much of the vowel normalization literature on consistent *relationships* of formants between vowel types across speakers (e.g., Joos, 1948; Nearey, 1978). For example, Nearey (1978) outlined a constant ratio hypothesis for vowels in which the ratio of the log F1 and F2 values between vowel categories should be constant across talkers, and can equivalently be expressed as a sliding template of vowel categories in the log F1-F2 space (e.g., Nearey & Assmann, 2007). Pattern uniformity extends this principle beyond vowel formants to apply more generally to phonetic implementation.

As a very broad constraint, pattern uniformity does not require any meaningful relationship between the internal structure of the surface segment and the corresponding phonetic targets (e.g., Hale & Reiss, 2000). Specifically, it does not place any constraints on how similar or distinct the phonetic targets of differing speech sounds should be. As an example, Figure 1.5 displays three possible patterns of phonetic targets giving rise to the measurable variable of VOT across the aspirated stop consonants, [p^h t^h k^h]. For the sake of argument, the phonetic target could be the intended duration of the glottal spreading gesture, timed to the stop release. Pattern uniformity would allow any template of targets provided all talkers converged on that pattern: increasing duration with more posterior places of articulation (Figure 1.5a), a longer duration for [t^h] than for either [p^h] or [k^h] (Figure 1.5b), or even decreasing duration with more posterior places (Figure 1.5c). It would therefore be entirely possible for a language to exhibit the pattern as in

[4] Note that the difference between phonetic targets would be equivalent to the ratio between phonetic targets after transforming it to a log scale.

Figure 1.5c, with decreasing VOT with more posterior places of articulation; however, very few, if any languages exhibit such a pattern (Cho & Ladefoged, 1999). Moreover, pattern uniformity in its ideal form would apply to the entire phonetic system, requiring, for example, a constant difference between phonetic targets giving rise to the F2 of [a] and the VOT of [k^h] across talkers. While pattern uniformity may still play a role in restricting variation among phonetic targets across talkers, at least within constrained natural classes, it does not make any restrictions on the degree of *similarity* and/or *separation* between phonetic targets. Moreover, it makes no references to the internal structure of the phonological surface segment in constraining phonetic implementation.

Figure 1.5. Pattern uniformity in the phonetic targets corresponding to the aspirated stop categories. The arrows reflect permissible variation across talkers.

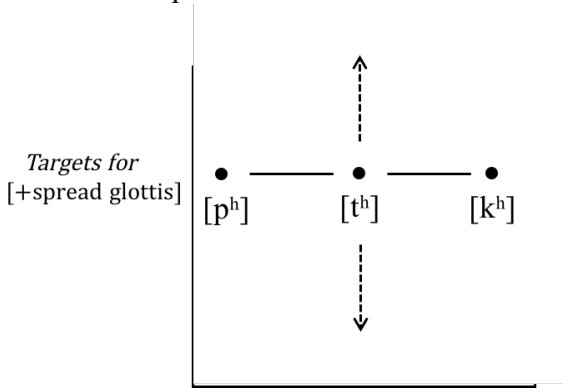


To account for consistent similarity and separation in phonetic realization, I am positing two constraints that are more specific instances of pattern uniformity that directly influence the mapping from distinctive features to phonetic targets: uniformity of target and uniformity of contrast. The first constraint of this set, uniformity of target can be defined as follows:

> **Target uniformity:** within a language *i* and speaker *j*, the phonetic targets τ_{ijk}^{t} corresponding to a phonological feature value [αF] are uniform for all phonological surface segments *k* that are specified [αF]

Target uniformity prioritizes *identity* between phonetic targets across a natural class of allophones with a shared distinctive feature value or a small set of related features.[5] Uniform phonetic targets across a natural class of segments would give rise to perfect covariation as talkers would differ only in a single dimension. For instance, the phonetic targets corresponding to the feature [+spread glottis] present in [p^h t^h k^h] may include the duration and magnitude of the glottal spreading gesture, the relative timing of the gesture to the oral constriction, as well as the amount and rate of airflow. As shown in Figure 1.6, this set of phonetic targets for each phonological surface segment would ideally be identical across each of these stops, resulting in strong covariation across talkers.

Figure 1.6. Target uniformity in the phonetic targets corresponding to [+spread glottis]. The arrows reflect permissible variation across talkers.



Uniformity in the mapping from phonetic features to phonetic targets has often been assumed in the literature, particularly in cases distinguishing 'automatic' from speaker-controlled phonetic actualizations of a phonetic target. The premise of several theoretical phonetics papers is that there should be a single phonetic target for each phonetic feature (e.g., Keating, 1984b; Ladefoged, 1988; Cho & Ladefoged, 1999).

[5] For instance, the phonetic implementation of [+spread glottis] could differ depending on whether the consonant is a stop ([-continuant]) or a fricative ([+continuant]). Part of the goal of the broader research program is to determine the appropriate set delineations that give rise to near-identical phonetic targets for a given feature value, but note that a primary objective of target uniformity is to minimize within-segment (vertical) context-sensitivity among features.

However, when automatic aerodynamic or physiological mechanisms cannot account for differences in the phonetic actualization of an assumed singular target, then segment-specific or context-sensitive phonetic targets have been introduced. The premise of a single phonetic target for each phonetic feature was likely inspired by the phonetic framework presented in Chomsky & Halle (1968), yet context independence (e.g., a one-to-one relationship between the feature value and its corresponding phonetic targets) was never explicitly required in that framework. In contrast, target uniformity overtly constrains the extent to which features can interact in determining the set of phonetic targets for a given phonological surface segment.

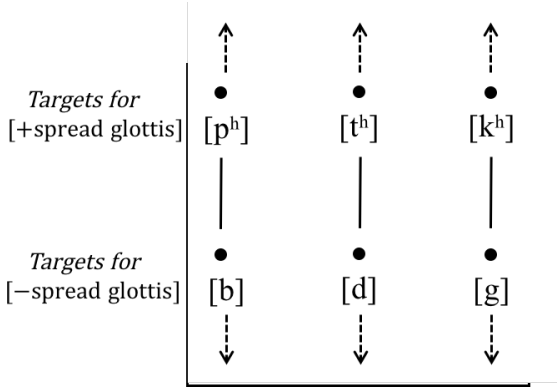
While target uniformity focuses on segments that *share* a distinctive feature value, the second constraint I am proposing, contrast uniformity, targets segments that contrast on a feature. This constraint has the following definition:

> **Contrast uniformity:** within a language *i*, for every speaker *j*, the difference between phonetic targets τ_{ijk}^{t} corresponding to contrasting values of a feature [F] is uniform for phonological surface segments k_1 and k_2 contrasting in feature [F]

Contrast uniformity, as its name implies, ensures that the distance between phonetic targets corresponding to a feature contrast is the same *across* talkers (as opposed to across contrasts) and is depicted in Figure 1.7. Note that target uniformity does not necessarily entail contrast uniformity: while speakers A and B could share the exact same phonetic target for [+F], speaker A could implement [-F] with uniformly low targets while speaker B could implement [-F] with uniformly high targets. The implementation of [-F] would be uniform *within a talker*, as required by target uniformity, yet the relationship between [+F] and [-F] *across talkers* would differ. Contrast uniformity

explicitly constrains the extent to which the phonetic targets corresponding to a feature differ across talkers.[6]

Figure 1.7. Contrast uniformity in the phonetic targets corresponding to [±spread glottis]. The arrows reflect permissible variation across talkers.



1.5 Statistical support for uniformity

With only the acoustic and articulatory data available for measurement, the underlying phonetic targets must be bridged to their observed physical instantiations. The following section presents several statistical methods that relate assumptions regarding the phonetic targets to observed acoustic measurements to assess the predictions and strength of the uniformity constraints. These methods include quantitative relationships between talker means of different phonetic categories (i.e., correlations and linear regression) and analysis of a linear mixed-effects model of phonetic variation.

[6] Contrast uniformity could have been defined as a uniform contrast between phonetic targets corresponding to contrasting values of a feature [F] within a talker. This definition would require that the difference between e.g., laryngeal targets for [pʰ] and [b] be the same as the difference between those for [kʰ] and [g]. Target uniformity, however, already entails this relationship: provided the targets for both [pʰ] and [kʰ], as well as [b] and [g] are uniform within a talker, then a uniform distance between segments contrasting in place would result. Note that this definition also does not predict correlations between [pʰ] and [b] across talkers, but would rather predict second-order correlations between the difference of [pʰ] and [b] and the difference of [kʰ] and [g].

1.5.1 Correlations

As discussed above, underlying structure in the phonetic targets can give rise to strong patterns of covariation along a single phonetic dimension across talkers. The strength to which two variables covary can be quantified with a correlation. The variables of interest here are the talker-specific means for phonetic category k_1 and phonetic category k_2.

Target uniformity would predict strong correlations of talker-specific means between segments that share a feature value for a phonetic correlate corresponding to the feature (e.g., [spread glottis] and VOT). A strong correlation, however, does not necessarily entail an underlying constraint of target uniformity, as the correlations must largely be due to identity in phonetic implementation. Without further analysis of the type of relationship between the category means, the correlations may reveal only the strength of pattern uniformity, or how interdependent the relationship is between phonetic category k_1 and phonetic category k_2.

Contrast uniformity would similarly predict strong correlations of talker-specific means between segments that contrast in a feature for a corresponding phonetic correlate. The presence of a correlation is indicative of an influence of contrast uniformity, as contrast uniformity requires only a constant difference or ratio of phonetic targets across talkers.

1.5.2 Simple linear regressions

While correlations reveal the strength of a linear relationship, simple linear regressions can uncover the type of relationship between two variables. In particular, the linear relationship could take the form of a constant difference between means ($y = \beta_0 +$

x), a constant ratio between means ($y = \beta_1 \cdot x$), or a combination of the two ($y = \beta_0 + \beta_1 \cdot x$). A simple linear regression estimates the best estimates of β_0 and β_1 to account for relationship between x and y, which correspond here to the talker-specific means for phonetic category k_1 and phonetic category k_2.

Mirroring the correlation coefficient, the extent to which two variables are linearly related through the linear regression can be assessed with the coefficient of determination, or R^2 value, which is simply the square of the correlation coefficient. Therefore, a high R^2 value indicates a systematic linear relationship between k_1 and k_2, and depending on the pair of categories examined, can indicate an influence of either target or contrast uniformity. The simple linear regression has particular value in assessing target uniformity, which in its ideal form would have an intercept (β_0) of 0 and a slope (β_1) of 1 in relating the phonetic targets of two categories with a shared feature value ($k_2 = k_1$). As only an acoustic or articulatory correlate of the target can be examined, the simple linear regression form could still provide evidence of target uniformity if any deviation from the expected linear form could plausibly be accounted for by an automatic biomechanical consequence of a uniform target.

1.5.3 Linear mixed-effects analysis

While the correlations and simple linear regressions shed light on the influence of each type of uniformity constraint, the analysis of target uniformity can still be strengthened. For target uniformity, the goal is to understand whether the phonetic target of differing phonetic categories can be derived primarily from a single phonological feature. To assess this, a linear mixed-effects regression analysis can be used to relate the

set of phonological features to variation in the phonetic target, as measured by a phonetic correlate.

A linear mixed-effects regression model is an extension of the simple linear regression presented above and assumes a linear relationship between a single predicted variable and a set of predictor variables separated into a fixed-effects component and a random-effects component. The mixed-effects model is also known as a multilevel or hierarchical linear model, as the random effects constitute a separate level to the model. The definitions of a fixed effect and a random effect vary depending on the source (e.g., Gelman & Hill, 2007), but very generally, fixed effects correspond to population variables of interest, whereas the random effects correspond to effects due to the experimental sample, or the underlying population. The best estimate of the predicted variable can be found through the linear combination of the fixed and random components.

The phonetics-phonology interface for an individual talker can be represented with the fixed effects component of the model only. For demonstration, the phonetic correlate will be the mid-frequency spectral peak ($Freq_M$), that has been shown to correspond to the constriction location of sibilant fricatives [s z ʃ ʒ] (further discussion and analysis presented in Chapter 3). The phonological surface segments for the class of sibilants can be decomposed into a place of articulation feature and a voice feature. If constriction location is indeed determined by the phonological feature for place of articulation in sibilants, [anterior], then other features present in the featural description of sibilants, i.e., [voice], should have no influence on the implementation of the constriction location. Assuming [anterior] and [voice] are binary-valued features with

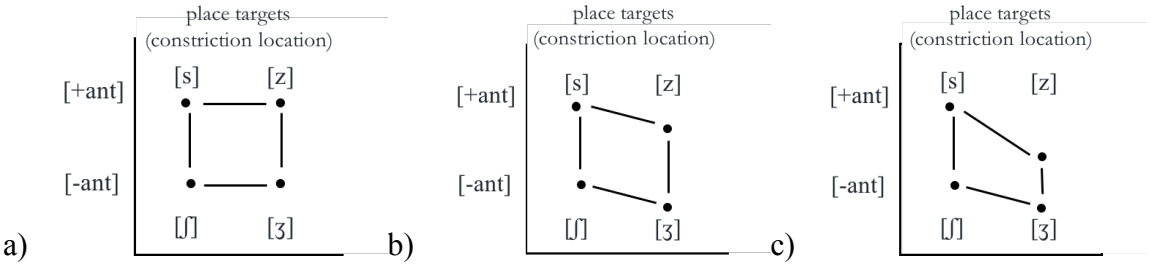
weights +1 and -1 corresponding to the feature specification of the segment, then a principle of target uniformity would predict that the beta weight of [voice] and the interaction between [anterior] and [voice] should be quite low, and substantially smaller than the beta weight for [anterior]. Essentially, the realization of $Freq_M$ should primarily be accounted for by the [anterior] feature with minimal influence from non-place features. This model and its predictions are presented below, and a graphical depiction is shown in Figure 1.8.

$$Freq_M \sim \beta_0 + \beta_1 * anterior + \beta_2 * voice + \beta_3 * anterior * voice$$

Predictions:
$|\beta_1| > |\beta_2|, |\beta_3|$

where anterior and voice are binary-valued features with weights +1 and -1 corresponding to the feature specification and each β is a fixed-effects weight

Figure 1.8. a) $Freq_M$ primarily reflects constriction location: uniformity in phonetic targets for the shared feature value of anteriority among coronal sibilant fricatives. b) Small, but uniform contribution of each value of [voice] in addition to the uniform contribution of each value of [anterior]. c) $Freq_M$ with a primary main effect of anteriority, secondary main effect of voice, and relatively weak interactions (context-sensitive or 'segment-specific' effects)



The model of the individual grammar can be extended to model phonetic variation in the larger population by adding a random-effects component representing talker variation. The fixed-effects component can now be interpreted as the population-level effects of the phonological features, and the random talker component can be interpreted

as the talker-specific deviations from the population model. Each z is drawn from a normal distribution centered at 0, with standard deviation σ_z.

$$Freq_M \sim \beta_0 + \beta_1 * anterior + \beta_2 * voice + \beta_3 * anterior * voice + z_0 + z_1 * anterior + z_2 * voice + z_3 * anterior * voice$$
$$\equiv$$
$$Freq_M \sim (\beta_0 + z_0) + (\beta_1 + z_1) * anterior + (\beta_2 + z_2) * voice + (\beta_3 + z_3) * anterior * voice$$
$$z \sim \mathcal{N}(0, \sigma_z)$$

Predictions:

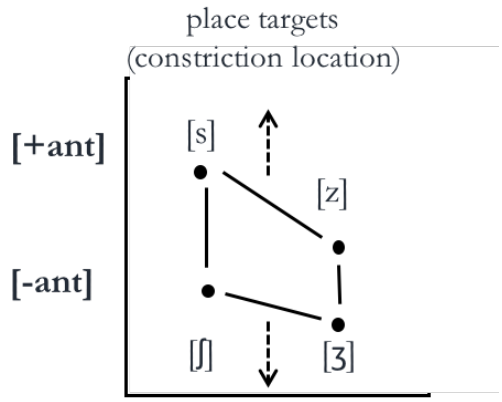
$|\beta_1| > |\beta_2|, |\beta_3|$

$\sigma_{z0} > \sigma_{z1}, \sigma_{z2}, \sigma_{z3}$

where anterior and voice are binary-valued features with weights +1 and -1 corresponding to the feature specification, each β is a population weight, and each z is a talker-specific 'random' effect.

If all talkers share similar phonetic grammars for phonetic implementation, then the largest source of variation across talkers should be in the intercept, z_0. This hypothesis corresponds to the constraint of pattern uniformity: regardless of how phonetic targets may be specified, the greatest variation across talkers should be in the exact positioning of the 'template' of targets in acoustic space (see Figure 1.9). The talker-specific intercept, z_0, dictates the position of this template along the dimension defined by $Freq_M$. In addition, contrast uniformity predicts minimal variation across talkers in the slope for place of articulation, which would indicate differences in the magnitude of separation between [+anterior] and [-anterior] realizations. The predictions of target uniformity parallel predictions of pattern uniformity in that the greatest variation should be in the intercept with minimal variation across talkers in the slopes for voice and the interaction between place and voice. If the effect of uniform phonetic targets from a secondary feature is minimal (e.g., voice on $Freq_M$), then talker variation is also expected to be minimal.

Figure 1.9. Pattern uniformity in the phonetic targets for sibilant fricatives.



1.6 Uniformity in adaptation

Listeners could employ prior knowledge of this structure when adapting to a novel talker, as experience with the phonetic realization of one sound could provide valuable information about how the same talker would realize other related sounds. The expectation that talkers should vary more in the intercept compared to segment-specific effects should constrain adaptation, facilitating generalization of talker-specific properties across phonetic categories. From the most general perspective, patterns of covariation indicate that talker-specific phonetic systems — which specify means and other parameters on many dimensions for each category — can be accurately represented in a space of relatively low dimensionality.

1.6.1 Patterned variation in cognitive models of adaptation

Uniformity in the pattern of acoustic-phonetic targets has been an underlying assumption in many proposed cognitive models of speaker normalization. Most prominently, vowel normalization procedures that incorporate information from across multiple vowels (‘extrinsic normalization’) rely on the assumption that all talkers have the same pattern of formants across all vowel categories. For example, the sliding

template model of vowel normalization (Nearey, 1978; Nearey & Assmann, 2007) directly instantiates pattern uniformity and assumes that an estimated talker-specific offset can be applied uniformly to all vowels. The model derives the talker-specific mean for each formant in log space from instances of all vowel categories and then applies mean subtraction to each category to standardize the vowel space. Nearey (1997) made the claim that much of talker adaptation is merely pattern recognition — that is, identifying the talker-specific offset from a normalized structure of speech sounds. Extrinsic vowel normalization techniques using *z*-scoring (Lobanov, 1971), range normalization (Gerstman, 1968), and vocal tract scaling (Nordström, 1976) also assume pattern uniformity among vowels. By using the entire vowel space for formant normalization, the underlying assumption is that talkers differ only in the absolute realization of formant values, but the pattern of vowel targets is otherwise constant.

Similar 'extrinsic' techniques have also been assumed in speaker normalization for acoustic-phonetic cues of fricatives. In the implementation of the C-CuRE model, the acoustic-phonetic properties of fricatives are normalized by subtracting the talker-specific mean taken over all fricatives (McMurray & Jongman, 2011). An underlying assumption of this model is that the fricative pattern for each acoustic-phonetic property is *uniform* across talkers, and that fricatives covary perfectly across talkers along each phonetic dimension.

Nielsen and Wilson (2008) proposed a hierarchical Bayesian model of adaptation with a multi-level representation of VOT grammars to account for perceptual generalization of talker-specific VOT from [p^h] to [k^h]. Critical in this model is an explicit representation of the systematic difference in VOT between [p^h] and [k^h]. The

VOT realizations from an individual are derived from a speaker-specific distribution of VOT generated by a population (or language-specific) distribution of VOT, as well as a superpopulation (universal) distribution of VOT. The VOT place-offset parameter is governed by the same hierarchy, such that talkers can vary in the overall difference between [p^h] and [k^h]. Talker adaptation results from the estimate of the novel speaker's distribution of VOT for the exposed category [p^h]; generalization thus arises as a direct result of the VOT place-offset parameter.

However, the normalization techniques discussed above are all offline methods of adaptation, in that they assume the listener has access to samples from all relevant segments (e.g., all vowels, fricatives, or stops). The models discussed for vowels and fricatives also assume that the acoustic-phonetic cue determines the grouping of segments. This may work out fine for these examples, particularly if done in an offline manner, but as will be discussed in Chapter 4, for online estimates of fricative spectral properties, the realization of [s] may be much more informative about the realization of [z] than for the realization of [v]. Mean subtraction, where the only estimate of the mean comes from [s], may result in unlikely estimates for featurally-distant fricatives.

1.6.2 Speaker adaptation in automatic speech recognition

Estimation of cross-category covariation and linear relations has proved fruitful in both off- and on-line speaker adaptation techniques in automatic speech recognition (e.g., Furui, 1980; Lasry & Stern, 1984; Leggetter & Woodland, 1994; Zavaliagkos et al., 1995). The motivating assumption behind these techniques is very much in line with the current proposal: the geometry of acoustic features for a given speech category ('phone') should be relatively constant across speakers. These relationships do not need to be re-

learned for each new speaker to the system. The two overarching techniques based on systematic relations in the acoustics are maximum a posteriori (MAP) methods and maximum likelihood linear regression (MLLR). MAP methods utilize a covariance matrix trained on a speaker-dependent system to constrain the relations between categories when encountering a new speaker (Cole et al., 1983; Lasry & Stern, 1984). MLLR methods instead translate the parameters of a base model for any speaker from a set of linear regressions also trained on a speaker-dependent system (Furui, 1980; Cox, 1995). In both cases, the relationships are generally learned in a speaker-dependent system and then applied to constrain or transform the models learned for a new speaker in a speaker-independent system.[7] As observation data from a new speaker is encountered, the system can simultaneously update the parameters of all categories with very minimal exposure (Furui, 1980; Zavaliagkos, et al., 1995).

These methods have been employed in Hidden Markov Model (HMM) systems with each state modeled as a Gaussian Mixture Model (GMM) of acoustic features. A 'phone', or sub-lexical unit, is typically comprised of three HMM states, which roughly correspond to the beginning, middle, and end of the phone. Both MAP and MLLR can be applied over all mixtures and states, but generating the covariance or linear relations for each mixture across each state can be unwieldy. Besides the computational cost, training data may be limited for certain states and speakers, in which case the estimates will not be robust. To reduce the computational cost and circumvent limitations of the data, MAP and MLLR techniques generally group states together into classes. In MAP techniques, the covariance matrix can be derived across phones, such that individual states within a

[7] Instead of transforming the model directly, many instances of MLLR, called constrained or feature-based MLLR (fMLLR) simply transform the observation data into the model space (Gales, 1998).

phone are modified to the same degree (e.g., Lasry & Stern, 1984; Zavaliagkos et al., 1995). In MLLR, the states are hierarchically clustered into a regression class tree, created through guided phonetic decisions (natural classes) or data-driven clustering, and an affine linear transformation is derived. For each new speaker to the system, the linear transformations are applied in a descending order until the data becomes insufficient (Leggetter & Woodland, 1995).

Acoustic covariation across talkers has been well developed for rapid speaker adaptation in automatic speech recognition. The use of phone sets and even phonetic natural classes has further advanced these techniques in minimizing the computational cost, while also pruning unreliable relations. However, the extension of these methods to cognitive models of talker adaptation is largely unexplored. As discussed in the following section, a few cognitive models capitalize on similar systematic relations across categories, but do not extend predictions of the model beyond a class of speech sounds or fail to relate the model to online adaptation.

1.6.3 Perceptual evidence for knowledge of structured relations

Evidence from studies of perceptual generalization suggest that listeners may indeed have prior knowledge of structured variation across speech sounds. Previous studies have demonstrated that listeners generalize talker-specific VOT across place of articulation (Eimas & Corbit, 1973; Theodore & Miller, 2010; Nielsen, 2011; cf. Clarke & Luce, 2005) and actively adjust the talker-specific vowel space after exposure to manipulated vowel formants (e.g., Ladefoged & Broadbent, 1957; Maye et al., 2008). The fact that listeners extrapolate talker-specific characteristics across speech sounds

demonstrates that listeners represent systematic relationships between these speech sounds, many of which reflect natural class structure.

However, the relative contributions of phonetic and general auditory mechanisms will need to be investigated, particularly for generalization of spectral properties across speech sounds. Comparable generalized adaptation effects can also be achieved with non-linguistic pre-cursors such as tones, and attributed to a general spectral contrast effect, which has been implicated as a low-level general auditory mechanism for adaptation (e.g., Mann, 1980; Lotto & Kluender, 1998; Holt, 2005; Laing et al., 2012).

The relevant generalization experiments and previously proposed explanations will be discussed in further detail in Chapters 2 and 4.

1.7 Outline

To evaluate the predictions of uniformity, I present several case studies of patterned variation in stop consonant VOT and sibilant fricative spectral shape (specifically $Freq_M$). Additionally, the predictions of phonetic covariation are analyzed in in several experiments investigating generalized adaptation to talker-specific VOT and fricative spectral shape. In Chapter 2, variation and covariation in VOT were examined across American English talkers in isolated speech and a large corpus of connected speech, as well as across children and cross-linguistically in a meta-analysis of previously reported language-specific VOT means. Furthermore, I examine perceptual knowledge of VOT covariation in a study of generalized adaptation. Chapter 3 extends the analysis of uniformity to sibilant fricatives across talkers of American English in multiple speech styles and Czech spontaneous speech. Chapter 4 examines listener knowledge of phonetic covariation in perceptual generalization of fricative spectral properties. For fricatives, the

phonetic covariation hypothesis was compared to a general auditory hypothesis based on spectral contrast and a cue-based normalization hypothesis. In the Conclusion, Chapter 5, I discuss the relation of uniformity to other well-known constraints and principles of phonetic realization, the research implications, limitations, and future research directions.

2 Chapter 2

2.1 Introduction

Considerable variability exists in the realization of stop consonant voice onset time (VOT) across languages and talkers within an individual language. As discussed in Chapter 1, this variability may be constrained, such that the phonetic implementation of one phonological segment is not *independent* of the implementation of a second phonological segment. The present chapter investigates the extent to which target and contrast uniformity influence the mapping from the laryngeal feature to laryngeal phonetic targets, as approximated by VOT in word-initial stop segments. This was evaluated across adult speakers of American English in isolated and connected speech, across child speakers of American English, and across languages in a meta-analysis of VOT. In addition, the implications of VOT covariation among stop consonants was investigated in generalized perceptual adaptation to talker-specific VOT. The following sections of the chapter introduction review the predictions of uniformity for VOT and its relation to underlying laryngeal distinctive features (section 2.1.1), and potential sources of VOT variation beyond the talker and language, which would need to be controlled for in order to isolate the talker- or language-specific contribution to VOT (section 2.1.2).

2.1.1 Uniformity in VOT

As an indicator of laryngeal voicing, VOT most directly corresponds to the laryngeal feature of the stop consonant. Several laryngeal features have been posited to account for differences in stop consonant contrasts and laryngeal settings cross-linguistically, including [voice], [spread glottis], [constricted glottis], [tense], and [lax], among others. The exact laryngeal feature employed in a language, however, should not

necessarily change the predictions of uniformity with respect to VOT. Target uniformity makes the prediction that the underlying phonetic targets for VOT should be uniform for all segments that share the corresponding phonological feature value [αF], whether that correspond to [+voice], [-voice], [+spread glottis], or any other laryngeal feature. Relatedly, for languages with a voicing contrast, contrast uniformity ensures that the differences between phonetic targets giving rise to VOT are uniform across segments that contrast in a laryngeal feature.

The phonetic realization of laryngeal features as measured by VOT can be roughly divided into three categories, defined by the relative timing and duration of the laryngeal gesture: negative VOT, short-lag VOT, and long-lag VOT. Negative VOT indicates that the onset of vocal fold vibration begins prior to the stop release; short-lag VOT indicates that the onset of vocal fold vibration begins shortly after the stop release and generally reflects voiceless unaspirated stops; long-lag VOT indicates that the onset of vocal fold vibration begins after the stop release, and there is a period of aspiration or lengthened frication which delays the onset of vocal fold vibration. Long-lag VOT corresponds most frequently to voiceless aspirated stops.

VOT presents an interesting case for target uniformity, as there are well-known differences in VOT across place of articulation. As discussed in the Introduction, the VOT of labial stops tends to be lower than that of their dorsal stop counterparts, a relation that holds for both negative and positive VOT values. The observed measurement, however, is not the same as the underlying phonetic target, and there are in fact several biomechanical, aerodynamic, and timing mechanisms that could give rise to differences in VOT across place of articulation even if the phonetic targets corresponding to a given

laryngeal feature value are uniform. These mechanisms are briefly summarized below, but in-depth explanations of the place differences in short- and long-lag stops can also be found in Maddieson (1997a) and Cho & Ladefoged (1999).

For short-lag stops, potential aerodynamic and physiological sources for place differences in VOT include the volume of the cavity both behind and in front of the constriction location (Hardcastle, 1973; Maddieson, 1997a), the movement of the articulators (Hardcastle, 1973; Kuehn & Moll, 1976; Maddieson, 1997a), and the extent of articulatory contact area (Stevens, 1998). These explanations critically assume that differences in VOT only arise from the interaction between biomechanical properties and an otherwise *uniform* laryngeal setting. For example, voicing can begin only when the pressure in the supraglottal cavity (the portion of the vocal tract above the glottis) is lower than the pressure in the subglottal cavity. A more posterior constriction location has a higher supraglottal pressure given the lower volume behind the constriction. The pressure behind the dorsal constriction takes longer to fall than the relatively lower pressure in the larger volume behind the labial and coronal constrictions. This would lead to a longer delay between the release and voicing onset for the dorsal than the labial or coronal stops.

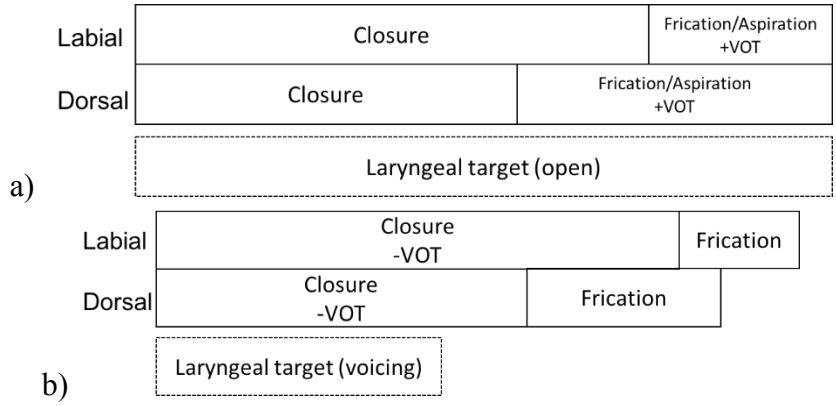
The volume in front of the constriction location will also contribute to the delay in achieving the appropriate transglottal pressure: the greater mass of air anterior to a dorsal constriction may result in greater obstruction of airflow than the air mass in front of the labial constriction (i.e., the ambient air). In addition, the tongue dorsum moves more slowly, and has a greater contact area, than the lips or tongue tip. These properties could

result in a longer period of frication subsequent to release of the stop constriction and thus a longer VOT for dorsals compared to labials or coronals.

For negative VOT, voicing tends to be sustained longer for labials than for dorsals. This observed difference could be accounted for by a uniform laryngeal target, with the actual duration of voicing modulated by independent effects of the physical constriction location. The smaller and less compliant surface area posterior to the dorsal stops relative to the labial stops could shorten the interval during which vocal fold vibration can occur (Ohala & Riordan, 1979).

An alternative explanation for place differences in VOT depends in large part on speech timing, in addition to the aerodynamic and biomechanical properties of articulation. In particular, the glottal opening gesture may be timed relative to the stop closure: as labials have a greater volume behind the point of constriction than dorsals, the closure duration is longer to build up sufficient supraglottal pressure to delay voicing. Assuming a fixed glottal opening gesture timed relative to the closure, the place differences in VOT duration would be inversely related to the place differences in closure duration. Provided the laryngeal gesture, whether glottal spreading or vocal fold vibration, is of a uniform duration and magnitude and timed relative to the beginning of the stop closure, then this independent difference plausibly account for the place effects on with negative, short-lag, and even long-lag VOT (Figure 2.1).

Figure 2.1. Diagram of place differences for the labial and dorsal aspirated stops given a) a uniform glottal spreading gesture and b) a uniform phonetic voicing target both timed relative to the onset of constriction. Figure adapted from Maddieson, 1997a, p. 622.



2.1.2 Sources of VOT variation

Cross-linguistically, phonologically voiceless stops have longer VOT than phonologically voiced stops in word-initial position (e.g., Lisker & Abramson, 1964).[8] Differences in VOT means across place of articulation have been extensively documented in the literature for a variety of languages. As discussed in the Introduction and in the preceding section, for voiceless unaspirated stops there is a general increase in VOT with more posterior places of articulation (Cho & Ladefoged, 1999), and for voiceless aspirated stops the VOT of [p^h] is reliably less than that of [k^h] (e.g., Peterson & Lehiste, 1960; Klatt, 1975; Zue, 1976). Regarding the relative ranking of [t^h] with respect to the other two aspirated stops, previous findings are inconsistent: while a few studies

[8] Throughout, 'voiceless' and 'voiced' are used as convenient and traditional terms to refer to the voiceless aspirated (fortis, long-lag) and unaspirated (lenis, short-lag) stops, respectively. We transcribe the latter as [b d g], even though these sounds are known to lack consistent phonetic voicing for many speakers at least in utterance-initial position (e.g., Lisker & Abramson, 1964; Davidson, 2016; but cf. Jacewicz et al., 2009; Hunnicutt & Morris, 2016). For discussion of the phonological representation of this contrast in AE and other languages, see for example Kingston & Diehl (1994) and Beckman et al. (2013).

We did not measure voicing during stop closure as this can take a variety of context-dependent forms, and need not be contiguous with the release of the stop, making negative (or lead) VOT values difficult to compare with positive (or lag) VOTs (e.g., Docherty, 1992; Möbius, 2004; Davidson, 2016). It could be that the presence, amount, or profile of closure voicing would correlate with positive VOT across talkers, but we leave this for future studies.

report a mean VOT of [t^h] between that of [p^h] and [k^h] (e.g., Peterson & Lehiste, 1960; Lisker & Abramson, 1964), other studies have found minimal differences between [t^h] and [k^h] in both American and British English (Suomi, 1980; Docherty, 1992; Yao, 2009).

In addition to voice and place features, numerous contextual, prosodic, lexical, and global factors also contribute to VOT variation. Longer VOTs are observed before high and tense vowels, particularly [i], for voiceless stops (Klatt, 1975; Port & Rotunno, 1979; Weismer, 1979; Flege et al., 1998; see also Nearey & Rochet, 1994 for Canadian English). At least for [t^h], VOT is subject to domain-initial strengthening effects and realized with a slightly longer VOT compared to utterance-medial [t^h] (Cho & Keating, 2009; see also Pierrehumbert & Talkin, 1992). The VOT of voiceless stops is also longer in monosyllabic words than in polysyllabic words (Klatt, 1975; Flege et al., 1998). Among lexical properties, more frequent words tend to have slightly shorter VOTs (Yao, 2009), and the VOT of word-initial voiceless stops is slightly longer in words with that have minimal pair neighbors beginning with voiced stops (e.g., Baese-Berk & Goldrick, 2009; Kirov & Wilson, 2012; Buz et al., 2016). Finally, like other durational phonetic measures VOT decreases at faster speaking rates (e.g., Miller et al., 1986; Kessinger & Blumstein, 1997, 1998; Allen & Miller, 1999; Allen et al., 2003; Theodore et al., 2009).

Significant variability in VOT has also been identified across talkers, even after talker differences in speaking rate have been taken into account (e.g., Allen et al., 2003; Theodore et al., 2009). Variability across talkers, particularly among the voiceless categories, can span tens of milliseconds, making this source one of the larger factors in VOT variation. Socioindexical factors, such as differences in dialect (e.g., Scobbie,

2006), gender (e.g., Smith, 1978; Swartz, 1992; Byrd, 1993; Whiteside & Irving, 1998), and age (e.g., Benjamin, 1982; Morris & Brown, 1994; Torre & Barlow, 2009; Kleinschmidt & Jaeger, submitted), as well as anatomical and physiological factors such as lung volume (Hoit et al., 1993) have all been implicated in talker-specific VOT values.

2.1.3 Outline

As reviewed in the Introduction, evidence for covariation of talker mean VOT among stop categories has been observed in previous studies (e.g., Koenig, 2000; Newman, 2003; Theodore et al., 2009); however, these have been limited to isolated American English speech, and with the exception of Theodore et al. (2009), have not analyzed talker-specific VOT patterns while taking into account the many other sources of VOT variation reviewed above. The present chapter investigated the production and perception of VOT covariation and the uniformity constraints in a variety of studies. VOT covariation was examined among all six stop consonants of American English first in isolated speech (section 2.2), and in a large multi-talker corpus of connected read speech, the Mixer 6 corpus (section 2.3). Importantly, these studies controlled for many other potential sources of variation that could give rise to patterns of covariation. In sections 2.4 and 2.5, patterns of VOT covariation were analyzed across children ages 2 to 5 in American English speech and across languages in a meta-analysis of VOT means. Finally, section 2.6 examines the extent to which individuals generalize VOT across stop place of articulation in perceptual adaptation to a novel talker.

2.2 Covariation of VOT in isolated speech

The goal of our first study was to replicate and extend previous findings of VOT covariation in isolated speech. Structured variability was explored through the

examination of (i) correlations, (ii) ordinal and linear relations among the talker-specific means, and (iii) a mixed-effects model. First, we assessed the strength of mutual predictability through correlations of stop means across talkers. The same analysis was performed on talker-specific means corrected for speaking rate. In addition, we examined whether the means and standard deviations of talker-specific VOT distributions covary.

Previous studies of place effects on VOT have focused primarily on ordinal rankings. We identified the rankings present in our data, but found that simple linear regressions of one stop mean against another to be more revealing. Finally, the VOT data was submitted to a mixed-effects linear regression model that included many of the predictors described in the introduction. The random effect estimates of such a model help to identify the major sources of variation across talkers.

2.2.1 Methods

2.2.1.1 Participants

Twenty-four students at Johns Hopkins University (13 female) participated in the experiment and received $10 or partial course credit. All participants were native speakers of American English. Data from eighteen of the participants were previously reported in Chodroff & Wilson (2014).

2.2.1.2 Procedure and measurements

Stop-initial CVC syllables were elicited in the carrier phrase "Say ___ again." The syllables were composed of the six stop consonants [p^h t^h k^h b d g] crossed with ten

vowels [i ɪ eɪ ɛ æ ʌ a ɔ oʊ u].[9] The final consonant was always the voiceless coronal stop. One CVC combination was omitted because it formed a taboo word.

Each syllable was assigned an orthographic form according to standard conventions for American English spelling, with the constraint that the consonant and vowel mappings were one-to-one for all stimuli regardless of lexical status. Participants completed five blocks, each syllable occurring once per block. This resulted in a maximum of 50 tokens per stop consonant and talker, except for [t^h], in which case there was a maximum of 45 tokens per talker. (Four participants completed only four blocks due to a programming error.) Stimuli were randomized within each block separately and presented with PsychoPy (Peirce, 2007). Each stimulus was displayed in the frame with a rhyming reference word, used to specify the intended pronunciation of the vowel spelling. The recordings were made in a sound-attenuated booth with a Shure SM58 microphone and Zoom H4n digital recorder with a sampling frequency of 48 kHz (16 bit). The experiment was self-paced and participants were given short breaks between blocks. A total of 6,776 tokens were analyzed (68 additional tokens were omitted due to pronunciation error).

Initial segmentation of the recordings was performed with the Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman, 2008). VOT boundaries for all word-initial stop consonants were then manually placed on the basis of waveform and spectrogram displays in Praat (Boersma & Weenink, 2015). VOT was defined as the duration of the interval from the beginning of the stop release to the start of periodicity in the waveform or a visible f0 track (whichever came first). This measure did not take into account any

[9] The contrast between /a/ and /ɔ/, represented orthographically in our materials by <O> and <AUGH>, may not have been present in the dialects of all of our speakers (e.g., Kurath & McDavid, 1961).

closure voicing, and as discussed in the introduction, is therefore more properly called positive (or lag) VOT. No attempt was made to distinguish among components of the release (i.e., transient, frication, and any following aspiration). In addition, local speaking rate was operationalized as the duration of the vowel in each trial (as in Theodore et al., 2009); this was determined from the manually-aligned stop release offset (equivalently, the vowel onset) and the vowel offset as marked by P2FA.

2.2.2 Results

Stop VOT means varied substantially across talkers: for example, the difference between the lowest and highest talker-specific values for [t^h] approached 100 ms (see Table 2.1). The distributions of talker means are shown as marginal histograms in Figure 2.2. The grand means for the voiceless stops were somewhat higher than figures previously reported for AE laboratory speech; we speculate that this reflects an overall slow speaking rate in the current experiment.

Table 2.1. Descriptive statistics of talker-specific VOT (ms) for each stop category in the isolated speech data. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Stop	Mean	SD	Range of Talker Means	Range of Talker SDs
p^h	89	27	46 – 139	12 – 27
t^h	98	28	57 – 156	10 – 26
k^h	99	24	67 – 137	11 – 20
b	13	5	11 – 20	2 – 8
d	21	7	14 – 32	3 – 10
g	28	10	19 – 42	4 – 13

2.2.2.1 Correlation analyses

The key finding was that the means of several stops were highly correlated across talkers. The correlations among voiceless stops were nearly perfect (r = 0.95 to 0.96; all

*p*s < 0.006), and moderate but significant correlations were observed among the voiced stops ([b - d]: $r = 0.54$, $p = 0.006$; [d - g]: $r = 0.56$, $p < 0.006$; [g - b]: $r = 0.56$, $p < 0.006$). Correlations between homorganic stop pairs failed to reach significance ($r = 0.18$ to 0.33, *ps* > 0.006). All of the correlations are reported in Table 2.2 and in Figure 2.2 together with best-fit linear regression lines.[10, 11]

Two additional analyses were performed to estimate the strength of the correlations in the larger population of AE talkers and to control for speaking rate variation. For each pair of stops separately, a confidence interval for the VOT correlation was estimated with a bootstrap procedure. In each of 1000 repetitions, a correlation was computed from a random sample (with replacement) of the talker-specific means for the two stops. The results of the repetitions were then combined to form a 95% confidence interval according to the bias-corrected and accelerated percentile (BCa) method (Efron, 1987). For instance, the bootstrap interval for [p^h] and [k^h] ranges from $r = 0.86$ to 0.99, suggesting that the point estimate ($r = 0.95$) did not arise from a handful of outliers (though the correlation in the population may be somewhat smaller).

The second analysis was performed on the residuals of a simple linear regression in which each VOT value was predicted from the corresponding speaking rate (operationalized as vowel duration). The residualized VOTs were then averaged by talker and stop category, just as before, and the correlations were recomputed. The magnitudes

[10] Throughout the chapter, the nominal alpha value of 0.05 was Bonferroni-corrected for multiple comparisons. For completeness, we present all relevant correlations even when they are non-independent; this redundancy is eliminated in the mixed-effects analysis reported further below.

[11] Previous work has also modeled VOT on the log scale given the non-linear perception of temporal properties and the large difference in variances between the voiced and voiceless categories (Volaitis & Miller, 1992; Kong, 2009; Sonderegger, 2015). The correlations of talker log VOT means, calculated as the mean of the logged VOTs, resulted in magnitudes comparable to the correlations of talker (linear) VOT means, but the pattern of significance did change (*r*s = [p^h – t^h] 0.96, [t^h – k^h] 0.97, [k^h – p^h] 0.96, *p*s < 0.006; [b – d] 0.51, [d – g] 0.50, [g – b] 0.51, [p^h – b] 0.18, [t^h – d] 0.36, [k^h – g] 0.17, n.s.).

of the correlations among voiceless stops did not deviate from the original magnitudes, demonstrating that differences among talkers in the realization of these sounds cannot be reduced to talker-specific speaking rates. Among the voiced stops and between homorganic pairs, the correlations increased considerably and reached significance (*voiced*: $r = 0.80 – 0.89$; *homorganic*: $r = 0.60 – 0.72$; all ps < 0.006). Differences in speaking rate thus appear to have obscured these relationships in the raw data. Bootstrap confidence intervals again indicated that these correlations were consistent in the population from which our speakers were sampled.[12]

The correlations among stop means suggest that variability is highly structured across talkers. Additional structure in phonetic realization may also exist between talker-specific means and standard deviations, as would be expected from general covariation of first and second moments for phonetic temporal measures (e.g., Byrd & Saltzman, 1998; Shaw et al., 2009; Turk & Shattuck-Hufnagel, 2014); indeed, increased temporal durations have been shown to correspond with greater variability throughout human motor behavior (Schmidt et al., 1979; Schöner, 2002). Significant correlations of the talker means and standard deviations were observed for all stops (*all*: $r = 0.90$), as well as for voiced stops ([b]: $r = 0.71$, [d]: $r = 0.76$, [g]: $r = 0.75$, all ps < 0.008). Moderate correlations were also observed for the voiceless stops; however, these failed to reach significance after correction for multiple comparisons ([p^h]: $r = 0.47$, $p = 0.02$; [t^h]: $r = 0.53$, $p = 0.008$; [k^h]: $r = 0.43$, $p = 0.04$). These correlations likely reflect restricted

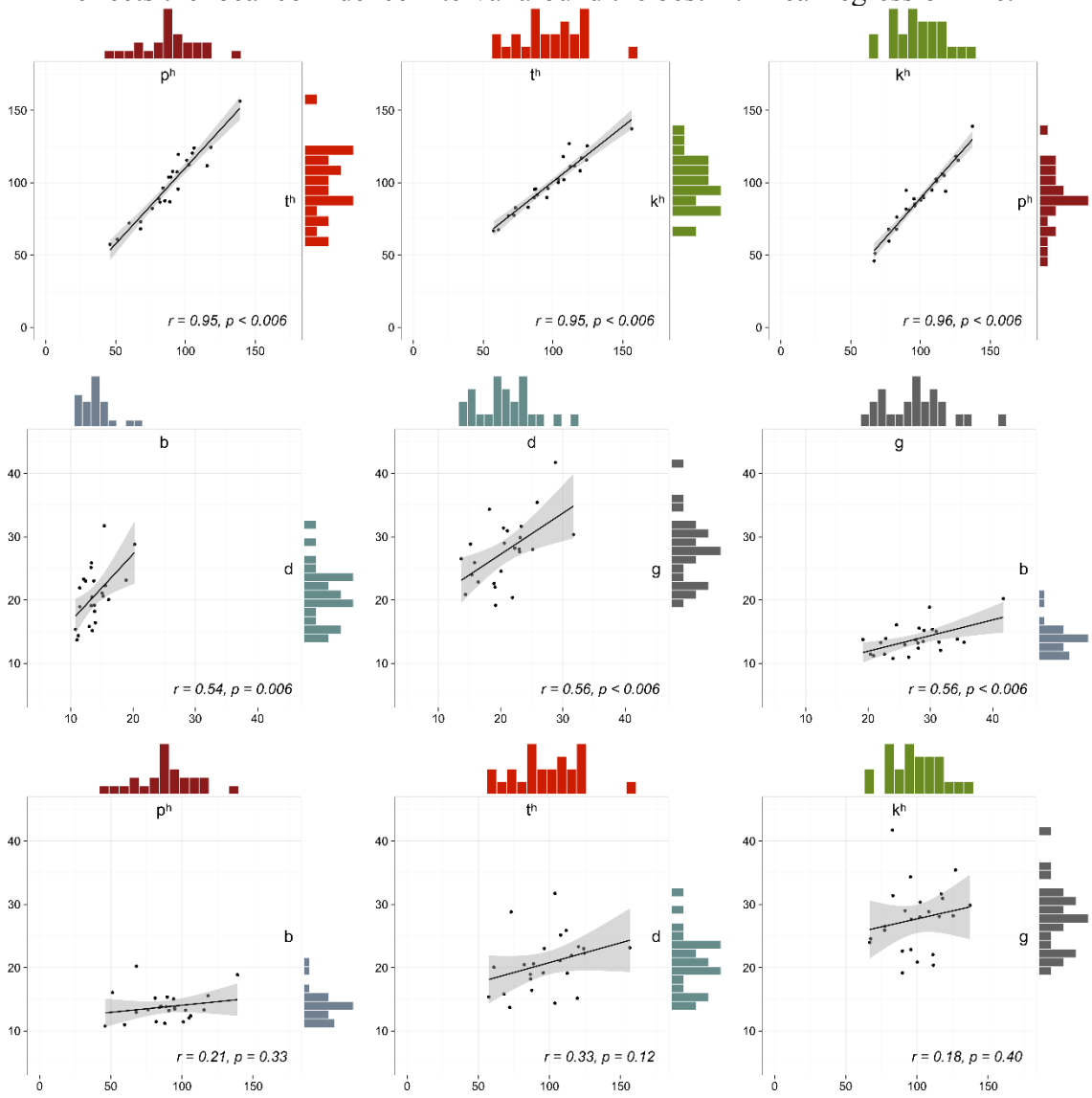
[12] This analysis residualized the dependent variable (VOT) against a predictor (speaking rate / vowel duration), and thus was not subject to the pitfalls of residualizing one predictor against another (Wurm & Fisicaro, 2014).

variation at the lower boundary for each voicing category: a lower bound at 0 ms for voiced stops and at the auditory boundary between the categories for the voiceless stops.

Table 2.2. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means for raw and residualized VOT (ms) in the isolated speech data.

	Raw VOT			***Residualized VOT***		
	Pearson's *r*	*p*-value	95% CI	Pearson's *r*	*p*-value	95% CI
p^h – t^h	0.95	< 0.006	[0.90, 0.98]	0.95	< 0.006	[0.89, 0.98]
t^h – k^h	0.95	< 0.006	[0.86, 0.98]	0.95	< 0.006	[0.88, 0.98]
k^h – p^h	0.96	< 0.006	[0.86, 0.99]	0.96	< 0.006	[0.88, 0.99]
b – d	0.54	0.006	[0.21, 0.77]	0.89	< 0.006	[0.69, 0.95]
d – g	0.56	< 0.006	[0.23, 0.78]	0.80	< 0.006	[0.54, 0.91]
g – b	0.56	< 0.006	[0.21, 0.84]	0.82	< 0.006	[0.56, 0.91]
p^h – b	0.21	0.33	[-0.42, 0.76]	0.72	< 0.006	[0.44, 0.90]
t^h – d	0.33	0.12	[-0.16, 0.61]	0.64	< 0.006	[0.25, 0.85]
k^h – g	0.18	0.40	[-0.25, 0.53]	0.60	< 0.006	[0.25, 0.82]

Figure 2.2. Variation and covariation of stop VOT means (ms) across talkers in the isolated speech data. Marginal histograms show variation in talker means. The top row shows correlations among the voiceless stops, the middle row among the voiced stops (note change of scale), and the bottom row within homorganic pairs. Gray shading reflects the local confidence interval around the best-fit linear regression line.



2.2.2.2 Ordinal and linear relations

Previous studies have generally considered the relationships among VOT means in terms of ordinal rankings (e.g., Peterson & Lehiste, 1960; Cho & Ladefoged, 1999). For comparison with these studies, we also assessed the ranking, and identified three predominant patterns across talkers: [b] < [d] < [g] < [pʰ] < [tʰ] < [kʰ] (11 talkers), [b] <

[d] < [g] < [p^h] < [k^h] < [t^h] (8 talkers), and [b] < [g] < [d] < [p^h] < [k^h] < [t^h] (3 talkers); two talkers exhibited other rankings. For all talkers and within both values of [voice], the mean dorsal VOT was longer than the mean labial VOT, consistent with cross-linguistic tendencies (e.g., Cho & Ladefoged, 1999). However, the relative ranking of coronal and dorsal means varied across talkers, with more variation among the voiceless than the voiced stops (see also Docherty, 1992; Yao, 2009).

The preceding correlations and ordinal rankings provide some information about systematic relations among talker-specific stop VOT means, but simple linear regressions can reveal additional structure. While the correlations indicate that stop-specific means are linearly related, this could take the form of a constant difference between means ($y = \beta_0 + x$), a constant ratio between means ($y = \beta_1 \cdot x$), or a combination of the two ($y = \beta_0 + \beta_1 \cdot x$). We performed a separate simple linear regression for each pair of stops, regressing the talker means of one stop against those of another.

Paralleling the correlation magnitudes, the proportion of variance accounted for by the regressions was largest for the voiceless stop pairs (adjusted R^2s > 0.50) and smallest for the voiced stop pairs and homorganic pairs (adjusted R^2s < 0.50). We will discuss only the model fits for the voiceless stops, but for completeness all models are reported in Table 2.3.

In predicting [k^h] from either [t^h] or [p^h], both the intercept and scaling factors were significant ([k^h ~ t^h]: $\beta_0 = 24.66$, $\beta_1 = 0.76$; [k^h ~ p^h]: $\beta_0 = 24.37$, $\beta_1 = 0.85$; all *p*s < 0.003). The linear fits inherently account for the ordinal rankings: [k^h] > [t^h], [p^h] is expected over much of the empirical range of VOT values; however, [t^h] and [p^h] also increase faster relative to [k^h], resulting in a point at which the ranking is reversed. For

[t^h] and [k^h], this point is within the reasonable range of values for isolated speech (103 ms). In the model predicting [t^h] from [p^h], only the scaling factor was significant, indicating a straightforwardly proportional relationship ([$t^h \sim p^h$]: $\beta_0 = 5.15$, $p = 0.43$, $\beta_1 = 1.05$, $p < 0.003$).

Table 2.3. Additive (β_0) and scalar (β_1) components of simple linear regressions on talker mean VOTs of one stop predicted from another. For each pair, the dependent variable is given first followed by the independent variable.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
$t^h \sim p^h$	5.15	0.43	1.05	< 0.003	0.91
$k^h \sim t^h$	24.66	< 0.003	0.76	< 0.003	0.90
$k^h \sim p^h$	24.37	< 0.003	0.85	< 0.003	0.92
d ~ b	6.06	0.23	1.06	< 0.003	0.28
g ~ d	14.22	0.004	0.65	0.005	0.26
g ~ b	10.00	0.09	1.28	0.004	0.29
$p^h \sim$ b	62.62	0.03	1.89	0.33	0.00
$t^h \sim$ d	63.36	0.009	1.70	0.12	0.07
$k^h \sim$ g	81.83	< 0.003	0.64	0.40	-0.01

Linear regression models like these have been employed in automatic approaches to speaker adaptation, as pairwise regressions between speech sounds and classes of sound allow for more rapid talker adaptation from limited talker-specific data (Furui, 1980; Cox, 1995). Strong linear relationships among talker-specific realizations of speech sounds could also have implications for cognitive models of adaptation, accounting for how listeners form expectations about the realization of unheard speech sounds after limited exposure.

2.2.2.3 Mixed-effects analysis

A mixed-effects linear regression model provided further statistical support for the findings reported above while allowing us to investigate additional factors known to influence VOT (Baayen et al., 2008). In addition to the factors already considered (i.e., the voice contrast, place of articulation, and speaking rate), the model included properties

of the following vowel that are known to condition VOT (i.e., vowel height and tenseness; Klatt, 1975; Port & Rotunno, 1979; Nearey & Rochet, 1994). While the manipulation of vowel properties was balanced across participants in our study, and therefore could not provide an alternative explanation for the speaker differences or the correlations among categories, it is important to identify the signature of phonetic covariation in mixed-effect models.

We analyzed the random effect component of the fit model to demonstrate that much of the variability in VOT across participants was due to differences in overall mean (intercept) and in the magnitude of the voicing contrast. Unlike the descriptive analyses reported above, the method of this section is more general: it can be employed for data sets in which vowel and other factors are not balanced across speakers provided there is sufficient data (e.g., in an analysis of spontaneous speech).

The model included fixed effects of phonological voice, place of articulation, speaking rate, vowel height, vowel tenseness, as well as the two-way voice × place, voice × rate, and height × tense interactions. All categorical factors were weighted effect coded to correct for slightly unequal sample sizes (Darlington, 1990; p. 246). The coding of the categorical variables was as follows, with contrast weighting reported in the parentheses: phonological voice (*voice*: voiceless = 1, voiced = -0.97), place of articulation (*poaCor*: coronal = 1, dorsal = 0, labial = -0.96; *poaDor*: coronal = 0, dorsal = 1, labial = -1), vowel height (*height*: high = 1, non-high = -0.41); vowel tenseness (*tense*: tense = 1, lax = -1.57). The continuous factor of speaking rate was *z*-scored using the mean and standard deviation (μ = 167 ms, σ = 43 ms) computed from all vowels collapsed across

participants. Similarly, the dependent variable (VOT) was centered at zero by subtracting the grand mean (μ = 57 ms) from each value.

The random effect for speaker included an intercept and slopes for voice, place, rate, and voice × place. While an attempt was made to include random slopes for vowel height and tenseness, these led to non-convergence and were removed. There was also a random intercept for syllable rime (VC portion), which is known to be a salient sublexical unit for English speakers (e.g., De Cara & Goswami, 2002).

The model revealed significant main effects of voice (*voice*: β = 37.30, t = 17.50) and place (*poaCor*: β = 1.48, t = 2.56; *poaDor*: β = 5.50, t = 9.55).[13] The effect of place of articulation here is the numerical counterpart of the ranking differences across places described earlier. Compared to the values that would be predicted from voice and place alone, coronal stops were significantly longer when voiceless than when voiced (*voice* × *poaCor*: β = 1.38, t = 2.79), whereas voiceless dorsal stops were significantly shorter (*voice* × *poaDor*: β = -1.46, t = -3.71).

There was also a main effect of speaking rate, and slightly shorter VOTs were found at faster speaking rates (*rate*: β = -2.46, t = -4.55). (The coefficient for speaking rate can be interpreted as the predicted change in VOT in milliseconds given a one standard deviation change in speaking rate.) Speaking rate interacted significantly with voice: as expected from their greater variability overall (see below), voiceless stops showed a stronger effect of rate than voiced stops (*voice* × *rate*: β = 0.39, t = 1.98). Vowel height, vowel tenseness, and their interaction did not reach significance (*height*: β = 0.37, t = 0.29; *tense*: β = 1.11, t = 1.82; *height* × *tense*: β = 1.36, t = 1.39).

[13] A t-value with magnitude greater than 2.0 was considered significant.

The random effect estimates can provide further insight into the major sources of talker variation. As shown in Table 2.4, the random intercept and the voice slope had the largest standard deviations, indicating differences across talkers in overall mean VOT and in the magnitude of separation between voiced and voiceless stops. In comparison, the variances for the other random talker slopes were much smaller (e.g., the variance of the voice slope was about four times that of either place effect). This is consistent with the finding of Theodore et al. (2009) that there are significant differences across talkers in the intercept, or overall mean, but not in the effect of place of articulation (for [p^h] and [k^h]).[14]

It is well known, and confirmed by our data, that there is greater VOT variation for voiceless stops than for voiced stops (see Figure 2.2; Dmitrieva et al., 2015). This presumably reflects both a relatively fixed auditory boundary between the voicing categories (e.g., Kuhl, 1981) and, in our study, the lower bound of zero on positive VOT measurements. Therefore, a speaker with a higher overall mean VOT is very likely to have a larger separation between voiced and voiceless stops (thus ensuring that the voiced stops lie below the boundary); and indeed, the random intercept and voice slope were tightly correlated ($r = 0.97$). While this might suggest that voiceless and voiced stops should simply be analyzed separately, the moderate correlations within homorganic pairs reported earlier indicate that some component of talker-specific VOT is shared by all of the stops.

[14] The model reported here performed significantly better than models with simpler random effect structures for talker as determined by log-likelihood ratio tests and Bayesian Information Criterion (BIC) comparisons. However, inclusion of additional factors beyond the intercept and voice gave diminishing returns in accounting for VOT variability. In comparison to a model with no talker-specific random effect, the BIC decreased by 5,293 for a model with a random intercept and voice slope for talker, but only by a further 150 units for the maximal random effect model reported in the main text.

Table 2.4. Standard deviations of the random effect components for talker in the maximal mixed-effects model.

Random effect for talker	SD
intercept	11.17
voice	10.40
poaCor	2.63
poaDor	2.63
speaking rate	2.26
voice × poaCor	2.16
voice × poaDor	1.63

2.2.3 Discussion

Despite substantial talker variation in VOT values, highly stable relations of talker means were observed among stop categories. These results are consistent with previous laboratory findings of correlations in talker means, but extend the findings to all six stop categories while also controlling for other sources of variability such as differences in speaking rate. The correlation and random effect analyses both provided evidence for the existence of strong positive linear relationships in talker VOT. In addition, there were also consistent ordinal rankings of stop VOT, with talkers predominantly exhibiting a lower mean VOT for labials than for dorsals within each voicing category. Yet, in describing the relation between VOT means, the linear relationships not only captured the ordinal rankings, but also accounted for the variability in the ranking of coronals and dorsals, and critically, quantified the magnitude of separation between VOT means.

The patterns of findings provided strong evidence in favor of target uniformity: correlations of talker mean VOT were quite strong among stops with a shared voice feature, especially among the aspirated stops and to a lesser extent among the unaspirated stops. While the linear regressions indicated differences in VOT means across place of articulation, these could be accounted for by an underlying uniform phonetic target,

perhaps in a glottal spreading gesture of equal magnitude and relative timing to the constriction. Regardless, the systematic place differences across talkers and restricted variation revealed that the phonetic targets corresponding to a shared laryngeal feature value may be highly similar.

Evidence for contrast uniformity was relatively weaker than that for target uniformity. The correlations among stops contrasting in the laryngeal specification were weak to moderate, and the R^2 of the linear regressions directly reflected those coefficients. While the greatest variation across talkers in the random effects component of the mixed-effects model was in the random intercept, there was nevertheless substantial variation in the random talker slope for voice. This variation indicated that talkers also differed in the separation between voiced and voiceless stop VOTs, counter to predictions of contrast uniformity. However, after correcting for speaking rate, many of the correlations strengthened considerably, especially among stops contrasting in the laryngeal feature. The role of contrast uniformity may have been obscured by speaking rate; whether this finding persists across other speech corpora will be evaluated in the following section.

These results established a high degree of structured variation among the VOT means of AE stops, and conformed in large part to the predictions of target uniformity. It is still unclear whether similar patterns would also be observed in the production of known lexical items in connected speech, as opposed to a controlled laboratory study of isolated speech. The following study addressed this question by examining patterns of talker VOT in a large corpus of read speech that contained a greater variety of prosodic and lexical factors, but otherwise matched sentential conditions for each talker. This

allowed for analysis of VOT as produced in a more natural and connected speech style, while also ensuring that talkers were producing approximately the same content.

2.3 Covariation of VOT in connected speech

Phonetic research has increasingly employed large connected speech corpora (e.g., Byrd, 1992; Cole et al., 2003; Yuan & Liberman, 2008). While laboratory conditions ensure a greater degree of control, speech corpora can provide great quantities of naturally-occurring speech. Large-scale corpus studies have been conducted for many aspects of speech, including but not limited to segmental realization (e.g., Byrd, 1992), coarticulatory and contextual effects (Keating et al., 1994; Gendrot & Adda-Decker, 2005; Bürki et al., 2011; Schuppler et al., 2011; Torreira & Ernestus, 2012; Elvin & Escudero, 2014; Yu et al., 2015), prosodic structure and speaking rate (e.g., Ostendorf et al., 2001; Kendall, 2009), and phonetic change over time (e.g., Fruehwald, 2013; Labov et al., 2013).

Many techniques originally developed for automatic speech recognition (ASR) have facilitated phonetic analysis of large corpora (e.g., Yuan & Liberman, 2008; Rosenfelder et al., 2011; Yoon & Kang, 2013). These include algorithms for extracting VOTs values (e.g., Das & Hansen, 2004; Yao, 2007; Sonderegger & Keshet, 2010), vowel formants (Evanini et al., 2009; Yao et al., 2010), and degrees of vowel nasalization (Yuan & Liberman, 2011), as well as for prosodic labeling (e.g., Wightman & Ostendorf, 1994; Hasegawa-Johnson et al., 2005; Gorman et al., 2011). With respect to VOT, large-scale analyses have examined population-level VOT distributions (Byrd, 1993), phonetic accommodation over time (Sonderegger, 2015), dialectal differences (Stuart-Smith et al., 2015), and effects of prosodic structure (Cole et al., 2007), among others.

The corpus employed in our analysis, the Mixer 6 corpus (Brandschain et al., 2010, 2013), is well-suited for the study of variation across talkers. The complete corpus contains speech from approximately 600 AE talkers recorded in one to three separate sessions. In each session, the participant completed an interview (15 minutes), transcript reading (15 minutes), and telephone call (10 minutes), and sessions were separated by at least two days. The corpus was collected primarily to support research in speaker recognition technologies; however, the read transcript portion was expressly added to foster basic scientific research on talker characteristics (for further details, see Brandschain et al., 2010 and Chodroff et al., 2016).

Other transcribed speech corpora can provide a large number of speakers (e.g., Switchboard: Godfrey et al., 1992; TIMIT: Garofolo et al., 1993), a large number of data points per talker (e.g., Buckeye Corpus: Pitt et al., 2005), or even the combination of these two (e.g., Wall Street Journal Corpus: Paul & Baker, 1992; LibriSpeech: Panayotov et al., 2015). A unique advantage of the Mixer 6 read speech portion is that it provides a large sample for each talker while holding constant prosodic, lexical, and syntactic/semantic factors. This allowed us to investigate talker variation at the level of phonetic categories without a major confound of sentential content.

The same set of analyses as presented in the preceding study were used to assess the extent of structured VOT variation in the connected speech study. Recall from section 2 that this includes a correlation analysis, an examination of the ordinal and linear relationships among talker means, and finally, an analysis of the talker-specific random effect variances in a linear mixed-effects model.

2.3.1 Methods

2.3.1.1 Corpus description

The following analysis employed an audited subset of the Mixer 6 read speech for 180 native AE talkers (102 female, 78 male). Each talker recorded three read speech sessions, resulting in approximately 45 minutes of speech. The script contained 335 selected sentences randomly drawn from utterances in the Switchboard corpus. The selected sentences were therefore naturally occurring and not selected for the research question at hand. Each selected sentence contained 1 to 17 words with a median of 7 words. Participants read the selected sentences in a fixed order in each session until 15 minutes had passed. The number of sentences completed and read correctly within each session ranged from 103 to 338 (median: 238; mean: 239).

All talkers in the present analysis were born in the United States: 83 were from Pennsylvania (57 from Philadelphia), 48 from other Northeast states, 18 from the Southeast, 14 from the Midwest, 11 from the West, and 6 from the Southwest. Talkers ranged in age from 18 to 86 years (median: 27 years).

2.3.1.2 Acoustic measurements

VOT measurements were extracted for all stops that appeared word-initially, in any utterance position, and that were followed immediately by vowels transcribed as bearing primary stress. Prior to measurement, reading and recording errors were removed with a combination of automatic and manual methods (for details see Chodroff et al., 2016). The cleaned transcripts were phonetically aligned to the corresponding audio using P2FA. AutoVOT (Sonderegger & Keshet, 2010, 2012) was then used to locate the onset of each stop release and the onset of the following vowel using pre-trained

statistical models. For voiceless stops, the temporal window for this analysis extended 30 ms before and 30 ms after the stop interval as marked by P2FA; for voiced stops, the P2FA interval was extended in both directions by 10 ms. The minimum VOT threshold, required by AutoVOT, was set to 15 ms for voiceless stops and 4 ms for voiced stops.

To estimate the accuracy of AutoVOT for this corpus, and following the same procedure as in section 2.1.2, we hand-measured the VOTs of a randomly selected subset of the stops (more than 3,000 tokens, or approximately 3% of the data). Comparison of the automatic and manual measurements yielded a root-mean-square deviation of 12.9 ms (somewhat larger than the 7.74 ms reported by Sonderegger & Keshet, 2010 for the Big Brother Corpus).[15] An additional 936 stops with VOTs equal to the minimum threshold, or with exceptionally long values, were hand-corrected.[16] Among the hand-corrected stops, tokens lacking visible stop bursts were excluded from all analyses (209 tokens omitted).

Measurements were taken from the boundaries placed by AutoVOT or, when available, the manually-placed boundaries. Because utterances in this corpus varied considerably in length and structure, we operationalized speaking rate for each one as the average word duration determined from the P2FA boundaries.

All words were retained in the analysis with the exception of *'to'* (which was highly frequent and subject to *wanna*-contraction and other phonetic reductions). VOT values 2.5 standard deviations above or below talker-specific category means were

[15] The root-mean-square deviation for each stop category was [p^h]: 7.3 ms, [t^h]: 16.3 ms, [k^h]: 6.3 ms, [b]: 2.2 ms, [d]: 2.9 ms, and [g]: 16.9 ms.

[16] AutoVOT provides the capability of training its statistical model on a user-supplied corpus. We trained on two-thirds of our manually-measured stops (1,488 voiceless, 990 voiced) and tested on the remaining third (755 voiceless, 489 voiced). The root-mean-square error of the resulting model (13.0 ms) was not superior to that of the pre-trained models.

excluded. This left a total of 88,725 measurements for analysis, with a median of 547 per talker (range: 296 – 741). The range and median number of tokens per talker and stop are given in Table 2.5 along with the total number of tokens per stop. These tokens are instances of 98 word types: 17 lexical items for [p^h], 14 for [t^h], 21 for [k^h], 18 for [b], 16 for [d], and 12 for [g].

Table 2.5. Range and median number of tokens per talker and stop category, and total number of tokens per stop category.

Stop	Range	Median	Total
p^h	44 – 100	77	13,517
t^h	17 – 77	46	8,218
k^h	46 – 114	82	14,619
b	42 – 117	80	14,661
d	58 – 184	131	23,086
g	52 – 118	82	14,763

2.3.2 Results

Talker-specific VOT means varied considerably within each stop category (Table 2.6). Within the voiceless stops, talker-specific means ranged from 28 ms to 78 ms for [p^h], from 40 ms to 96 ms for [t^h], and from 36 ms to 79 ms for [k^h]. For the voiced stops, the range in talker-specific VOT was limited by the minimum positive VOT and the voicing boundary; however, talker-specific means still differed by up to 19 ms (Table 2.6). The grand mean VOTs for the voiceless stops were comparable to figures reported in previous studies of read and spontaneous speech (e.g., Byrd, 1993; Yao, 2007), but overall shorter than those observed in isolated speech (e.g., Lisker & Abramson, 1964).

Table 2.6. Descriptive statistics of talker-specific VOT (ms) for each stop category in the connected speech data. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Stop	Mean	SD	Range of Talker Means	Range of Talker SDs
p^h	51	9	28 – 78	11 - 35
t^h	61	9	40 – 96	9 - 34
k^h	56	8	36 – 79	11 - 30
b	8	2	6 – 14	2 - 8
d	14	3	8 – 22	4 - 13
g	17	3	9 – 28	6 - 15

2.3.2.1 Correlation analyses

As shown in Table 2.7 and

Figure 2.3, correlations among the voiceless stop consonants were particularly strong, with coefficients ranging from $r = 0.77$ to 0.83 (all ps < 0.006). Among the voiced stops, talker means were significantly correlated between [b] and [g] ($r = 0.49$, $p < 0.006$), as well as [d] and [g] ($r = 0.33$, $p < 0.006$), but not between [b] and [d] ($r = 0.07$, $p = 0.33$). Correlations between homorganic stops were also significant for coronals and dorsals (*coronal*: $r = 0.53$; *dorsal*: $r = 0.43$, ps < 0.006; cf. *labial*: $r = 0.15$, $p = 0.05$).[17, 18]

The same pattern of significance emerged in the correlations of residualized talker means which were obtained after removing the effect of speaking rate on VOT with a simple linear regression (Table 2.7). While speaking rate was measured here as mean

[17] The fact that [p^h] - [b] showed the lowest correlation among homorganic stops (see Table 2.7) could reflect a limitation of our method; [b] is the stop most amenable to phonetic voicing, and a higher correlation may emerge when positive and negative VOTs are measured.

[18] The correlations of talker log VOT means had comparable magnitudes and the same pattern of significance as the correlations of linear VOT means (rs = [p^h – t^h] 0.79, [t^h – k^h] 0.71, [k^h – p^h] 0.79, [d – g] 0.39, [g – b] 0.43, [t^h – d] 0.54, [k^h – g] 0.44, ps < 0.006; [b – d] 0.07, [p^h – b] 0.18, both n.s.).

word duration per utterance, the same pattern of significance was also realized when speaking rate was measured as following vowel duration.[19]

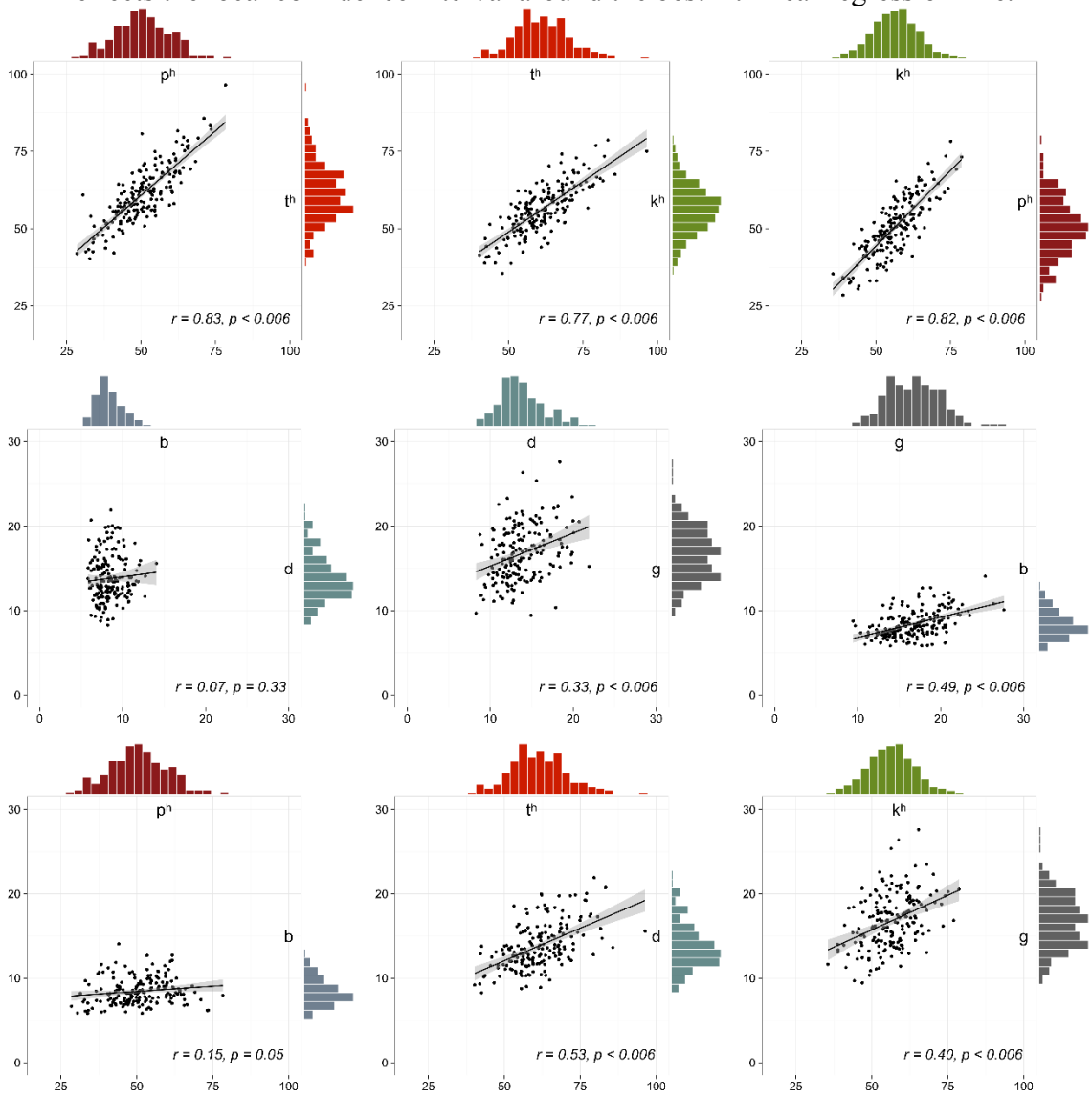
Consistent with previously observed temporal patterns in speech (and other motor behaviors), strong positive correlations were also found between talker-specific means and standard deviations. Means and standard deviations were significantly correlated in an analysis of all stops together ($r = 0.90$). Moderate correlations were present for each of the voiceless stops ([p^h]: $r = 0.57$, [t^h]: $r = 0.47$, [k^h]: $r = 0.51$, ps < 0.008). For the voiced stops, strong correlations were observed within [b] and [d], and a moderate correlation was observed for [g] ([b]: $r = 0.79$, [d]: $r = 0.76$, [g]: $r = 0.47$, ps < 0.008).

Table 2.7. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means for raw and residualized VOT (ms) in the connected speech data.

	Raw VOT			***Residualized VOT***		
	Pearson's *r*	*p*-value	95% CI	Pearson's *r*	*p*-value	95% CI
p^h – t^h	0.83	< 0.006	[0.77, 0.88]	0.81	< 0.006	[0.74, 0.86]
t^h – k^h	0.77	< 0.006	[0.71, 0.82]	0.75	< 0.006	[0.67, 0.80]
k^h – p^h	0.82	< 0.006	[0.77, 0.86]	0.80	< 0.006	[0.74, 0.85]
b – d	0.07	0.33	[-0.05, 0.19]	-0.03	0.65	[-0.15, 0.09]
d – g	0.33	< 0.006	[0.20, 0.46]	0.20	0.008	[0.05, 0.33]
g – b	0.49	< 0.006	[0.36, 0.59]	0.41	< 0.006	[0.28, 0.53]
p^h – b	0.15	0.05	[-0.01, 0.30]	-0.11	0.17	[-0.27, 0.18]
t^h – d	0.53	< 0.006	[0.43, 0.63]	0.40	< 0.006	[0.28, 0.52]
k^h – g	0.40	< 0.006	[0.29, 0.50]	0.27	< 0.006	[0.14, 0.39]

[19] The same pattern of significance found for the entire set of talkers was present within the female and male subgroups. Correlations among voiceless stop VOTs ranged from $r = 0.80$ to 0.85 for female talkers and from $r = 0.74$ to 0.78 for males. Among the voiced stops, correlations ranged from $r = 0.25$ to 0.58 for females and from $r = 0.36$ to 0.50 for males. Relations between homorganic stops were also similar (female: $r = 0.18$ to 0.47; male: $r = 0.42$ to 0.47).

Figure 2.3. Variation and covariation of stop VOT means (ms) across talkers in the connected speech data. Marginal histograms show variation in talker means. The top row shows correlations among the voiceless stops, the middle row among the voiced stops (note change of scale), and the bottom row within homorganic pairs. Gray shading reflects the local confidence interval around the best-fit linear regression line.



2.3.2.2 Ordinal and linear relations

As in the isolated speech data, three rankings were predominant: [b] < [d] < [g] < [pʰ] < [kʰ] < [tʰ] (113 talkers), [b] < [d] < [g] < [pʰ] < [tʰ] < [kʰ] (31 talkers), or [b] < [g] < [d] < [pʰ] < [kʰ] < [tʰ] (27 talkers). Other patterns were observed for 9 talkers. For five talkers, the order was [b] < [g] < [d] < [pʰ] < [tʰ] < [kʰ], and for four talkers, [tʰ] was

marginally shorter than [p^h]. For all but three talkers and within each voicing category, the mean labial VOT was shorter than the mean dorsal VOT. In all cases, the mean VOTs for the voiceless stops were greater than the voiced.

As in the isolated speech study, the linear relationships between stop means were explored with simple regression models predicting the talker mean VOT of one stop from another. The additive and scalar factors for all pairwise linear regression models are provided in Table 2.8. The best fits, in which the proportion variance accounted for exceeded 0.50, were among the voiceless stops. In each of these models, both the intercept and scaling factor were significant, indicating a combination of additive and proportional factors in the relationship between talker means ([$t^h \sim p^h$]: $\beta_0 = 19.09$, $\beta_1 = 0.83$; [$k^h \sim t^h$]: $\beta_0 = 16.25$, $\beta_1 = 0.65$; [$k^h \sim p^h$]: $\beta_0 = 21.11$, $\beta_1 = 0.70$; all ps < 0.001). These linear fits reflect that the fact that the differences in VOT means for [t^h] and [p^h] as well as [k^h] and [p^h] become smaller as the mean of [p^h] increases, and that all but the lowest VOT means for [t^h] tend to be higher than those of [k^h] (see Figure 2.3). For connected speech, these models provide the best-fitting linear description of how knowledge of one talker-specific mean could be generalized to the other voiceless stops.

Table 2.8. Additive (β_0) and scalar (β_1) components of simple linear regressions on talker mean VOTs of one stop predicted from another. For each pair, the dependent variable is given first followed by the independent variable.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
$t^h \sim p^h$	19.09	< 0.003	0.83	< 0.003	0.69
$k^h \sim t^h$	16.25	< 0.003	0.65	< 0.003	0.60
$k^h \sim p^h$	21.11	< 0.003	0.70	< 0.003	0.68
d ~ b	12.76	< 0.003	0.13	0.33	0.00
g ~ d	11.35	< 0.003	0.39	< 0.003	0.10
g ~ b	8.25	< 0.003	1.01	< 0.003	0.24
$p^h \sim b$	43.08	< 0.003	0.89	< 0.05	0.02
$t^h \sim d$	35.84	< 0.003	1.84	< 0.003	0.28
$k^h \sim g$	35.57	< 0.003	1.00	< 0.003	0.16

2.3.2.3 Mixed-effects analysis

The model included all of the fixed effects considered for isolated speech (section 2.2.3): voice, place of articulation, speaking rate, vowel height and tenseness, as well as the two-way voice × place, voice × rate, and height × tense interactions. (Recall that all measured stops appeared before vowels bearing primary stress, therefore effects of different stress levels or of following non-syllabic approximants could not be investigated.) In addition, there were fixed effects of the position of the word in the utterance, number of syllables in the word, and word frequency.

To accommodate unequal sample sizes, weighted effect coding was used for the categorical variables. Voice had two levels (*voice*: voiceless = 1, voiced = -0.69), and place of articulation had three levels, corresponding to two contrasts with labial as baseline (*poaCor*: coronal = 1, dorsal = 0, labial = -1.13; *poaDor*: coronal = 0, dorsal = 1, labial = -1.08). As described in the methods, speaking rate was the average word duration per utterance defined by the P2FA boundaries. This predictor was *z*-scored across all talkers (μ = 242 ms, σ = 59 ms). Additional binary factors were vowel height (*height*: high [i ɪ u ʊ] = 1, non-high [æ ɛ ei ʌ ɚ a ɔ oʊ oi ai aʊ] = -0.67) and vowel tenseness

(*tense*: tense [i ei ʌ a ɔ oʊ u oi ai aʊ] = 1, lax [ɪ ɛ æ ɚ ʊ] = -2.15). Position of the word (utterance position) was coded as one of five categories: initial, medial, final, pre-pausal, or post-pausal. Tokens that were utterance-medial but preceded or followed by a decoded silence were labeled respectively as post-pausal and pre-pausal. For P2FA to decode a segment as silence, the duration of the segment must be at least 30 ms long. Medial position served as the baseline level (*posInit*: initial = 1, medial = -0.10, else = 0; *posFinal*: final = 1, medial = -0.17, else = 0; *posPrePaus*: pre-pausal = 1, medial = -0.05, else = 0; *posPostPaus*: post-pausal = 1, medial = -0.04, else = 0). Number of syllables per word was categorized into three levels: monosyllabic, disyllabic, and polysyllabic (> two syllables), and the monosyllabic level served as baseline (*syllDi*: disyllabic = 1, polysyllabic = 0, monosyllabic = -0.39; *syllPoly*: disyllabic = 0, polysyllabic = 1, monosyllabic = -0.15). Lexical frequency was calculated as the log SUBTLEX frequency (Marian et al., 2012). The dependent variable (VOT) was centered by subtracting the overall mean (μ = 30 ms) from each value.

The model also included random effects of talker and word. The random effect structure for talkers included an intercept and slopes for voice, place, speaking rate, and the voice × place interaction. Attempts were made to include additional factors, but this resulted in non-convergence. The random effect of word included an intercept only.

Significant main effects emerged for voice (*voice*: β = 24.99, t = 29.07) and place (*poaCor*: β = 1.95, t = 2.40; *poaDor*: β = 1.99, t = 2.55). The interaction between voice and the first place contrast did not reach significance (*voice* × *poaCor*: β = 1.52, t = 1.74), but there was a significant interaction between voice and the second place contrast, indicating a smaller difference between the voiced and voiceless dorsals than would have

been predicted by voice and place independently (*voice* × *poaDor*: β = -4.00, *t* = -5.00). These interactions reflect the difference in ranking of places of articulation within each voicing category: while VOT increases with more posterior place among voiced stops, there is little difference in VOT between coronals and dorsals among voiceless stops. Significantly shorter VOTs corresponded to faster speaking rates (*rate:* β = 1.40, *t* = 16.87), but this was modulated by a significant interaction between voice and rate (*voice* × *rate*: β = 1.20, *t* = 16.41). The effect of rate was augmented for voiceless stops and essentially negated for voiced stops.

Vowel height and tenseness did not reach significance (*height:* β = 1.56, *t* = 1.88; *tense:* β = 0.43, *t* = 0.96); however, there was a significant interaction between height and tenseness indicating that VOT before high tense vowels was significantly longer (β = 1.07, *t* = 2.71). There were significant main effects of utterance position: the VOTs of stops in utterance-initial, pre-pausal, and post-pausal positions were significantly longer than the mean (*posInit*: β = 3.50, *t* = 17.88; *posPrePaus*: β = 0.81, *t* = 3.00; *posPostPaus*: β = 2.78, *t* = 11.84), whereas VOTs of utterance-final stops were significantly shorter (*posFinal*: β = -1.36, *t* = -9.43). The number of syllables was not significant, but the two effects trended in the expected directions. VOTs of disyllabic and polysyllabic words were generally shorter than the VOT of monosyllabic words (*syllDi:* β = -1.15, *t* = -1.04; *syllPoly:* β = -2.72, *t* = -1.39). Finally, higher lexical frequency was associated with a decrease in VOT (β = -1.89, *t* = -2.37).

Analysis of the talker random effects revealed that the intercept and the slope for voice had the largest standard deviations (Table 2.9). This indicates that the talkers differed most in their overall mean VOT values and in the degree of separation between

voiced and voiceless stops. However, the intercept and voice slope were highly correlated (r = 0.91), reflecting the fact that the measured means for voiced stops fell between our minimum threshold (4 ms) and the natural voicing category boundary.

Table 2.9. Standard deviations of the random effect components for talker in the maximal mixed-effects model.

Random effect for talker	SD
intercept	3.68
voice	4.25
poaCor	1.84
poaDor	1.77
speaking rate	0.74
voice x poaCor	1.45
voice x poaDor	1.62

2.3.3 Discussion

The patterns observed in connected speech paralleled those in isolated speech. The mean VOTs of stops were highly correlated, especially within each of the two voicing categories. In addition, there were moderate to strong correlations of talker-specific means and standard deviations for each stop. The magnitudes of the correlations were comparable to those in the previous study, but all reached significance in the connected speech analysis. As the Mixer 6 corpus contained many tokens (n = 88,725) and talkers (n = 180), strong statistical power may have led to an increase in the type I error rate (i.e., false positives). However, this concern was addressed with bootstrap confidence intervals, each of which provides a range of population correlations that is not associated with any null-hypothesis statistical test.

Interestingly, the strength of the correlations in isolated speech increased substantially after correcting for following vowel duration, particularly among the voiced stops and between homorganic stops. No improvement, however, was seen in the

connected read speech when either average word duration or following vowel duration were used to estimate speaking rate. Aspects of the isolated speech study, such as the homogeneous repetition of similarly-structured syllables, may have resulted in greater similarity in the realization of stop consonants, and thus stronger correlations after rate correction. Alternatively, it may be that strong correlations are indeed present in connected speech, but harder to estimate given the greater contextual and speaking rate variability.[20] In addition, the connected speech study depended on automatic alignment not only of the VOT, but also the individual words in each utterance and the following vowel durations necessary for the speaking rate measurement. Improved precision of these alignments may reveal stronger relations of VOT like those observed in isolated speech after rate correction. These differences notwithstanding, it is striking that a strong pattern of VOT covariation was present among the voiceless stops in spite of the many sources of variation in the connected speech corpus.

Systematic rankings of mean VOT were also observed in both studies, with the notable exception of variation in the talker-specific ranking of [t^h] and [k^h]. Specifically, in the connected speech study, there was a strong tendency for talkers to exhibit a slightly greater VOT for [t^h] in comparison to [k^h]. Many previous studies have focused on systematic rankings at the population level, and typically report a greater VOT for [k^h] in comparison to [p^h]. The present study observed a strong tendency for the ranking of [p^h] < [k^h], consistent with previous findings, and little difference between the means of [t^h]

[20] An additional analysis in which VOT was residualized not only with speaking rate (vowel duration) but also vowel height, vowel tenseness, the interaction between height and tenseness, number of syllables, utterance position of the word, and lexical frequency (described in section 3.2.3) resulted in significant correlations of talker means across all stop pairs; however, the change in magnitude was less substantial than in the laboratory speech analyses (*r*s = [p^h – t^h] 0.83, [t^h – k^h] 0.78, [k^h – p^h] 0.82, [b – d] 0.25, [d – g] 0.35, [g – b] 0.53, [p^h – b] 0.29, [t^h – d] 0.58, [k^h – g] 0.40, all *p*s < 0.006).

and [k^h] within or across talkers in both studies. Among the voiced stops, the overwhelming majority of speakers had increasing VOT with more posterior places of articulation ([b] < [d] < [g]). Ordinal rankings are not as informative, however, as linear fits: even consistent ordinal ranking does not entail a linear relation (as any magnitude of separation between VOT means could be consistent with a given ranking), and ordinal rankings are entailed by linear relations (within particular lower and upper limits). In almost all estimated fits between the voiceless stop VOTs, both the additive and scaling factors were significant, indicating that the difference between VOT means varied systematically. The exception in this case was the estimated fit between [t^h] and [p^h] in the isolated speech, for which only the scaling factor was significant.

In both studies the mixed-effects linear models revealed large variation across talkers in the grand mean VOT (intercept) for all six stops and in the degree of separation between voiced and voiceless stops (voice slope). Considerably less variation was observed in the realization of VOT across stop place of articulation. The mixed-effects model also accounted for other important sources of variation in the realization of VOT. For both isolated and connected speech there were significant effects of speaking rate on the VOT of voiceless stops, and while vowel height and tenseness failed to reach significance individually, a significant interaction was revealed for connected speech, implicating longer VOTs in the context of high tense vowels, [i] and [u] (see also Nearey & Rochet, 1994). In connected speech, utterance position was also a significant factor: compared to the average, stops in utterance-initial, post-pausal, and pre-pausal positions had longer VOTs, whereas stops in utterance-final position had shorter VOTs. The significant VOT lengthening found for utterance-initial stops is consistent with previous

findings of domain-initial strengthening at the beginning of the utterance (e.g., Cho & Keating, 2009). VOT tended to be shorter in polysyllabic than in monosyllabic words, however, this effect failed to reach significance. There was also a significant decrease in VOT with higher lexical frequency.

As in the isolated speech study, the evidence for target uniformity was strong, particularly for the aspirated stops, whereas evidence in support of contrast uniformity was quite weak. The primary difference between the two speech styles was in the strength of the correlations among the unaspirated and homorganic stops, which were much weaker in the present study than in the isolated speech study, even after correcting for speaking rate differences.

The large-scale analysis implemented here contributes to the understanding of VOT variation and covariation in a speech corpus with a greater number of talkers, larger variety of contextual, prosodic, and lexical factors, and greater amount of data than is typically collected in a laboratory experiment. Despite some measurement error, the automated alignment with P2FA and AutoVOT yielded a pattern that corresponded closely with the findings for isolated speech. Overall, the methods and analyses employed in this section extend our understanding of structured VOT realization to a connected speech style, and more generally advance research in corpus-based phonetics.

2.4 Child VOT production

Patterns of variation and covariation of VOT across adult American English speakers as observed in the previous studies revealed evidence for uniformity in a mature phonetic grammar. Uniformity may also have an early presence in language acquisition, or may develop over time, reaching maturity only in the adult grammar. The following

study investigated systematic relations in stop consonant VOT as produced by children aged 2 to 5 to determine whether uniformity influenced the phonetic grammar in acquisition.

Stop consonant production has been argued to follow a universal pattern of acquisition in that infants first develop the ability to produce unaspirated stop consonants. Whalen et al. (2007) examined the babbling of English and French infants between 9 and 12 months of age. In adult production, English stop consonants contrast in the presence or absence of aspiration, whereas French stop consonants do not have aspiration, and contrast in the presence or absence of voicing. Regardless of the exposure language, all infants produced unaspirated stop consonants with a short VOT. Nevertheless, the VOT of infants acquiring English was longer than the VOT of infants acquiring French, and this difference increased with the age of the child.

The aspiration involved in the production of word-initial English stops typically does not develop until approximately 2 years old (Port & Preston, 1972). Macken & Barton (1980) reported that the age at which the voicing contrast is acquired can range from just over 1 year old to up to almost 3 years old (Barton, 1976; Velten, 1943; Major, 1976; Smith, 1973), and most children acquire the contrast by 2;6 years old (Zlatin & Koenigsknecht, 1976; Gilbert, 1977). Several studies of American English have observed relatively longer VOT means for voiceless aspirated stops in children between 2 and 4 years old compared to adult VOT values (Gilbert, 1977; Menyuk & Klatt, 1975; Smith, 1978; Barton & Macken, 1980). In fact, Barton & Macken (1980) have posited an 'overshoot' phase in VOT production to account for the long VOT means reported for

children around 3 to 4 years old. In contrast, Koenig (2000) found no difference at least between 5 year olds and adults in their mean VOT values for [p^h] and [t^h].

Child production of VOT in voiceless aspirated stops has been characterized by a high degree of variability relative to adult productions, which may persist until puberty (Eguchi & Hirsh, 1969; Ostry et al., 1984; Ohde, 1985; Koenig, 2000). Koenig (2000) conducted a thorough examination of the source of this variation, and argued that children may have underdeveloped control over the laryngeal and aerodynamic factors involved in aspirated stop production, such as management of the glottal abduction degree, vocal fold tension, and transglottal pressure and flow. This contrasts with a likely alternative explanation in that variation arises from the complex timing control between glottal and supraglottal articulations. Evidence for the 'laryngeal management' argument comes from the fact that the duration not only of the aspirated stops (VOT) but also of [h] was more variable in 5 year olds than adults. In addition, the standard deviations of [h] duration and stop VOT for [p^h] and [t^h] were correlated across children. As [h] does not have a supraglottal articulation and [h] appears to pattern with the aspirated stops, these findings oppose the complex timing explanation.

To some extent, dependencies among stop categories have also been observed in child VOT. The well-documented VOT ranking across place of articulation was also found in unaspirated stops produced by English- and French-babbling infants (Whalen et al., 2007). Furthermore, Koenig (2000) found significant correlations of talker-specific VOT medians and maxima between [p^h] and [t^h] across 5 year olds and adult talkers (median VOT $r = 0.78$; maximal VOT $r = 0.79$, ps < 0.05); however, the extent to which this correlation held across the 5 year olds alone was not reported.

The following section examines the predictions of target uniformity in the VOT of [t^h] and [k^h] across children ages 2 to 5 years old. In addition to analysis of the VOT parameters for each age group, a correlation analysis, simple linear regression, and mixed-effects model were implemented to determine the degree to which children converge on a uniform phonetic target in the production of the [+spread glottis] stops. These analyses shed light on when structure may begin to emerge in the phonetic realization of stop consonants.

2.4.1 Methods

The analysis employed the PhonBank Paidologos Corpus of English-speaking Children's Productions, which contains isolated speech produced by 81 children (40 female) aged 2;0 to 5;11 (Edwards & Beckman, 2008). The transcript for each recording was aligned to the audio with the Penn Forced Aligner (Yuan & Liberman, 2008). The resulting word-initial stop boundaries were then extended by 40 ms in each direction to create an interval of analysis for AutoVOT. Of the word-initial, voiceless aspirated stop consonants, only [t^h] and [k^h] were elicited in the corpus and therefore available for analysis. All stop boundaries were manually corrected to align with the burst release, marked by the transient in the waveform, and the onset of voicing, marked by the start of periodicity in the waveform or the presence of the voice bar in the spectrogram. Any instance in which the child did not produce the intended stop-initial prompt was excluded from analysis; however, as long as the child produced a recognizable pronunciation of the prompt, then the stop consonant was retained as the *intended* stop category. For example, if the transcriber happened to perceive an intended [k^h] as [t^h], this was still coded as [k^h].

In total, there were 2,057 stops for analysis, and a median of 12 tokens of [t^h] (range: 5 to 17) and 14 tokens of [k^h] per child (range: 7 to 18). There were 29 unique words: 14 beginning with [t^h] and 15 beginning with [k^h]. (The word 'twisted' was excluded from analysis.)

2.4.2 Results

The population VOT means and standard deviations for [t^h] and [k^h] for each age group are reported in Table 2.10, along with the range of talker-specific means and standard deviations. The combined VOT means for [t^h] and [k^h] were overall high, but nevertheless lower than the adult VOT means observed in isolated speech (section 2.2.2). As shown in

Table 2.11, talker-specific means and standard deviations were also moderately correlated for both [t^h] and [k^h], again paralleling previous findings in adult VOT production (section 2.2.2.1). Furthermore, the pattern of these results is present as young as 2 years old, and comparable for each age group.

Table 2.10. Means and standard deviations in milliseconds for each age group and overall.

	[t^h]			**[k^h]**		
Age	Mean (SD)	Range of Talker Means	Range of Talker SDs	Mean (SD)	Range of Talker Means	Range of Talker SDs
2	97 (28)	49 – 150	15 – 58	96 (24)	59 – 143	22 – 93
3	89 (21)	61 – 136	17 – 65	86 (14)	61 – 124	21 – 59
4	80 (15)	49 – 128	17 – 46	79 (19)	48 – 123	10 – 83
5	87 (25)	53 – 161	14 – 63	84 (23)	44 – 153	9 – 45
Combined	88 (23)	49 – 161	14 – 65	86 (21)	44 – 153	9 – 93

Table 2.11. Correlations of talker means and corresponding standard deviations.

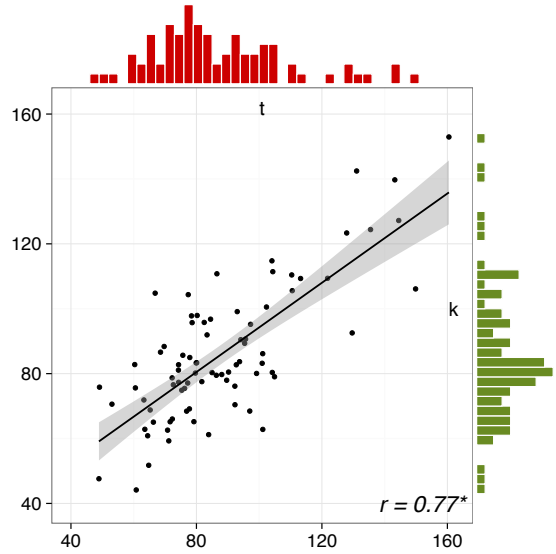
Age	[t^h]	[k^h]
2	0.64, $p < 0.001$	0.62, $p < 0.001$
3	0.65, $p < 0.01$	0.41, $p = 0.07$
4	0.45, $p < 0.05$	0.60, $p < 0.01$
5	0.52, $p < 0.05$	0.71, $p < 0.001$
Combined	0.64, $p < 0.001$	0.62, $p < 0.001$

In addition, the talker-specific mean VOTs were highly correlated between [t^h] and [k^h] across 2 to 5 year olds, as shown in Table 2.12 and Figure 2.4. The correlation was slightly lower than the adult VOT correlation for [t^h] and [k^h] in isolated speech reported in section 2.2.2.1 ($r = 0.98$), but still strong at $r = 0.77$. A strong correlation was observed within each age group except for the 4 year olds, in which only a moderate correlation was observed. It is unclear whether the relatively weaker correlation for the 4 year olds is systematic across talkers of that age, or simply an artifact of this particular dataset.

Table 2.12. Correlations of [t^h] vs. [k^h] for each age group and overall.

Age	[t^h] – [k^h]
2	0.79
3	0.73
4	0.59
5	0.81
Combined	0.77 [0.62, 0.85]

Figure 2.4. Variation and covariation of VOT means (ms) across talkers. Marginal histograms show variation in talker means. Each point is a talker-specific mean. The asterisk indicates that the correlation reached significance ($p < 0.025$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



Variation in VOT was analyzed with a linear mixed-effects model with fixed effects of the place of articulation, speaking rate, vowel height, vowel tenseness, age, gender, and the interaction between height and tenseness. In addition, the random effects component for the talker included a random intercept and place slope. All fixed effects were weighted-effect coded in the following manner: place of articulation (*coronal* 1, *dorsal* -0.85), age (*age2*: two = 1, five = -0.96; *age3*: three = 1, five = -0.92; *age4*: four = 1, five = -1.05), gender (female = 1, male = -0.98), vowel height (high [i ɪ u ʊ] = 1, non-high [eɪ ɛ ʌ oʊ ɔ a] = -0.51), and vowel tenseness (tense [i eɪ ʌ u oʊ ɔ a] = 1, lax [ɪ ɛ ʌ ʊ] = -2.96). Speaking rate was coded as the following vowel duration, as estimated from the P2FA boundaries (e.g., Theodore et al., 2009). Both the continuous factors of speaking rate and VOT were centered on the grand mean (speaking rate: 174 ms, VOT: 86 ms).

While place of articulation was not significant (β = -0.51, t = -0.27), there was a significant effect of speaking rate: longer vowel durations corresponded to longer VOTs

when compared to the VOT at mean vowel duration (β = 0.04, *t* = 5.32). A slow speaking rate may also reflect a degree of hyperarticulation in the child speech. Vowel height did not have a significant effect on VOT alone (β = 4.62, *t* = 1.86); however, vowel tenseness and the interaction between height and tenseness were significant (*tenseness*: β = 3.44, *t* = 2.85; *height* × *tenseness:* β = 3.14, *t* = 2.18). These results indicate that VOTs were longer when the stop consonant preceded tense vowels, and specifically the high tense vowels, [i] and [u]. In comparison to the mean, 2 year olds had significantly longer VOTs (β = 10.43, *t* = 2.84), whereas 4 year olds had significantly shorter VOTs (β = -7.87, *t* = -2.25). The VOT of 3 year olds did not differ significantly from the mean (β = 0.15, *t* = 0.04). In addition, there was no significant effect of gender (β = -0.15, *t* = -0.07). Analysis of the talker random effects revealed that standard deviation of the talker intercept was much larger than the standard deviation of the random slope for place (sd intercept: 17.87; sd slope: 3.56), indicating that even young talkers differed most in their overall mean VOT values as opposed to in any separation between [t^h] and [k^h].

2.4.3 Discussion

The VOT means for [t^h] and [k^h] as produced by children aged 2 to 5 years old do not differ qualitatively from adult VOT means in isolated speech (section 2.2.2). A future analysis, however, should consider differences that could arise once the effect of speaking rate is taken into account for both the adult and child VOT means. Consistent with many previous studies, however, the standard deviations were much higher for children than adults for both [t^h] and [k^h] VOTs (e.g., Koenig, 2000). Specifically, the median standard deviations for the children were higher than the maximum observed adult standard deviations for isolated speech (child median [t^h] SD: 29 ms, adult

maximum [t^h] SD: 26 ms; child median [k^h] SD: 30 ms, adult maximum [k^h] SD: 20 ms; data from section 2.2.2). Overall, the 2 year olds had higher VOTs and the 4 year olds shorter VOTs than average. As in adult VOT patterns, both speaking rate and a following high, tense vowel had significant influences on VOT.

Strong correlations of talker mean VOT between [t^h] and [k^h] were observed for each age group, but these were not quite as high as the correlations observed in adult isolated speech, which were at or above $r = 0.95$. The relatively lower correlation coefficient for child than adult speech could be due to the higher variability present in child VOT production. Determining whether variability in production could affect the correlation of talker means should be addressed in a future analysis via data simulation. In addition, the correlation was relatively weaker for the 4 year olds in comparison to all other age groups. Analysis of additional data from the 4-year-old population could address this concern.

Lastly, the mixed-effects model revealed no significant difference between the VOTs of [t^h] and [k^h]. This could be an indicator of target uniformity, but it is not clear whether a single timing between the supraglottal constriction and glottal spreading gesture would give rise to the same VOT. For a uniform timing relationship to give rise to the same VOT across [t^h] and [k^h], then the closure duration for each place of articulation would have to be approximately the same. Regardless, the differences in mean VOT between [t^h] and [k^h] were minimal and non-significant. Taken together with the strong linear relationship between the two categories, these findings provide evidence for target uniformity: for segments sharing the same laryngeal feature value, a high

degree of similarity was observed in the acoustic correlate of the underlying phonetic targets.

2.5 Covariation of VOT across languages

The previous sections supported an influence of target uniformity among stop consonants with a shared feature value in American English adult and child speech. As a constraint on the phonetic grammar, uniformity should also exert a clear influence on the phonetic patterning of speech sounds cross-linguistically. Provided there is a common laryngeal feature value among stop segments, target uniformity ensures a high degree of similarity among those targets, regardless of how they are specified. The present study examined the strength of VOT covariation and the predictions of target uniformity on the realization of a shared laryngeal feature value in a meta-analysis of approximately 60 languages.

2.5.1 Methods

A large-scale literature review was conducted to identify previously reported VOT values from a variety of languages. For a list of the studies from which data was collected, please see the Appendix (Table 6.1). The focus of the current data collection was on studies reporting adult, L1 speech patterns. Studies which examined child speech, L2 speech, or bilingual speech were excluded from the analysis. VOT data points from a total of 58 languages and 24 different language families were collected. There were 70 unique primary sources from which these were obtained. Language family and genus data were identified through the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013). If the language was missing from WALS, the language family and genus were located separately.

For comparison across studies, the original laryngeal classifications were categorized into one of three types: short-lag, long-lag, and voiced. The difference between short-lag and long-lag was defined by VOT: for each set of stop segments differing only in place of articulation (e.g., derived from the same study, language, and context, and sharing the same laryngeal specification), if the longest VOT value was between 0 and 50 ms, then the stops were coded as short-lag.[21] If the longest VOT value was above 50 ms, then the stops were coded as long-lag. Note that ejective stops were retained in the analysis (n=44) with all but two stops coded as long-lag. Finally, sets of stops that contained a negative VOT value were coded as voiced. Stop consonant place of articulation was broadly coded as either labial, coronal, or dorsal using standard assumptions regarding terminology for place of articulation (e.g., bilabial corresponded to labial place, dental and alveolar to coronal place, and velar to dorsal place).[22]

In total, there were 277 labial stops, 243 coronal stops, and 305 dorsal stops. Table 2.13 presents the number of stop pairs for each laryngeal specification entered in the correlation analyses. Multiple sets of data points were present for many of the languages represented in the meta-analysis; these were retained for a more complete picture of cross-linguistic and cross-talker variation. The list of language families, represented languages, and number of instances is presented in Table 2.14.

[21] While a VOT of 50 ms may in some cases reflect moderate aspiration, this value was chosen as the cut-off point in part to increase the range of variation for the correlation analysis among stops with short, positive VOT values. It was also a fairly medial value among the positive VOTs.
[22] Uvular stops were not retained in the present analysis.

Table 2.13. Number of VOT pairs with each laryngeal specification and in total.

Pair	Short-lag	Long-lag	Voiced	Total
labial – coronal	100	84	38	222
coronal – dorsal	100	100	39	239
dorsal – labial	120	105	40	265

Table 2.14. The language families, represented languages within each language family, and the number of stops per language family.

Language Family	Represented Languages	Number of stops
Afro-Asiatic	*Dahalo, Hebrew*	15
Altaic	*Turkish*	12
Austro-Asiatic	*Remo*	3
Austronesian	*Tsou, Yapese*	20
Chapacura-Wanham	*Wari'*	5
Dravidian	*Tamil, Telugu*	46
Eskimo-Aleut	*Aleut (Eastern), Aleut (Western)*	6
Ijoid	*Defaka*	5
Indo-European	*Armenian (Eastern), Bengali, Catalan, Danish, Dutch, English, French, Gaelic, Greek, Hindi, Italian, Norwegian, Polish, Portuguese (Brazilian), Portuguese (European), Serbian, Spanish, Swedish*	363
Japanese	*Japanese*	12
Kartvelian	*Georgian*	18
Korean	*Korean*	39
Mayan	*Tzutujil*	6
Muskogean	*Chickasaw*	5
Na-Dene	*Apache (Western), Hupa, Navajo, Tlingit*	45
Nakh-Daghestanian	*Udi*	6
Niger-Congo	*Bowiri, Shekgalagari, Zulu*	41
Oto-Manguean	*Mazatec (Jalapa)*	6
Quechuan	*Quechua (Bolivian), Quechua (Cuzco), Quichua*	18
Salishan	*Montana Salish*	20
Sino-Tibetan	*Burmese, Cantonese, Galo, Hakka, Khonoma Angami, Mandarin*	95
Tai-Kadai	*Tai Khamti, Thai*	22
Ticuna	*Ticuna*	6
Uralic	*Hungarian*	11

2.5.2 Results

Substantial cross-linguistic variation was observed in the realization of VOT, but the observed variation for one place of articulation was not independent of the variation for other places of articulation (Figure 2.5). Rather, the place-specific VOT means were almost perfectly correlated across languages. With all three laryngeal classifications included, correlations ranged from $r = 0.97$ to $r = 0.98$ (ps < 0.001). The correlations remained strong even when calculated only over stops with positive VOT values (labial-coronal: $r = 0.94$, coronal-dorsal: $r = 0.91$, labial-dorsal: $r = 0.92$, ps < 0.001). Table 2.15 presents the correlation coefficients within each type of laryngeal specification. Covariation was quite strong among the long-lag and voiced stops, and moderate among the short-lag stops. The correlations between short-lag stops may have been weakened by the truncated VOT range.

Figure 2.5. Variation and covariation of VOT means (ms) across languages. Marginal histograms show variation in language means. Each point is a language-specific mean. Blue points correspond to voiced stops, green points to short-lag stops, purple points to long-lag stops. The asterisk indicates that the correlation reached significance ($p < 0.001$). Gray shading reflects the local confidence interval around the best-fit linear regression line.

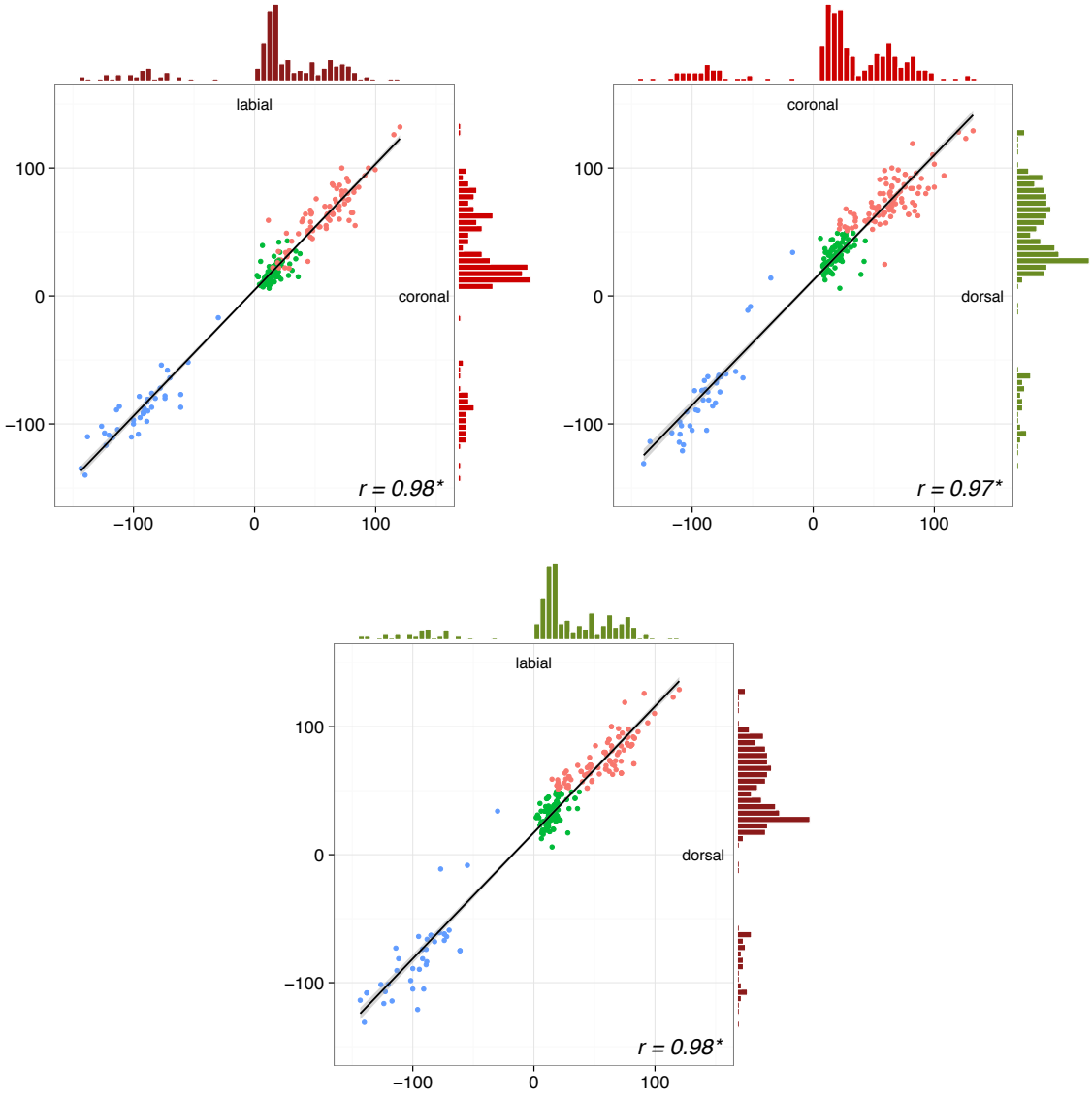


Table 2.15. Correlations of VOT means within each voicing type. All *p*s < 0.001.

Pair	Short-lag	Long-lag	Voiced
labial – coronal	0.55	0.88	0.88
coronal – dorsal	0.47	0.77	0.93
dorsal – labial	0.47	0.83	0.81

The correlation analysis revealed a strong linear association between VOT means; the form of this relationship was further investigated with simple linear regressions. Three simple linear regression were fit to all VOT means with one place of articulation serving as the explanatory variable and the second place of articulation as the dependent variable. As expected from the high correlation coefficients, each of the linear fits had an adjusted R^2 of 0.95 or higher, indicating that 95% of the variation or more could be accounted for by the model (cor ~ lab: adj. $R^2 = 0.97$, dor ~ cor: adj. $R^2 = 0.95$, dor ~ lab: adj. $R^2 = 0.95$). The linear fits strongly suggested that the relationships between VOT means were purely additive given the proximity of the scaling factor (β_1) to unity (cor ~ lab: $\beta_0 = 4.70$, $\beta_1 = 0.98$, dor ~ cor: $\beta_0 = 12.35$, $\beta_1 = 0.98$, dor ~ lab: $\beta_0 = 17.28$, $\beta_1 = 0.99$, ps < 0.001). The values of the estimated intercepts indicated that the coronal VOT mean could be estimated by adding approximately 5 ms to the labial VOT mean, and that the dorsal VOT mean could be estimated by adding approximately 17 ms to the labial VOT mean and 12 ms to the coronal VOT mean.

As the cluster of voiced VOT means may have deceptively magnified the strength of the linear relationship, three additional linear regressions were modeled using only the positive VOT data. The adjusted R^2s were between 0.80 and 0.88 (cor ~ lab: adj. $R^2 = 0.89$; dor ~ cor: adj. $R^2 = 0.82$; dor ~ lab: adj. $R^2 = 0.85$). The relationship between labial and coronal means was primarily additive, with an offset of approximately 5 ms and a scaling factor close to unity (cor ~ lab: $\beta_0 = 5.44$, $\beta_1 = 0.97$, ps < 0.001). For each of the stop pairs, both the additive and scalar factors contributed in modeling the relationship and indicated that the difference between coronals and dorsals, and between labials and dorsals decreased as the overall VOT value increased, at least within the range of positive

VOT values observed in the dataset (dor ~ cor: $\beta_0 = 18.11$, $\beta_1 = 0.85$, ps < 0.001; dor ~ lab: $\beta_0 = 20.99$, $\beta_1 = 0.89$, ps < 0.001).

2.5.3 Discussion

Strong linear relationships of VOT means were identified among stop consonants with a shared laryngeal feature value across languages from 24 diverse language families. These relations held across stop consonants with a variety of laryngeal classifications, and across both positive and negative VOT values. The linear regressions revealed an additive relationship between stop place of articulation when all VOT values were included. The predicted mean VOT differences with respect to the labial place of articulation were about 5 ms and 12 ms for the coronal and dorsal places, respectively. Among the positive VOT values, the predicted mean VOT difference between any two places of articulation was determined by both additive and scalar factors, which largely indicated that within the range of observed VOTs, the difference between VOT means generally decreased with higher VOT values.

The observed covariation among VOT means reveals a strong influence of target uniformity on phonetic grammars universally. Within a language, stop consonants that share a laryngeal feature value are realized in a highly systematic manner that demonstrates interdependence not only between the phonological segments but also the phonetic realizations. Specifically, the similarity between VOT values, along with the biomechanical explanations of the place differences, reveals a constraint of uniformity on the mapping from a shared laryngeal feature value to the phonetic target.

A uniformity constraint that limits variation across stop targets sharing a laryngeal feature value restricts theoretically permissible variation in the phonetic grammar (e.g.,

Ladefoged, 1988; Cho & Ladefoged, 1999). Following Ladefoged (1988), Cho & Ladefoged (1999) posited context-sensitive phonetic targets in which a VOT value would be specified for the combination of the laryngeal and place features, especially for long-lag stops. Without any further restriction, context-sensitivity in the phonetic implementation of segments can allow for independent laryngeal targets for each place of articulation. The limited variation across VOT values within a language, and especially the high degree of *covariation* across languages, reveals that the phonetic targets corresponding to a shared feature are *not* independent of one another. Target uniformity explicitly constrains the phonetic targets of segments with a shared laryngeal value to be highly similar or uniform, thus accounting for the observed linear dependencies across a highly diverse set of languages.

2.6 Generalized perceptual adaptation to talker-specific VOT

Prior knowledge of the linear relationships between stop categories may facilitate perceptual adaptation to and generalization of talker-specific pronunciations across phonetic categories. If a listener has heard a novel talker produce [p^h] but not [k^h], prior knowledge of how the VOT of [p^h] and [k^h] covary may allow the listener to form reasonable expectations about the talker's VOT for [k^h]. In essence, the means of stops for which the listener has little talker-specific evidence can be 'read off' the regression lines, as depicted in Figure 2.2 and Figure 2.3. Knowledge of phonetic covariation would likely be most beneficial when direct evidence regarding the talker's speech is limited. Evidence from a single phonetic category could be used to update talker-specific parameters for the perceived category, as well as many related categories simultaneously, improving both the speed and precision of adaptation.

Some evidence that knowledge of VOT correlations plays a role in talker adaptation has been provided by previous studies of perceptual generalization and phonetic imitation. Eimas & Corbit (1973) established that repeated exposure to a long VOT value of either [p^h] or [t^h] resulted in an upwards shift in listeners' VOT voicing boundary for the trained place of articulation, as well as the unheard place of articulation. In this scenario, the listener may have inferred a precise estimate of either a relatively high mean or low standard deviation for the trained place of articulation, but critically, this knowledge generalized across place of articulation. Moreover, the direction of the shift is consistent with the positive correlation between [p^h] and [t^h].

Evidence for perceptual generalization across stop place of articulation has also been identified with lexically-induced perceptual learning. In a lexical decision task, Kraljic & Samuel (2006) exposed listeners to words with an ambiguous /t/-/d/ sound in medial position. The lexical properties of the word biased listeners to either a [t^h] or [d] interpretation of the VOT value. Listeners not only shifted the /t/-/d/ VOT boundary to accommodate the relatively low VOT [t^h] or relatively high VOT [d], but also generalized this shift to a [p^h]-[b] continuum.

Relatedly, Theodore & Miller (2010) demonstrated that listeners transfer acoustic-phonetic detail from one place of articulation to another at a talker-specific level. Listeners were trained on two talkers who differed only in their mean VOT for [p^h], one talker with characteristically "short" VOTs and one with "long" VOTs. After exposure to the [p^h]-initial stimuli, listeners could identify in a two-alternative forced choice task that a long VOT for [k^h] was more characteristic of the talker with the long VOT for [p^h], and

correspondingly, the short VOT for [k^h] was more characteristic of the talker with the short VOT for [p^h].

Finally, listeners generalized a talker's characteristically long VOT from [p^h] to [k^h] in phonetic imitation, without any prior exposure to that talker's [k^h] (Nielsen, 2007, 2011). After exposure to a lengthened VOT for [p^h], participants lengthened their VOT not only for [p^h], but also for [k^h]. Conversely, a reduced VOT for [p^h] did not result in any imitation to either the trained or untrained place of articulation. Generalization may have been inhibited by the natural lower limit of VOT for English voiceless stop consonants, as the reduced VOT may impinge too greatly on the stop voicing boundary. A similar lack of perceptual generalization has been observed when the VOTs of /p/ and /k/ were reduced to values typical of voiceless unaspirated stops as in French and Spanish (Clarke & Luce, 2005). In this case, listeners failed to associate the lowered VOT of /p/ or /k/ to the talker at hand, let alone generalize the lowered VOT to the unheard stop category.

While these studies provide critical evidence of generalized VOT adaptation, there were several limitations to these studies, including extensive exposure to the novel talker, limited stimulus variability, or highly exaggerated VOT differences. In both Eimas & Corbit (1973) and Kraljic & Samuel (2006), the exposure phase was relatively short (Eimas & Corbit: two minutes of 120 repetitions; Kraljic & Samuel: 40 critical exposure items), but involved only a single VOT value. Theodore & Miller (2010) alternated exposure and testing phases, but the results did not reveal the time course of generalization. By the end of the experiment, listeners had been heard 80 instances of [p^h] for each talker (2 days x 10 familiarization and training phases x 4 stimuli per talker).

Moreover, the study involved only two unique VOT values per stop category and talker, and the difference between the short- and long-VOT talker was quite exaggerated (e.g., 88 ms for the short VOT talker vs. 183 ms for the long VOT talker). The observed generalization may therefore have been due to a highly salient experimental manipulation. The exposure stimuli in Nielsen (2011) had variable and natural VOT values (e.g., mean = 112 ms, SD = 12 ms), but listeners also heard 80 instances of the talker's [p^h] prior to the test phase.

The present study examined prior listener knowledge of VOT covariation among the aspirated stops while addressing many of the limitations of previous studies. In particular, the study employed more natural and variable stimuli, and generalized adaptation was examined after minimal exposure to the talker to investigate the time course of adaptation.

2.6.1 Methods

2.6.1.1 Participants

Forty-eight participants were recruited from the Johns Hopkins University undergraduate community and were divided into two groups (Test [k^h] and Test [p^h]) and two conditions within each group (Train Long VOT or Train Short VOT). There were 12 participants in each condition (Train Long – Test [k^h]: 11 female; Train Short – Test [k^h]: 8 female; Train Long – Test [p^h]: 8 female; Train Short – Test [p^h]: 7 female). An additional four participants completed a similar version of the experiment, but were excluded after changes were made to the experiment design. All participants were compensated with course credit.

2.6.1.2 Stimuli

To address limitations of previous generalization studies, stimuli were generated from natural VOT distributions with substantial variability. All stimuli were created from careful-speech productions of CVC syllables in a previously collected laboratory corpus with 24 talkers sampled at 48 kHz (Chodroff & Wilson, 2014). The syllables were composed of one of six stop consonants [p t k b d g] crossed with 9 vowels [i eɪ ɛ æ ʌ ɑ ɔ oʊ u/ and a final /t/. A gamma density, which closely approximates the shape of natural VOT distributions, was fit to each of the voiceless stop tokens from two of the male talkers: one with naturally long VOT values and one with naturally short values. The designation of short and long VOTs was determined relative to the observed stop means across talkers in the laboratory speech corpus of 24 talkers. Each voiceless stop mean for the long VOT talker was within the third quartile of talker means, and for the short VOT talker within the first quartile of talker means. The observed Gaussian and gamma parameters for the two selected talkers are provided in Table 2.16.

Table 2.16. Gaussian and gamma parameters fit to the isolated laboratory speech productions of the long and short VOT talkers. The mean and standard deviation (SD) were used to select the two talkers, and the shape and rate of the gamma distribution were used to generate VOT values. All values are in milliseconds.

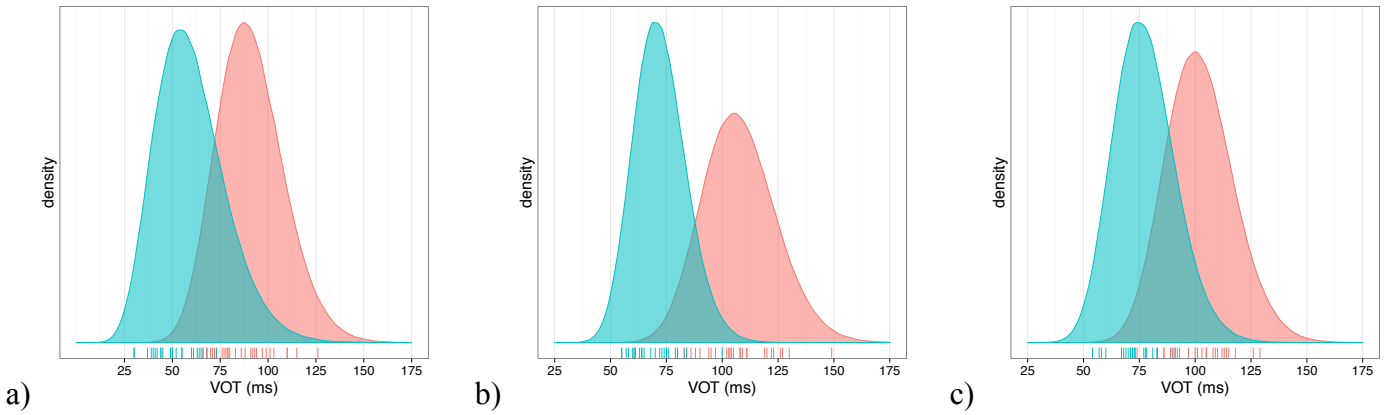
	Long VOT talker		Short VOT talker	
Stop	Mean (SD)	Shape (rate)	Mean (SD)	Shape (rate)
[p^h]	91 (17)	26.96 (0.30)	60 (18)	10.66 (0.18)
[t^h]	108 (16)	43.08 (0.40)	72 (12)	37.87 (0.53)
[k^h]	102 (15)	44.86 (0.44)	77 (14)	31.28 (0.40)

For each voiceless stop and each vowel, three repetitions of the first male talker's productions (nicknamed “Mike” in the experiment) were selected. The stop duration was manipulated to match values randomly generated from the long and short gamma distributions for each stop consonant. The duration was manipulated by removing or

copying portions of the aspiration period. To avoid clipping, all splices preserved continuity in the waveform and were made at zero-crossings. In addition, repeated portions of the aspiration period were varied in duration to avoid periodicity in the aspiration segment. The result was 162 stimuli (3 stops x 9 vowels x 3 repetitions x 2 VOTs).

For any given production, we ensured that the short VOT value was indeed shorter than the long VOT value; however, we tried to adhere to the randomly generated values as best as possible. In the Test [k^h] group, the differences between the long and short VOT [k^h]s ranged from 2 to 60 ms with a mean difference of 29 ms. For the Test [p^h] group, the differences between the test [p^h]s ranged from 7 to 84 ms with a mean difference of 36 ms. Note that there was overlap in the VOT values of the short and long exposure conditions; however, participants were exposed to only one of those distributions.

Figure 2.6. Gamma distributions fit to the short (blue) and long (red) VOT talkers. The randomly generated VOT values for the corresponding conditions are plotted in the rug below the distributions. In a) are the distributions for [p^h], in b) the distributions for [t^h], and in c) the distributions for [k^h].



2.6.1.3 Procedure

There were four conditions in the experiment. For each condition, two of the aspirated stops ([p^h t^h] or [t^h k^h]) were selected as exposure categories and both assigned relatively long or short VOT levels. Within a trial, one instance of each of the training stops with the appropriate VOT level was presented, and then the participant performed a two-alternative forced choice task for the untrained stop (i.e., [k^h] or [p^h]). The choices differed only in VOT (long vs. short), and participants were asked to select the one that "sounded most like Mike". The two-alternative forced choice task resembled the design in Theodore & Miller (2010) with two major differences: exposure and testing alternated within a single trial, rather than alternating in blocks, and participants learned about a single talker's voice, as opposed to two talkers' voices. The vowel category was held constant within a trial, and testing stimuli were counterbalanced such that the long VOT option was presented first in exactly half of the trials. Separating the two exposure stimuli was a 1500 ms ISI; between exposure offset and test onset, a 2000 ms ISI, and between the two test stimuli, a 1000 ms ISI. There was a period of 1500 ms between trials. Each participant completed 6 blocks with 27 unique trials per block. Instructions were presented by the experimenter during the first trial which was self-paced by the participant. The experiment was presented using PsychoPy (Peirce, 2007) in a sound-attenuated booth and Sennheiser HD 518 headphones.

2.6.2 Results

Inspection of the results revealed a strong bias to respond with the first stimulus regardless of choice order or condition (Yeshurun et al., 2008; Garcia-Perez & Alcala-Quintana, 2011). Two analyses were performed to determine whether, despite this bias,

participants showed generalization of the VOT level to the untrained stop. The logistic mixed-effects analysis predicted the response to the first test option, and included an intercept (representing the bias to choose the first option), effect of exposure group (long: +1, short: -1), effect of the test VOT ratio (log(VOT #1/VOT#2)), and the interaction between condition and VOT ratio.[23] The interaction between condition and VOT ratio indicates whether the first option was congruent (positive) or incongruent (negative) with the training stops (e.g., when exposure involved long VOT [p^h t^h] and the long VOT [k^h] was the first option, then selecting the first option would be congruent with exposure). There was also a random intercept for the base word (VC portion of syllable) and a random intercept and VOT ratio slope for participant.

For both the Test [k^h] and Test [p^h] groups, there was a significant bias to respond with the first test item as revealed in the positive intercept (Test [k^h]: $\beta_0 = 0.36$ $p < 0.001$; Test [p^h]: $\beta_0 = 0.27$ $p < 0.05$). Despite this bias, listeners significantly generalized talker VOT in both groups (Test [k^h]: $\beta_{condxvot.ratio} = 0.28$ $p < 0.001$; Test [p^h]: $\beta_{condxvot.ratio} = 0.36$ $p < 0.001$). In the Test [k^h] model, the main effects of the log VOT ratio and condition did not reach significance (*vot.ratio*: $\beta = -0.09$ $p = 0.10$, *cond*: $\beta = 0.01$ $p = 0.87$). In the Test [p^h] model, the log VOT ratio did reach significance, indicating a slight bias to choose the first option when it was long, regardless of condition ($\beta = 0.13$ $p < 0.05$), but condition alone did not significantly affect whether a listener chose the first or second option ($\beta = 0.20$ $p = 0.05$).

[23] The difference of the log VOT values (equivalent to the log of the VOT ratio) provided the best quantitative account of the congruency effect on the choice responses, but similar results were also observed with the VOT difference.

Generalization occurred rapidly, and a significant interaction between condition and VOT ratio was observed for each group after the first block, or 54 exposure stimuli (Test [k^h]: $\beta_{condxvot.ratio} = 0.25$ $p < 0.01$; Test [p^h]: $\beta_{condxvot.ratio} = 0.43$ $p < 0.001$). In the Test [k^h] model with only one block of exposure, neither the log VOT ratio or condition had a significant effect on the choice option (*vot.ratio*: $\beta = -0.09$ $p = 0.10$, *cond*: $\beta = 0.01$ $p = 0.87$). However, in the Test [p^h] model with only one block of exposure, both the log VOT ratio and condition significantly affected the choice option: when the first VOT was long, listeners were more likely to choose the first option, and if the listeners were in the long VOT exposure group, they were also more likely to choose the first option (*vot.ratio*: $\beta = 0.23$ $p < 0.05$, *cond*: $\beta = 0.26$ $p < 0.05$).

These findings of generalized adaptation were supported by a second analysis assessing response bias (log β) and sensitivity (*d'*) for each participant (Wickens, 2002). Consistent with the results from the logistic mixed-effects models, participants had a significant bias to choose the first option (Train Long – Test [k^h]: 0.35 $p < 0.05$; Train Short – Test [k^h]: 0.41 $p < 0.01$; Train Long – Test [p^h]: 0.60 $p < 0.01$; Expose Short – Test [p^h]: 0.34 $p < 0.01$). Sensitivity to the difference between long and short VOTs for the test stop was also significantly different from chance ($d' = 0$) in all but the Train Short – Test [p^h] condition (Train Long – Test [k^h]: 0.22 $p < 0.01$; Train Short – Test [k^h]: 0.41 $p < 0.01$; Train Long – Test [p^h]: 0.53 $p < 0.001$; Train Short – Test [p^h]: 0.26 $p = 0.06$).

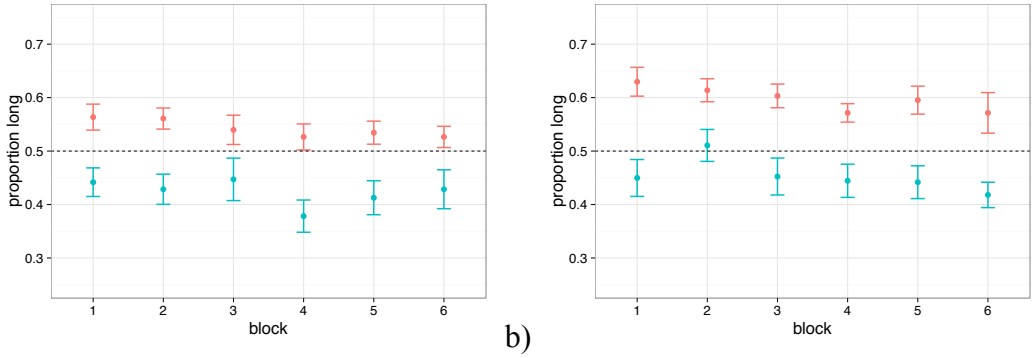
We also tested the hypothesis that listeners responded in a manner consistent with the *ordinal rankings* of the stop categories ([p^h] < [k^h]), but not necessarily the linear relationship between stop VOT means. This hypothesis makes the strongest predictions

for the Train Short – Test [k^h] and Train Long – Test [p^h] conditions because the ordinal and linear predictions are most likely to differ within each trial. Specifically, the linear relationship hypothesis predicts that after exposure to short [p^h t^h], listeners should choose the short [k^h] VOT, whereas the ordinal ranking hypothesis predicts that listeners should choose any [k^h] VOT that is longer than the observed [p^h] VOT.

For the Train Short – Test [k^h] group, the ideal response is the long [k^h] option if the VOT of the exposure [p^h] is longer than the VOT of the short [k^h]. Otherwise, both the long and short VOT [k^h] options are equally good and are chosen at chance (50/50). Given these criteria, listeners should choose the long VOT option *minimally* 54% of the time in the experiment. However, the observed rate of selecting the long option in the Train Short – Test [k^h] condition was significantly different and critically *less* than the expected rate at only 42% ($\chi^2(1) = 118.9$ $p < 0.001$).

For the Train Long – Test [p^h] group, the ideal response is the short [p^h] option if the VOT of the exposure [k^h] is shorter than the VOT of the long [p^h] option. Listeners would then be expected to choose the short [p^h] option minimally 61% of the time. The short option in the Train Long – Test [p^h] condition was selected 40% of the time, which as before, was significantly different and substantially less than the expected rate given the ordinal ranking hypothesis ($\chi^2(1) = 393.2$ $p < 0.001$). Listeners chose the long [p^h] option more often than would be expected if the listener was trying to ensure that the VOT of the selected [p^h] was shorter than the observed [k^h] VOT. This response pattern thus suggests that listeners may be exploiting knowledge of the *linear* VOT relationships between stop categories.

Figure 2.7. a) Proportion long response in the Test [k^h] VOT group. b) Proportion long response in the Test [p^h] VOT group. Long VOT exposure conditions are in red and short VOT exposure conditions in blue. Error bars reflect ±1 standard error of the proportion.



2.6.3 Discussion

Listeners generalized talker-specific VOT across stop place of articulation for both the long and short VOT distributions. The present findings are consistent with Theodore & Miller (2010), which also observed generalization of relatively long and short VOT values in a perceptual task with more extreme VOT manipulations. Previous studies have found no perceptual learning or generalization when listeners were asked to produce a shortened VOT (e.g., Nielsen, 2011) or voiceless VOT values were more consistent with unaspirated stops (e.g., Clarke & Luce, 2005). It may be that shortening VOT is more difficult than lengthening VOT in production, the VOT values must be identified as standard for the language, or some combination of factors. Nevertheless, the present findings provide support of perceptual generalization of VOT values that are relatively short for isolated speech productions.

In addition, generalization was also observed with less exaggerated and more variable VOT values than several of the previous VOT generalization experiments. The manipulation was therefore more representative of natural talker variation, but

commensurately weaker adaptation effects were obtained relative to Theodore & Miller (2010), in which listeners selected the VOT value congruent with exposure approximately 80% of the time (cf., 55%-60% before bias-correction in the present experiment). The data indicate substantial perceptual noise (see also Kronrod et al., 2012) and a strong response bias. Remaining questions include how much noise is present in VOT perception and how the observed response bias relates to perceptual noise. It is also unclear why there was a strong primacy bias when other sequential two-interval experiments have found recency biases (Yeshurun et al. 2008; Garcia-Perez & Alcala-Quintana 2011).

Finally, perceptual generalization occurred rapidly in the present experiment, with listeners significantly more likely to select the VOT congruent with exposure after only a single block. A block was comprised of 27 trials and lasted approximately six minutes. This sheds light on some of the perceptual mechanisms underlying rapid, online adaptation.

These results are consistent with prior listener knowledge of VOT covariation or knowledge of the linear relationships between stop categories. In particular, the logistic mixed-effects models revealed that listeners employed information about both the exposure and test VOTs in categorization. The perceptual results, therefore, cannot be reduced to biased guessing or condition-independent VOT preferences. Listeners also did not simply have a dispreference for test VOTs that disobeyed the rank order (VOT [p^h] < VOT [k^h]) relative to exposure VOTs predicted less generalization in the Train Short – Test [k^h] and Train Long – Test [p^h] conditions as both test VOT options generally obeyed the ranking. Rather, listeners were more likely to choose the VOT value that was

consistent with the exposure VOTs according to the linear relationships between the stop categories.

These results provide evidence in support of an account that listeners exploit knowledge of VOT covariation (or linear relationships) in the population to generalize relatively small talker-specific effects from two of the aspirated stops to the third one. A central finding is that perceptual adaptation occurs rapidly, with listeners tuning expectations about a talker's voice after brief exposure (e.g., Morton et al., 2015). These findings have substantial implications for cognitive models of talker adaptation, as listeners may initially rely on structured variation across categories to refine a talker-specific model. Information about covariation and linear relations has already proved fruitful in on-line automatic speaker adaptation (e.g., Lasry & Stern, 1984; Cox, 1995; Zavaliagkos et al., 1995) and has been incorporated to a certain extent in other cognitive models of talker adaptation and perceptual generalization (e.g., Nielsen & Wilson, 2008; McMurray & Jongman, 2011; Pajak et al., 2013; cf., Johnson, 2005; Kleinschmidt & Jaeger, 2015). Further research will be required to further refine our understanding of how precisely listeners encode dependencies among phonetic categories, and how perceptual noise may interact with the encoding and implementation of structured variation in generalized adaptation.

2.7 General discussion

Patterns in talker- and language-specific VOT were examined to investigate the predictions of target and contrast uniformity on the phonetic realization of the laryngeal feature (e.g., [voice] or [spread glottis]) within stop consonants. While target and contrast uniformity are assumed to operate on the mapping from phonological segments to

phonetic targets, VOT was used as an acoustic correlate of the laryngeal phonetic target(s). Within and across languages, there was relatively strong covariation of mean VOT among stop consonants; however, the relatively stronger covariation among stop consonants with a *shared* laryngeal feature value reveals that target uniformity may provide a more precise account of the data than the more general constraint of pattern uniformity.

In addition to the strong correlations, supporting evidence for target uniformity came from the minimal variation in VOT across place of articulation: the observed place differences could be accounted for by automatic mechanisms given a uniform phonetic target for a laryngeal feature. The deviation from a fixed VOT pattern across place of articulation was quite minimal across talkers and languages, further indicating that target uniformity has a substantial influence on an individual phonetic grammar, regardless of the language. While pattern uniformity could provide additional support to target uniformity in constraining the phonetic targets, the minimal and highly predictable differences across place of articulation would not necessarily be predicted by pattern uniformity alone. Note that the predictions of pattern uniformity would also be met even if the place difference between [p^h] and [k^h] for example was quite large, but talkers and languages nevertheless maintained this same difference.

For American English and across languages, there were strong correlations of talker mean VOT among the aspirated (long-lag) stops, but only moderate correlations among the unaspirated (short-lag) stops. This may suggest that target uniformity has a greater influence on the [+spread glottis] stops than the [-spread glottis] stops. The relatively lower correlations among unaspirated stops may be attributable to the truncated

range, and thus smaller variance compared to aspirated stops. Nevertheless, there were only minimal deviations in VOT across place of articulation among the unaspirated stops within a talker / language, indicating that target uniformity may have exerted some influence. Moreover, pairs of unaspirated stop means across languages conformed to the same linear relationship as the long-lag and voiced stop means, suggesting that these means may be governed by the same principle as the long-lag and voiced stop means. Target uniformity thus appears to play a role in shaping the phonetic targets underlying VOT for all sets of stop consonants that share a laryngeal feature value.

While examined only in adult AE speech, the evidence for contrast uniformity was considerably weaker than that for target uniformity. The correlations between stop consonants contrasting in the laryngeal feature ranged from weak to moderate, and there was substantial variability across talkers in the separation between voiced and voiceless stops, as evidenced by the high standard deviation of the random talker slope for voice in the mixed-effects model. These findings suggest that the phonetic implementation of phonological surface segments may not be influenced by contrast uniformity, or that contrast uniformity plays a diminished role, minimally for AE stop consonants.

As a constraint on the phonetic grammar of individual talkers, uniformity is expected to apply universally, regardless of the language, and may also be present early on in acquisition. The child language study demonstrated that by 2 years of age, strong linear relationships were already present within speech production, which is consistent with an influence of target uniformity on the mapping from distinctive features to phonetic targets. An alternative explanation is that children successfully mimic a single talker, thus retaining the relationship between means. Distinguishing these two accounts

in child language may be difficult; however, future research should nevertheless examine whether patterns of covariation and minimal place differences are present prior to 2 years of age in infant babbling (e.g., Whalen et al., 2007). In particular, a study of this sort could examine a potentially earlier influence of uniformity, and may also shed light on whether systematic relations emerge even when children acquiring English still lack adult-like aspiration. Additionally, the cross-linguistic study found that for stops with a shared laryngeal feature value, place differences in VOT means were not only minimal, but also highly predictable, indicating an underlying pressure for a uniform phonetic target across place of articulation cross-linguistically.

Finally, previous studies of perceptual adaptation provided support for knowledge of VOT covariation, but had examined generalization only after considerable exposure to a new talker. The final section of this chapter demonstrated that listeners generalized talker-specific characteristics of VOT after brief exposure, using more natural and variable stimuli than had previously been tested. Perceptual generalization of VOT across place of articulation may derive from knowledge of pattern or target uniformity; alternatively, listeners could directly track VOT covariation among stops. Regardless, the presence of covariation of talker-specific VOT means is readily used by listeners in perceptual adaptation.

One clear extension of our study would be to carefully investigate VOT covariation in spontaneous speech. As a preliminary step, AutoVOT was used to extract measurements for all of the word-initial prevocalic stops of 38 talkers from the Buckeye corpus (Pitt et al., 2005). Unlike the Mixer 6 corpus, the content of the Buckeye corpus is not matched across talkers. In spite of the much greater variation in prosodic, lexical, and

syntactic contexts, talker means were again found to be highly correlated after removal of outliers across all voiceless stop pairs and between [b] and [g] (e.g., [p^h - t^h]: $r = 0.82$; [t^h - k^h]: $r = 0.83$, $p < 0.001$; [k^h - p^h]: $r = 0.81$, all ps < 0.006; [b - g]; $r = 0.43$, $p < 0.01$). While further examination of the patterns in this and other spontaneous speech corpora is certainly warranted, we tentatively conclude that strong correlations, at least for aspirated stops, will be found in essentially any speech style.

The systematic relations observed among stop categories may be present not only for VOT, but also for other acoustic-phonetic cues to stop consonant place and voice. Research is currently underway to investigate talker systematicity in stop consonant spectral center of gravity (Blumstein & Stevens, 1979; Chodroff & Wilson, 2014), f0 (Haggard et al., 1970; Ohde, 1984; Whalen et al., 1990; Kong & Edwards, 2016), relative amplitude (Repp, 1979; Ohde & Stevens, 1983), and following vowel duration (Summerfield, 1981; Allen & Miller, 1999). Systematicity in closure duration and prevoicing, and their respective relations to positive VOT, also warrant further investigation. Additional research is necessary to determine whether these relations exist for acoustic-phonetic cues among other natural class such as fricatives (see Chapter 4), nasals, and liquids.

Structured VOT variation could potentially be one reflection of talker differences in domain-initial strengthening (Fougeron & Keating, 1997; Cho & Keating, 2001) or other types of hyperarticulation (e.g., Lindblom, 1990). If talkers vary in the degree of strengthening due to prosodic boundaries, and the effect of strengthening on VOT is similar for all stops that have the same laryngeal specification, the correlations observed here would be predicted. Note that this analysis crucially assumes a form of target

uniformity (i.e., talker-specific prosodic effects would have to apply uniformly to all stops within each voicing category). Talker-specific VOT values would then reflect the talker's degree of hyperarticulation and be expected to correlate with other measures of domain-initial strengthening (see Bang & Clayards, 2016 for related research). In this way, a small number of prosodic (or hyperarticulation) variables would account for many idiosyncratic aspects of a talker's phonetic system. Listeners could then adapt to a talker by estimating these higher-level variables, jointly inferring the means and other parameters of many phonetic categories along multiple dimensions.

2.8 Conclusion

This chapter investigated the predictions of uniformity in the production and perception of stop consonant VOT. Strong evidence was observed for a constraint of target uniformity on the implementation of stop consonants across adult AE speakers in multiple speech styles, across 2- to 5-year-old AE speakers and across a range of languages. The strong linear relationships between VOT means and the minimal VOT differences across place of articulation are highly consistent with target uniformity in that phonetic targets corresponding to segments with a shared laryngeal feature value are constrained to be nearly identical, or *uniform*, across all segments that share that feature value. Finally, listeners generalize a talker's characteristic VOT across stop place of articulation, and do so early on in adaptation. VOT generalization could derive from direct knowledge of how the uniformity constraints shape the phonetic grammar, or from statistical learning of VOT dependencies across talkers.

3 Chapter 3

3.1 Introduction

Variation in the phonetic realization of sibilant fricatives has been extensively documented across languages, dialects, and talkers. Few studies, however, have examined whether there are dependencies among the realizations of different sibilants in the speech of a language community or individual talker. This chapter investigates uniformity in the phonetic implementation of place of articulation, specifically the [anterior] feature, in the sibilants of American English and Czech. The predictions of pattern, target, and contrast uniformity were tested in three multi-talker corpora of American English differing in speech style (isolated laboratory speech, connected read speech, and spontaneous interview speech), as well as in a multi-talker Czech spontaneous speech corpus.

3.1.1 Phonetic correlate of place of articulation

The uniformity constraints are assumed to operate on the mapping from the distinctive features (e.g., [anterior]) to the phonetic targets (e.g., constriction location). As the phonetic target cannot be measured directly, an acoustic property that correlates with the target must be selected. While there may be several components of the target that corresponds to the [anterior] feature in fricatives, a central phonetic property is the articulatory location of the constriction (or obstruction). Because the raw acoustic speech signal convolves information from the filter, which is principally determined by constriction location (as well as tongue shape), and the vocal source (e.g., vocal fold vibration in voiced fricatives), finding a measure that tracks the constriction *per se* can be difficult. This difficulty is not limited to acoustic correlates: even a 'direct' articulatory

measurement of constriction location would still only approximately reveal the underlying phonetic target.

Previous research has identified several acoustic-phonetic measures that correlate to varying degrees with place of articulation in obstruents generally and fricatives in particular. These measures include the spectral center of gravity (COG) along with the higher spectral moments of variance, skewness and kurtosis (e.g., Forrest et al., 1988; Jongman et al., 2000), spectral peak (e.g., Hughes & Halle, 1956), and the mid-frequency peak ($Freq_M$; Koenig et al., 2013), among others (e.g., formant transitions and spectral slope; e.g., Delattre et al., 1962). The COG and spectral peak are defined as the weighted mean and mode of the distribution of energy across the frequency spectrum, respectively. While these two measures have been widely used in previous studies of fricative place, they do not cleanly separate components of the spectrum due to the filter and the source (see Koenig et al., 2013); for example, COG can be lowered by harmonics of the fundamental frequency in voiced fricatives.[24]

One way in which the influence of the source can been minimized is through high-pass filtering of the spectrum, which removes low-frequency energy due to vocal fold vibration. Estimates of the appropriate cut-off for filtering range from 300 Hz (e.g., Maniwa et al., 2009; Holliday et al., 2015), which will typically exclude the fundamental frequency, to higher values in the F2 region, which will not only exclude f0, but also several of its harmonics (e.g., 1000 Hz as in Tabain, 2001 and 1720 Hz as in Li et al.,

[24] The spectral estimation technique has also come into question in the analysis of sibilant fricatives. The periodogram, derived form a discrete Fourier transform (DFT), has been criticized for being more prone to estimation error than multitaper spectral analysis (Blacklock, 2004). However, Reidy & Beckman (2012) reported no significant difference between sibilant spectral estimates derived separately from DFT and multitaper techniques.

2007). Nevertheless, harmonics of the fundamental exist throughout the frequency range, and low-pass filtering will not completely remove higher contributions of the source. Furthermore, other aspects of the source such as airflow rate can affect energy in higher frequencies, and increased vocal effort can shift the entire energy distribution upwards slightly (Zue, 1976; Koenig et al., 2013). Thus, while COG and spectral peak do primarily indicate place of articulation, they are still quite affected by these confounding influences of the source.

The mid-frequency peak ($Freq_M$) has been proposed as an alternative and more precise acoustic measure of fricative place (Koenig et al., 2013; Shadle et al., 2014). $Freq_M$ has been employed in phonetic studies of [s], where it was defined as the peak frequency between 3000 and 7000 Hz. Previous research has shown that $Freq_M$ is inversely correlated with the tongue constriction location; more precisely, it reflects the resonances of the vocal tract cavity anterior to the constriction location (Shadle et al., 2016), where the length of the cavity is itself a function of the location of constriction. This measure is also known to be relatively unaffected by the source properties such as voicing and vocal effort.

This chapter adopts $Freq_M$ as the best available phonetic correlate of fricative place (with the general caveat that no phonetic measure can be perfectly identified with a phonetic target). As both alveolar and post-alveolar sibilants were analyzed (e.g., [s] and [ʃ]), $Freq_M$ was defined as the peak frequency between 3000 and 7000 Hz for the alveolars and between 2000 and 6000 Hz for the post-alveolars. It was desirable to lower the frequency range for the post-alveolar sibilants because of their longer anterior cavities and lower resonance frequencies. However, the category-specificity of the measurement

does introduce some complexities (particularly for perceptual implications of the current findings); for this reason, and for purpose of comparison with previous related studies, we also report spectral COG after high-pass filtering at a cut-off value of 550 Hz (Koenig et al., 2013).

3.1.2 Sources of variation in sibilant spectral shape

Variation in the spectral properties of sibilants arises from several sources including the phonetic category of the sibilant, coarticulation with nearby sounds, and speech style, as well as cross-linguistic, dialectal, and cross-talker differences. The cross-linguistic and talker differences were reviewed in Chapter 1 (sections 1.1 and 1.2). This section reviews the influence of the other sources of variation on sibilant spectral properties.

As alluded to above, spectral properties of sibilants can signal both place of articulation and laryngeal contrasts. Spectral energy is concentrated at higher frequencies for alveolar fricatives in comparison to post-alveolars (Jongman et al., 2000) because a smaller cavity anterior to the constriction location results in concentration of energy higher along the frequency spectrum (Shadle, 1985; Stevens, 1998). Previous studies have reported spectral peak locations for alveolar sibilants between both 3.5 and 5 kHz (Behrens & Blumstein, 1988), and even up to 6 and 8 kHz (Jongman et al., 2000). These contrast substantially with the peak locations for post-alveolar sibilants, reported to be between 2 and 4 kHz (Hughes & Halle, 1956; Behrens & Blumstein, 1988).

With respect to the voicing dimension, previous studies have found that voiced fricatives have greater energy in lower frequencies (e.g., Silbert & de Jong, 2008; Jongman et al., 2000). Hughes & Halle (1956) specifically reported a strong excitation

below 700 Hz attributable to vocal fold vibration that is never found in voiceless fricatives. Significant effects of the voice contrast have been found in both COG and spectral peak (Jongman et al., 2000); however, as predicted by the observation by Hughes & Halle (1956), when frequency bands below 750 Hz have been removed from analysis, the COG of voiced fricatives becomes numerically but not significantly lower than that of voiceless fricatives (Silbert & de Jong, 2008).

While spectral properties are defining features of sibilant categories, they are nevertheless affected by coarticulation and indexical (talker) properties. These influences have been studied not only with respect to acoustic-phonetic realization, but also for the effects that they induce on perceptual discrimination and identification.

Spectral properties of sibilant fricatives vary substantially as a function of the following vowel and other contextual sounds. Soli (1981) observed a higher spectral peak before the front vowel [i] than before back vowels [a] and [u] for all four sibilant fricatives ([s z ʃ ʒ]) in English. Within the back vowels, there was a strong effect of lip rounding such that that prominent frequencies in the vicinity of F2 were lower before round vowels; this resulted in a lower overall COG for sibilants preceding [u] compared to those before [a] (see also, Hughes & Halle, 1956; Yeni-Komshian & Soli, 1981; Shadle & Scully, 1995). The effects of vowel coarticulation have generally been found to be stronger for alveolar fricatives such as [s] than for post-alveolars such as [ʃ] (Nittrouer et al., 1989; Tabain, 2001). Silbert & de Jong (2008) identified a significant effect of syllable-internal position such that fricatives in onset position exhibited higher COGs than those in coda position. (In Silbert & de Jong's study, COG values were also numerically higher for fricatives

that had semantic focus compared to unfocused productions, but this difference did not reach significance.)

In addition to contextual influences, spectral properties of sibilants also vary according to speech style. Maniwa et al. (2009) reported that alveolar sibilants produced in a clear speech style had a significantly higher spectral peak and COG than those said in a more conversational style. In this study, speech in both styles involved producing isolated syllables: conversational tokens were elicited by having the talker read the syllables in a casual style, while the clear-speech tokens were elicited by having the talker correct syllables which were reported to be misperceived by a computer program. The clear speech effect on COG extended to the labiodental ([f v]) and interdental ([θ ð]) fricatives; however, no differences were observed across clear and conversational styles in the post-alveolar sibilants ([s ʃ]). As was found for coarticulation, the post-alveolars appear to be relatively invariant across linguistic contexts.

Talker differences are a major source of variability in fricative spectra. For example, as discussed in Chapter 1 (section 1.2) the COG distributions of [s] and [ʃ] are quite distinct within each talker but overlap across talkers (Newman et al., 2001). Such talker differences partially reflect anatomical parameters including the shape and size of the vocal tract, tongue, and incisors. However, there is substantial variability in the phonetic implementation of sibilants that cannot be fully reduced to anatomical differences, as demonstrated by studies of cross-linguistic and socioindexical differences in the realization of these sounds.

Cross-linguistic differences in the realization of sibilants were reviewed in Chapter 1 (section 1.1). Socioindexical properties that contribute to differences within a

language include gender, sexual orientation, and socioeconomic status. For example, female talkers generally have a higher COG or peak [s] than male talkers of the same language. This gender difference has been found for American English (Strand & Johnson, 1996; Flipsen et al., 1999; Podesva & Van Hofwegen, 2014), Canadian English (Heffernan, 2004), upper-class speakers of Glaswegian English (Stuart-Smith et al., 2003), British English (Levon & Holmes-Elliot, 2013), and from "mixed" English talkers (Australia, North America, UK: Fuchs & Toda, 2010). Fuchs & Toda (2010) found that both English and German females articulated [s] with a more anterior constriction location than male talkers. However, in that study the gender difference in spectral peak was observed only for English; the failure to find the same effect in German may be attributable an overall wider constriction width for that language.

Critically, differences in the spectral realization of sibilants that covary with gender cannot be wholly reducible to independent anatomical differences. While females have shorter vocal tracts than male talkers on average (Schwartz, 1968), this dimorphism is found primarily posterior to typical constriction locations for sibilants (Strand, 1999). As already noted, the acoustic measures that differentiate place of articulation in these sounds largely reflect the size of the oral cavity anterior to the constriction (as well as turbulence created at the upper and lower teeth). Fuchs & Toda (2010) explicitly controlled for palate size and length, and nevertheless found that females had a more fronted articulation of [s] than males in both English and German language groups.

Moreover, if the gender difference was entirely due to anatomy we would expect comparable differences between males and females in the sibilant realization across dialects and languages. However, the size of the gender difference for [s] varies across

speech communities, and even appears to be minimal or absent in some (e.g., working-class speakers of Glaswegian English: Stuart-Smith, 2003; German: Fuchs & Toda, 2010 — but note the articulatory difference). Specifically, Levon & Holmes (2013) found a greater difference in the COG of [s] between working class males and females from Essex, UK in comparison to upper class males and females from Chelsea, UK. Such variation would not be expected given a purely anatomical explanation.

Beyond gender differences, perceived and self-identified sexual orientation are also associated with moderately different phonetic realizations of sibilant fricatives. Linville (1998) showed that gay males had a higher spectral peak and longer duration for [s] than straight males, and that listeners' perception of a talker's sexual orientation (gay or straight) was correlated with these measures. Munson et al. (2006) also observed that knowledge of the talker's sexual identity can shift the [s]-[ʃ] perceptual boundary, in the direction predicted by Linville (1998); however, this study reported an effect for speech from female talkers. In accord with Linville (1998), Campbell-Kibler (2011) found that a male talker was more likely to be perceived as gay and effeminate when [s] and [z] were fronted compared to when the sounds had medial or backed articulations.

Finally, talker age—and even social traits such as preference for rural or urban living—can also condition the phonetic realization of sibilants. Podesva & Van Hofwegen (2014) showed that the COG of [s] was correlated with age among talkers with an orientation towards country living in Redding, California. Older country-oriented talkers had a lower COG [s] than younger country-oriented talkers, who exhibited an [s] COG more comparable to that observed among talkers with an orientation towards the town.

3.1.3 Predictions of uniformity

Cross-linguistic and sociophonetic studies have demonstrated that the same sibilant fricative category can be implemented with highly varied phonetic targets, even when the fricative inventory is held constant across sociolects. In principle, it would be possible for the phonetic targets of different categories to vary independently. For example, one talker could implement a German-like [s], an English-like [z], and a Japanese-like [ʃ], whereas a second talker could implement an English-like [s], a Japanese-like [z], and a German-like [ʃ]. However, the patterns of phonetic covariation reviewed in the Introduction (section 1.3) suggest that the realization of related sounds may be more restricted than would be expected from independent implementation of each category.

The present chapter examined the predictions of target and contrast uniformity (see Introduction, section 1.5) with respect to [anterior] targets for the sibilant fricatives ([s z ʃ ʒ]) as measured by $Freq_M$. Three primary statistical methods were used to assess the empirical support for uniformity. First, we examined the extent and pattern of covariation with linear correlations of mean $Freq_M$ values across talkers. Second, we analyzed simple linear regressions between talker means of contrasting categories to determine the type of relationship between means (e.g., identity, additive, scalar, etc.). Third, a linear mixed-effects model of token-specific $Freq_M$ values was evaluated to examine the relative magnitudes of the fixed effects corresponding to two phonological features ([anterior], [voice], and their interaction [anterior]x[voice]), as well as the relative variances of the talker random effect components (i.e., intercepts and random slopes for the phonological features).

The target uniformity constraint applied to place of articulation should render the effect of the non-place feature ([voice]) minimal. This should be reflected in a strong correlation of talker-specific $Freq_M$ means between [s] and [z] (which are [+anterior]), as well as between [ʃ] and [ʒ] (which are [-anterior]). Indeed, perfect adherence to the target uniformity principle would result in equality of the talker means for each of these two pairs, modulo measurement artifacts, which can be assessed in the simple linear regression. In the linear mixed-effects model, the coefficient for [anterior] should be substantially larger in magnitude than that of [voice] or the interaction between [anterior] and [voice]. That is, the effect of [anterior] alone should account for the majority of the variation in the mid-frequency peak, as there should be identical constriction targets for both [+anterior] sibilants, and separately for both [-anterior] sibilants. With respect to the random effects, variation across talkers should be found primarily in the grand mean (talker intercept) and perhaps in the size of the [anterior] effect (talker [anterior] slope), with minimal variance for the [voice] slope.

The predictions of contrast uniformity are first, that there should be strong correlations of talker mean $Freq_M$ between [s] and [ʃ], as well as [z] and [ʒ]. Second, in the mixed-effects model, the variation across talkers should be found primarily in the grand mean, or random talker intercept, with minimal talker-specific effect of the [anterior] feature. Specifically, talkers should not stray far from the overall population difference between [+anterior] and [-anterior].

3.1.4 Outline

The predictions of target and contrast uniformity were tested in three multi-talker corpora of American English, each with a distinct speech style (isolated laboratory

speech, connected read speech, and spontaneous interview speech), and in a multi-talker spontaneous speech corpus of Czech. There were several motivations for analyzing a variety of speech corpora. First, analysis of two separate languages allowed a preliminary investigation into the cross-linguistic nature of the findings. Second, the replicability of the findings for American English was assessed in multiple speech styles, ensuring that (for example) the effect is not limited to isolated speech under laboratory conditions. In addition, the American English corpora varied in the number of talkers, representation of the sibilant fricatives, and recording details such as sampling rate. For sibilants, which can contain substantial high frequency energy, sampling rate can affect measures such as COG; our primary measure of $Freq_M$, confined to fall within the 2000-7000 Hz frequency range across all sibilants examined here, is essentially invariant across rates at or above 16 kHz. In summary, by using multiple speech corpora we could assess whether covariation patterns hold across various speech styles, segmental contexts, talker groups, and across the full sibilant inventory ([s z ʃ ʒ]) in two unrelated languages.

3.2 Covariation of $Freq_M$ in American English isolated speech

The predictions of target and contrast uniformity were first tested in a corpus of fricative-initial syllable productions from 22 native speakers of American English. The corpus contained a relatively equal number of tokens for all four sibilant fricatives [s z ʃ ʒ] in matched segmental contexts. The high degree of control in the stimuli contributed to the goal of isolating the potential sources of variation and covariation primarily due to talker differences.

3.2.1 Methods

3.2.1.1 Participants

Twenty-two participants (15 female) were recruited at New York University for an experiment on non-native consonant cluster production and perception, in which one of the tasks was a fricative-initial syllable production task. All participants were native speakers of American English. Participants were given a small monetary compensation for participating.

3.2.1.2 Materials and procedure

Participants recorded fricative-initial consonant-vowel-consonant (CVC) syllables in isolation as distractor items in an experiment that mainly focused on perception and production of non-native consonant clusters. All recordings were made with a Zoom H4n digital recorder and an Audio-Technica ATM-75 head-mounted condenser microphone in a sound-attenuated booth at a sampling frequency of 44.1 kHz. The CVC syllables were composed by fully crossing the fricatives [ð θ f v s ʃ z ʒ] with the vowels [i ɪ eɪ ɛ æ a ɔ oʊ ʊ u ʌ], and [t] (Jongman et al., 2000). Two [ʃ]-initial combinations were excluded due to their sensitive nature. Only syllables beginning with the sibilant fricatives [s ʃ z ʒ] were considered for analysis. In many cases, participants could not readily interpret the orthographic mapping for [ʒ] and [ð]. Two participants (1 female, 1 male) did not produce any instances of [ʒ]. Data for from these participants was retained in deriving the means and standard deviations for [s z ʃ] and correlations that did not involve [ʒ], but they were excluded from the linear mixed-effects model reported below.

Each trial in the experiment consisted of three parts. First, a pre-recorded multi-syllabic nonce word was played over the speakers. Then, the fricative-initial CVC

syllable was displayed on the monitor in a standardized orthographic form. Finally, the participant produced the CVC syllable followed by the original multi-syllabic nonce word. There were 12 unique presentation orders, and each CVC syllable was presented 2 to 3 times. This resulted in a total of 1,890 sibilants for analysis, with the median number of tokens per talker and category ranging from 20 to 24 (Table 3.1).

Table 3.1. Range and median number of tokens per talker and fricative, and total number of tokens per fricative in American English isolated speech.

Fricative	Range	Median	Total
s	16 – 25	24	522
z	15 – 25	24	511
ʃ	14 – 22	20	442
ʒ	10 – 24	22	415

3.2.1.3 Data preparation

A phonetic segmentation of the CVC syllables was obtained with the Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman, 2008). All boundaries were manually corrected to align to the onset and offset of the fricative. The onset corresponded to the start of frication, and the offset of the fricative was defined as the offset of frication. This often coincided with the onset of periodicity in the vowel, but for instances when periodicity and frication overlapped, the boundary was placed after the frication ended.

3.2.1.4 Acoustic analysis

A multitaper spectrum was estimated from the middle 50% of each extracted sibilant (tapers = 8, time bandwidth = 4.0; Blacklock, 2004) and used to measure the mid-frequency peak ($Freq_M$; Koenig et al., 2013; Shadle et al., 2014). As discussed earlier, $Freq_M$ was defined as the frequency with the largest amplitude between 3000 and 7000

Hz for the alveolar sibilants, and between 2000 and 6000 Hz for the post-alveolar sibilants. This adjustment for [ʃ] was additionally motivated by a strong floor effect when the same frequency range (3000 – 7000 Hz) was employed for all fricatives.

The mid-frequency peak was originally defined on a linear scale (Hz). We also converted each $Freq_M$ value to equivalent rectangular bandwidth (ERB) units, which approximate the non-linear frequency representation of the auditory system (Glasberg & Moore, 1990). Finally, spectral COG was measured in Hz from a multitaper spectrum after high-pass filtering at 550 Hz (Koenig et al., 2013).

3.2.2 Results

The population $Freq_M$ means for the sibilants of each place of articulation were quite similar to each other, indicating highly comparable constriction locations regardless of voicing status (Table 3.2). The population standard deviation of talker means was larger for [s] and [z] than for [ʃ] and [ʒ]; however, it should be noted that on the ERB scale, this ranking was reversed. The range of variation in average $Freq_M$ across talkers was substantial for each sibilant, with the extreme values differing by approximately 1500 to 2000 Hz. Talker-specific standard deviations also ranged considerably; we examined whether there was a linear relationship between talker-specific means and standard deviations but found significant correlations for the post-alveolar fricatives only ([s]: $r = -0.32$, 95% CI: [-0.64, 0.10], $p = 0.14$; [z]: $r = -0.31$, 95% CI: [-0.70, 0.13], $p = 0.16$; [ʃ]: $r = 0.56$, 95% CI: [0.25, 0.71], $p < 0.01$; [ʒ]: $r = 0.72$, 95% CI: [0.48, 0.86], $p < 0.001$). The negative correlations for the [+anterior] sibilants indicate smaller standard deviations at higher $Freq_M$ values, but this may reflect a slight bias in the automatic measurement, particularly for female talkers, some of whom may actually have several

$Freq_M$ values above 7000 Hz (i.e., the negative correlation may reflect a ceiling effect). The lower strength and variation in directionality of these correlations also stands in contrast to the moderate to strong and consistently positive correlations observed between talker means and standard deviations in many temporal phonetic measures (e.g., Byrd & Saltzman, 1998; Shaw et al., 2009; Turk & Shattuck-Hufnagel, 2014).[25]

Table 3.2. Descriptive statistics for each sibilant in American English isolated speech. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Measure	Fricative	Mean	SD	Range of Talker Means	Range of Talker SDs
$Freq_M$ (Hz)	s	5968	555	4861 – 6721	296 – 1167
	z	5948	527	4818 – 6720	254 – 1264
	ʃ	3304	490	2488 – 3969	134 – 1218
	ʒ	3251	462	2514 – 3951	157 – 1144
$Freq_M$ (ERB)	s	30.56	0.88	28.77 – 31.72	0.41 – 1.94
	z	30.52	0.84	28.68 – 31.72	0.35 – 2.17
	ʃ	25.28	1.30	22.96 – 26.97	0.36 – 2.50
	ʒ	25.13	1.22	23.06 – 26.84	0.53 – 2.38
COG (Hz)	s	7947	1130	6115 – 10200	404 – 1252
	z	7589	956	5714 – 8721	418 – 2066
	ʃ	4253	668	3275 – 5215	182 – 501
	ʒ	4071	606	2982 – 4981	147 – 843

As shown in Table 3.3 and Figure 3.1, talker $Freq_M$ means were highly correlated for [s] and [z] ($r = 0.88$, $p < 0.001$) and for [ʃ] and [ʒ] ($r = 0.84$, $p < 0.001$). Moderate numerical correlations were also observed for [s] and [ʃ], and for [z] and [ʒ], but these did not reach significance ([s - ʃ]: $r = 0.56$, [z - ʒ]: $r = 0.31$, ps > 0.01). Furthermore, while the correlations remained strong and significant between the sibilants of the same place within each gender ([s - z] *female*: $r = 0.84$, *male*: $r = 0.86$; [ʃ - ʒ] *female*: $r = 0.77$, *male*: $r = 0.80$, ps < 0.001), correlations within each gender weakened between sibilants sharing

[25] Throughout the dissertation, correlations are described using modifiers based on recommendations in Evans (1996): a 'moderate' correlation means the coefficient is between 0.40 and 0.59, a 'strong' correlation means the coefficient is above 0.59, and a 'weak' correlation describes a coefficient below 0.40.

the same voicing specification ([s - ʃ] *female*: $r = 0.50$, *male*: $r = 0.21$; [z - ʒ] *female*: $r = 0.22$, *male*: $r = -0.25$, $ps > 0.01$). This suggests that the overall correlation may simply reflect a difference in phonetic implementation across genders, as opposed to across individuals.

Table 3.3. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means in American English isolated speech.

Measure	Fricative Pair	All	Female	Male
$Freq_M$ (Hz)	s – z	0.88* [0.71, 0.94]	0.84* [0.55, 0.92]	0.86* [-0.60, 0.96]
	ʃ – ʒ	0.84* [0.48, 0.95]	0.77* [0.10, 0.97]	0.80* [-1.00, 0.98]
	s – ʃ	0.56 [0.22, 0.79]	0.50 [-0.09, 0.85]	0.21 [-0.69, 0.85]
	z – ʒ	0.31 [-0.24, 0.71]	0.22 [-0.61, 0.75]	-0.25 [-0.96, 0.56]
$Freq_M$ (ERB)	s – z	0.88* [0.72, 0.94]	0.84* [0.51, 0.93]	0.85$^+$ [0.03, 0.95]
	ʃ – ʒ	0.88* [0.55, 0.96]	0.82* [0.20, 0.97]	0.86$^+$ [-1.00, 0.99]
	s – ʃ	0.56$^+$ [0.18, 0.80]	0.53 [-0.03, 0.84]	0.20 [-0.81, 0.86]
	z – ʒ	0.32 [-0.19, 0.69]	0.26 [-0.45, 0.75]	-0.23 [-0.94, 0.69]
COG (Hz)	s – z	0.92* [0.82, 0.96]	0.88* [0.66, 0.95]	0.98* [0.78, 0.99]
	ʃ – ʒ	0.93* [0.78, 0.97]	0.89* [0.57, 0.97]	0.94* [0.14, 0.98]
	s – ʃ	0.53 [0.11, 0.76]	0.51 [-0.04, 0.81]	0.10 [-0.93, 0.85]
	z – ʒ	0.44 [-0.05, 0.72]	0.35 [-0.51, 0.70]	0.31 [-1.00, 0.92]

* = $p < 0.001$, $^+$ = $p < 0.01$

Simple linear regressions predicting each talker's mean $Freq_M$ for one sibilant from the same talker's mean value for another sibilant were performed to gain further understanding of the preceding correlations. The way in which two means covary across talkers could be additive, proportional, or involve a combination of additive and scalar

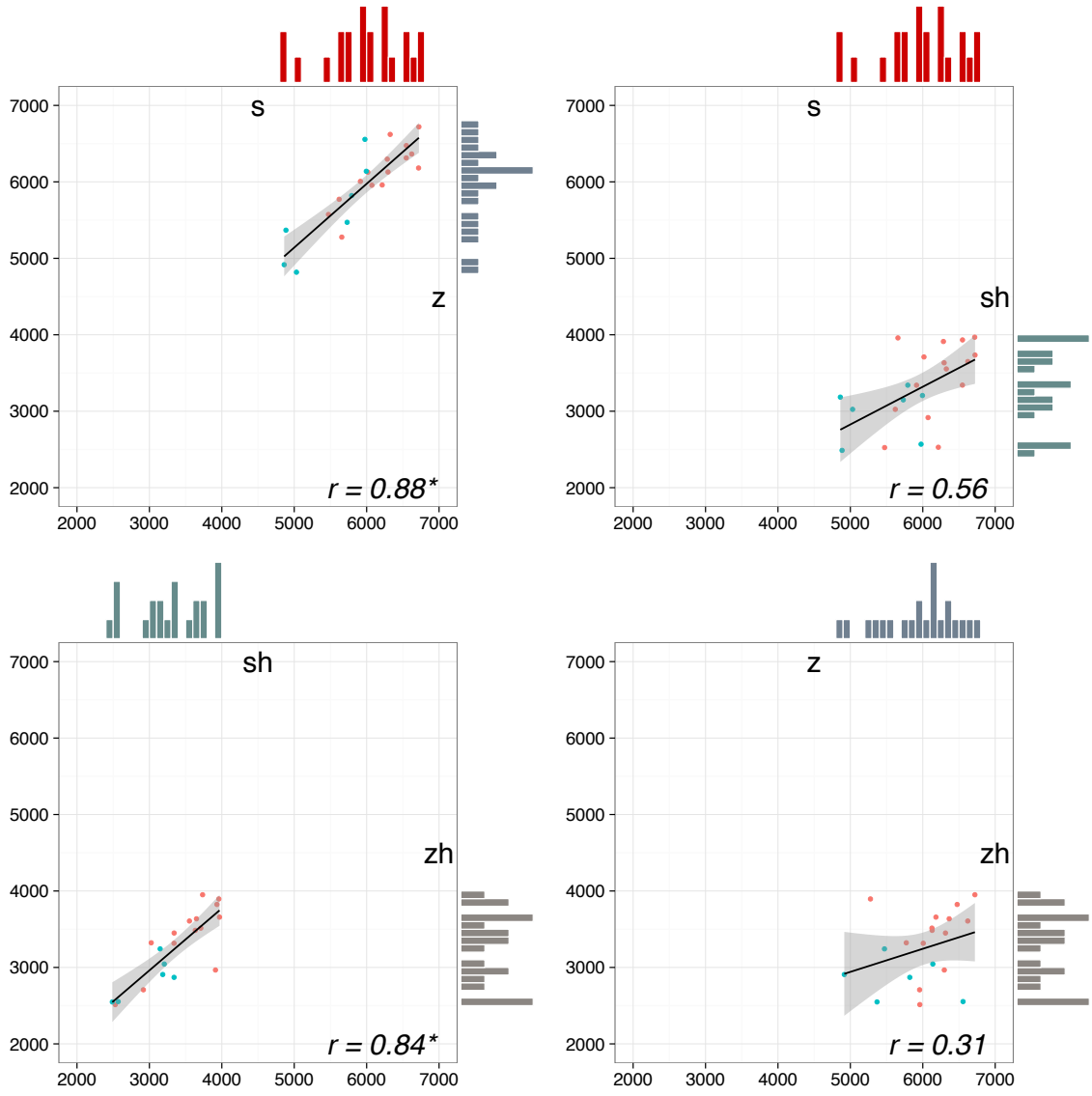
factors. Target uniformity in particular predicts that the relationship between the underlying targets for sibilants with a shared place feature should be identical, which would be most clearly supported by additive estimates of zero and scalar estimates of one. The fit coefficients for pairwise linear regression models are reported in Table 2.8.

Table 3.4. Additive (β_0) and scalar (β_1) components of simple linear regressions of mean $Freq_M$ values for sibilant pairs in American English isolated speech. For each pair, the talker-specific mean of the first sibilant was the dependent variable predicted from the talker-specific mean of the second sibilant.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
s ~ z	460	0.50	0.93	< 0.001	0.76
ʃ ~ ʒ	555	0.22	0.86	< 0.001	0.68
s ~ ʃ	3882	< 0.001	0.63	< 0.01	0.28
z ~ ʒ	4983	< 0.001	0.32	0.18	0.05

The best fits, in which the proportion variance accounted for exceeded 0.50, were among the homorganic sibilants. The predictions of target uniformity were borne out through additive factors that did not significantly differ from zero and significant scalar values close to unity. The numerical offsets between talker means (e.g., β_0 = 460 Hz for predicting $Freq_M$ of [s] from that of [z]) could potentially be accounted for by phonetic voicing: in spite of our efforts to select a measure that diagnoses only place of articulation, voicing may have lowered the $Freq_M$ of [z] and [ʒ] relative to their voiceless counterparts.

Figure 3.1. Variation and covariation of sibilant $Freq_M$ (Hz) across talkers in American English isolated speech. Marginal histograms display variation in talker means. Each point represents a talker-specific pair of means and is color-coded to specify talker gender (red = female, blue = male). The asterisk indicates that cases in which correlation reached significance ($p < 0.025$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



Next, a linear mixed-effects model of the token-level $Freq_M$ values was used to examine sibilant realization at the level of the language and across individuals. Previous research has shown that the spectral properties of sibilants vary not only due to place and voice features, but also according to coarticulation and socioindexical variables as

reviewed in section 3.1.2. Other sources of variation beyond the idiosyncrasies of the talker could potentially obscure systematic relationships of talker mean $Freq_M$ among the relevant speech sounds, though this is of less concern for the balanced laboratory sample analyzed here than for subsequent analyses of corpus data.

The model included fixed effects of place of articulation, voice, following vowel height and backness, gender, and the two-way interactions of place and voice, vowel height and backness, place and gender, and voice and gender. Previous research has implicated place, voice, gender, and the interactions between place and voice, as well as place and gender, as significant sources of variation in sibilant spectra. Effects of the following vowel have been established in the literature, but these have generally been attributed to vowel roundness, and specifically coarticulation with [u] (e.g., Mann & Repp, 1980; Soli, 1981; Whalen, 1981). As suggested by Linker (1982), coarticulation with vowel rounding may be modulated by vowel height and backness. To our knowledge, vowel height, vowel backness (separately from rounding), and the interaction between voice and gender have not been directly examined in the literature. The latter interaction could plausibly arise due to differences in the fundamental frequency (f0) of voiced fricatives and the resulting harmonic differences.

All categorical factors were weighted effect coded to correct for slightly unequal sample sizes (Darlington, 1990: 246; te Grotenhuis et al., 2016). The contrast weighting of the categorical variables was as follows: place of articulation (*place*: +anterior = 1, –anterior = -1.21), voice (*voice*: voiceless = 1, voiced = -1.04), vowel height (*height*: high = 1, non-high = -0.41), vowel backness (*backness*: front = 1, non-front = -0.91), and gender (*gender*: female = 1, male = -2.11). The dependent variable ($Freq_M$) was centered

at zero by subtracting the grand mean (μ = 4737 Hz) from each value prior to analysis. Talker random effects included an intercept and slopes for place, voice, and their interaction, and the model included a random intercept for stimulus (i.e., syllable).

There was a large and significant effect of place of articulation on $Freq_M$ (β = 1203.72, *t* = 20.61).[26] The effect of voice and the interaction between place and voice were not significant, indicating minimal influence of [voice] on constriction location targets (*voice*: β = 44.73, *t* = 1.12; *place x voice*: β = -28.79, *t* = -0.80). While the effect of vowel height was not significant (β = -99.45, *t* = -1.64), sibilants had significantly lower $Freq_M$ values before back vowels than before front vowels (β = 298.27, *t* = 7.44). The interaction between vowel height and backness was also significant, indicating that the $Freq_M$ of a sibilant preceding the high back vowel [u] (and possibly [ʊ]) was significantly lower than back vowels in general (β = 130.96, *t* = 2.08). The effect of gender also reached significance, revealing a higher $Freq_M$ for female talkers than for male talkers (β = 180.59, *t* = 3.66). Gender was not significantly modulated by interactions with either place or voice (*place* x *gender*: β = 35.35, *t* = 1.09; *voice* x *gender*: β = 14.52, *t* = 1.08).

Inspection of the random effects component can provide further insight into the structure of talker variation. The random talker intercept had the largest standard deviation, reflecting the substantial differences across talkers in overall mean $Freq_M$ (Table 3.5). The slope for place of articulation also varied considerably across talkers, suggesting that talkers also differ in the overall degree of separation between [+anterior] and [-anterior] sibilants on this phonetic dimension. This runs contrary to the predictions

[26] A *t*-value with magnitude greater than 2.0 was considered significant.

of contrast uniformity, according to which the size of the contrast should be (approximately) identical across talkers. Talker slopes for voice and the interaction between place and voice had much smaller variances: while talkers differ in their phonetic realization of sibilant place overall, and in the size of the anterior contrast, they are quite similar in showing minimal effects of the non-place feature.

Table 3.5. Standard deviations of talker random effects in the maximal mixed-effects model of $Freq_M$ in American English isolated speech.

Random effect for talker	SD
intercept	330
place	222
voice	71
place x voice	61

3.2.3 Discussion

Talker-specific means and standard deviations for $Freq_M$ varied considerably for each sibilant. This is consistent with previous research on talker variation in the COG of American English [s] and [ʃ], in which the means, standard deviations, and contrast of [s] and [ʃ] were shown to vary within the language community (Newman et al., 2001). Importantly, sibilants of the same place of articulation do not vary independently: the means of homorganic sibilants covary across talkers with minimal effect of [voice] on the constriction location, as measured by $Freq_M$.

The structure that emerged from our analyses provides strong evidence for target uniformity (and thus also for the more general notion of pattern uniformity). First, $Freq_M$ correlations were strongest among sibilants that shared the place feature (i.e., [s] - [z] and [ʃ] - [ʒ]). Second, simple linear regressions were approximately consistent with an invariant identity relation among the homorganic sibilants across talkers. Third, the fixed effects component of the mixed model revealed that place of articulation had the largest

effect on sibilant $Freq_M$. In contrast, the effect of voice did not reach significance. Fourth, the analysis of the random effects component in the mixed model revealed that the largest source of variation across talkers was in the overall intercept, with only the place effect varying to a comparable extent. In sum, we have provided multiple lines of evidence that sibilants sharing the same abstract place feature (here, [+anterior] vs. [-anterior]) are realized with near-identical place of articulation targets within each talker.

Evidence for contrast uniformity was, in comparison, relatively weak. The correlations of heterorganic $Freq_M$ means across talkers failed to reach significance, indicating that the difference between phonetic targets of contrasting features was not consistent across talkers. Furthermore, there was considerable variation in the talker-specific slope for place of articulation, which also suggests that talkers differ in the degree of separation between the phonetic place targets of alveolar and post-alveolar sibilants.

3.3 Covariation of $Freq_M$ in American English connected speech

While isolated speech provides tremendous control over the speech context, the fact that the syllables in the previous study were repeated several times and produced with careful articulation may have given rise to covariation that is atypical of speech in general. The present experiment examined the predictions of each of the uniformity constraints in a large corpus of connected read speech from 180 talkers of American English. The goal of this study was to determine whether the patterns observed in isolated speech were preserved in a larger and less homogenous group of talkers and for sibilants produced in a relatively natural and variable set of linguistic contexts.

3.3.1 Methods

3.3.1.1 Corpus description

The following analysis used data from an audited subset of read speech in the Mixer 6 Corpus (Brandschain et al., 2010; Brandschain et al., 2013; Chodroff et al., 2016). The read speech subset contained speech from 180 native AE talkers (102 female, 78 male) with approximately 45 min of speech per talker; further description of the corpus can be found in Chapter 2, section 2.3.1.1.

3.3.1.2 Data preparation and acoustic analysis

Word-initial and word-medial [s z ʃ] in prevocalic position were extracted from a time-aligned segmental transcript generated from the cleaned read speech transcripts using P2FA. Note that, unlike the laboratory study reported above, the voiced post-alveolar fricative [ʒ] could not be included here due to its extreme rarity in the corpus.[27] Each sibilant was measured using the same methods described earlier (see section 3.2.1.4). Within each sibilant category, tokens 2.5 standard deviations above or below their talker-specific $Freq_M$ mean were identified as outliers and excluded from the analysis. This resulted in a total of 55,304 tokens. The median number of tokens per talker was not the same for the three categories: in particular, there were many more instances of [s], which is among the most frequent consonantal sounds in English lexical items generally (Table 3.6; Hayden, 1950; Mines et al., 1978).

[27] [ʒ] is rare in English per standard phoneme counts (Hayden, 1950; Mines et al., 1978).

Table 3.6. Range and median number of tokens per talker and fricative, and total number of tokens per fricative in American English connected speech.[28]

Fricative	Range	Median	Total
s	110 – 314	223.5	39,431
z	21 – 44	33	6,006
ʃ	30 – 84	54	9,867

3.3.2 Results

The population $Freq_M$ means for the [+anterior] sibilants were comparable to each other and, as expected, greater than the population mean for [ʃ] (Table 3.7). This same pattern held for population standard deviation, which was larger for [s] and [z] than for [ʃ]. The range of $Freq_M$ variation within each sibilant category was substantial across talkers, spanning approximately 3000 Hz. Talker-specific standard deviations also ranged considerably but not in a way that tracked the means; there were only moderate to moderately weak correlations between talker-specific means and standard deviations ([s]: r = -0.32, 95% CI: [-0.45, -0.15]; [z]: r = -0.32, 95% CI: [-0.46, -0.17]; [ʃ]: r = 0.49, 95% CI: [0.35, 0.60], ps < 0.001). As in the isolated speech study, the negative correlations of the [+ anterior] sibilants indicate smaller standard deviations at higher $Freq_M$ values, again possibly a ceiling effect driven by female speakers.

[28] There were no word-initial instances of [z], again reflecting phonotactic frequencies in English. For each of the other sibilants, a sizable percentage (greater than 25%) of tokens occupied each word position.

Table 3.7. Descriptive statistics for each sibilant in American English connected speech. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Measure	Fricative	Mean	SD	Range of Talker Means	Range of Talker SDs
$Freq_M$ (Hz)	s	5656	731	3573 – 6753	198 – 1365
	z	5735	718	3713 – 6856	127 – 1439
	ʃ	3181	501	2178 – 5341	108 – 1334
$Freq_M$ (ERB)	s	30.05	1.23	26.09 – 31.76	0.27 – 2.38
	z	30.17	1.20	26.41 – 31.90	0.17 – 2.70
	ʃ	24.95	1.36	21.85 – 29.64	0.39 – 3.01
COG (Hz)	s	5328	588	3814 – 6587	284 – 1490
	z	4124	901	2010 – 6254	452 – 2069
	ʃ	3689	586	2429 – 5239	224 – 893

As shown in Table 3.8 and Figure 3.2, a near-perfect correlation of talker mean $Freq_M$ was observed between [s] and [z] ($r = 0.96$, $p < 0.001$), and a moderately strong correlation was observed between [s] and [ʃ] ($r = 0.68$, $p < 0.001$). The correlation remained strong between [s] and [z], which share [+anterior] place, within both gender groups (*female*: $r = 0.92$; *male*: $r = 0.92$, ps < 0.001). The correlation for [s] and [ʃ], which contrast on [anterior] but share the [-voice] features, was also significant within both genders albeit reduced in magnitude relative to the correlation magnitude in the population (*female*: $r = 0.49$; *male*: $r = 0.38$, ps < 0.001).

Table 3.8. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means) in American English connected speech.

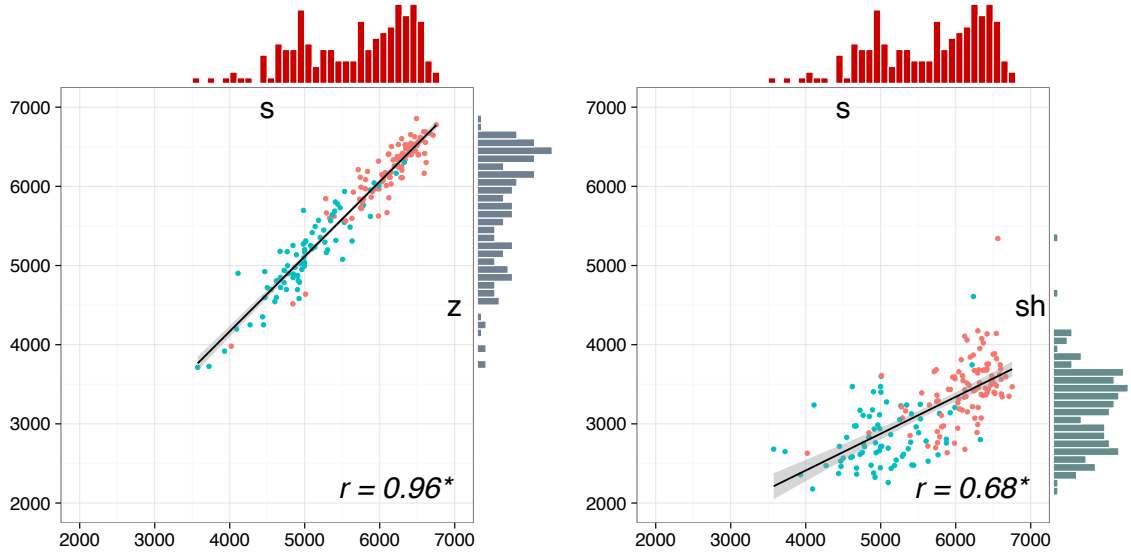
Measure	Fricative Pair	All	Female	Male
$Freq_M$ (Hz)	s – z	0.96* [0.95, 0.97]	0.92* [0.85, 0.96]	0.92* [0.86, 0.95]
	s – ʃ	0.68* [0.60, 0.74]	0.49* [0.35, 0.58]	0.38* [0.16, 0.60]
$Freq_M$ (ERB)	s – z	0.96* [0.95, 0.97]	0.92* [0.84, 0.96]	0.93* [0.88, 0.96]
	s – ʃ	0.69* [0.61, 0.75]	0.50* [0.34, 0.60]	0.38* [0.19, 0.58]
COG (Hz)	s – z	0.62* [0.52, 0.70]	0.63* [0.50, 0.73]	0.57* [0.38, 0.72]
	s – ʃ	0.54* [0.41, 0.63]	0.26* [0.05, 0.44]	0.52* [0.36, 0.63]

As shown in Table 3.9, the simple linear regression between the talker-specific $Freq_M$ means for [s] and [z] indicate that, within each talker, these two sounds have nearly identical phonetic realizations on this dimension. As indicated by the high R^2 value, the linear fit was very good. Additionally, the additive value was small and non-significant, and the scalar value was near unity. In contrast, the fit between [s] and [ʃ] was poor; both additive and scalar values were significant, though the model indicated a primarily additive relationship in which the $Freq_M$ of [s] was approximately 2500 Hz greater than the $Freq_M$ of [ʃ] across talkers.

Table 3.9. Additive (β_0) and scalar (β_1) components of simple linear regressions of mean $Freq_M$ values for sibilant pairs in American English connected speech. For each pair, the talker-specific mean of the first sibilant was the dependent variable predicted from the talker-specific mean of the second sibilant.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
s ~ z	34	0.78	0.98	< 0.001	0.93
s ~ ʃ	2507	< 0.001	0.99	< 0.001	0.46

Figure 3.2. Variation and covariation of sibilant $Freq_M$ means (Hz) across talkers in American English connected speech. Marginal histograms show variation in talker means. Each point represents a talker-specific pair of means and is color-coded to specify talker gender (red = female, blue = male). The asterisk indicates that the correlation reached significance ($p < 0.025$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



As in the laboratory study, a linear mixed-effects model was used to examine the sources of variation in sibilant $Freq_M$ within the population and among individuals. The model was identical in structure, but because the dataset lacked instances of [ʒ] and sibilants before the high, back vowel [u], we did not include the interactions between place and voice or following vowel height and backness. The contrast weighting of the categorical variables was as follows: place of articulation (*place*: +anterior = 1, –anterior = -4.60), voice (*voice*: voiceless = 1, voiced = -8.21), vowel height (*height*: high = 1, non-high = -0.24), vowel backness (*backness*: front = 1, non-front = -0.33), and gender (*gender*: female = 1, male = -1.32). The dependent variable ($Freq_M$) was centered at zero by subtracting the grand mean (μ = 5238 Hz) from each value prior to analysis. The random effect structure had a talker intercept and slopes for place and voice, and a random intercept for word.

There were significant effects of place and voice on $Freq_M$ (*place*: β = 453.83, *t* = 57.61; *voice*: β = 48.74, *t* = 9.21). Vowel backness did not reach significance (β = -31.91, *t* = -1.14) but there was a significant main effect of vowel height: $Freq_M$ was elevated in sibilants preceding high vowels (β = 222.19, *t* = 5.77). In addition, female talkers had significantly higher sibilant $Freq_M$ than male talkers (β = 436.22, *t* = 15.78), and this effect was modulated by a significant interaction between place and gender (β = 38.72, *t* = 7.06). This interaction is consistent with female talkers having a somewhat larger place contrast than male talkers. The interaction between voice and gender did not reach significance (β = 2.26, *t* = 1.63).

The random talker component of the mixed-effects model was also analyzed to determine the primary aspects of talker variation. Table 3.10 shows the standard deviation in the talker intercept and slopes for place and voice. The random intercept had the largest standard deviation, indicating that talkers primarily varied in the overall $Freq_M$ mean, but otherwise maintained highly comparable patterns of $Freq_M$ across sibilant categories. In comparison to the intercept, the standard deviations of the talker slopes of place and voice were substantially smaller: the standard deviation of the intercept was almost five times larger than the standard deviation of the talker anteriority slope and almost 24 times larger than that of the talker voice slope.

Table 3.10. Standard deviations of talker random effects in the maximal mixed-effects model of $Freq_M$ in American English connected speech.

Random effect for talker	SD
intercept	425
place	83
voice	18

3.3.3 Discussion

As in isolated speech, talkers varied considerably in the overall $Freq_M$ means and standard deviations for each sibilant category. Consistent with the predictions of the general notion of pattern uniformity, moderate to strong covariation of mean $Freq_M$ was observed among the sibilant categories. In addition, the mixed-effects model revealed the greatest variation in the realization of $Freq_M$ in the random talker intercept, indicating that talkers primarily differed in the overall deviation from the population pattern. While other aspects of anatomy may nevertheless contribute to the overall variation across talkers, $Freq_M$ was only weakly correlated with talker height within each gender, providing indirect evidence against an account attributing variation to vocal tract length differences (rs = -0.27 to -0.17 for male speakers and rs = -0.14 to 0.03 for female speakers, all ps > 0.01).

The patterns of covariation revealed additional information relevant for the predictions of target and contrast uniformity. In particular, the correlation of talker mean $Freq_M$ between [s] and [z] was quite strong (r = 0.96), and as demonstrated in the mixed-effects analysis, [voice] had only a minimal influence on the actualization of constriction location, as measured by $Freq_M$. A moderately strong correlation of talker mean $Freq_M$ was observed between [s] and [ʃ]; however, this pattern weakened substantially when the data was split by gender, indicating that within each cluster of talker means, there was relatively little systematicity across talkers.

3.4 Covariation of $Freq_M$ in American English spontaneous speech

In addition to the isolated and connected speech styles, we also examined the predictions of uniformity on the mapping from [anterior] to constriction location in the

Buckeye Corpus of Spontaneous Speech (Pitt et al., 2007). The corpus contains interview speech from 40 native talkers of American English, and in contrast to the Mixer 6 corpus, contains a sufficient number of all four sibilant categories for analysis. As the speech is naturally occurring, the relative number of tokens varies across sibilant categories, contexts, and talkers; however, many of these differences were brought under statistical control in the mixed-effects analysis below. Spontaneous speech like that in the Buckeye corpus is of interest because it represents the most common and natural speech style, and because it is known to be highly variable.

3.4.1 Methods

3.4.1.1 Corpus description

The Buckeye Corpus contains speech produced by 40 native speakers of American English from the Columbus, Ohio area (Pitt et al., 2007). The talker demographics were counterbalanced for gender and age, such that there were 20 female, 20 male, 20 "young" (under age 30), and 20 "old" (over age 40) talkers; all speakers were white and middle to upper class. Each talker was interviewed in a quiet room for 30 to 60 minutes on current local issues, and was naïve to the true purpose of the recording until after the interview had concluded. The original recordings were sampled at 48 kHz, but downsampled to 16 kHz for distribution. A word-level and phone-level transcription and alignment were provided with the corpus. Further details regarding the corpus can be found in Pitt et al. (2007).

3.4.1.2 Data preparation and acoustic analysis

The sibilants in the Buckeye corpus were analyzed with the same acoustic measurements and statistical methods as in the previous experiments (see section

3.2.1.4). Tokens greater than 2.5 standard deviations above or below their talker-specific $Freq_M$ means were excluded from analysis as outliers. In total, there were 17,722 sibilants for analysis. As in the Mixer 6 corpus, [s] was over-represented relative to each of the other sibilants; [ʒ] was present here but rarely (see Table 3.11).

Table 3.11. The range and median number of tokens for each sibilant category per talker in American English spontaneous speech. The final column indicates the total number of tokens analyzed per sibilant category.

Fricative	Range	Median	Total
s	120 – 514	304.5	12,359
z	18 – 100	52	2,258
ʃ	44 – 161	87.5	3,710
ʒ	2 – 27	9	395

3.4.2 Results

The population means and standard deviations for each sibilant are presented in Table 3.12. In the spontaneous speech data, the population standard deviations for [s] and [z] were larger than those for [ʃ] and [ʒ]; in contrast to the connected speech data, this pattern was also obtained on the ERB scale. Talker means and standard deviations also ranged considerably. The pattern of correlation between talker mean and standard deviation for each sibilant category was highly comparable to the previous studies, in that there were only weak negative correlations for the alveolar sibilants, but moderate, positive correlations for the post-alveolar sibilants ([s]: $r = -0.28$, 95% CI: [-0.59, 0.10], $p = 0.07$; [z]: $r = -0.15$, 95% CI: [-0.53, 0.20], $p = 0.36$; [ʃ]: $r = 0.43$, 95% CI: [0.18, 0.64], $p < 0.01$; [ʒ]: $r = 0.57$, 95% CI: [0.21, 0.79], $p < 0.001$).

Table 3.12. Descriptive statistics for each sibilant in American English spontaneous speech. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Measure	Fricative	Mean	SD	Range of Talker Means	Range of Talker SDs
$Freq_M$ (Hz)	s	5320	849	3390 – 6451	306 – 1194
	z	5270	845	3579 – 6591	273 – 1218
	ʃ	3296	540	2357 – 4695	233 – 823
	ʒ	3175	558	2348 – 4495	6 – 1138
$Freq_M$ (ERB)	s	29.45	1.53	25.54 – 31.35	0.43 – 2.27
	z	29.36	1.53	25.88 – 31.54	0.39 – 2.25
	ʃ	25.22	1.39	22.49 – 28.45	0.69 – 2.14
	ʒ	24.84	1.45	22.50 – 27.96	0.02 – 2.97
COG (Hz)	s	5248	737	3274 – 6465	313 – 774
	z	4777	860	2803 – 6462	419 – 1647
	ʃ	3800	570	2948 – 4863	253 – 536
	ʒ	3410	579	2283 – 4724	43 – 1315

As shown in Table 3.13 and Figure 3.3, a near-perfect correlation of talker mean $Freq_M$ was observed for the [+anterior] sibilants [s] and [z] ($r = 0.97$, $p < 0.001$), and a very strong correlation was found for the [-anterior] sibilants [ʃ] and [ʒ] ($r = 0.90$, $p < 0.001$). The correlations remained strong between homorganic sibilants within each gender ([s - z] *female*: $r = 0.87$, *male*: $r = 0.96$; [ʃ - ʒ] *female*: $r = 0.86$, *male*: $r = 0.71$, ps < 0.001). Talker mean $Freq_M$ was moderately correlated between the voiceless fricatives [s] and [ʃ] ($r = 0.77$, $p < 0.001$) and the voiced fricatives [z] and [ʒ] ($r = 0.72$, $p < 0.001$), but the strength of the correlations was quite weak within each gender ([s - ʃ] *female*: $r = 0.58$, *male*: $r = 0.66$, ps < 0.01; [z - ʒ] *female*: $r = 0.37$, *male*: $r = 0.50$, ps > 0.01). Interestingly, the decrease in correlation magnitude for the non-homorganic sibilants was not as severe in the spontaneous speech as in the other speech styles, particularly between [s] and [ʃ].

Table 3.13. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means) in American English spontaneous speech.

Measure	Fricative Pair	All	Female	Male
$Freq_M$ (Hz)	s – z	0.97* [0.96, 0.98]	0.87* [0.55, 0.95]	0.96* [0.91, 0.98]
	ʃ – ʒ	0.90* [0.82, 0.95]	0.86* [0.68, 0.93]	0.71* [0.26, 0.84]
	s – ʃ	0.77* [0.63, 0.85]	0.58^{+} [0.14, 0.80]	0.66^{+} [0.45, 0.88]
	z – ʒ	0.72* [0.61, 0.82]	0.37 [-0.09, 0.64]	0.50 [-0.06, 0.77]
$Freq_M$ (ERB)	s – z	0.97* [0.95, 0.99]	0.83* [0.49, 0.94]	0.96* [0.93, 0.98]
	ʃ – ʒ	0.89* [0.80, 0.94]	0.84* [0.60, 0.92]	0.73* [0.45, 0.90]
	s – ʃ	0.78* [0.65, 0.86]	0.58^{+} [0.15, 0.78]	0.66^{+} [0.30, 0.83]
	z – ʒ	0.73* [0.60, 0.82]	0.36 [-0.11, 0.63]	0.52^{+} [-0.04, 0.81]
COG (Hz)	s – z	0.88* [0.74, 0.94]	0.69* [0.31, 0.87]	0.96* [0.90, 0.98]
	ʃ – ʒ	0.90* [0.82, 0.93]	0.79* [0.62, 0.88]	0.83* [0.64, 0.91]
	s – ʃ	0.78* [0.62, 0.87]	0.38 [-0.15, 0.63]	0.70^{+} [0.27, 0.89]
	z – ʒ	0.60* [0.28, 0.78]	0.16 [-0.36, 0.55]	0.67^{+} [0.13, 0.90]

* = $p < 0.001$, $^{+}$ = $p < 0.01$

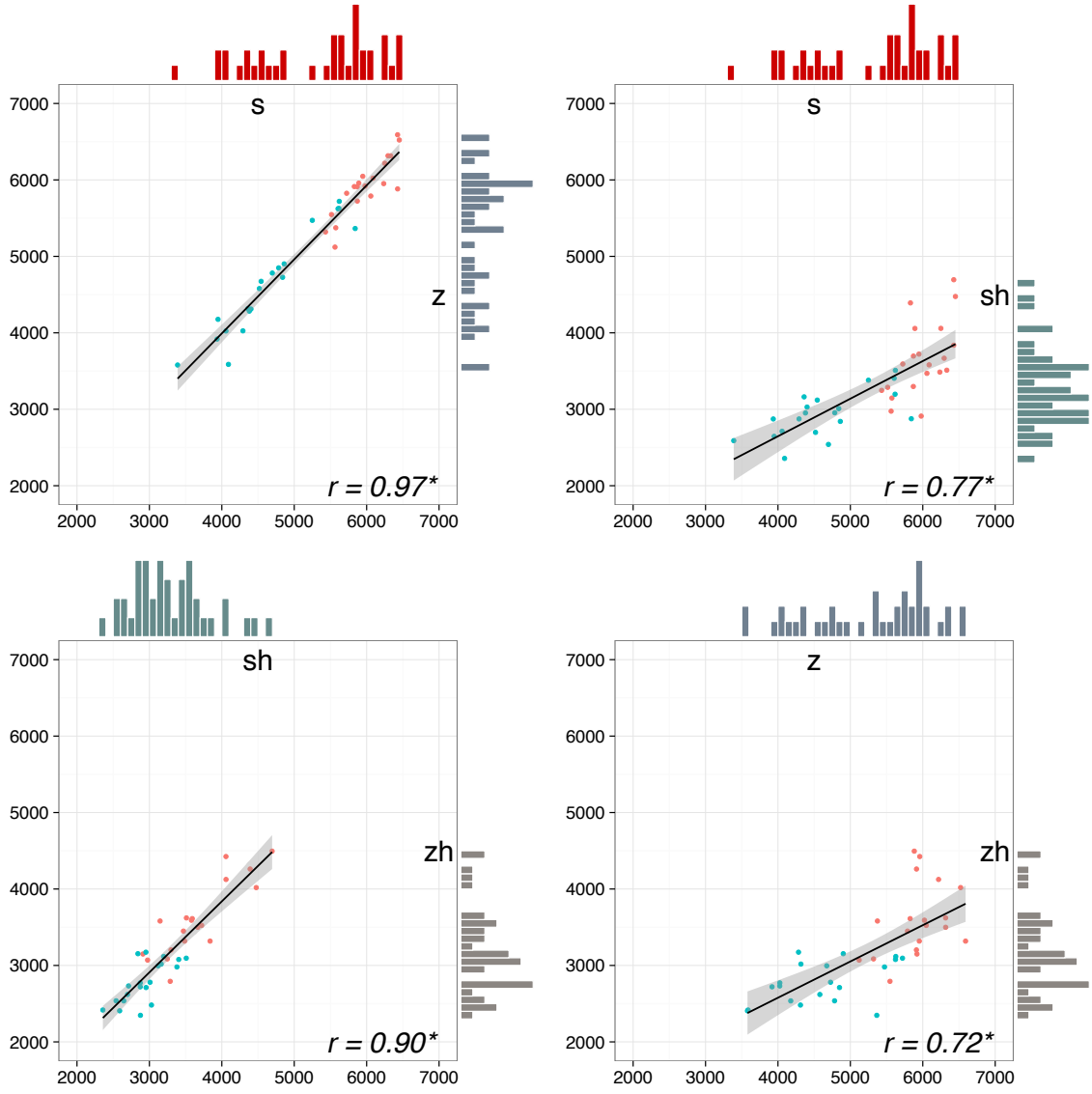
Simple linear regressions predicting the talker mean $Freq_M$ for one sibilant from another were performed as before (Table 3.14). The model fits all exceeded an R^2 of 0.50; however, the fits between sibilants with the same place of articulation were much higher than those between sibilants contrasting in place of articulation. The fitted intercept and scalar values predicting [s] from [z] corresponded to the predictions of target uniformity: the intercept was non-significant and quite small ($\beta_0 = 159$, $p = 0.43$), and the slope was significant and close to unity ($\beta_1 = 0.98$, $p < 0.001$). The fit between [ʃ] and [ʒ] revealed a slightly larger offset between the voiced and voiceless post-alveolar

sibilants. In this case both the intercept and scalar value were significant: the intercept was approximately 500 Hz ($\beta_0 = 498$, $p < 0.05$), and the scalar value was close to but slightly below one ($\beta_1 = 0.88$, $p < 0.001$). Within the range of $Freq_M$ values allowed by our measurement procedure (2000 – 6000 Hz), the separation between the means for [ʃ] and [ʒ] predicted by the linear regression never exceeds 260 Hz.

Table 3.14. Additive (β_0) and scalar (β_1) components of simple linear regressions of mean $Freq_M$ values for sibilant pairs in American English spontaneous speech. For each pair, the talker-specific mean of the first sibilant was the dependent variable predicted from the talker-specific mean of the second sibilant.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
s ~ z	159	0.43	0.98	< 0.001	0.95
ʃ ~ ʒ	498	< 0.05	0.88	< 0.001	0.81
s ~ ʃ	1306	< 0.05	1.22	< 0.001	0.59
z ~ ʒ	1745	< 0.01	1.11	<0.001	0.51

Figure 3.3. Variation and covariation of sibilant $Freq_M$ means (Hz) across talkers in American English spontaneous speech. Marginal histograms show variation in talker means. Each point represents a talker-specific pair of means and is color-coded to specify talker gender (red = female, blue = male). The asterisk indicates that the correlation reached significance ($p < 0.025$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



As in the previous analyses, the sibilant $Freq_M$ tokens were submitted to a linear mixed-effects model. The model was identical in structure to the model reported in the laboratory experiment (section 3.2.2). The contrast weighting of the categorical variables was as follows: place of articulation (*place*: +anterior = 1, –anterior = -3.56), voice

(*voice*: voiceless = 1, voiced = -6.06), vowel height (*height*: high = 1, non-high = -0.52); vowel backnesss (*backness*: front = 1, non-front = -1.21), and gender (*gender*: female = 1, male = -0.90). The dependent variable ($Freq_M$) was centered at zero by subtracting the grand mean (μ = 4828 Hz) from each value.

The results largely paralleled the findings reported in the previous sections of this chapter. The model revealed a large, significant effect of place on $Freq_M$ (β = 442.20, t = 27.32), and a small, but significant effect of voice (β = 17.12, t = 4.46). While vowel height was not significant (β = 7.74, t = -0.58), there was a significant effect of vowel backness, such that $Freq_M$ was higher for sibilants preceding front vowels than back vowels (β = 53.12, t = 6.71). In contrast to isolated speech, the interaction between height and backness did not reach significance here (β = -1.36, t = -0.12). Female talkers had significantly higher $Freq_M$ values than male talkers (β = 643.04, t = 8.79) and, as in the connected speech style, the interaction between place and gender indicates that female talkers have a larger place contrast than male talkers (β = 72.37, t = 4.35). The interaction between voice and gender did not reach significance (β = 2.88, t = 0.92).

As shown in Table 3.15, the random intercept had the largest standard deviation, indicating substantial differences across talkers in overall mean $Freq_M$, and the standard deviations of the random talker slopes were much smaller: the standard deviation of the intercept was over four times larger than the standard deviation of the talker place slope, approximately 25 times larger than that of the talker voice slope, and 27 times larger than the standard deviation of the interaction slope.

Table 3.15. Standard deviations of talker random effects in the maximal mixed-effects model of $Freq_M$ in the American English connected speech.

Random effect for talker	SD
intercept	459
place	99
voice	18
place x voice	7

3.4.3 Discussion

Considerable variation was observed across talkers in the means and standard deviations of $Freq_M$ for each sibilant category in the spontaneous speech style. Patterns of structured variation among sibilants emerged in the spontaneous speech data in a way that closely mirrored the patterns in isolated and connected speech. Consistent with previous findings for spontaneous speech, the overall standard deviation across talkers was higher here than in either the isolated or connected speech styles, as was the range of talker-specific standard deviations.

The strong correlations of talker mean $Freq_M$ between each of the sibilants indicates an influence of pattern uniformity, and more specifically, target uniformity, on sibilant implementation in the phonetic targets underlying $Freq_M$. The variation in the random intercept for talker was considerably greater than the variation in talker slopes for place, voice, or the interaction for place and voice. In direct support of target uniformity, the correlations between categories with a shared place feature were stronger and more robust within each gender than those between categories that contrasted in place of articulation. While the fixed effect of voice was significant in the mixed-effects model, indicating a role of the voice feature in determining the phonetic target for constriction location, the measurement may not perfectly reflect the targets, and critically, the

contribution of the voice feature was almost 25 times *smaller* than that of the phonological place feature ($\beta_{voice} = 17$; $\beta_{place} = 442$).

The findings did not provide a high degree of support for contrast uniformity: the correlations of talker mean $Freq_M$ between sibilants contrasting in place were largely due to a separate cluster of means for each gender; while the correlations between [s] and [ʃ] within each gender reached significance, they weakened in magnitude from the overall correlation, and the correlations between [z] and [ʒ] within gender were not significant. Estimates of the random effects of talker indicated that the greatest source of variation was the intercept, and the variation in talker slope for place of articulation was much larger than that of the voice slope. This again shows sizable variation in the separation between the anterior and posterior sibilant fricatives along the $Freq_M$ dimension. Contrast uniformity, which enforces a consistent contrast size across talkers, may therefore have a relatively weak influence on this aspect of phonetic implementation. Talkers may merely ensure that a sufficient contrast exists among sibilants that differ in place, rather than attempting to make the contrast maximal or of some other consistent magnitude.

3.5 Covariation of $Freq_M$ in Czech spontaneous speech

The pattern of results from American English have provided strong evidence for target uniformity in the phonetic implementation of sibilant place. However, the predictions of both target and contrast uniformity extend beyond English, and therefore warrant investigation in additional languages. For example, the relatively weaker support for uniformity of contrast, in comparison to uniformity of target, could be an idiosyncrasy of English due to the somewhat marginal status of [ʒ]. Alternatively, this may be a more

general asymmetry that holds across languages that have different sibilant inventories and frequency distributions.

This section investigates the predictions of target and contrast uniformity in the sibilant fricatives of Czech. Czech has a full place by voice contrast in its sibilant inventory, and there are many native words beginning with all four sibilants. This allows for a complete examination of both uniformity of target ([s] - [z]; [ʃ] - [ʒ]), as well as uniformity of contrast ([s] - [ʃ]; [z] - [ʒ]). A large multi-talker corpus of spontaneous Czech, the Nijmegen Corpus of Casual Czech, was kindly made available to us for this purpose by Mirjam Ernestus and her colleagues (Ernestus et al., 2014).

Czech is a West Slavic language, spoken predominantly in the Czech Republic and natively by approximately 10 million people (Dankovičová, 1997a). The corpus employed here contains speech from individuals native to Prague and the surrounding Central Bohemian Region. The talkers therefore represent the Bohemian dialect, which is native to approximately 6 million people in the western part of the country (Šimáčková et al., 2012).

In addition to the sibilant fricatives [s z ʃ ʒ], Czech also has the sibilant affricates [tʃ dʒ] and the non-sibilant fricatives [f v x ɦ]. There has been some debate regarding the retroflex status of Czech [ʃ] (e.g., Zygis, 2003; Hamann, 2004). An x-ray tracing from Skaličková (1974) indicates a flat and retracted tongue body, which may be consistent with retroflexion (Hamann, 2002b, 2004); however, x-ray tracings in Palková (1994) and Polland & Hala (1926:23) show a domed tongue shape with the articulation clearly made by the tongue body as opposed to the tongue tip. Additional evidence that the articulation is non-retroflex comes from the phonological distribution. Post-alveolar sibilants that are

realized with a retroflex articulation in other Slavic languages never occur before /i/; however, Czech has no constraint against this sequence (Zygis, 2003).

As the study also references the vowels following sibilants, I will briefly present the Bohemian Czech vowel system. Bohemian Czech has ten monophthongal vowels [ɪ e a o u i e: a: o: u:] and three diphthongs [ou au eu]. The monophthongal vowel system is often described with five vowel qualities that contrast in duration; however, the phonetic contrast between the phonemically long and short high front vowel in Bohemian Czech is primarily spectral as opposed to temporal (Dankovičová, 1997a; Šimáčková et al., 2012).

3.5.1 Methods

3.5.1.1 Corpus description

The analysis of Czech fricatives was accomplished with the Nijmegen Corpus of Casual Czech, which contains spontaneous speech from 60 native speakers of Czech from Prague and the surrounding Central Bohemia region (Ernestus et al., 2014). There were 30 female and 30 male talkers, with ages ranging from 19 to 26 years old. The corpus was collected from informal, spontaneous conversations from 10 groups of three female friends and 10 groups of three male friends. Within each group, there were two naïve talkers, who were unaware of the recording, and one confederate, who was aware of the recording and ensured continuity in the conversation. Each group was recorded for 90 min, resulting in a total of 30 hours of speech. All recordings were sampled at 44.1 kHz.

Orthographic transcriptions of the recordings were provided with the corpus recordings. Both speech and non-speech events were transcribed and manually segmented by utterance (average of 2 seconds of speech) with a start and end time.

Further details pertaining to the corpus collection and corresponding transcription can be found in Ernestus et al. (2014).

3.5.1.2 Corpus preparation

Word-level transcripts accompanied each of the recordings and have had separate track for each speaker. The word-level transcript was aligned to the audio with Czech acoustic models trained using the Kaldi ASR toolkit (Povey et al., 2011). The output of this process was a time-aligned phone- and word-level transcript. The process for deriving this output is as follows: The original transcripts were audited to standardize the transcription. Incomplete or mispronounced words were removed from the transcript and skipped during training and final analysis. A Czech pronunciation dictionary was developed using the words from the transcript, which were then converted into phones with a rule-based grapheme to phoneme conversion script. Czech has a highly phonetic and systematic orthography, making the orthographic to phonetic mapping fairly transparent (Caravolas & Volín, 2001). The pronunciation lexicon also allowed for standard alternations such as obstruent devoicing in utterance-final position (Šimáčková et al., 2012). Words with obvious English orthography and/or pronunciation and a few acronyms were manually corrected to the canonical pronunciation.

The audited transcripts were then used to train acoustic triphone models from which the time alignments were extracted. The models were trained using MFCC features in a standard HMM-GMM training recipe including monophone training and alignment, speaker adaptation, and triphone training and alignment.

3.5.1.3 Acoustic analysis

Word-initial and word-medial [s z ʃ ʒ] in prevocalic position were considered for analysis. As in the previous studies, $Freq_M$ was measured from a multitaper spectrum estimated from the middle 50% of each sibilant. Outliers beyond 2.5 standard deviations of the talker-specific $Freq_M$ mean for each sibilant category were excluded from analysis, resulting in a total of 51,982 sibilants. As shown in Table 3.16, the median number of tokens per talker was 359.5 [s], 122 [z], 69.5 [ʃ], and 213 [ʒ].

Table 3.16. Range and median number of tokens per talker and fricative, and total number of tokens per fricative in Czech spontaneous speech.

Fricative	Range	Median	Total
s	130 – 897	359.5	24,159
z	45 – 353	122	8,239
ʃ	24 – 178	69.5	4,799
ʒ	45 – 576	213	14,785

3.5.2 Results

Within each sibilant category, the variation in talker mean $Freq_M$ was extensive, ranging over approximately 3000 Hz for the alveolar fricatives and 1500 Hz for the post-alveolar fricatives (Table 3.17). There was also considerable variation in the talker-specific standard deviations, and the extent to which the mean and standard deviation were mutually predictable varied by sibilant category. The correlation between these two parameters was virtually non-existent for [s] ($r = 0.03$, $p = 0.83$, 95% CI: [-0.25, -0.30]), weak for [z] ($r = 0.38$, $p = 0.003$, 95% CI: [0.12, 0.59]), but quite strong for [ʃ] ($r = 0.77$, $p < 0.001$, 95% CI: [0.59, 0.87]), and for [ʒ] ($r = 0.85$, $p < 0.001$, 95% CI: [0.74, 0.92]).

Table 3.17. Descriptive statistics for each sibilant in Czech spontaneous. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Measure	Fricative	Mean	SD	Range of Talker Means	Range of Talker SDs
$Freq_M$ (Hz)	s	5362	731	3736 – 6538	419 – 1307
	z	5099	729	3745 – 6399	359 – 1513
	ʃ	3195	492	2351– 4473	163 – 977
	ʒ	3065	426	2379 – 4022	255 – 1087
$Freq_M$ (ERB)	s	29.55	1.26	26.47 – 31.45	0.62 – 2.34
	z	29.03	1.29	26.49 – 31.22	0.84 – 2.78
	ʃ	24.97	1.31	22.50 – 28.03	0.57 – 2.23
	ʒ	24.59	1.15	22.60 – 26.90	0.94 – 2.72
COG (Hz)	s	6574	1101	4646 – 8951	407 – 1653
	z	4979	1230	2837 – 7741	1025 – 2663
	ʃ	4027	668	2818 – 5382	264 – 970
	ʒ	3494	612	2450 – 4805	306 – 1225

Correlations of talker mean $Freq_M$ between sibilant categories were quite strong. As shown in Table 3.18 and Figure 3.4, near-perfect correlations were observed between homorganic fricatives ([s – z]: $r = 0.94$, [ʃ – ʒ]: $r = 0.95$, ps < 0.001). In addition, strong correlations were observed between fricatives contrasting in anteriority ([s – ʃ]: $r = 0.71$, [z – ʒ]: $r = 0.72$, ps < 0.001); however, the strength of these correlations was likely due to the bimodal distribution of talker $Freq_M$. When considering each gender separately, the correlations between homorganic sibilants remained strong ([s – z] *female*: $r = 0.85$, [s – z] *male*: $r = 0.86$, [ʃ – ʒ] *female*: $r = 0.92$, [ʃ – ʒ] *male*: $r = 0.79$, ps < 0.001), but the correlations between sibilants contrasting in place decreased substantially ([s – ʃ] *female*: $r = 0.27$, [s – ʃ] *male*: $r = 0.41$, [z – ʒ] *female*: $r = 0.30$, [z – ʒ] *male*: $r = 0.31$, ps > 0.001). This suggests that the correlations between sibilants contrasting in anteriority largely arose from the presence of two clusters of talker-specific means (male and female), in comparison to there being a consistent relationship between individual talker means.

Table 3.18. Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means for $Freq_M$ (Hz) in Czech spontaneous. For each fricative pairing, correlations are provided first for all talkers together, then within each gender category.

Measure	Fricative Pair	All	Female	Male
$Freq_M$ (Hz)	s – z	0.94* [0.91, 0.96]	0.85* [0.63, 0.92]	0.86* [0.70, 0.93]
	ʃ – ʒ	0.95* [0.92, 0.96]	0.92* [0.86, 0.95]	0.79* [0.65, 0.89]
	s – ʃ	0.71* [0.55, 0.79]	0.27 [-0.08, 0.53]	0.41 [0.11, 0.69]
	z – ʒ	0.72* [0.59, 0.81]	0.30 [-0.10, 0.55]	0.31 [-0.02, 0.55]
$Freq_M$ (ERB)	s – z	0.93* [0.89, 0.95]	0.80* [0.55, 0.90]	0.84* [0.66, 0.93]
	ʃ – ʒ	0.94* [0.90, 0.96]	0.92* [0.86, 0.95]	0.79* [0.65, 0.89]
	s – ʃ	0.71* [0.55, 0.80]	0.23 [-0.12, 0.50]	0.44 [0.17, 0.67]
	z – ʒ	0.73* [0.59, 0.82]	0.34 [-0.04, 0.63]	0.31 [-0.02, 0.55]
COG (Hz)	s – z	0.85* [0.77, 0.91]	0.75* [0.54, 0.88]	0.50+ [0.08, 0.77]
	ʃ – ʒ	0.92* [0.80, 0.95]	0.90* [0.79, 0.94]	0.76* [0.52, 0.90]
	s – ʃ	0.76* [0.62, 0.84]	0.37 [0.05, 0.62]	0.62* [0.30, 0.81]
	z – ʒ	0.84* [0.75, 0.89]	0.51+ [0.19, 0.73]	0.69* [0.47, 0.83]

* = $p < 0.001$, $^{+}$ = $p < 0.01$

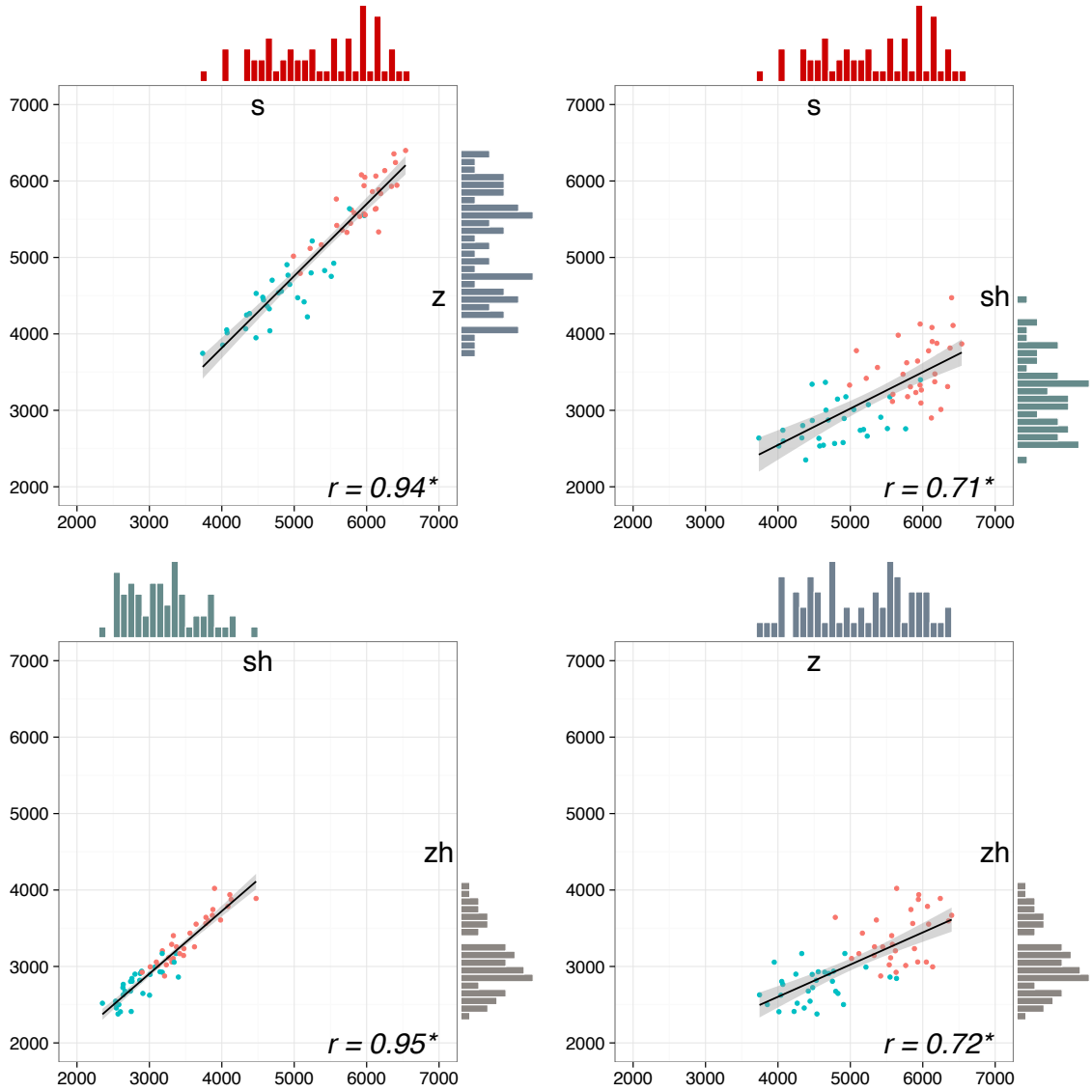
In Table 3.19 are the additive and scalar factors for the pairwise linear regressions, predicting the talker mean $Freq_M$ for one sibilant from another. As in the American English spontaneous speech, the best model fits as measured by the adjusted R^2 were among the homorganic pairs. In predicting [s] from [z], both the additive and scalar components were significant. The offset between [s] and [z] was approximately 540 Hz ($\beta_0 = 540$, $p < 0.05$), and the scalar component was quite close to unity ($\beta_1 = 0.95$, $p < 0.001$). While target uniformity predicts identity between these pairs in the underlying

targets, the measured $Freq_M$ may have been slightly lower for [z] due to a small interaction with low-frequency energy due to phonetic voicing. Between [ʃ] and [ʒ], the additive component was negative but not significant ($\beta_0 = -156$, $p = 0.31$), and the scalar factor was just over 1 ($\beta_1 = 1.09$, $p < 0.001$). Within the $Freq_M$ range (2000 – 6000 Hz), the model predicted a maximum difference of 384 Hz between [ʃ] and [ʒ] and only marginal differences at the low end of the scale.

Table 3.19. Additive (β_0) and scalar (β_1) components of simple linear regressions of mean $Freq_M$ values for sibilant pairs in Czech spontaneous speech. For each pair, the talker-specific mean of the first sibilant was the dependent variable predicted from the talker-specific mean of the second sibilant.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
s ~ z	540	< 0.05	0.95	< 0.001	0.95
ʃ ~ ʒ	-156	0.31	1.09	< 0.001	0.89
s ~ ʃ	1999	< 0.001	1.05	< 0.001	0.49
z ~ ʒ	1312	< 0.01	1.24	<0.001	0.51

Figure 3.4. Variation and covariation of sibilant $Freq_M$ means (Hz) across talkers in Czech spontaneous speech. Marginal histograms show variation in talker means. Each point represent a talker-specific pair of means and is color-coded to specify the talker gender (red = female, blue = male). The asterisk indicates that the correlation reached significance ($p < 0.025$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



Variation in sibilant $Freq_M$ within the population and among individuals was assessed using a linear mixed-effects model that was nearly identical in structure to that

reported in sections 3.2.2 and 3.4.2. The present model included main effects of vowel height, backness, and rounding instead of the interaction between height and backness.

As in the previous models, categorical factors were weighted effect coded according to their sample size. The contrast weighting of the categorical variables had the following values: anteriority (*place*: +anterior = 1, –anterior = -1.65), voice (*voice*: voiceless = 1, voiced = -1.26), vowel roundness (*round*: round = 1, non-round = -0.10), vowel height (*height*: high = 1, non-high = -0.34); vowel backness (*backness*: front = 1, non-front = -3.39), and gender (*gender*: female = 1, male = -1.07). The dependent variable ($Freq_M$) was centered at zero by subtracting the grand mean (μ = 4474 Hz) from each value.

The model revealed a significant effect of place (β = 830.74, t = 37.61), as well as a small, but significant effect of voice (β = 94.19, t = 8.17). The interaction between place and voice also reached significance and indicated a slightly larger separation between the alveolar sibilants than between the post-alveolar sibilants (β = 24.04, t = 3.45). There were also significant main effects of vowel rounding, height, and backness (*round*: β = -388.26, t = -18.58, *height*: β = -93.80, t = -7.49, *backness*: β = 37.74, t = 8.53). A significantly higher $Freq_M$ was observed among female talkers than among male talkers (β = 467.50, t = 11.35), and female talkers had a significantly larger place contrast than male talkers (β = 87.20, t = 4.22). The interaction between voice and gender did not reach significance (β = -4.97, t = -0.51); however, the three-way interaction between anteriority, voice, and gender was significant and likely reflected the slightly smaller contrast between the post-alveolar sibilants than between the alveolar sibilants among male talkers (β = -13.13, t = -2.54).

Consistent with the previous patterns of talker random effects, the random intercept for talker had the largest standard deviation (Table 3.20), indicating that talkers primarily differed in the overall offset from the population $Freq_M$ pattern across sibilants. In comparison to the intercept, the standard deviations for the random talker slopes for place, voice, and the interaction between place and voice were considerably smaller, yet the standard deviation in the talker slope for anterior was sizable, indicating that talkers also varied in the separation of $Freq_M$ between the anterior and posterior sibilants.

Table 3.20. Standard deviations of talker random effects in the maximal mixed-effects model of $Freq_M$ in Czech spontaneous speech.

Random effect for talker	SD
intercept	328
place	164
voice	72
place x voice	35

3.5.3 Discussion

Overall, the findings within Czech spontaneous speech strongly mirrored those of American English spontaneous speech. Substantial variation was observed in the talker means and standard deviations of $Freq_M$ for each sibilant category, yet the variation was not independent for each category. Rather, the talker means of $Freq_M$ were highly correlated among the sibilants. As with American English, the strongest correlations were observed between sibilants with a shared place of articulation, and while the correlations between sibilants contrasting in place were strong across the population, they were largely attributable to systematic variation across gender, as opposed to across individuals. The mixed-effects model also revealed a very similar pattern of fixed effects in that the effect of place accounted for a large amount of the variation in $Freq_M$ with a

significant, but very modest contribution of the voice effect. The analysis of the random effects component for talker again revealed that greatest source of variation across talkers was in the overall sibilant mean $Freq_M$, but with notable differences in the degree of contrast between the two places of articulation. This pattern of results is consistent with a constraint of target uniformity on the phonetic implementation of the [anterior] feature of sibilant fricatives. The role of contrast uniformity on phonetic implementation may be weak, if at all present.

3.6 General discussion

Highly comparable findings were observed across multiple speech styles within American English and in spontaneous speech in Czech. Talkers varied considerably in the phonetic implementation of sibilant fricatives, and this variation was also highly structured among the sibilant categories. Substantial evidence was observed for target uniformity, which constrains the degree of within-segment context-sensitivity in the phonetic implementation of the phonological surface segment. Specifically, the acoustic-phonetic correlate of the constriction location, $Freq_M$, was nearly identical for sibilants that shared the same value of the [anterior] feature. This was demonstrated in part through strong correlations of talker mean $Freq_M$ between [s] and [z], as well as between [ʃ] and [ʒ], and from the mixed-effects analysis: there was only a modest influence of [voice] in comparison to [anterior] on the realization of $Freq_M$. In addition, there was minimal variation across talkers in the effect of [voice] on $Freq_M$, indicating that talkers largely conformed to the population pattern. While the effect of [voice] was small, it did reach significance for many of the models. This could be due either to the acoustic measurement or the underlying phonetic target, which counter to target uniformity, would

indicate marginal context-sensitivity in the mapping from distinctive feature values to phonetic targets.

Evidence for contrast uniformity, however, was relatively weak in each of these studies. The random talker component of the mixed-effects model showed that while the primary dimension of variation was in the overall mean $Freq_M$, talkers nevertheless varied in the place slope, or degree of separation between [+anterior] and [-anterior] sibilants. In addition, the correlations of talker means between [s] and [ʃ], as well as [z] and [ʒ], were largely driven by two clusters of talker means corresponding to gender. Within each gender, the contrast between the sibilants was not consistent across talkers. The contrast between sibilants may nevertheless be governed by other principles: there may be upper- and lower- bounds on the degree of separation, but the present findings did not reveal a uniform implementation of this contrast across talkers.

The more general constraint of pattern uniformity could also account for some of the findings, but its influence would be quite similar to target uniformity, as the correlations are strongest between segments that share an [anterior] feature value. Nonetheless, the primary dimension of variation across talkers, as indicated in the mixed-effects models, is in the grand mean, and variation in the degree of separation between place and voice is comparatively less than that of the intercept.

Qualitatively, our findings are compatible with Maniwa et al. (2009), in which the population mean $Freq_M$ is highest in isolated clear speech (e.g., [s]: 5968 Hz), lower in connected read speech ([s]: 5656 Hz), and lowest in spontaneous speech (e.g., [s]: 5320), but this pattern is much stronger for the alveolar sibilants than for the post-alveolar sibilants (isolated speech [ʃ]: 3304 Hz; connected speech [ʃ]: 3181 Hz; spontaneous

speech [ʃ]: 3296 Hz). The Czech spontaneous speech sibilant means are also quite comparable to the American English spontaneous speech means, indicating that there may not be large differences in the phonetic implementation of sibilants between these two languages. Nevertheless, this has not yet been substantiated by statistical comparison. In a future analysis, we plan to standardize the sampling rates across corpora to allow for numerical comparison of the sibilant $Freq_M$ values.

Further research should also examine additional sociolects of American English, as well as additional languages. One intriguing line of research would be to determine whether the target uniformity also applies to processes of sound change. For instance, Stuart-Smith (2016) reported ongoing change in the phonetic realization of [s] within Glaswegian vernacular; is it the case that [z] follows suit? In this sense, target uniformity resembles the notion of *parallel shifts* posited in Fruehwald (2013) and (2017), which indicates that sound change tends to target distinctive feature values. So far, this has only been examined in vowels.

Analyzing additional sociolects and languages may also uncover cases in which the phonetic implementation of sibilant place differs substantially from American English or Czech. There may exist languages which violate the uniformity constraints more so than others. Based on the present data, however, the prediction is that languages should generally abide by target uniformity with few violations of its predictions. In comparison, contrast uniformity may play little role in the phonetic mapping from phonological features to phonetic targets. This remains to be seen in further studies of sibilants and other speech sounds within particular languages and cross-linguistically.

4 Chapter 4

4.1 Introduction

A central issue in speech perception concerns the processes and mechanisms underlying talker adaptation. In spite of substantial acoustic-phonetic variation across people (one aspect of the well-known lack of invariance between acoustic signals and intended speech categories), listeners adapt rapidly to the speech patterns of novel talkers (e.g., Norris et al., 2003; Clarke & Garrett, 2004; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006; Bradlow & Bent, 2008; Maye et al., 2008). There are likely many mechanisms involved in rapid and general talker adaptation, including processes of intrinsic normalization, distributional learning, top-down modulation, and extrinsic normalization. The existence of such mechanisms is typically supported by carefully-controlled laboratory speech perception experiments, and they have, to varying degrees, been incorporated into theories and computational models of adaptation.

Intrinsic normalization makes use of information internal to the speech sound, such as the distance between formants and the relationships between formants and fundamental frequency. This type of normalization likely underlies the ability of listeners to identify isolated vowels well above chance (e.g., Ainsworth, 1975; Strange et al., 1976), and has been formalized in models of talker adaptation by encoding dependencies among segment-internal phonetic properties such as F1 and f0 (e.g., Nearey, 1978; Syrdal & Gopal, 1986). Listeners can also adapt to talkers over time through distributional learning: they update expectations regarding category-specific parameters after exposure to the talker's distribution of cues (e.g., Clayards et al., 2008; Kleinschmidt & Jaeger, 2015). Knowledge of talker identity can also influence speech categorization and

expectations about talker acoustics in a top-down manner (e.g., Strand & Johnson, 1996; Najafian et al., 2014; Kleinschmidt & Jaeger, 2015; Tatman, 2016). Explicit knowledge about the talker's gender or accent can modulate a listener's expectations about the acoustic category boundaries between minimally different sounds.

An additional mechanism linked to talker adaptation is extrinsic normalization, in which listeners combine information from multiple speech sounds to form a general model of a talker's speech pattern. Listeners achieve greater accuracy in vowel categorization with exposure to multiple vowel categories from the same talker (e.g., Ainsworth, 1975; Assmann et al., 1982), and more generally transfer talker-specific phonetic properties across speech sounds. For example, several studies have demonstrated that learning a talker's characteristic stop VOT values transfers across different places of articulation (see Chapter 2; Eimas & Corbit, 1973; Theodore & Miller, 2010; Nielsen, 2011; cf. Clarke & Luce, 2005). Similarly, listeners actively adjust talker-specific vowel spaces and consonant boundaries after exposure to manipulated formants in a preceding auditory context (e.g., Ladefoged & Broadbent, 1957; Mann, 1980; Lotto & Kluender, 1998; Holt, 2005; Laing et al., 2012). Thus adaptation to a new talker is not entirely a process of learning the acoustic-phonetic properties of individual sounds: it also involves generalization of talker-specific characteristics across classes of related sounds.

Generalized perceptual adaptation has been attributed to a variety of sources, including vocal tract length normalization, compensation for coarticulation, and general auditory mechanisms such as normalization with the long-term average spectrum (LTAS). An alternative account examined in the present series of experiments is that listeners have prior perceptual knowledge of phonetic covariation among speech sounds.

Knowledge of phonetic covariation could be derived via knowledge of pattern, target, and/or contrast uniformity constraints on phonetic implementation. Alternatively, it could result from directly tracking covariation of phonetic properties among speech sounds across talkers. In either case, prior knowledge of phonetic covariation makes unique predictions regarding the expected patterns of generalized adaptation.

The present study examined whether listeners exploit covariation among fricatives (see Chapter 3) to generalize talker-specific spectral distributions from one fricative to another. Many previous studies have examined perceptual learning in fricatives, and have established that listeners can readily adapt their perception of these sounds to novel accents or talkers. To our knowledge, however, no previous study has examined whether listeners transfer talker-specific spectral properties across fricative categories, particularly after exposure to speech sounds that are perceptually unambiguous. The hypothesis that knowledge of covariation supports generalized adaptation to fricatives was compared with a general auditory hypothesis based on spectral contrast and a cue-based normalization hypothesis, each of which is described in further detail below.

4.1.1 Perceptual adaptation to fricatives

Listeners show remarkable flexibility in their perception of speech sounds. Perceptual learning of dialect- or talker- specific acoustic distributions has been established for many sound categories including vowels (e.g., Maye et al., 2008), fricatives (e.g., Norris et al., 2003; Kraljic & Samuel, 2005), and stop consonants (e.g., Allen & Miller, 2004; Kraljic & Samuel, 2006). Both top-down and bottom-up processes contribute to this flexibility. For example, a category boundary can be retuned to

accommodate an acoustically ambiguous fricative that is disambiguated by lexical knowledge (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006) or by visual input (e.g., Bertelson et al., 2003), or to accommodate talker-specific acoustic distributions shifted relative to the population or a second talker (e.g., Allen & Miller, 2004; Maye et al., 2005; Theodore & Miller, 2010).

A large number of experiments have focused explicitly on perceptual adaptation to fricatives. For instance, listeners adjust their perceptual boundary of an ambiguous [s]-[f] sound depending on whether it was embedded in lexical items biasing [s] interpretations or [f] interpretations (Norris et al., 2003; Eisner & McQueen, 2005; McQueen et al., 2006). The adjusted perceptual boundary generalizes to words that were not heard in the initial exposure period (McQueen et al., 2006), and this type of perceptual learning can be long-lasting: listeners can maintain the same categorization of inherently ambiguous fricatives after a 25-minute break (Kraljic & Samuel, 2005) and even after a full night's sleep (Eisner & McQueen, 2006).

Perceptual adaptation to the spectral properties of fricatives is also highly specific to a given talker and category combination (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). Eisner & McQueen (2005) found that perceptual learning affected [s]-[f] categorization only when the fricatives specifically were produced by the same talker in exposure and at test, even if the following vowel was produced by a novel talker. No effect was obtained when the exposure and test fricatives were produced by clearly different talkers. Additional evidence for talker-specificity in fricative perceptual learning was found in Kraljic & Samuel (2007). Listeners received exposure to a male and female voice with an acoustically ambiguous [s]-[ʃ] sound. Critically, the sound was

disambiguated in a lexically-biasing context, but in opposite directions for the male and female voice. At test, listeners shifted their perceptual boundary in accordance with the exposure voice, i.e., in opposite directions. This indicates that listeners maintained distinct representations of the fricative as produced by the two speakers.

4.1.2 Perceptual generalization and spectral contrast

While evidence from perceptual adaptation and learning indicates that listeners are highly sensitive to talker-specific fricative acoustics, it remains to be seen whether listeners transfer what they learn about a new talker beyond the particular fricative categories presented in exposure. The lexical disambiguation studies reviewed above have demonstrated a shift in the category boundary between two sounds (e.g., [s]-[ʃ]), but cannot speak to whether listeners adjusted their representations of both sounds or only the one supported by top-down lexical knowledge.

Several studies have examined generalization of talker-specific spectral properties for other sound classes. Most notably, Ladefoged & Broadbent (1957) demonstrated that listeners interpret a talker's vowel space relative to the preceding speech context. A vowel identified as [ɪ] with no preceding context was overwhelmingly more likely to be heard as [ɛ] following a carrier phrase in which vowel F1 was systematically lowered. Similarly, perceptual identification of a vowel changed from [ʌ] to [æ] following a phrase in which F2 was lowered. The authors argued that the formant structure of a vowel *relative* to an overall pattern of vowel realization was more important to identification than absolute formant values. Comparable results were found in Maye et al. (2008), in which listeners generalized a talker's lowered F2 in the front vowel [ɪ] to perception of the back vowel

[ʊ], without any prior exposure to the talker's back vowel pronunciations; this generalization persisted one to three days after exposure to the talker's front vowels.

More recent findings indicate that perceptual 'generalization' can also arise from more general properties of the preceding context (e.g., Watkins & Makin, 1994, 1996; Holt, 2005, 2006; Sjerps et al., 2011; Laing et al., 2012). In particular, both speech and non-speech acoustic precursors have been shown to influence speech perception. One prominent account is that auditory sound processing is affected by *spectral contrast*, such that high frequency components of a stimulus are enhanced in the context of low frequency precursors and vice versa (e.g., Lotto & Holt, 2006). In essence, exposure to greater energy in a particular frequency range or band temporarily decreases sensitivity to that part of the spectrum, and thus comparatively enhances other frequency components.

Early support for the spectral contrast proposal came from a perceptual phenomenon that was previously attributed to compensation for coarticulation. When they categorize stimuli ranging from [da] to [ga] along an F3 continuum, listeners reported more [ga] responses when the preceding syllable is [al] than when it is [aɹ] (Mann, 1980). This could occur because listeners attribute the coronal acoustics of the stop to coarticulation with the preceding lateral liquid, and effectively 'subtract' this coarticulatory influence when categorizing the stop. This articulatorily-based account can also be reframed as an acoustic-phonetic account in which listeners correct for the influence of [l], which has a higher F3 than [ɹ], on the third formant (F3) of the following stop. As [d] has a higher F3 than [g], removing the effect of the relatively high F3 in [l] could result in a greater number of [g] responses. Importantly, the effect on [d]-[g] categorization was subsequently replicated with a series of sine wave tone precursors that

had a high mean frequency in the F3 region (2300 Hz), mimicking the effect of [al], or a low mean frequency (1800 Hz), similar to [aɹ] (e.g., Lotto & Kluender, 1998; Holt, 2005). This suggest that compensation for coarticulation may be rooted in general auditory processes rather than phonetic knowledge and computations specifically.

Spectral contrast effects have also been linked to talker adaptation: by tracking the long-term average spectrum (LTAS) of a speech or non-speech signal, listeners could adjust their phonetic boundaries using spectral contrast (e.g., Lotto & Kluender, 1998; Lotto et al., 2003; Holt, 2005). Together with the alternative account of compensation for coarticulation effects, this is part of a broader hypothesis that the object of speech perception is auditory rather than articulatory.

Importantly, spectral contrast should be operative only to the extent that the precursor has frequency components in the range that is relevant for the phonetic categorization under question. For instance, Laing et al. (2012) found a significant contrast in the proportion of [ga] responses following sine wave tones differing in mean frequency, but only when the tones were within the F3 frequency range (1656 – 3070 Hz). When the tones varied within the F1 frequency range (125 – 870 Hz), there was no significant differences in [da]-[ga] categorization. Because the LTAS differed under both conditions, the lack of perceptual normalization in the F1-manipulation condition suggests that listeners instead track relative energy in different frequency bands rather than a single average value. Additional evidence that the manipulated frequency range must be within the range relevant for phonetic distinctions was found in Sjerps et al. (2011).

4.1.3 Present study and predictions

The present study investigated whether listeners generalize talker-specific spectral characteristics from one fricative to another. Specifically, [s]-[ʃ] categorization was tested in a series of experiments that manipulated spectral center of gravity (COG) for several types of context sounds: [z], [v], speech-shaped noise, and alternating presentations of speech and noise (Experiments 1 – 4). In addition, we examined the influence of the delay between exposure and test in two additional experiments (Experiments 5 – 6). The predictions of several accounts of generalized adaption—spectral contrast, cue-based normalization, and phonetic covariation—were considered in light of the adaptation patterns found experimentally.

The spectral contrast account makes the following predictions: *high* frequency energy in a preceding sound should enhance *low* frequency energy present in a subsequent sound (and vice versa), shifting perception *contrastively*. Adaptation should occur only when precursor sounds have energy in frequency ranges that are relevant for perception (discrimination or categorization) of target stimuli. Critically, non-speech precursors should elicit the same effects as matched speech contexts (e.g., Lotto & Kluender, 1998; Holt, 2005, 2006; Laing et al., 2012).

The cue-based normalization account is based on quantitative models of extrinsic normalization that employ either mean subtraction or z-scoring across multiple phonetic categories for a single phonetic cue (e.g., for vowels: Lobanov, 1971; Nearey, 1978; for fricatives: McMurray & Jongman, 2011). In this account, members of a natural class of sounds can be characterized by a common set of acoustic/auditory *cues* (e.g., formants for vowels, burst spectra and formant transitions for stops, spectral shape for fricatives), and

cue values for each sound in a class are represented relative to a cue-specific mean (and, some accounts, standard deviation). An assumption underlying mean subtraction in particular is that talkers share a sufficiently *uniform pattern* of cues for each natural class, such as vowels or fricatives, such that talkers differ primarily in the overall offset of this pattern on the cue dimension (perhaps due to differences in vocal tract length or other aspects of vocal tract anatomy). Adaptation involves determining the talker-specific statistical distribution for each cue and shifting the expected values of *all* members of the class accordingly (i.e., by mean subtraction or z-scoring).

The phonetic covariation account is motivated by the observation that members of a natural class have acoustic-phonetic distributions that *covary* across talkers to varying degrees. As demonstrated in the following section, talker mean COGs for [s] and [z] are highly correlated (see also Chapter 3), but talker means for [s] and [v] are not. Listeners could infer talker-specific parameters in a way that takes into account this pattern of partial covariation. For example, if a listener hears a talker with a relatively high COG for [z], this would license the inference that the talker will also have a relatively high value for [s]. In contrast, exposure to a high COG for [v] should give rise to little if any inferences regarding the talker's [s]. This is in stark contrast to cue-based normalization, which by definition assumes that all sounds pattern together equally across talkers.

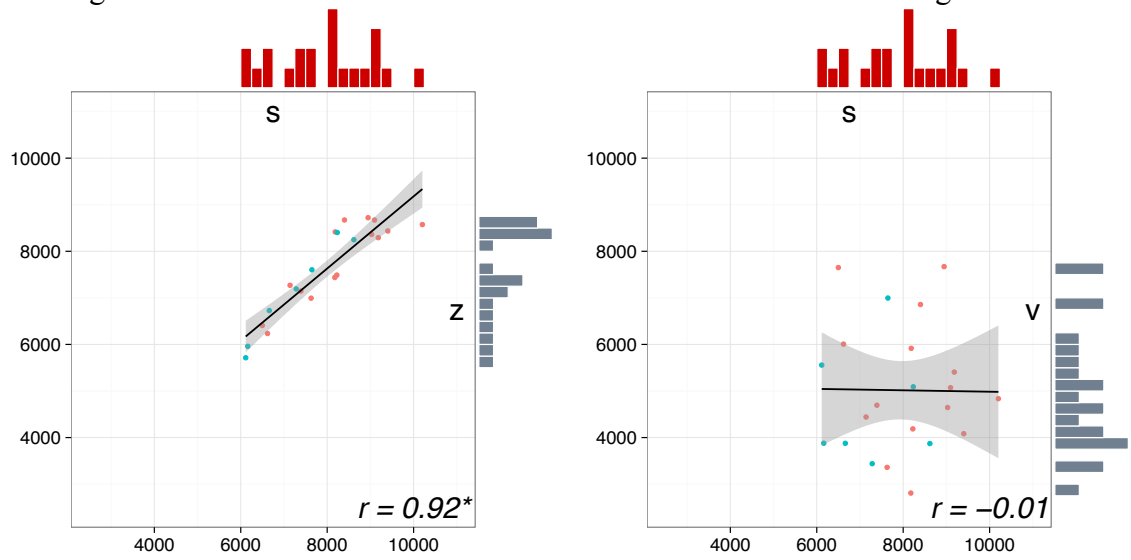
4.2 Fricative spectral correlations

The following studies were designed primarily based on the patterns of COG covariation among American English fricatives in the NYU corpus of isolated speech presented in Chapter 3. As described in section 3.2.1, the corpus was recorded in a laboratory setting with talkers producing CVC syllables composed of an initial fricative

[θ ð f v s z ʃ ʒ] crossed with 10 vowels [i ɪ eɪ ɛ æ ʌ a ɔ oʊ u]. The COG was measured from a multitaper spectrum taken over the middle 50% of the manually-aligned fricatives after high-pass filtering at 550 Hz. High-pass filtering removed any low frequency energy due to vocal fold vibration, resulting in a spectrum that largely reflects filtering by the constriction and anterior vocal tract cavity.

The talker mean COG was calculated for each fricative category, and the correlations among the fricatives were calculated. As shown in Figure 4.1, the correlation of talker COG means for [s] and [z] was quite high ($r = 0.92$, $p < 0.001$) but nonexistent for [s] and [v] at $r = -0.01$ ($p = 0.96$). Note that the correlation between [ʃ] and [z] was only moderate at $r = 0.54$ ($p = 0.009$) and the correlation between [ʃ] and [v] was weak at $r = 0.14$ ($p = 0.53$). Because of this, we will focus primarily on the distinct predictions made by the correlations of [z] and [v] with [s]. This correlation pattern was also found in the fricative productions of 180 talkers from the Mixer 6 corpus ([s – z]: $r = 0.62$, $p < 0.001$; [s – v]: $r = 0.09$, $p = 0.22$; [ʃ – z]: $r = 0.22$, $p = 0.002$; [ʃ – v]: $r = -0.10$, $p = 0.19$; see Chapter 3, section 3.3.1 for details of the corpus).

Figure 4.1. Variation and covariation of COG means (Hz) across talkers in the isolated speech. Marginal histograms show variation in talker means. Each point is a talker-specific pair of means and is color-coded to specify the talker gender (red = female, blue = male). The asterisk indicates that the correlation reached significance ($p < 0.001$). Gray shading reflects the local confidence interval around the best-fit linear regression line.



4.3 Experiment 1: Exposure to [z]

The first experiment tested whether listeners transfer talker-specific spectral characteristics from [z] to [s], and if so, whether this occurred early after exposure. Listeners heard [z]-initial syllables with a relatively high or low COG [z] and categorized members of an [s]-[ʃ] continuum. Exposure and categorization were combined as two parts of a single trial, so that the time course of any effect on continuum perception could be investigated. All three accounts of talker adaptation considered here make the prediction that listeners should generalize from [z] to [s].[29]

Spectral contrast: The distribution of energy in the high and low COG [z] may have a low-level auditory influence on the perception of energy in the frequency range

[29] Listeners may alternatively generalize talker-specific detail from [z] to the [s]-[ʃ] boundary, or generalize talker-specific detail from [z] to both [s] and [ʃ]. These alternatives are quite plausible, but cannot be distinguished in the present experiments. Alternative methods, such as goodness ratings for each category separately, may shed light on exactly which representations are affected by exposure.

relevant for the [s]-[ʃ] contrast, and shift perception contrastively. Specifically, both [z] and [s] are characterized by a higher concentration of energy than [ʃ]. If [z] and an ambiguous [s]-[ʃ] are heard in succession, then perception should shift contrastively: a high COG [z] should result in a greater number of [ʃ] percepts, as the high-frequency components of the stimulus are perceptually dampened, whereas a low COG [z] should result in a greater number of [s] percepts, as the high-frequency components are relatively unaffected (and lower components are suppressed).

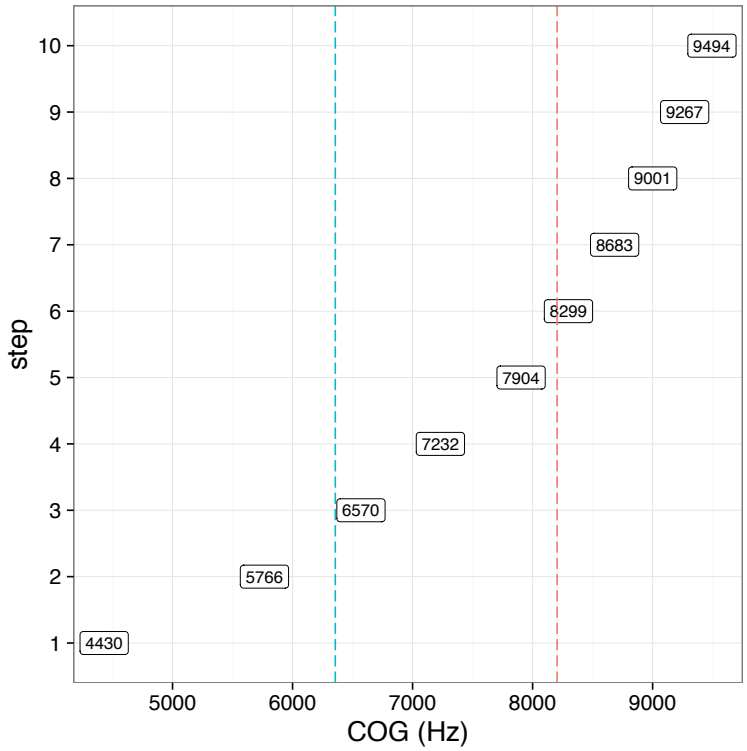
Cue-based normalization: The cue-based normalization account predicts that listeners accumulate talker-specific statistics for each fricative cue. After exposure to a certain number of instances of high COG [z], listeners should infer a high talker-specific mean for the COG cue and accordingly project a higher value for [s]; exposure to sufficiently many low COG [z] instances should result in a lower mean estimate and accordingly a lower value for [s]. The prediction is that the [s]-[ʃ] boundary for listeners in the high COG exposure condition should be higher than for listeners in the low COG exposure condition.

Phonetic covariation: For the particular fricatives examined in this experiment, the phonetic covariation and cue-based normalization accounts are practically equivalent. The cue-based normalization account assumes perfect covariation among speech sounds, and covariation is empirically quite high for [s] and [z]. A linear regression fit to the talker COG means for [s] and [z] in the laboratory data described in section 4.2 accounts for a high proportion of the variance ($R^2 = 0.88$) and had the following form:

$$\mu_s = 728.96 + 0.932\mu_z$$

Given the constructed mean [z] COGs of 8021 Hz and 6038 Hz (see section 4.3.1.2), this linear fit predicts an expected [s] COG mean of 8204 Hz for the high [z] COG talker and 6356 Hz for the low [z] COG talker (Figure 4.2; construction of the [s]-[ʃ] continuum is described in section 4.3.1.2).

Figure 4.2**.** The COG (Hz) of each member of the [s]-[ʃ] continuum, bandpass-filtered between 550 Hz and 10,000 Hz. The red line shows the predicted mean for [s] given the high COG exposure, and the blue line corresponds to the predicted mean for [s] given the low COG exposure condition. Predictions were determined by a linear regression fit to talker means in the American English laboratory data.



4.3.1 Methods

4.3.1.1 Participants

Twenty-eight participants (21 female) were recruited from the Johns Hopkins University undergraduate community. All were native speakers of American English. Twenty-two were monolingual and six were bilingual (Cantonese, German, Hebrew,

Korean, Mandarin, and Spanish). One participant reported having a speech impediment but no hearing impairment. All participants were compensated with partial course credit.

4.3.1.2 Stimuli

Exposure stimuli: [z]-initial syllables. The exposure stimuli were [z]-initial CVC syllables created by concatenating natural recordings. All recordings were selected from a corpus of CVC syllables from 22 native speakers of American English (15 female), recorded at New York University and sampled at 44.1 kHz (see section 4.2 for further detail). The corpus from which the stimuli were created contained CVC syllables composed of an initial fricative [θ ð f v s z ʃ ʒ] crossed with 10 vowels [i ɪ eɪ ɛ æ ʌ a ɔ o u]; the syllable always ended in [t]. From this corpus, one female speaker was selected for having a high COG [z], one female speaker was selected for having a low COG [z], and two female speakers were identified with relatively neutral COGs in their realization of [s ʃ z].

The syllable bodies (VC portion) of the exposure stimuli were selected from the two neutral COG speakers. As in the experiment proper, we will refer to these speakers as Meg and Kim. For each speaker and each of the 10 vowels in the corpus, we selected a [z]-initial CVC syllable with a medial [z] COG. The VC portion of the syllable was extracted at zero-crossings ([z] excluded) and normalized to 65 dB.

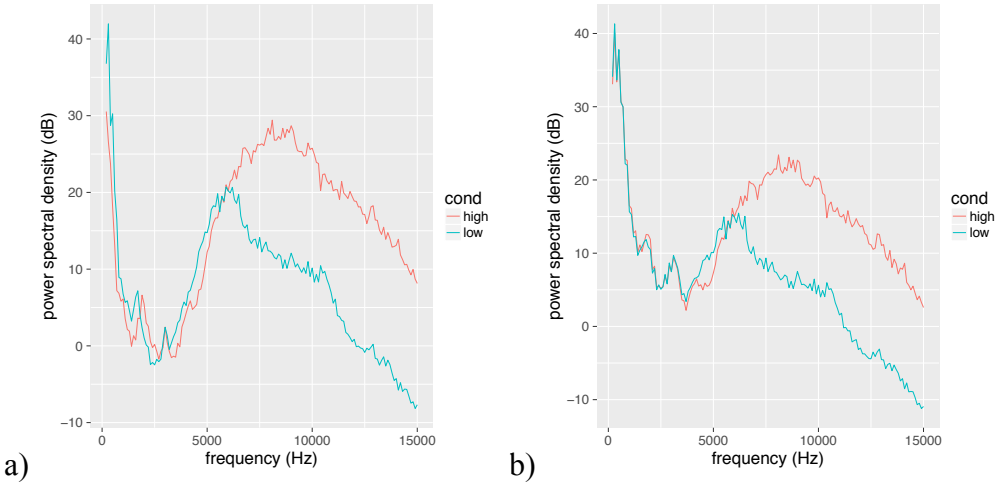
The critical manipulation was in the COG of the [z]. For each vowel, the highest COG [z] that could naturally be appended to the VC portion was selected from the speaker with overall high [z] COGs. The same process was carried out for the female speaker with relatively low COGs across sibilants, but instead selecting the lowest COG [z] with natural concatenation. The mean COG of the high [z] was 8021 Hz with a

standard deviation of 481 Hz, and the mean COG of the low [z] was 6038 Hz with a standard deviation of 731 Hz. The COG values for each selected [z] are presented in the Appendix.

The [z] durations were reduced to the common shared value of 85 ms. The original durations of the selected [z] segments ranged from 98.7 ms to 164.1 ms for the high COG speaker and 87.6 ms to 148.3 ms for the low COG speaker. The retained 85 ms portion of the [z] was designed to sound as natural as possible when appended to the VC body. Consideration was also taken to ensure an amplitude trajectory continuous with the beginning of the VC portion. Throughout this procedure, all cuts were made at zero-crossings in the waveform, and the amplitude of each [z] was normalized to 65 dB.

The high and low COG [z]s were concatenated with the vowel-matched VC portions from both Meg and Kim. The stimuli were then tapered at the beginning over a period of 50 ms (targeting the [z] portion), and 20 ms of silence was appended to both ends. Altogether, there were four sets of 10 [z]-initial stimuli: Meg-high COG, Meg-low COG, Kim-high COG, Kim-low COG. The LTAS of the high and low COG [z]s and full syllables (CV portion) are shown in Figure 4.3.

Figure 4.3. Long-term average spectra of a) the high and low COG exposure [z]s and b) the high and low COG exposure [z]s together with the following vowels.



Categorization stimuli: [s]-[ʃ] continua. An 11-point continuum was synthesized using Bark scale interpolation between endpoints corresponding to [s] and [ʃ] (Winn, 2014). The endpoints were generated from white noise with specifications for three spectral peak locations, their slopes, and their relative amplitudes. The three peaks of the [s] endpoint were located at 2500 Hz, 6000 Hz, and 10000 Hz with respective peak slopes of 25 dB/oct, 55 dB/oct, and 55 dB/oct. The relative amplitude of the 1st to 2nd peak was -25 dB and from the 3rd to the 2nd point, 20 dB. For the [ʃ] endpoint, the peaks were at 1700 Hz, 3500 Hz, and 7000 Hz with respective peak slopes of 35 dB/oct, 45 dB/oct, and 40 dB/oct. The relative amplitude of the 1st to the 2nd peak was 5 dB and from the 3rd peak to the 2nd peak was -4 dB. The peak values for each endpoint were estimated from natural productions. All durations were 150 ms, with a rise time of 110 ms and a fall time of 30 ms. The intensity of the [s]-[ʃ] segment was then scaled to 65 dB. The highest COG endpoint was excluded, resulting in 10 steps in the continuum. The spectral shapes of the first, middle, and final segments of the continuum are plotted in Figure 4.4.

Members of the [s]-[ʃ] continuum were then appended to both [it] and [ut] VC syllable bodies produced by Meg and Kim. The VC tokens were selected from natural recordings of 'seat', 'sheet', 'suit', or 'shoot'; these were chosen primarily on the basis of fluency and naturalness, and for having a relatively neutral fricative COG. For Meg, the VC portions came from recordings of 'seat' and 'shoot', and for Kim, the VC portions came from recordings of 'sheet' and 'suit'. The VC portion was extracted at zero crossings and scaled to 65 dB. The [s]-[ʃ] segments were appended to the onset, and 20 ms of silence was added to each end of the stimulus.

Figure 4.4**.** Long-term average spectra of the low endpoint (step 1), high endpoint (step 10), and middle point (step 5) of the [s]-[ʃ] continuum.

4.3.1.3 Procedure

The main manipulation of the experiment was [z] exposure condition (high vs. low [z] COG). Each participant was exposed to high [z] COG for one speaker (Meg or Kim) and low [z] COG for the other speaker. For example, a participant who received the high [z] COG condition for Meg first was exposed to the low [z] COG condition for Kim second. Speaker order and exposure order were fully counterbalanced across participants.

Each trial consisted of exposure followed by a single categorization. For the exposure, a single [z]-initial syllable was presented twice with a 1500 ms ISI. The speaker's name and a spelling of the intended (non)word were simultaneously presented on the screen (e.g., "Listen to Meg say the word ZATE…"). There was a 1000 ms delay between the exposure and categorization. Listeners were then asked to categorize the initial fricative of a single syllable from one of the [s]-[ʃ] continua in a two-alternative forced choice task. The response options were 'S' and 'SH'. The ITI from [s]-[ʃ] categorization to the beginning of the following trial was 1500 ms. Altogether, there were six blocks of twenty trials for each speaker/exposure condition. Within each block, the ten [z]-initial syllables were presented in random order twice, and each of the twenty total [s]-[ʃ] continuum members was presented once.

The first trial served as practice in which the experimenter guided the participant through the structure of the exposure and categorization phases. Listeners were informed that the exposure words would always begin with the sound 'z', and that some of the words would be familiar and others would be novel. Additionally, they were instructed to listen closely to and get to know the speaker's voice.

4.3.2 Results

Responses were analyzed with a logistic mixed-effects model, with a binary dependent variable (1 = [s] response, 0 = [ʃ] response). The model included fixed effects of condition, following vowel, continuum step, and speaker, and interactions between condition and vowel and between condition and speaker. There was also a random

intercept for participant.[30] The coding for the categorical predictors was as follows: condition (high = 0.5, low = -0.5), test stimulus vowel ([u] = 0.5, [i] = -0.5), and speaker (Meg = 0.5, Kim = -0.5). Continuum step was converted to a numeric predictor scaled to have mean zero and standard deviation 1.

There was a significant main effect of condition, indicating that participants were *less* likely to respond [s] after exposure to *high* COG [z] ($\beta = -1.30$, $p < 0.001$; Figure 4.5). Consistent with previous perceptual findings, the following vowel also had a significant effect on categorization, with an [s] response less likely in the context of [i] than [u] ($\beta = 2.01$, $p < 0.001$). As expected, the step number in the COG continuum was significant, with higher steps receiving more [s] responses ($\beta = 6.66$, $p < 0.001$). The effect of speaker and the interactions between condition and speaker, and between condition and vowel, were not significant (*speaker*: $\beta = -0.15$, $p = 0.20$; *cond × vowel*: $\beta = 0.28$, $p = 0.24$; *cond × speaker*: $\beta = -0.63$, $p = 0.57$). The non-significant interaction between condition and vowel is demonstrated in Figure 4.5, which shows that for both the [i] and [u] continua, the separation in response curves between the two COG conditions was comparable to the separation observed in the aggregate. In addition, the effect of condition had a significant influence on categorization after only one block of exposure ($\beta = -1.33$, $p < 0.001$).

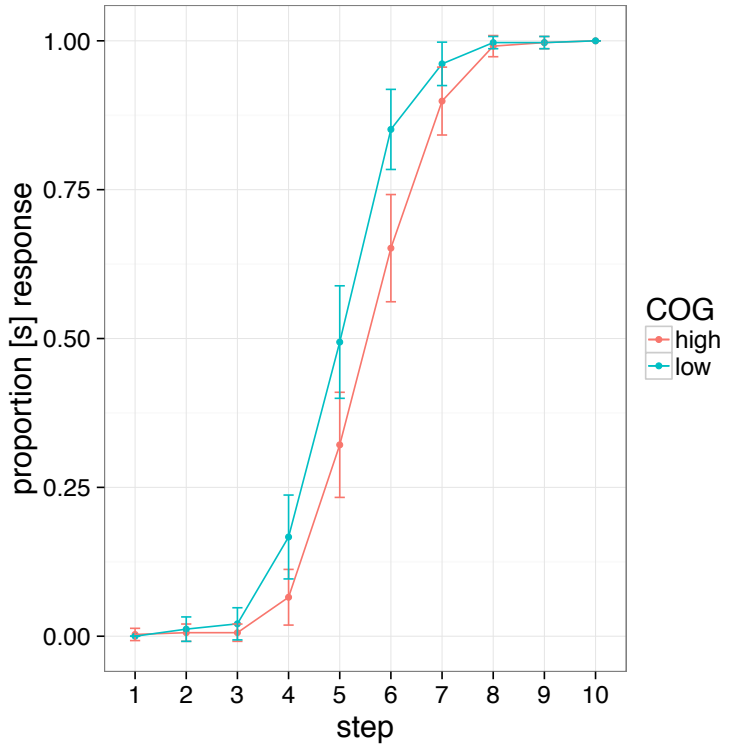
To further understand the acoustic basis of the adaptation effect, we also compared a set of models differing only in the composition of the condition effect. The effect of condition was replaced with various acoustic measurements of the [z]-initial

[30] A few alternative models were also considered, but failed to converge. These included a model with full interactions between all four effects, and any model with both an interaction in the fixed-effects and random slopes by condition or vowel for participant.

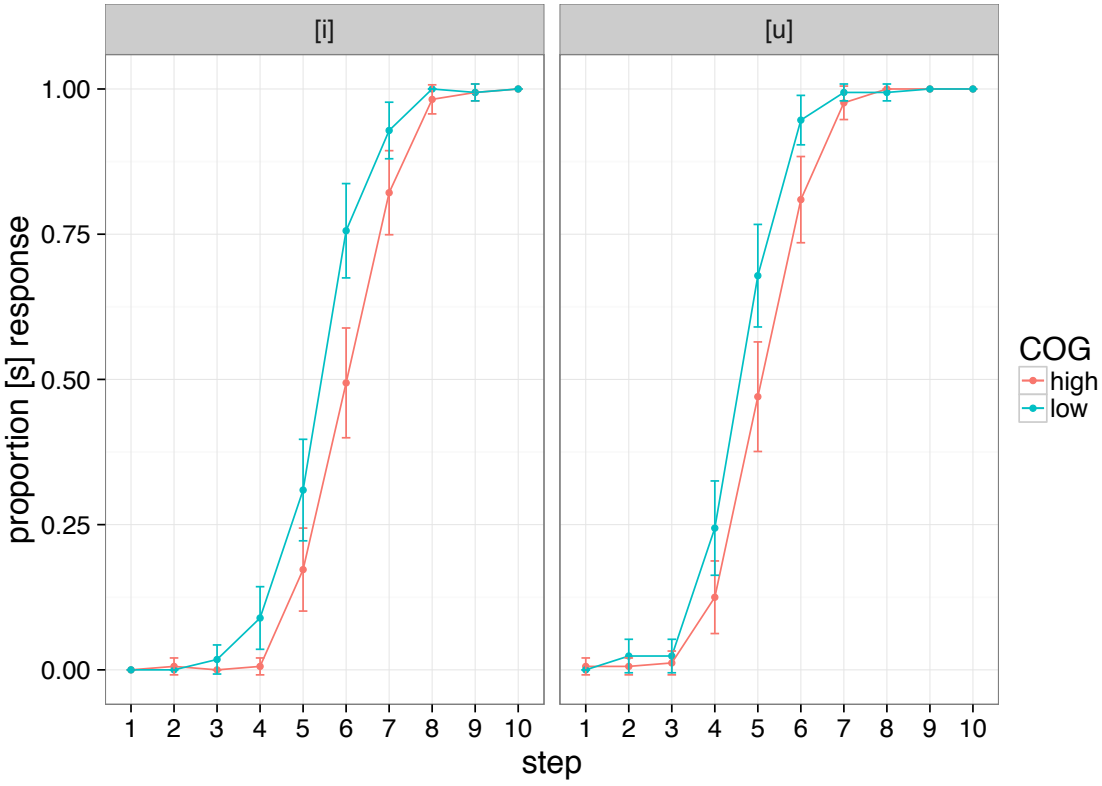
syllable in each trial. Four acoustic measures of the exposure stimuli were tested: the COG of the [z], the COG of the [z] after high-pass filtering at 550 Hz, the COG of the full CV portion of the syllable, and the COG of the full CV portion after high-pass filtering at 550 Hz. All measures were made after low-pass filtering at 10 kHz. This value was selected by approximation of the frequency range relevant for speech perception: Stelmachowicz et al. (2001) observed that optimal performance in the categorization of fricatives spoken by a female speaker required a bandwidth of 9 kHz (larger bandwidths were not tested), thus 10 kHz was chosen as an estimate of the upper frequency bound for sibilant perception. The acoustic measures for the effect of condition were centered on the grand mean, and coding of other factors was the same as in the original model.

The models were compared using the Bayesian Information Criterion (BIC), which assesses the goodness of model fit with a penalty for free parameters. The preferred model has the lowest BIC. The model with the COG of the [z] after high-pass filtering returned the lowest BIC of 2084, followed by the COG of the [z] without high-pass filtering (2109), the COG of the full CV portion (2128), and the COG of the full CV portion after high-pass filtering (2136). We speculate that the model with high-pass filtered [z] COG would perform best as it most directly reflects aspects of the articulation and acoustics that are similar for [s] and [z] (i.e., the front cavity resonance).

Figure 4.5. Proportion [s] response following exposure to the high and low COG [z] stimuli a) for the [i] and [u] [s]-[ʃ] continua combined and b) for each continuum.



a)



b)

4.3.3 Discussion

The results indicated that listeners generalized a spectral distribution from [z] to [s], and that generalization occurred early during exposure. Moreover, the acoustic-phonetic model of the data revealed that the high-pass filtered COG of the [z] accounted for the results better than other COG measurements of the stimuli (e.g., without high-pass filtering or spectra calculated over the entire syllable). Two primary inferences arise from this: first, listeners may be preferentially weighting information from the [z] in generalizing to [s]; second, and consistent with the first, [z] contains greater energy in the mid-frequency range (e.g., approximately 2000 to 7000 Hz) than the vocalic portion of the syllable, and this is the frequency range most relevant for [s]-[ʃ] perception. These findings are consistent with the spectral contrast (though perhaps not LTAS normalization that gives all frequencies equal weight), cue-based normalization, and the account of talker adaptation founded on phonetic covariation.

4.4 Experiment 2: Exposure to [v]

Experiment 2 examined the differing predictions of the cue-based normalization and phonetic covariation accounts. In particular, if listeners track talker-specific distributional information at the level of fricative cues, exposure to any fricative should affect perception of any other fricative through the shared cue mean (and other moments). However, the mean COG of [v] does not covary with the mean COG of [s] or [ʃ] across talkers, even though the speech sounds are all fricatives (and of course produced with the same vocal tracts). If listeners exploit perceptual knowledge of phonetic covariation, then listeners should not generalize talker-specific COG from [v] to [s] or do so less than was found for [z] exposure. A lack of generalization could also be consistent with a spectral

contrast effect, as the spectrum of [v] does not contain high energy in the frequency range relevant for [s]-[ʃ] categorization (i.e., the mid-frequency range; see Figure 4.6). The present experiment followed the same design as the preceding one, but replaced the high and low COG [z] exposure with exposure to high and low COG [v].

4.4.1 Methods

4.4.1.1 Participants

A separate set of 28 participants (13 female) from the Johns Hopkins undergraduate community completed Experiment 2. Twenty-seven of the participants were native speakers of American English, and one participant was a native speaker of Mandarin, but fully fluent in English. Of the 28 participants, 21 were monolingual and 7 were bilingual (Japanese, Korean, Mandarin, Marathi, and Spanish). All participants were compensated with partial course credit.

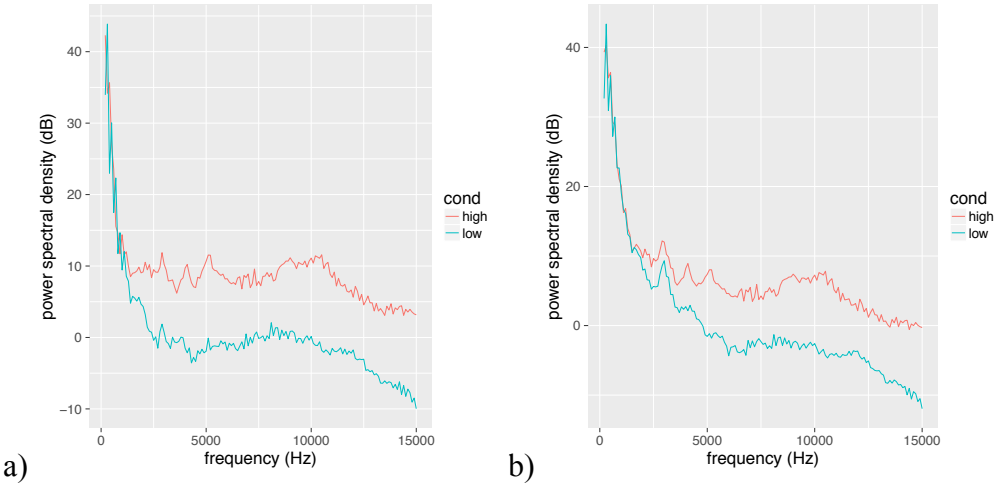
4.4.1.2 Stimuli

Exposure stimuli: [v]-initial syllables. The procedure for creating the [v]-initial stimuli closely followed that for the [z]-initial stimuli. Each stimulus was composed of a high or low COG [v] concatenated with a VC syllable body from two different speakers, Meg and Kim. All recordings were selected from the corpus of fricative-initial CVC syllables described in section 4.2. For the [v] portion of the stimulus, two female speakers, one with a relatively high COG [v] and one with a relatively low COG [v], were identified and their recordings of [v] extracted. For the VC portion of the stimulus, recordings from the same two female speakers described in section 4.2 with relatively neutral COG values were used, allowing the same [s]-[ʃ] continua to be used as the speech targets.

As before, the syllable bodies were extracted from [v]-initial words produced by Meg and Kim. There were 10 unique VC portions for each speaker, each with a different vowel ([i ɪ eɪ ɛ æ ʌ a ɔ oʊ u]). All splices were made at zero crossings, and the amplitude was normalized to 65 dB.

A high COG [v] and low COG [v] were chosen for each vowel type. We strived to select a [v] that preceded the same vowel type of the syllable being created in the original recording. In cases when this was not possible, we chose a neighboring vowel. Each [v] was truncated to 85 ms and ramped in intensity over the first 30 ms for a more natural sound. The amplitude was then normalized to 65 dB. The high and low COG [v]s were then concatenated with the vowel-matched VC portions from Meg and Kim, and 20 ms of silence was appended to both ends. The LTAS of the high and low COG [v]s and full syllables (CV portion) are shown in Figure 4.6.

Figure 4.6. Long term average spectra (LTAS) of a) the high and low COG [v]s and b) the high and low COG [v]s with the following vowel.



Categorization stimuli. The stimuli presented for categorization were members of the same [s]-[ʃ] continua used in Experiment 1 (section 4.3.1.2).

4.4.1.3 Procedure

Experiment 2 followed the same procedure as Experiment 1 except that [z]-initial syllables were replaced with [v]-initial syllables (section 4.3.1.3).

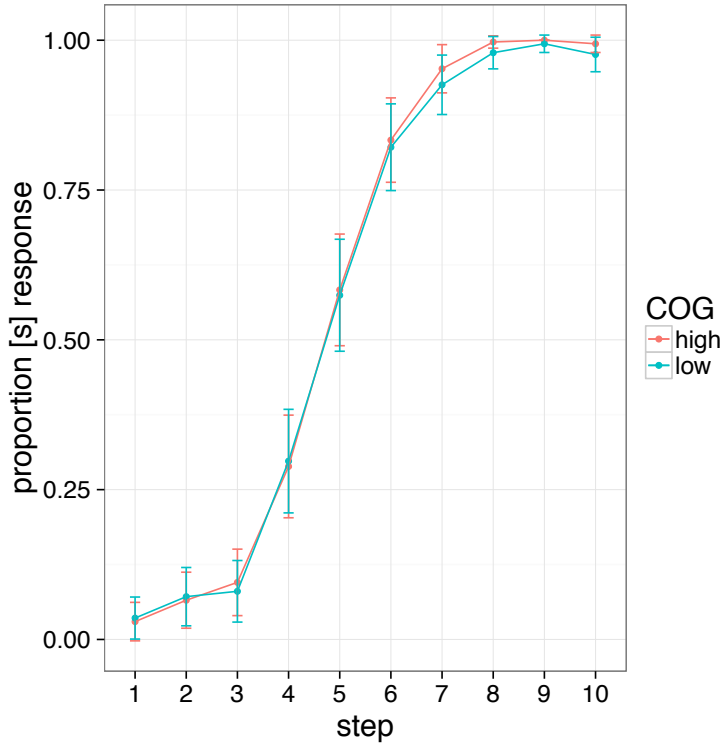
4.4.2 Results

4.4.2.1 [v] Exposure

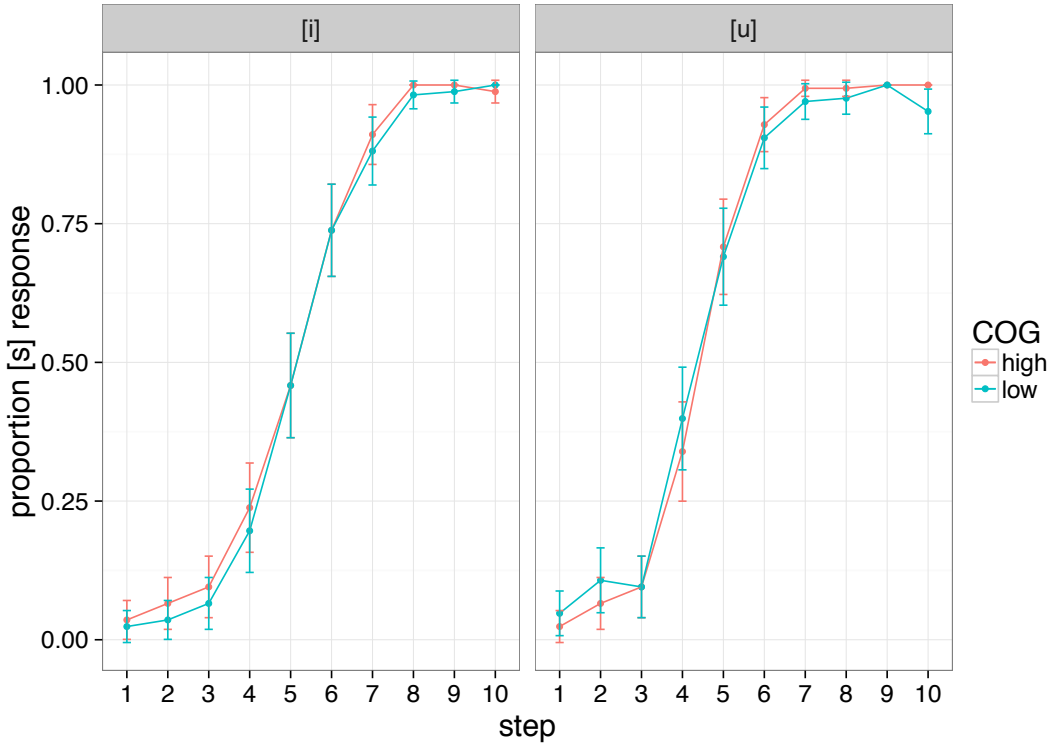
The categorization responses were submitted to a logistic mixed-effects regression model with the same structure as in the analysis of Experiment 1 (section 4.3.2). The model predicted the probability of [s] response from the condition, vowel, speaker, trial number, and step, as well as the interactions between condition and vowel and condition and speaker. The model included a random intercept for participant.

The main effect of condition failed to reach significance, indicating that the COG of the [v] exposure syllables did not significantly influence the categorization response ($\beta = 0.13$, $p = 0.19$; Figure 4.7a). As before, there were significant effects of vowel ($\beta = 1.05$, $p < 0.001$) and continuum step ($\beta = 4.29$, $p < 0.001$), but no significant effect of speaker ($\beta = 0.02$, $p = 0.85$). Finally, the condition × vowel and condition × speaker interactions were not significant (*cond × vowel*: $\beta = -0.23$, $p = 0.22$; *cond × speaker*: $\beta = 0.91$, $p = 0.42$). Note the lack of separation between the high and low COG response curves in the aggregate and for the [i] and [u] [s]-[ʃ] continua separately (Figure 4.7b).

Figure 4.7. Proportion [s] response following exposure to the high and low COG [v] stimuli a) for the [i] and [u] [s]-[ʃ] continua combined and b) for each continuum.



a)



b)

4.4.2.2 [z] and [v] exposure comparison

We wanted to demonstrate that the response pattern in Experiment 2 was significantly different from that of Experiment 1. The data from both experiments was submitted to a combined logistic mixed-effects model with fixed effects of experiment ([z] or [v] exposure), condition, vowel, step, and speaker, the interaction between experiment and condition, and a random intercept for participant. Because the interactions between condition and vowel and condition and speaker did not reach significance in either of the experiments, these were not excluded from the model. There was a significant effect of the experiment, indicating that participants were less likely to select [s] after exposure to [z] ($\beta = -1.03$, $p < 0.01$), and a significant effect of condition, in which participants were less likely to select [s] in the high COG condition ($\beta = -0.42$, $p < 0.001$). Critically, the interaction between experiment and condition was significant and revealed that the effect of condition was entirely driven by the [z] exposure experiment ($\beta = -1.15$, $p < 0.001$). There were also significant effects of vowel and continuum step (*vowel*: $\beta = 1.40$, $p < 0.001$; *step*: $\beta = 5.11$, $p < 0.001$), but no main effect of speaker ($\beta = -0.05$, $p = 0.51$).

4.4.3 Discussion

Listeners apparently did not generalize talker-specific spectral properties from [v] to [s] or the [s]-[ʃ] contrast: there was a significant difference in the effect of the COG manipulation between Experiment 1, with exposure to [z], and Experiment 2, with exposure to [v]. This pattern of results runs counter to the predictions of a cue-based normalization account, as listeners do not combine spectral evidence from all fricatives equally when adapting to a novel talker. There is no evidence that talker-specific COG

for [v] substantially influences the listener's expectation of the talker's COG for [s] (or [ʃ]). In contrast, these findings are consistent with the phonetic covariation account: the generalization observed in the first experiment could reflect listener knowledge of the strong covariation of [s] and [z] across talkers, and the absence of covariation between [s] and [v] licenses no generalization in the present experiment. In short, listeners may know that [s] and [v] are statistically independent across talkers in a way that [s] and [z] clearly are not. The results are also consistent with the spectral contrast account, which specifies that there be energy in the *relevant* frequency range for categorization. As [v] contains very little mid to high frequency energy, the perception of [s] and [ʃ] is largely unaffected by exposure to high vs. low variants of the labial fricative.

4.5 Experiment 3: Exposure to speech-shaped noise

The results of both previous experiments are consistent with listener knowledge of phonetic covariation among fricatives. Nevertheless, the within-trial alternation of exposure and test stimuli may have induced general spectral contrast effects that could plausibly give rise to the same 'generalization' from [z] to [s]. The spectral contrast account makes the additional prediction that appropriately constructed non-linguistic exposure items should yield the same pattern. We tested the predictions of this account using the same design and procedure as Experiment 1, replacing the [z]-initial exposure with white noise that was matched in duration, amplitude, and the LTAS to the [z]-initial syllables.

4.5.1 Methods

4.5.1.1 Participants

An additional 28 participants (16 female) from the Johns Hopkins undergraduate community completed Experiment 3. Twenty-seven participants were native speakers of American English, and one was a native speaker of Greek but spoke American English fluently. Twenty-six were monolingual, one was bilingual (Spanish), and two were trilingual (Italian and Spanish; German and Greek). Participants received partial course credit for completion of the experiment.

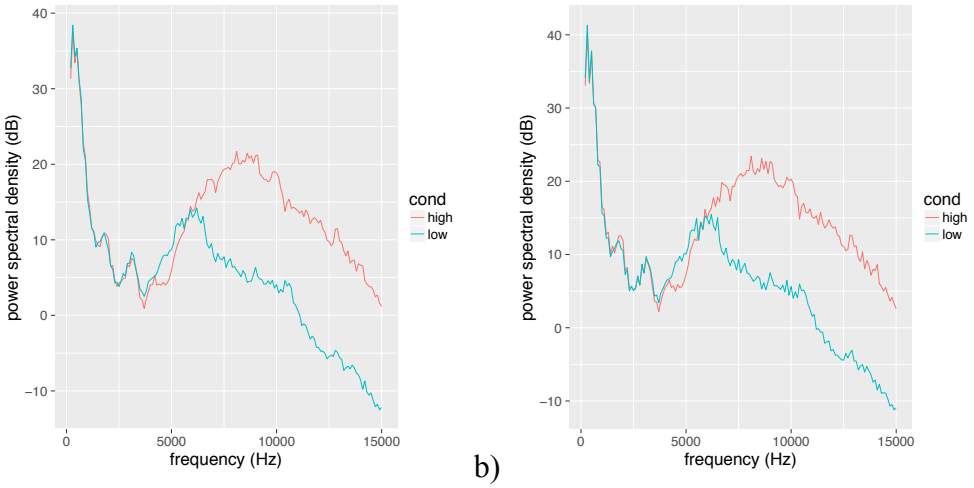
4.5.1.2 Stimuli

Exposure stimuli: noise. For each [z]-initial stimulus, a white noise version was created that matched the CV portion of the stimulus in duration, amplitude, and LTAS (Figure 4.8). The final [t] of the original syllable was primarily comprised of silence for the closure period and had minimal amplitude for the final release of the [t]. Because of this, the CV portion was the most perceptually salient portion of the syllable. The noise signal was tapered at each end over a period of 50 ms and then matched in amplitude to the CV portion of the corresponding speech syllable. Silence was appended to the resulting noise that was matched in duration to the final [t] of the original syllable. Finally, 20 ms of silence was appended to each end of the signal.

The LTAS-matched white noise stimuli were generated in Praat by shaping a white noise stimulus by the LTAS of the corresponding speech stimulus (Winn, 2014). Between the offset of the noise exposure stimulus and the onset of the [s]-[ʃ] test syllable, there was 1040 ms of silence (this included the 20 ms silence at the end of the noise stimulus and 20 ms silence at the beginning of the test).

Categorization stimuli. The categorization stimuli were the [s]-[ʃ] continua used in the previous two experiments (see section 4.3.1.2).

Figure 4.8. Long-term average spectra (LTAS) of a) the high and low COG white noise matched in LTAS to the [z]-initial stimuli and b) the original high and low COG [z]-initial stimuli (CV portion).



4.5.1.3 Procedure

The procedure in Experiment 3 followed the same structure as in Experiment 1 except that [z]-initial exposure stimuli were replaced with high or low COG noise stimuli. Listeners received either the high or low COG noise for the first speaker and the opposite COG noise stimuli for the second speaker. Speaker order and exposure order were fully counterbalanced. (Note that speakers are relevant here only for the categorization test items; no reference was made to speakers regarding the noise exposure stimuli.)

The exposure phase of each trial began with a single noise stimulus presented twice (1500 ms ISI). As in Experiment 1, listeners then categorized the initial fricative of an [s]-[ʃ] test stimulus in a two-alternative forced choice task. Between the exposure and categorization phase, there was a 1000 ms ISI, and the ITI was 1500 ms. In contrast to Experiment 1, there were four blocks of 20 trials as opposed to six blocks. Given the early presence of the effect in Experiment 1 and the fact that participants would be

listening to static noise that they might find annoying, we judged that a shorter experiment would be effective and preferable. Within each block, the ten unique noise stimuli were presented in random order twice, and the twenty [s]-[ʃ] test stimuli were each presented once. The noise and test stimulus pairing was randomized after each round of the four conditions (two speakers, two orders of COG level).

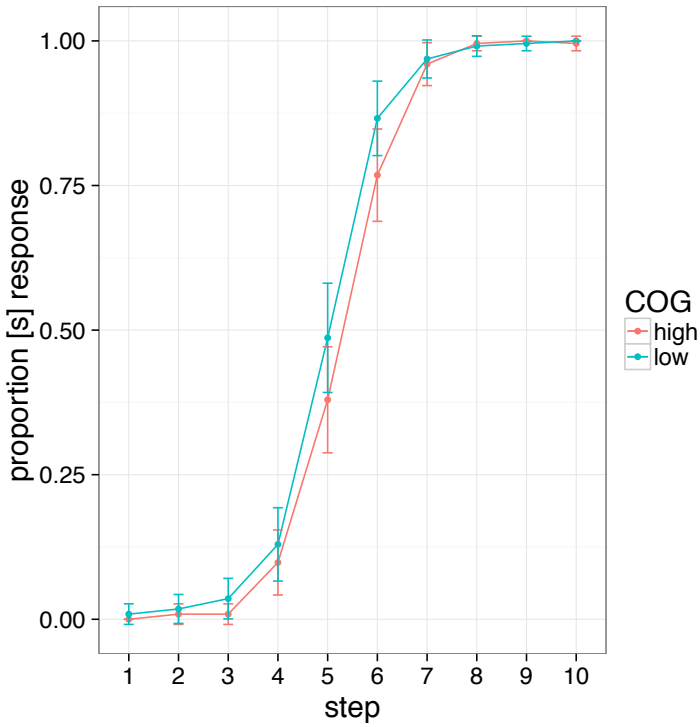
The first trial again served as practice in which the experimenter guided the participant through the structure of the exposure and categorization phases. Listeners were told they would be listening to a new speaker (either Meg or Kim), but they would first hear two identical non-speech sounds. Listeners were instructed to listen closely to both the sounds and the speaker's voice.

4.5.2 Results

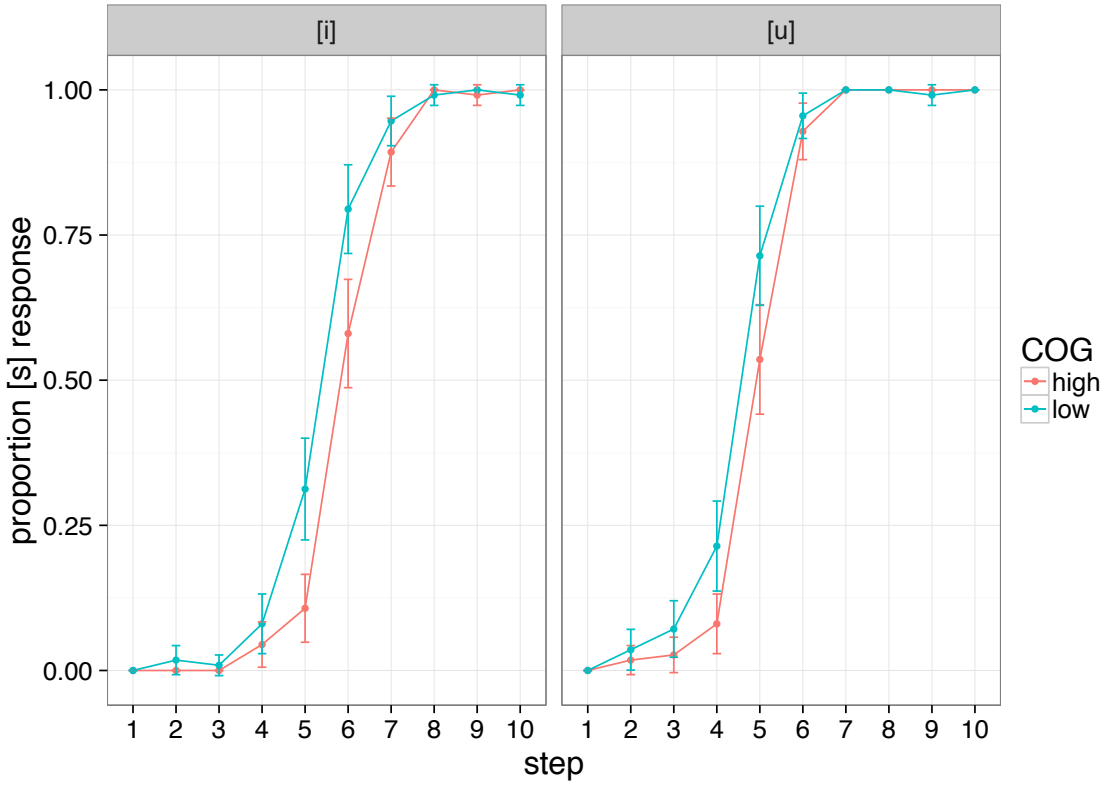
4.5.2.1 Noise exposure

The results were analyzed with a logistic mixed-effects model having the same structure as in Experiments 1 and 2. Probability of [s] response was predicted from fixed effects of the condition, vowel, continuum step, and speaker, and the interactions between condition and vowel and condition and speaker. Paralleling the pattern of results from the corresponding model in Experiment 1 ([z]-exposure), there were significant effects of condition ($\beta = -1.02$, $p < 0.001$), vowel ($\beta = 2.00$, $p < 0.001$), step ($\beta = 6.35$, $p < 0.001$), and speaker ($\beta = 0.58$, $p < 0.001$), but no significant interaction between condition and vowel ($\beta = 0.30$, $p = 0.29$) or condition and speaker ($\beta = 0.06$, $p = 0.94$). The effect of condition was also significant within the first block of exposure ($\beta = -1.06$, $p < 0.001$). The proportion [s] response following exposure to the high and low COG white noise stimuli is shown in Figure 4.9.

Figure 4.9. Proportion [s] response following exposure to the LTAS-matched white noise a) for the [i] and [u] [s]-[ʃ] continua combined and b) for each continuum.



a)



b)

4.5.2.2 [z] and noise exposure comparison

The pattern of results in the [z] and noise exposure experiments were qualitatively parallel. However, there may have been a stronger effect of condition from exposure to speech than from white noise exposure. To test this, the data from both the [z] and white noise exposure experiments were analyzed together with a logistic mixed-effects model having fixed effects of experiment ([z] vs. white noise), condition, vowel, speaker, continuum step, the interaction between experiment and condition, and a random intercept for participant. Consistent with the pattern of results from the individual experiments, there was a significant effect of the continuum step ($\beta = 6.49$, $p < 0.001$), condition and vowel were also significant (*cond*: $\beta = -1.15$, $p < 0.001$; *vowel*: $\beta = 2.00$, $p < 0.001$), and there was no main effect of speaker ($\beta = 0.15$, $p = 0.10$). Contrary to the suggestion above, the effect of experiment was not significant: that is, the categorization pattern with exposure to white noise did not differ significantly from that with exposure to the [z]-initial stimuli ($\beta = -0.25$, $p = 0.47$). Furthermore, the interaction between condition and experiment was not significant, indicating that the effect of condition was not enhanced by either speech or white noise exposure ($\beta = -0.23$, $p = 0.22$).

4.5.3 Discussion

The LTAS-matched white noise stimuli had a significant and statistically indistinguishable effect on [s]-[ʃ] categorization as the [z]-initial stimuli. This pattern thus provides strong evidence for the spectral contrast account, as both linguistic and non-linguistic exposure stimuli with energy in the relevant frequency range for categorization had comparable effects on the perception of coronal fricatives. Spectral contrast is also consistent with the results of the [v]-exposure experiment, given that the spectral

distributions of coronal and labial fricatives overlap minimally and hence should not interact contrastively. Because the categorization shift observed in the present experiment is uniquely predicted by spectral contrast (within the set of alternative accounts that we consider), at least on parsimony grounds alone, the parity between [z] and noise-matched adaptors casts doubt on the phonetic covariation account of perceptual 'generalization' in the first experiment. The following experiments addressed first, whether listeners make preferential use of linguistic information when alternating with white noise, and second, whether listeners have knowledge of phonetic covariation that plays a role when the potential for spectral contrast effects on perception are minimized (e.g., in non-local environments).

4.6 Experiment 4: Exposure to alternating speech-shaped noise and [z]

Significant effects of linguistic and non-linguistic exposure on fricative categorization were found in Experiments 1 and 3. The goal of this experiment was to determine the relative weighting of these two types of exposure. The experiment had a structure similar to that of the preceding experiments, in which exposure alternated with [s]-[ʃ] categorization within each trial. However, in this case each exposure consisted of [z]-initial syllables and white noise with opposing COG levels.

Specifically, prior to each instance of categorization, half of the participants heard two repetitions of a high COG [z] syllable followed by the low COG noise, whereas the other half heard two repetitions of the low COG [z] syllable followed by the high COG noise. The LTAS of each stimulus type thus contrasted across the two groups, but the LTAS of the entire preceding exposure signal was equal. If participants make greater use of linguistic than non-linguistic stimuli in speech adaptation, then the [s]-[ʃ] boundaries

should differ according to the COG of the speech exposure items. If participants make greater use of the non-linguistic stimuli (or of the stimulus immediately preceding categorization), categorization boundaries should differ according to the noise COG. Finally, if listeners do not preferentially treat linguistic or non-linguistic stimuli, but rather track the statistics of the full exposure sequence, then speech and noise should effectively cancel one another out, leading to no difference in the boundary location across the two conditions.

4.6.1 Methods

4.6.1.1 Participants

Twenty-eight additional participants (12 female) from the Johns Hopkins undergraduate community completed Experiment 4. Twenty-seven of the participants were native speakers of American English, and one participant was a native speaker of Cantonese, but grew up also speaking English. Fifteen participants were monolingual, seven were bilingual (Arabic, Korean, Mandarin, Portuguese, and Spanish), and two were trilingual (Hindi and Tamil; Cantonese and Mandarin). All participants received partial course credit for participation.

4.6.1.2 Stimuli

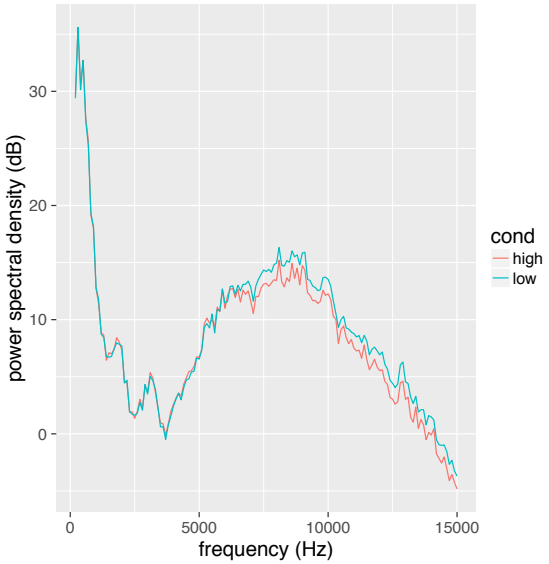
Exposure stimuli. The [z]-initial syllables and LTAS-matched noise stimuli described in sections 4.3.1.2 and 4.5.1.2 were concatenated to form a sequence of speech and noise that alternated in the direction of the COG manipulation. For each speaker, the high COG [z] stimuli were paired with the low COG noise, and the low COG [z] stimuli were paired with the high COG noise. The sequence of speech and noise was repeated

twice for a total of four presentations (i.e., the entire sequence at the beginning of a trial was speech – noise – speech – noise).

There was 20 ms of silence at the beginning of each exposure sequence, 500 ms of silence between each member of the sequence, and 1040 ms of silence in between the final noise stimulus and the onset of the [s]-[ʃ] test stimulus (this included the 20 ms silence at the end of the noise stimulus and 20 ms silence at the beginning of the test). The noise stimuli were revised slightly from earlier such that the amplitude of the noise matched the amplitude of the CV portion of the corresponding speech stimuli after tapering at both two ends. This resulted in a slight increase in the intensity of the noise. Otherwise, all aspects of the original speech and noise stimuli were the same as in Experiments 1 and 3. As shown in Figure 4.10, the long-term average spectra of the full exposure sequences in the high and low COG speech conditions were nearly indistinguishable.

Categorization stimuli. The [s]-[ʃ] continua were the same as those in the previous experiments (section 4.3.1.2).

Figure 4.10. Long-term average spectra of the alternating high and low COG speech and contrasting white noise stimuli.



4.6.1.3 Procedure

As there were two sets of speech and noise stimuli differing only in the COG level for each speaker, the high and low COG manipulations for the speech and noise stimuli could be fully crossed. Listeners received the high or low [z] COG stimulus interleaved with the opposite COG level noise stimulus for the first speaker; for the second speaker, the COG levels for the speech and noise were switched. Speaker order and exposure order were counterbalanced across participants.

Within each trial, exposure was comprised of two repetitions of the [z]-initial syllable followed by the noise stimulus. As described in the stimulus preparation section, there was 500 ms of silence between each presentation. The pairing of the [z]-initial syllable and the noise stimulus was constant: the noise stimulus corresponded to the [z]-initial syllable with the same VC portion, but had the opposite [z] COG level. The speaker's name and intended (non)word were presented on the screen during the audio

presentation (e.g., “Listen to Meg say the word ZATE, followed by a brief sound…”). Immediately following exposure, participants categorized the initial fricative of a randomly-selected [s]-[ʃ] test stimulus. Between the exposure and categorization phase there was a 1000 ms delay. There was a total of four blocks of twenty trials each. Each [z]-initial syllable was presented twice within each block and the 20 [s]-[ʃ] test stimuli were presented once. The exposure and test stimulus pairing was randomized for each round of the four conditions (two speakers, two orders of COG level).

To adjust participants to the speech and noise alternation, there were two practice exposure trials with speech and LTAS-matched noise generated from an unrelated voice. The words ‘BIRD’ and ‘PINK’ were selected from the American Spoken Lexicon Corpus (Seidl-Friedman et al., 1999) for their relatively high lexical frequency, similarity in structure to the test syllables, and because they did not contain any fricative consonants or word-initial coronals. The corresponding noise stimulus was matched in LTAS, duration, and amplitude to each syllable excluding the final stop consonant. The structure of the practice exposure sequence was otherwise identical to exposure in the critical trials.

In addition to these two practice items, listeners were again guided through the initial trial by the experimenter. Listeners were told that the speaker’s words would always begin with the sound ‘z’, and that some of the words would be familiar and others would be novel. Additionally, they were instructed to try to listen closely to the speaker’s voice and the sounds.

4.6.2 Results

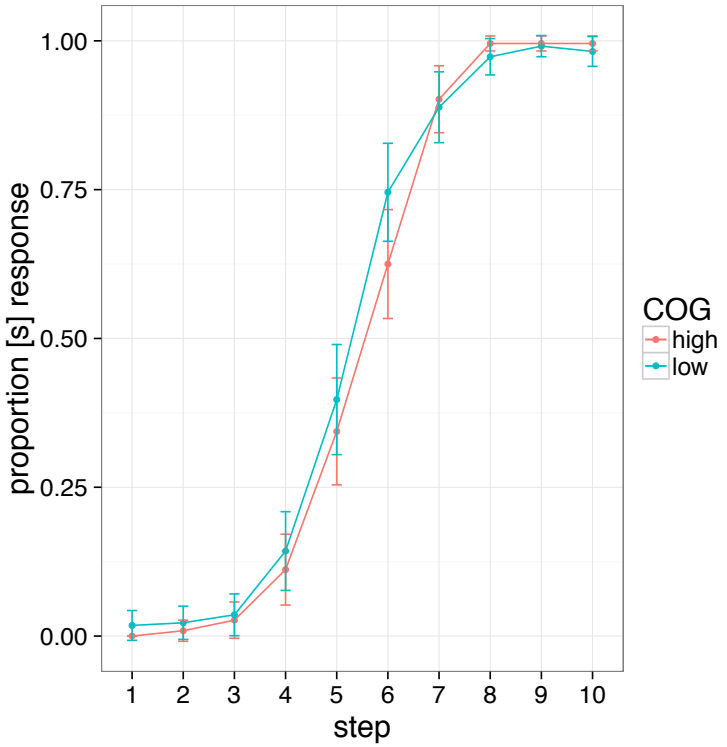
The results were analyzed in the same manner as for the previous experiments, with a logistic mixed-effects model predicting the probability of [s] response from the condition, vowel, continuum step, speaker, and interactions between condition and vowel and condition and speaker. The coding of condition was 0.5 for the high COG [z]s and low COG noise, and -0.5 for the low COG [z] and high COG noise.

The model revealed a significant effect of condition, indicating that the influence of the [z] was significantly stronger than the influence of noise on the response pattern ($\beta = -0.31$, $p < 0.001$; Figure 4.11). However, closer inspection of the data showed that the effect was largely driven by a single participant.[31] Excluding that participant's data, the effect of condition failed to reach significance, but trended in the direction consistent with preferential weighting of speech ($\beta = -0.20$, $p = 0.09$). This trend was obtained even though the noise always immediately preceded the categorization stimulus. The model also revealed significant effects of the vowel ($\beta = 1.59$, $p < 0.001$) and continuum step ($\beta = 4.54$, $p < 0.001$).[32] The main effect of speaker and the interactions between condition and vowel and condition and speaker did not reach significance (*speaker*: $\beta = 0.11$, $p = 0.34$; *cond x vowel*: $\beta = -0.16$, $p = 0.49$; *cond x speaker*: $\beta = -0.42$, $p = 0.64$).

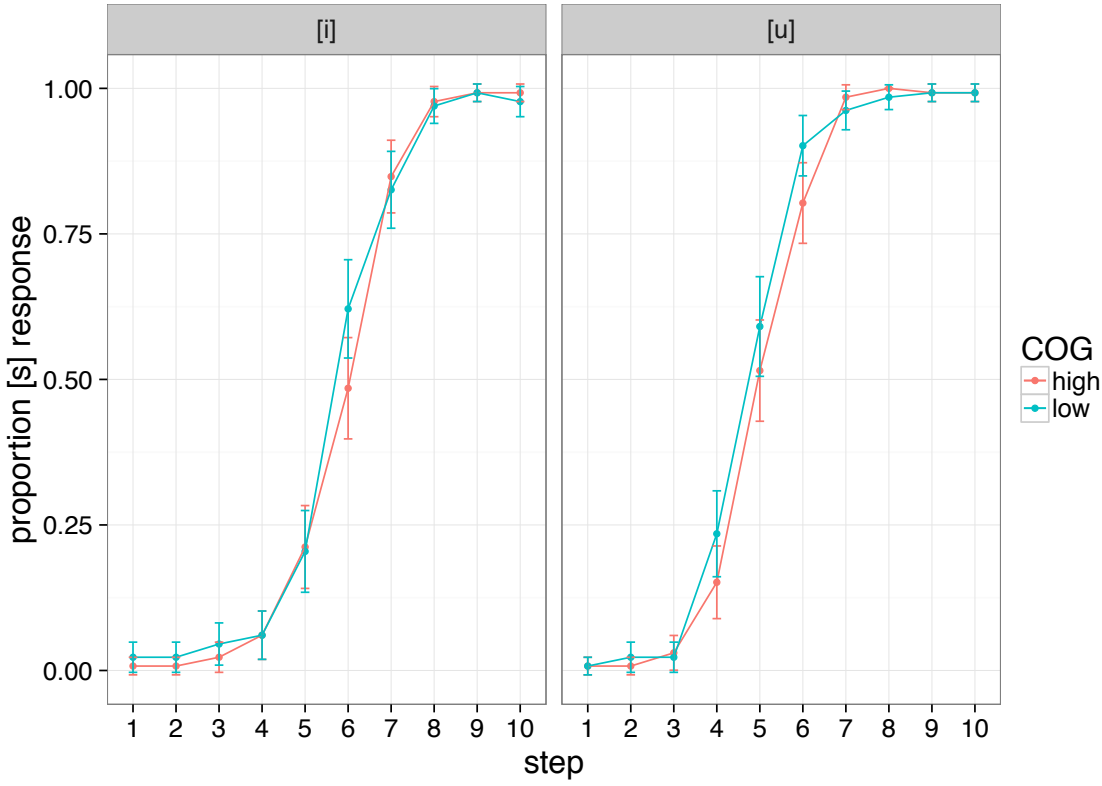
[31] Due to convergence issues, it was often difficult to include additional random effect structure for participant beyond the intercept. For this dataset, an additional model was fit without the interaction between condition and speaker, but a random intercept and slope of condition for participant. In that model, the effect of condition only trended towards significance ($\beta = -0.32$, $p = 0.09$). It should, however, be noted that the direction was consistent with preferential treatment for the speech, and this trend was obtained even though the noise directly preceded categorization.

[32] The values reported here come from the model derived from the full dataset.

Figure 4.11. Proportion [s] responses following exposure to alternating speech and white noise a) for the [i] and [u] [s]-[ʃ] continua continua and b) for each continuum.



a)



b)

4.6.3 Discussion

The null effect of the speech COG condition (with the exception of one anomalous participant) suggests that the alternating linguistic and non-linguistic exposure stimuli had equal and opposite effects on categorization. The auditory influence of and linguistic information in the speech stimuli did not override the auditory influence of the non-speech stimuli. This could provide further evidence for the spectral contrast account, according to which any precursor with relevant frequency components is expected to have the same perceptual influence.

The results of this experiment do bear on the window of temporal integration that supports spectral contrast. The representation that induces contrast cannot have been computed from the immediately preceding auditory stimulus, as this would have predicted an effect of high vs. low COG noise like that found in Experiment 3. Instead, it seems plausible that contrast was computed over all four exposure stimuli (or minimally the last two). This accumulative effect is similar to the finding of Holt (2005) that the average COG of an entire series of preceding tones best accounted for spectral contrast effects on [d]-[g] categorization (cf. the COG of the tone immediately preceding the test stimulus). Averaging apparently occurs even with 500 ms intervening between each successive exposure stimulus.

4.7 Experiment 5: Delayed categorization

The preceding experiments indicate that low-level auditory effects have an immediate and strong influence on speech perception. To determine if and how listeners learn talker-specific characteristics of speech, we conducted two additional experiments in which the exposure phase was separated from the categorization phase by

approximately 15 minutes. Participants heard the high or low COG [z]-initial stimuli in the first phase of the experiment, then performed a visual one-back repetition task for 14 minutes, and lastly completed the [s]-[ʃ] categorization. The separation between exposure and test probes the limits of a pure spectral contrast account, as it is highly unlikely that any generalization from [z] to [s] (or indeed any perceptual effect) across a 14-minute delay could be attributed to low-level auditory adaptation. In addition, a subset of the participants was also exposed to ocean noise during the intervening period, which should have minimized any auditory influence from the exposure stimuli. While spectral contrast certainly seems to have clear immediate effects on speech perception, generalization from talker-specific characteristics of [z] to [s] over an extended period would more plausibly be accounted for by a learned representation of the talker-specific [z], which could then be used to generalize to [s] in the way supported by phonetic covariation.

4.7.1 Methods

4.7.1.1 Participants

A total of 46 participants completed the delayed categorization experiment with either silence between exposure and test or intervening ocean noise between exposure and test. For the silent delay, 30 participants (22 female) were recruited from the Johns Hopkins University undergraduate community. All participants spoke English fluently, and 28 were native English speakers. One participant moved to the United States at age 11, but had attended an English-speaking international school, and another moved to the United States at age 4, and English was his dominant language. Thirteen participants were bilingual or a heritage speaker of another language (Cantonese, Korean, Mandarin, Spanish, Urdu, Yoruba) and one participant was trilingual (Mandarin and Spanish).

For the ocean noise delay, 16 participants (10 female) were recruited from the JHU undergraduate community. All were native speakers of English. Six participants were bilingual (Arabic, Mandarin, Spanish) and one participant fluently spoke three languages in addition to English (Hindi, Mandarin, and Urdu).

4.7.1.2 Stimuli

Exposure stimuli. This experiment employed the [z]-initial syllables described in section 4.3.1.2. To increase the amount of variability in exposure, the duration of each stimulus was manipulated by a factor of 0.85 and 1.25 using the overlap-add algorithm in Praat. This resulted in a total of 30 unique stimuli per speaker. The fricative of one syllable was perceived more as [s] after lengthening, and was thus shortened by approximately 15 ms. All stimuli were then scaled in intensity to 65 dB.

Categorization stimuli. The same stimuli from the [s]-[ʃ] continua described in section 4.3.1.2 were used in this experiment.

Images. A total of 200 images were employed in the experiment and depicted objects against a white background in a 512 x 512-pixel format. The visual stimuli were originally created for an object memory experiment in Ferrara et al. (2015) and based on an object memory experiment described in Brady et al. (2008) and Konkle et al. (2010).

4.7.1.3 Procedure

In contrast to the previous experiments, the present experiment grouped all exposure stimuli into one phase and all categorization stimuli into a separate phase. Between exposure and test, there was a 14-minute intervening period, during which participants performed a one-back repeat detection task with visual stimuli in silence

(silent delay) or while listening to ocean noise that gradually faded away (ocean noise delay).

During the exposure phase, there was a continuous sequence of randomly presented [z]-initial syllables. As before, participants were in either the high- or low-COG exposure condition and heard only one speaker's voice: Meg or Kim. Participants were instructed to get to know the speaker's voice and to press the space bar if the same exact recording was played twice in a row. There was a total of six blocks of 30 unique stimuli that repeated sequentially with a probability of 0.1. This corresponded to a repetition about once every 10 trials. No repetition was permitted within the first two trials or within two trials of a repeated recording. The inter-trial interval was 1.5 s. The exposure phase lasted approximately 7 minutes.

In between exposure and test, there was a visual one-back repeat detection task. This immediately followed the exposure phase with no interruption from the experimenter. In the instructions, participants were reminded to remember the speaker's voice. The 200 images described above were presented consecutively in random order with a display rate of 3 s, an inter-trial interval of 0.800 s, and a probability of sequential repetition equal to 0.1. If an image was presented twice in a row, participants were instructed to press the space bar. As in the preceding exposure phase, no repetition occurred within the first two trials or within two trials of a repeated image. This procedure was loosely based on an experiment described in Konkle et al (2010). For the participants with the ocean noise delay, ocean noise was played in the background during this phase and decreased linearly in volume from trial 2 to trial 180. The final 20 trials

were silent. For the participants with the silent delay, no audio accompanied the task. The distractor phase lasted approximately 14 minutes.

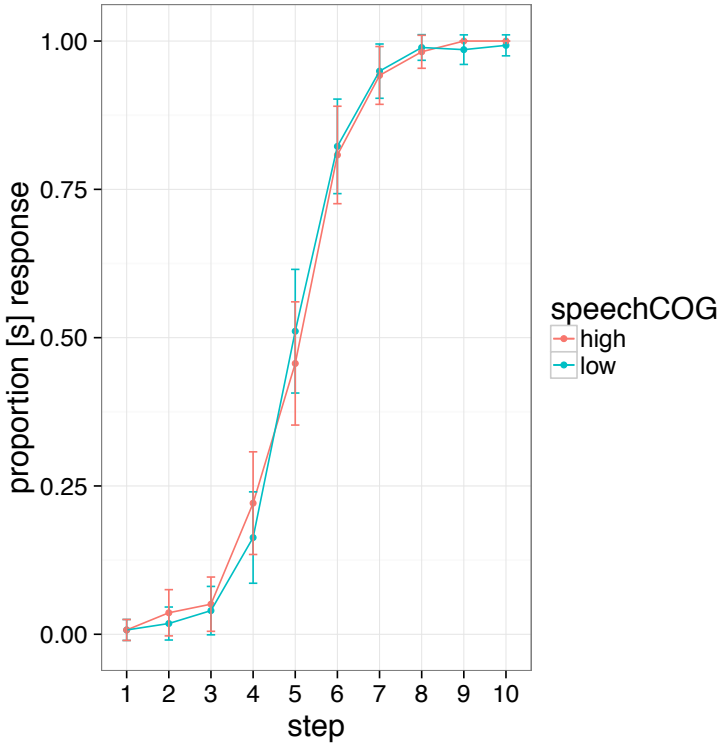
After the visual repeat detection task, the categorization phase began. In each trial, a single test stimulus was presented and participants categorized whether the speaker (Meg or Kim) said a word beginning with [s] or [ʃ] by clicking on the appropriately marked box ('S' or 'SH'). The inter-trial interval, timed from the participant's response to the onset of the following syllable, was 1.5 s. The categorization phase lasted approximately 4 minutes.

4.7.2 Results

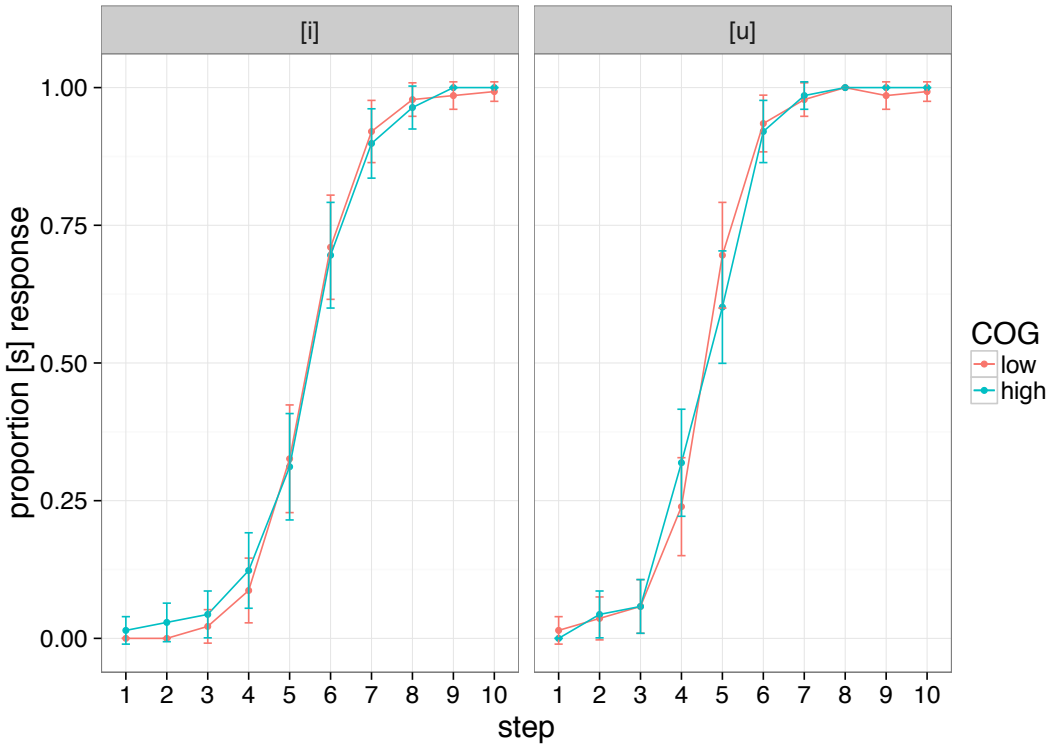
As in the preceding experiments, the responses were analyzed with a logistic mixed-effects model with fixed effects of the condition, experiment delay type (silence or ocean), vowel, and step, as well as interactions between condition and experiment and condition and vowel, and a random intercept for participant. Contrary to expectation, there was no main effect of condition or experiment delay type, suggesting that listeners did not generalize the COG of the [z] to categorization with the intervening delay and that there were no major differences in the response pattern between intervening silence and ocean noise (*cond*: $\beta = 0.03$, $p = 0.94$; *delay*: $\beta = 0.58$, $p = 0.13$). There was also no significant interaction between condition and experiment delay, indicating that the effect of condition was not modulated by the intervening audio, or lack thereof ($\beta = 0.10$, $p = 0.90$). There were, however, significant effects of vowel and step (*vowel*: $\beta = 1.55$, $p < 0.001$; *step*: $\beta = 5.06$, $p < 0.001$), but no significant interaction between condition and vowel (*cond x vowel*: $\beta = -0.11$, $p = 0.61$). As shown in Figure 4.12, the response curves

for the two conditions were nearly identical, indicating no effect of condition in this experiment.

Figure 4.12. Proportion of [s] response following exposure to the high and low COG [z] stimuli and a 14-minute intervening period with either silence or ocean noise a) for the [i] and [u] [s]-[ʃ] continua combined and b) for each continuum.



a)



b)

4.7.3 Discussion

No significant difference was observed in the categorization of the [s]-[ʃ] stimuli between the high and low COG [z] exposure conditions after 14 minutes of intervening silence or ocean noise. Interestingly, however, there was no main effect of ocean noise on categorization, which may have been expected under a spectral contrast account. These results could indicate that any apparent generalization may be reducible to spectral contrast effects which were too distant to have any significant effect on the categorization. Alternatively, listeners may still learn about the talker-specific spectral characteristics of fricatives and even generalize these properties across different fricative categories, but there may be limiting factors to the experiment.

First of all, the boundary curves observed in categorization are relatively steep, indicating that there may not be a sufficient number of ambiguous stimuli to observe a difference between the high and low COG conditions. Second, the [s]-[ʃ] continuum was designed in a manner consistent with the predictions for the expected [s] COG given the exposure [z] COGs. As will be discussed in the next section, listeners may be more attuned to other characterizations of the fricative spectrum than the energy-weighted mean frequency, or COG.

4.8 Experiment 6: Delayed categorization (high-ambiguity continuum)

The preceding studies pointed towards strong general auditory effects on perceptual adaptation; however, these studies also made the assumption that perceptual knowledge of phonetic covariation among fricatives could be approximated via the COG of fricative spectra. While COG is commonly used in phonetic research and reduces the dimensionality of a spectrum for convenient comparisons, there are several other

measures of the spectrum that may be more relevant for fricative perception than the overall average. In particular, Koenig et al. (2013) demonstrated that the mid-frequency peak ($Freq_M$) of the spectrum strongly reflects the front cavity resonance critical for determining the place contrast. Furthermore, findings presented in Chapter 3 indicate that talker-specific $Freq_M$ means are almost perfectly correlated between [z] and [s] in the laboratory speech data. In accordance with the predictions of target uniformity, it may be that the feature most relevant for talker-specific generalization from [z] to [s] is a uniform constriction location, and in this respect $Freq_M$ may be a more appropriate acoustic measure than spectral COG for approximating listeners' representation of phonetic covariation.

A linear regression was fit to the talker-specific $Freq_M$ data in the laboratory speech ($R^2 = 0.76$) and had the following form:

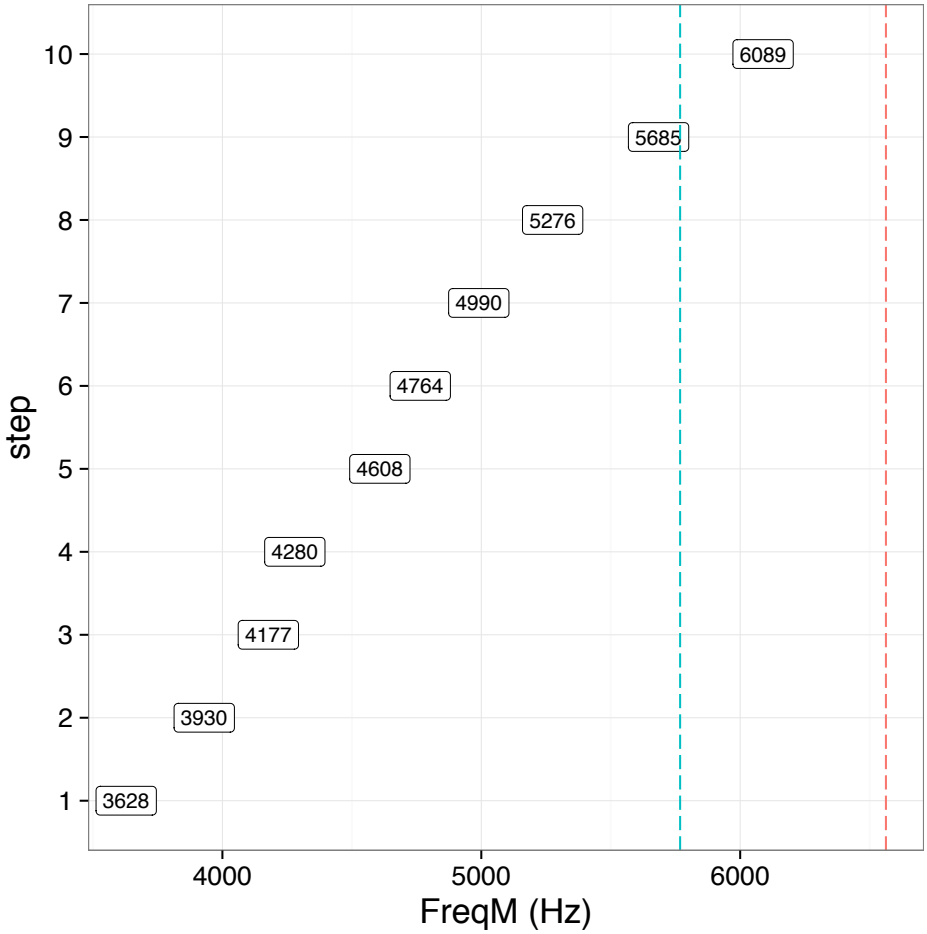
$$\mu_s = 439.30 + 0.930\mu_z$$

Given the mean $Freq_M$ of the talker-specific [z]s in the exposure stimuli (high: 6586 Hz; low: 5730 Hz), the predicted mean $Freq_M$ for [s] would be 6564 Hz for the high COG condition and 5768 Hz for the low COG condition. As shown in Figure 4.13, the predicted $Freq_M$ mean for the high COG condition is beyond any of the continuum stimuli, and the predicted $Freq_M$ mean for the low COG condition also falls high within the continuum members. Therefore, the [s]-[ʃ] continuum may not actually have been representative of the expected $Freq_M$ realization for either talker.

The following study extended the previous delayed categorization experiment, and presented a preliminary test of the $Freq_M$ hypothesis. First, the experiment employed the most ambiguous portion of the [s]-[ʃ] continuum and supplemented it with additional

tokens in the designated (lower) range. Second, in addition to having high and low COG exposure groups, a third group of participants received no exposure to the talker prior to categorization. If listeners learned anything at all about the talker in exposure, then their boundary may differ substantially from listeners with no prior exposure. Given the high predicted $Freq_M$ means for both the high and low COG [z] talker, the boundary for both exposure groups would likely be shifted towards the canonical [s] in comparison to the boundary for listeners with no exposure.

Figure 4.13. The $Freq_M$ (Hz) over the entire fricative of each member of the [s]-[ʃ] continuum, bandpass-filtered between 550 Hz and 10,000 Hz. The red line corresponds to the predicted mean for [s] given the high exposure condition. The blue line corresponds to the predicted mean for [s] given the low exposure condition.



4.8.1 Methods

4.8.1.1 Participants

A total of 51 participants were recruited from the JHU undergraduate community for the delayed categorization experiment with ambiguous categorization stimuli; however, three participants were excluded after reporting that they had heard [s] at some point during the exposure phase. Therefore, data from 48 participants (29 female) was analyzed. Participants received one of three conditions: high COG [z] exposure, low COG [z] exposure, or no exposure to the speaker.

For the high and low COG [z] exposure, there were 32 participants (22 female). All participants spoke English fluently and all but four spoke English as their first language. Two participants learned Korean as their L1, one learned Amharic, and another learned Mandarin. Including these four, 17 participants were bilingual, trilingual, or heritage speakers of another language (Amharic, Cantonese French, Hindi, Korean, Mandarin, Romanian, Spanish, Urdu).

For the no exposure condition, there were 16 participants (7 female). Fifteen of the participants were native speakers of English, and one was a native speaker of Filipino, but spoke English fluently. Including this participant, seven participants were bilingual, trilingual, or heritage speakers of another language (Cantonese, Hindi, Filipino, Greek, Serbian, Spanish, Mandarin).

4.8.1.2 Stimuli

Exposure stimuli. This experiment employed the [z]-initial exposure stimuli described in section 4.3.1.2.

Categorization stimuli. The [s]-[ʃ] continua were modified such that the stimuli ranged over the ambiguous region identified in the previous experiments. Specifically, the acoustic parameters that defined former step 3 and step 7 served as the endpoints of a new 11-point interpolation. As before, the segments were 150 ms in duration with a rise time of 110 ms and a fall time of 30 ms. The three main spectral points of the [s] endpoint were at 2226 Hz, 5039 Hz, and 8909 Hz with respective slopes of 28 dB/oct, 52 dB/oct, and 50.5 dB/oct. The relative amplitudes of the first and third peak compared to the second peak were -16 dB and 12.8 dB. For the [ʃ] endpoint, the peaks were located at 1908 Hz, 4068 Hz, and 7729 Hz and had respective slopes of 32 dB/oct, 48 dB/oct, and

44.5 dB/oct. The relative amplitudes of the first and third peak compared to the second peak were -4 dB and 3.2 dB. Each segment was scaled to 65 dB. The segments were appended to the VC portions ([it] and [ut]) described in section 4.2.1.2, and 20 ms of silence was added to each end. While 11 points were generated, the most [ʃ]-like segment was excluded, resulting in 10 steps per continuum.

4.8.1.3 Procedure

The same procedure was used as in Experiment 5 (section 4.7.1.3). In addition to a high and low [z]-exposure condition, a third of the participants received no [z] exposure and thus performed only the image recognition and [s]-[ʃ] categorization tasks.

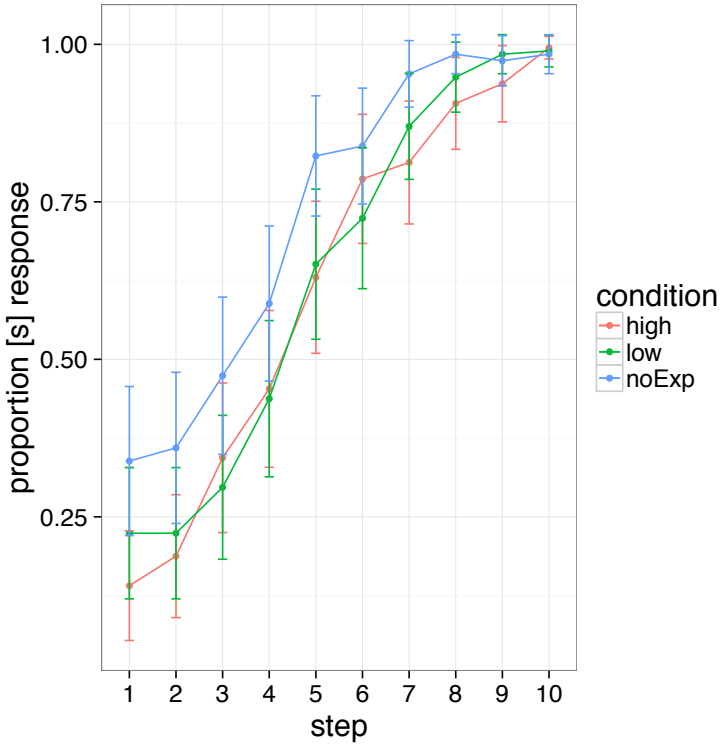
4.8.2 Results

A logistic mixed-effects model was used to analyze the response pattern. The model had fixed effects of the condition (*cond.high:* no exposure 0, high 0.5, low -0.5; *cond.noExp*: no exposure 0.5, high 0, low -0.5), vowel, and step, as well as interactions between condition and vowel, and a random intercept for participant. As in the preceding experiment, participants were no less likely to respond [s] after exposure to the high COG [z] compared to the average (*cond.high*: $\beta = -1.18$, $p = 0.17$); however, there was a significant effect of overall exposure in that participants were more likely to respond [s] after no exposure compared to the average of all participants (*cond.noExp*: $\beta = 1.71$, $p < 0.05$). This suggests that both the high and low COG [z] conditions shifted listeners' boundaries towards [s]. This may be related to the predictions of the mid-frequency peak, which may have greater perceptual relevance for tracking sibilant properties.

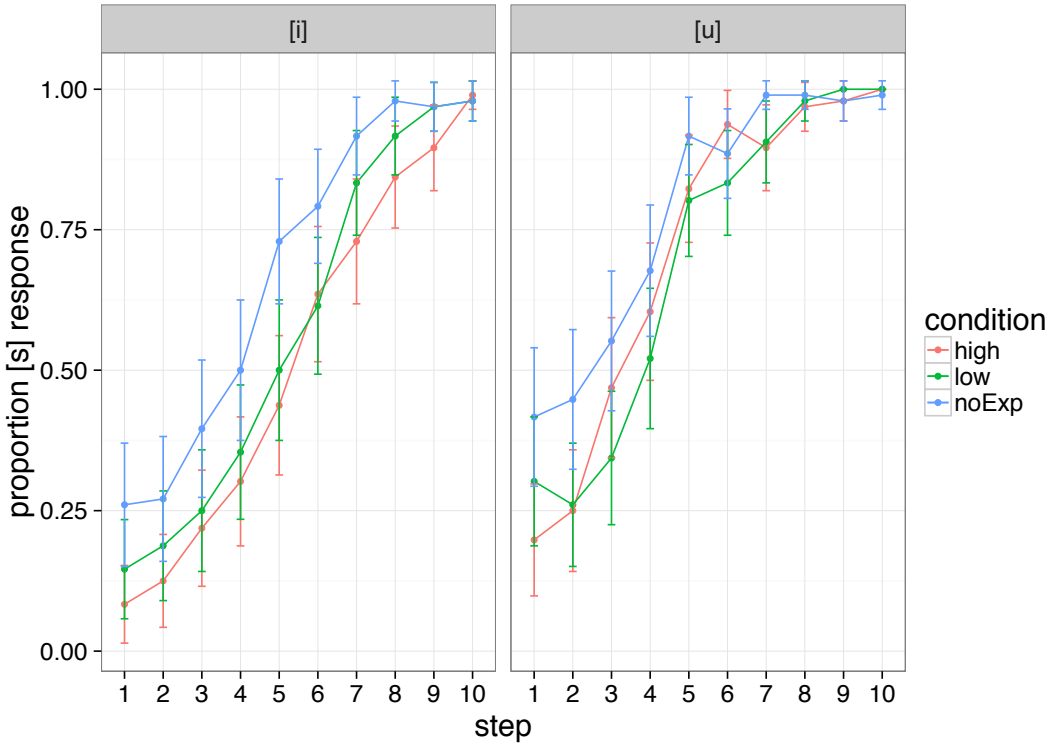
In addition, there were significant effects of vowel and step (*vowel*: $\beta = 1.47$, $p < 0.001$; *step*: $\beta = 2.60$, $p < 0.001$). The interaction between the high COG [z] exposure and

vowel also reached significance, indicating a greater separation in [s] response rates between [i] and [u] for participants in the high COG [z] condition than on average (*cond.high x vowel:* $\beta = 0.80$, $p < 0.01$). The interaction between no exposure and vowel revealed a trend towards a smaller separation in [s] response rates between [i] and [u] for the no exposure condition; however, this did not reach significance (*cond.noExp x vowel*: $\beta = -0.42$, $p = 0.10$).

Figure 4.14. Proportion [s] response following exposure to the high and low COG [z] stimuli, or after no exposure, after a 14-minute intervening period a) for the high-ambiguity [i] and [u] [s]-[ʃ] continua combined and b) for each continuum.



a)



b)

4.8.3 Discussion

The results of this experiment revealed a significant difference in the [s] response rate between speech exposure and no speech exposure. Furthermore, the shift of the category boundaries is congruent with the expected direction given the talker mean $Freq_M$ for [s] predicted by the linear fit to the talker population. Specifically, the [s]-[ʃ] boundary on this continuum should be quite high for both the high and low exposure conditions, which was indeed the case. Moreover, the observed boundaries for the high and low exposure conditions were higher than the boundaries given no exposure. This suggests that listeners transferred spectral properties of the [z] to [s] across a 14-minute period. As studies of spectral contrast focus on the highly local effects, it appears unlikely that a spectral contrast account could explain these results. Nevertheless, an additional study is necessary to determine whether listeners also generalize spectral properties of LTAS-matched noise across this length of time. These findings provide preliminary evidence that listeners extract talker-specific spectral properties of sibilants and generalize these to unheard speech sounds. Furthermore, the direction of the boundary shifts points to the mid-frequency peak as a potentially relevant auditory cue to sibilant categorization.

4.9 General discussion

This series of experiments provided strong evidence in support of spectral contrast effects in speech perception, as well as preliminary evidence that listeners may employ knowledge of phonetic covariation in generalized adaptation to a novel talker's voice. The findings of Experiment 2 indicate that cue-based normalization accounts do not reflect perceptual behavior, which has implications for several models of talker

normalization that perform mean subtraction (e.g., Nearey & Assmann, 2007; McMurray & Jongman, 2011), and models of talker adaptation more generally (e.g., Kleinschmidt & Jaeger, 2015). For instance, the C-CuRe model achieves talker normalization of fricative COG through subtraction of the talker mean COG from all 8 fricative categories [s z ʃ ʒ θ ð f v] prior to token categorization (McMurray & Jongman, 2011). Although COG may adequately characterize [s] and [v], there is no systematic relationship of the talker mean COG between these two categories. Moreover, listeners are sensitive to this relationship, and do not transfer spectral characteristics of [v] to [s].

Generalized adaptation, however, does occur in certain instances, and perceptual mechanisms related to spectral contrast, as well as phonetic covariation may play a role. The first four experiments analyzed generalized adaptation in local contexts. The observed transfer from one sound to another could largely be accounted for by a spectral contrast account: both speech and matched non-speech stimuli, with sufficient energy in the frequency range relevant for [s]-[ʃ] categorization, elicited comparable adaptation effects. Phonetic covariation could also underlie generalized adaptation in non-local contexts, or over a lengthened period (~14 minutes).

Note, however, that the present experiments do not fully distinguish spectral contrast from phonetic covariation in this case. While spectral contrast has been distinguished from very low-level auditory influences, it is still assumed to rely on relatively local influences. For example, Holt & Lotto (2002) found that spectral contrast effects can persist over a 1.3 second delay, which was already longer than expected under the assumption that spectral contrast may be a low-level automatic effect. Nevertheless, listeners could still transfer non-linguistic auditory information over this period.

Additional research is required to determine whether such delayed transfer is also observed with non-speech stimuli.

A few insights were gained from the delayed categorization experiments. First, the pattern of generalization in the [s]-[ʃ] categorization aligned more neatly with predictions of spectral mid-frequency peak ($Freq_M$) covariation as opposed to spectral COG covariation. $Freq_M$ more accurately tracks the place of articulation than COG, providing a phonetically-based motivation for the observed transfer effect, but this spectral peak may also be prominent in general auditory mechanisms of adaptation, particularly for [s]-[ʃ] categorization.

Second, generalized adaptation over a delay was not nearly as strong as that observed in the rapid alternation between exposure and test, and could be detected only with high-ambiguity [s]-[ʃ] stimuli and critically in comparison to a control group with no prior exposure to the talker's voice. In Experiment 1, there was a clear separation between the high and low COG [z] stimuli with low-ambiguity [s]-[ʃ] stimuli, but no separation between the high and low COG exposure response curves. Perceptual noise introduced during the long interval between exposure and test may have diminished the effect. Alternatively, transfer over a lengthened period may arise from a separate mechanism (e.g., phonetic covariation) than spectral contrast.

Additional research is necessary to disentangle a few confounds in the present design. First, the continuum employed in this experiment was designed based on predictions of COG covariation among fricative properties; however, the pattern of generalization suggested that listeners may be more attune to the mid-frequency peak as opposed to the COG. Designing a new continuum based on the predictions of $Freq_M$ may

clarify questions regarding the strength and time course of generalized adaptation, regardless of whether transfer occurs due to spectral contrast or phonetic covariation. Second, we plan to test whether listeners also transfer spectral properties (e.g., $Freq_M$) of non-speech stimuli across a 14-minute period, as they do for speech stimuli. This could be analyzed be replacing the [z]-initial stimuli in the delayed categorization experiment (Experiment 6) with the LTAS-matched white noise stimuli. Finally, further research is necessary to determine whether listeners were amenable to spectral contrast effects from white noise because of its turbulent source or whether tonal stimuli would elicit comparable effects to sibilant fricatives.

4.10 Conclusion

The present study investigated the mechanisms behind generalized adaptation to talker-specific spectral properties in fricative consonants. The findings revealed that spectral contrast had a strong effect on perceptual generalization, particularly when the following conditions were met. First, the exposure and test stimuli were adjacent to one another, and second, the exposure stimulus contained sufficient energy in the relevant frequency range for the [s]-[ʃ] contrast (e.g., the mid frequency range, estimated to be around 2000 to 7000 Hz). Effects of the exposure condition also persisted with a 14-mintue period between exposure and test. Additional research is necessary to distinguish whether spectral contrast can persist over such a period, or whether perceptual knowledge of phonetic covariation may better account for non-local generalized adaptation.

5 Conclusion

In this thesis, I proposed three uniformity constraints that operate at the phonetics-phonology interface: pattern, target, and contrast uniformity. Pattern uniformity, as a very general constraint on phonetic implementation, requires a similar structure of phonetic targets across talkers. As more specific instances of pattern uniformity, target and contrast uniformity directly influence the mapping from distinctive features to phonetic targets. Target uniformity requires similar (or identical) phonetic realization of a distinctive feature value, whereas contrast uniformity requires a comparable phonetic difference in sounds that contrast in a feature across talkers.

The predictions of target and contrast uniformity, in particular, were evaluated in the realization of stop consonant VOT and sibilant mid-frequency peak ($Freq_M$) across languages and talkers. Converging evidence from multiple statistical methods, speech corpora, and languages denoted a strong role of target uniformity on the phonetic implementation of stops and sibilants, whereas evidence for contrast uniformity was relatively weak. As a more general constraint behind covariation, pattern uniformity could also account for many of the observed patterns and could still apply if a language were to violate target uniformity, even marginally.

Not only did target uniformity account for patterns of covariation specific to adult speakers of American English, but it could also account for VOT covariation observed in the speech of children, and across a large and diverse set of languages. These studies revealed the scope of uniformity in shaping the phonetic grammar: the constraint is not specific to an adult English grammar. Rather, it has an early and universal influence on phonetic implementation.

The studies of generalized perceptual adaptation demonstrated that listeners may also exploit knowledge of phonetic covariation among speech sounds in online adaptation to novel talkers. Transfer of talker-specific detail from one category to another was observed for talker-specific stop VOT, as well as for talker-specific $Freq_M$ from [z] to [s]. These findings are consistent with prior knowledge of phonetic covariation among speech sounds, which may be learned either through statistical learning of the exact acoustic relationships, or through knowledge of uniformity in the phonetic realization of distinctive features. While many of the findings involving spectral generalization were consistent with a phonetic covariation account, there were also strong effects of general auditory spectral contrast effects in that white noise matched to the spectral properties of speech could elicit the same response pattern in categorization.

The following sections discuss several topics in phonetics and phonology that bear on uniformity. In particular, I examine its relation to anatomical and physiological influences on phonetic targets in section 5.1, to perceptual dispersion in section 5.2, and economy in section 5.3. Additionally, uniformity has implications for topics in sociolinguistics such as 'bricolage' and relates to notions of social coherence and parallel sound changes (section 5.4). In section 5.5, I consider how a principle of uniformity and its associated statistical diagnostics could potentially be used to "reverse engineer" phonological structure, and in section 5.6, I discuss some cases in which target uniformity appears to be violated and the implications thereof. Finally, I conclude with a discussion of future directions in section 5.7.

5.1 Uniformity, anatomy, and physiology

To a certain extent, covariation of spectral properties (e.g., vowel formants, fricative spectral shape) can be attributed to talker-specific anatomical properties, such as the length and shape of the vocal tract, that have a direct physical relation to resonant frequencies. While physiological and aerodynamic accounts have been offered for VOT differences across place of articulation in unaspirated stops, extension of such mechanical explanations to aspirated stops has been vexed (see Hoole, 1997 and Cho & Ladefoged, 1999 for extensive reviews).

In addition, cross-linguistic and sociolinguistic variation demonstrates that the same phonological surface segment can be implemented with highly varied phonetic targets, even when the phonological inventory is held constant across sociolects. Theoretically, it could be possible for there to be independent phonetic targets for each surface segment, regardless of its featural relationships to other segments. Drawing on an example given previously, it would be physiologically possible for one talker to produce a Japanese-like [s] and an English-like [z] (which differ in both spectral COG and spectral dynamics; Li et al., 2007; Reidy, 2016), and for another talker to reverse this pattern. Similarly, from a strictly articulatory perspective, it would be possible for a talker of American English to systematically produce [p^h] with a VOT that is long relative to the population average but [k^h] with a VOT that is relatively short. However, the systematic relationships across talkers documented in this dissertation indicate that the phonetic variation of related speech sounds is more restricted than would be observed given independent implementation of each segment.

5.2 Uniformity and perceptual dispersion

Perhaps the most widely invoked constraint on phonetic systems (aside from anatomical limitations on possible speech sounds) is that of *perceptual dispersion* (e.g., Liljencrants & Lindblom, 1972; Lindblom, 1986; Flemming, 2004). The pressure to maintain sufficient perceptual distance between contrasting categories could potentially account for some of our findings, but we argue that it is insufficient in accounting for the full pattern of results.

In particular, perceptual dispersion should enhance differences between phonetic targets. Dispersion could underlie the VOT covariation of some voiceless stop pairs, such as [p^h] and [k^h], because VOT potentially serves as a secondary perceptual cue for place of articulation (as suggested by Cho & Ladefoged, 1999: 220). Talkers who have relatively long means for [p^h] could ensure that the putative VOT cue for place remains reliable by also having longer means for [k^h]. Critically, however, not all observed correlations among voiceless stops correspond to consistent differences: [t^h] and [k^h] are highly correlated but have similar (and to a degree inconsistently ordered) VOT means across talkers; this covariation arguably shows that the two stops are *less dispersed* within each talker than would be expected from contrast preservation alone.

Similarly, values of $Freq_M$ for sibilants sharing an [anterior] value (e.g., [s] and [z], as well as [ʃ] and [ʒ]) were, in many cases, statistically equivalent, with no significant effect of [voice] on the realization of $Freq_M$. In principle, perceptual distinctiveness could prefer greater contrast in the phonetic realization of these sibilants than is observed along this dimension. Thus a dispersion-theoretic approach to our findings has significant

limitations, especially as it pertains to the phonetic implementation of a shared feature value (i.e., target uniformity effects).

Conversely, the weaker effects of contrast uniformity *could* be reducible to trade-offs between perceptual distinctiveness and articulatory ease: across talkers, the relationship between phonetic targets corresponding to contrasting values of a feature was less consistent than between targets of a shared feature value. Perceptual dispersion may simply require sufficient contrast, which would not necessarily lead to a consistent difference across talkers. The observed, albeit weak covariation observed could be attributable to uniformity effects or a standard setting for contrast required by perceptual distinctiveness.

5.3 Uniformity and economy

Target uniformity, when perfectly enforced, reduces the number of unique targets that appear in the phonetic representations of a given talker. Even imperfect enforcement minimizes the variety of targets that the talker must plan and execute. In this sense, target uniformity could be considered a possible subcase of economy, in which the grammar favors a low-dimensional representation of phonetic structure (Keating, 1984; Clements, 2003). Economy has long been discussed with respect to its role in shaping the phonological and phonetic systems of individual talkers and languages, and has several subcases. For instance, economy of gesture, or motor economy, favors simple articulations, which can be quantified by the number of contributing gestures specified for each sound (e.g., Lindblom, 1983, 1990). Another sense of economy, implicated in phoneme systems, is 'feature economy' or the related notion of a 'maximum utilization of available features' (e.g., Clements, 2003; Ohala, 1979, 1980).

While target uniformity does reduce the number of targets specified within an inventory, it does not necessarily reduce the gestural complexity internal to a segment. Provided that the mapping from feature value to phonetic target is uniform across the relevant set of segments, then the phonetic target could be a complex set of gestures and the constraint proposed here would be satisfied. Target uniformity is more closely related to, yet still distinct from, feature economy. Feature economy requires features to be recombined into a maximal set of phonemically contrasting segments (subject to limitations by other constraints). Target uniformity requires the feature combinations generated by the phonology, whatever these may be, to be realized with an economical system of feature-to-target mappings.

This last sense of economy appears to have been long, albeit implicitly, assumed by phoneticians. For example, in their discussion of VOT differences across stop place of articulation, Hoole & Pouplier (2015) remark that "speakers and languages may try to get as much mileage as possible out of a fairly constant duration of the glottal abductory-adductory cycle (cf. Weismer 1980; Shipp 1982)" (p. 142). The uniformity constraint codifies the notion of maximal reuse of phonetic targets across segments with shared feature values.

5.4 Bricolage, coherence, and parallel shifts

The realization of a linguistic variable is in part determined by socioindexical properties inherent to or conveyed by the talker, including gender, socioeconomic status, age, regional identity, personality traits, as well as more temporary stances such as emotion and level of formality. In principle, a speaker may piece together a collection of linguistic variables to express social identity and meaning, referred to as *bricolage*

(Eckert, 2008; Zimman, 2017). While bricolage permits independence among variables, a given indexical property could also govern a set or cluster of linguistic variables, giving rise to *social coherence* or *cohesion* (Guy, 2013; Guy & Hinskens, 2016).

For example, /t/-releasing in word-final position (as in the word 'cat') is associated with a set of social variables and meaning. On the one hand, it may signal a high degree of education, and on the other, it may signal annoyance, among several other variables such as Britishness, elegance, prissiness, and exasperation (Eckert, 2008:469). This variable could be expressed independently of others. Alternatively, a chosen indexical property could govern /t/-releasing together with a set or cluster of other linguistic variables.

The notion of social coherence specifically highlights *dependencies* among linguistic variables. These dependencies may reflect arbitrary clustering for the purpose of conveying socioindexical properties, or may be due to structural or grammatical relations among variables such as a shared syntactic or phonetic feature. Because target uniformity enforces dependencies of the latter kind, it offers an explanation for some kinds of observed coherence that might otherwise be considered socially arbitrary.

Target uniformity also places important limitations on bricolage. A talker is not free to select independent values for all phonetic variables. Instead, talker-specific values for one phonetic variable sharply limit the values for others. Specifically, target uniformity requires that certain phonetic variables are yoked across segments by virtue of shared phonological features. For example, a talker's realization of the place of articulation feature [s] is paralleled in the realization of [z] to a high degree: these two variables are not socioindexically independent, but rather cohere for a principled reason.

In this respect, uniformity also relates to the notion of *parallel shifts* in sound change (Fruehwald, 2013, 2017). The essential idea of Fruehwald's proposal, like target uniformity, is that phonetic realization operates on phonological features rather than entire segments. Changes in phonetic realization can thus induce parallel changes in several featurally-related segments. This view converges with the framework of structured phonetic variation presented in this dissertation, and with the evidence for target uniformity in particular. This principle affect phonetic implementation within each individual talker, and thus also has clear implications for expected patterns in language variation and change.

5.5 Reverse engineering structure

The statistical methodology adopted here has revealed striking patterns in the phonetic realization of stops and fricatives. In particular, the predictions of target uniformity were largely confirmed in strong, sometimes near-perfect covariation of talker-specific realizations among segments with a shared phonological feature value. In addition, mixed-effects models were used to assess the contribution of each phonological feature to a given phonetic dimension, while also taking into account other known sources of phonetic variation. In each case, phonological features other than the one primarily expressed by a phonetic cue (e.g., place of articulation features for VOT and [voice] for $Freq_M$) had only marginal contributions within and across talkers.

If indeed target uniformity has a strong influence on phonetic implementation, then the statistical methods presented here may be able to assess hypotheses about feature inventories and specification. Assuming that the phonetic target of a hypothesized feature specification has a close acoustic (or articulatory) correlate, strong covariation of the

target should be observed among the segments that share the specification. Similarly, there should be relatively little influence of co-occurring, but unrelated phonological specifications on the relevant aspect of phonetic realization.

5.6 Deviations from target uniformity

The constraints proposed in this dissertation are intended to be violable and may therefore conflict with other constraints on phonetic implementation. The evidence in this dissertation points towards a strong role of target uniformity, but there exist several findings that appear to violate target uniformity to varying degrees. In the following section, I discuss a sampling of cases in which the mapping from a distinctive feature value to a phonetic target may not be uniform for all segments: intrinsic f0 and duration in vowels, VOT differences among aspirated stops, and distinct glottal spreading gesture in voiceless stops and fricatives.

One explanation for apparent differences among segments with a shared feature is that the relevant phonetic targets are in fact uniform, but an acoustic or articulatory interaction affects the measured values of the target's correlate. However, it is also likely that target uniformity partially yields to competing constraints on phonetic implementation such as perceptual dispersion and articulatory ease.

5.6.1 Intrinsic f0

Intrinsic f0 refers to a pattern in which the fundamental frequency (f0) of high vowels such as [i] and [u] is higher than for low vowels such as [a] (e.g., Mohr, 1971; Whalen & Levitt, 1995). While the size of the effect is small (approximately 4 to 25 Hz greater for high vowels; Ohala & Eukel, 1987), and can vary across languages, a difference in f0 across vowels of different height is nevertheless consistently observed

(Whalen & Levitt, 1995). One explanation for intrinsic f0 would be that the phonetic targets controlling the rate of vocal fold vibration are different for high and low vowels. As tonal languages show, it would be possible for separate phonetic targets to exist in the production of vowels: talkers of even non-tonal languages could in principle employ a high(er) tone target for high vowels and a low(er) tone target for low vowels.

An alternative explanation, suggested by the term 'intrinsic f0', is that the pitch differences arise from an interaction between the different phonetic targets for the vowel height feature and a uniform phonetic target for pitch. In fact it has been argued that the higher f0 of high vowels arises as an automatic consequence of the raised tongue body, which in turn raises the hyoid bone and causes increased tension on the laryngeal system (e.g., Lehiste, 1970; Ohala, 1972). This suggests that the phonetic pitch target for vowels may not vary by vowel height; rather, the f0 effect may arise only in the articulatory interaction of this invariant target with the targets that realize different tongue heights.

Some debate has surrounded the increase in tension on the vocal folds, particularly with respect to the cricothyroid muscle. A few studies have found increased activity of this muscle for high vowels, suggesting an intentional enhancement of the f0 difference across vowels (e.g., Autesserre et al., 1987; Honda & Fujimura, 1991). However, additional findings from Whalen et al. (1999) revealed that the correlation between CT activity and observed f0 varied considerably across talkers and vowels, indicating that the previously reported patterns of this muscle could not be used as conclusive evidence of deliberate enhancement. The overwhelming prevalence of this pattern cross-linguistically, and even among pre-linguistic infants (Whalen et al., 1995), suggests that this case may be a strong candidate for target uniformity of the laryngeal

settings across vowel segments. As with many of the cases to be presented, even if target uniformity is violated, the observed differences are consistently small, indicating some pressure to maintain a high degree of similarity between the laryngeal settings for the high and low vowels.

5.6.2 Intrinsic vowel duration

Intrinsic vowel duration refers to the fact that low open vowels such as [a] have longer average durations than high close vowels such as [i] or [u]. As low vowels require greater articulatory movement in jaw opening, the durational difference could arise from a uniform phonetic target (Lindblom, 1967). Specifically, if the duration and amplitude of the force input to the jaw muscle were held constant for all vowels, the "sluggishness" of the jaw muscle would result in a longer vowel duration for the low vowels than for the high vowels. The abstract phonetic target would be the same, and only in the physical articulation would the observed timing difference arise.

However, in an articulatory study of vowel production, Westbury & Keating (1980) found that force input to the jaw had not only greater amplitude but also longer duration for low [a] than for high [i]. The fact that talkers accentuate force input beyond that required for an acoustic difference in duration suggests that these two vowels do not share a single target on the dimensions relevant for duration, but rather that the target is affected by the segment-internal [high] and [low] specifications. This would be a clear violation of perfect uniformity, though a relatively minor one: for example, durational differences among vowels of different heights are presumably small relative to those among vowels at different speaking rates (or among short and long vowels in languages with vowel length contrasts).

5.6.3 VOT of aspirated stops

While most languages exhibit an increase in VOT with more posterior stops, there are nevertheless exceptions to this rank ordering (see Chapter 2). Previous studies on English aspirated stops have reported considerable variation in the relative ordering of [t^h] and [k^h] (Chapter 2; Docherty, 1992; Yao, 2009), both across languages and talkers. Articulatory evidence from English production also indicates a longer glottal opening gesture for [t^h] than for [k^h], suggesting that the phonetic target for the [+spread glottis] feature is *not* uniform for each segment (Cooper, 1991; Hoole & Pouplier, 2015).[33] Instead, the presence of [CORONAL] appears to interact with the duration of the glottal spreading gesture. The [CORONAL] feature has a relatively unmarked status, and coronals may enjoy somewhat greater freedom of phonetic implementation than other segments. Nevertheless, the observed variation is quite minimal, especially considering the otherwise consistent cross-linguistic patterns. While there may be some context-sensitivity between [+spread glottis] and [CORONAL], the overall patterns suggest a strong influence of target uniformity on the phonetic implementation of [+spread glottis].

5.6.4 Glottal spreading gesture of stops and fricatives

Uniformity in the phonetic targets for the voicing feature shared among stops and fricatives has been suggested based on highly comparable durations in intervocalic position (Weismer, 1980). However, for word-initial voiceless stops and fricatives, Hoole (2006) found that longer glottal gestures and an earlier onset of abduction relative to occlusion for fricatives in comparison to stops (see also Munhall et al., 1985; Löfqvist &

[33] Deviations from the presumed universal ranking have also been observed in Dahalo and Navajo (Cho & Ladefoged, 1999). In Dahalo, the average VOT for unaspirated [t] was greater than the VOT for [k] ([t]: 42 ms, [k]: 27 ms), and in Navajo, the VOT for unaspirated [t] was lower than both unaspirated [p] and [k] ([p]: 12 ms, [t]: 6 ms, [k]: 45 ms).

McGarr, 1987). The earlier onset of abduction in fricatives than in stops may be required in order to minimize "laryngeal resistance to air flow: and to facilitate "the build-up of oral pressure necessary for driving the noise source" (Löfqvist & McGarr, 1987; p. 399). In addition, talkers may avoid an early onset of abduction in stop consonants so as not to pre-aspirate (Hoole, 2006). While uniformity may still be realized in the intervocalic position, the articulatory evidence from word-initial position indicates that the phonetic target for the laryngeal feature in part depends on manner of articulation.

This case raises the interesting and more general question of *which* segment internal features have greater influence on phonetic targets. That is, if there is context-sensitivity in phonetic targets, are there certain features that induce context-sensitivity more than others? Manner features may have a strong influence on the phonetic targets associated with place (e.g., constriction location) and voice (e.g., laryngeal settings), while place and voice may minimally affect the targets of one another.

5.6.5 Discussion

The cases presented in this section demonstrate observed differences — either in the acoustic or articulatory instantiation or in the phonetic targets — when target uniformity would otherwise require identity. The examples reviewed here are by no means exhaustive; other violations include differences tongue configuration between oral and nasal [i] ([i] and [ĩ]) in American English (Carignan et al., 2011). (For differences in the tongue body articulation of paired oral and nasal vowels in other languages, see also Shosted et al., 2012 and Carignan, 2013). This articulatory modification results in a stable realization of F1 for these two vowels, which as described in Carignan et al. (2011), may prevent neutralization of the contrast between [ɪ] and [ɪ̃] on the F1 dimension. For [a] and

[ã], the oral configuration was relatively stable while F1 differed. Consistent with the previous explanation, the marginal raising of F1 due to nasalization does not endanger any contrasts with [a] (e.g., with [ə]), and therefore may not require any modification of the oral articulators.

Cases such as this one suggest that phonetic targets are susceptible to competing pressures from other constraints such as perceptual distinctiveness and articulatory ease. Constraint interaction of this sort could potentially be modeled in a constraint weighting or harmonic grammar framework with an associated cost incurred for uniformity violations (e.g., Legendre et al., 1990; Flemming, 2001).[34] The prediction of a constraint interaction model would be that, in spite the observed deviations from target uniformity, the differences between segments with shared feature specifications are smaller than would be possible given entirely *independent* targets for each segment.

5.7 Future directions

Structured variation in the form of phonetic covariation across talkers serves as an organizing principle for investigating phonological structure, the mapping between phonetics and phonology, and the mechanisms involved in perceptual adaptation and generalization. The evidence for the proposed uniformity constraints, and particularly uniformity of target, is quite strong, and the constraints make predictions about many aspects of speech production and perception. As discussed above, empirically-observed covariation patterns can simultaneously inform our understanding of the production system, including the interaction of several constraints on phonetic implementation and

[34] Uniformity violations could be calculated with a squared difference metric (e.g., Flemming, 2001): that is, any difference between two phonetic targets which should otherwise be uniform could be penalized by squaring the difference, such that greater distances would correspond to more extreme violations of the constraint.

the relationship between phonetic and socioindexical variables. In addition, empirical covariation can inform predictions about perceptual generalization of talker-specific characteristics across speech sounds.

How broadly uniformity applies remains to be seen on several dimensions. First, it will be critical to understand which natural classes are subject to strong uniformity effects. For example, is the [-anterior] feature of affricates such as [tʃ] and [dʒ] required to be uniform in its phonetic realization with fricatives such as [ʃ] and [ʒ], and similarly, do all [+anterior] obstruents ([t], [d], [s], and [z]) have targets that are measurably drawn together by uniformity within the speech of individual talkers? Or does uniformity apply strongly only within classes defined by a manner feature, as suggested by the previous discussion of glottal spreading in stops and fricatives? In addition, future research is necessary to determine whether uniformity applies as strongly to other features and segments as it does to the stops and sibilants investigated in the present dissertation.

In a similar vein, questions also remain as to how universally uniformity applies for a given feature-target mapping, how early patterns that are consistent with uniformity arise in phonetic acquisition, and how uniformity influences language variation and change across dialects. The potential for future research in cross-linguistic research, phonetic acquisition, and sociophonetic study is immense. Uniformity should also be considered in light of other, often conflicting constraints on the grammar such as perceptual distinctiveness and articulatory ease.

In its current formulation, uniformity is assumed to apply on the phonetic implementation of phonological surface segments that contain distinctive feature values. The proposed constraints, however, could be slightly modified to govern the mapping

from underlying segments to phonetic targets directly (e.g., Liberman, 2017a), or to apply within a single level of gestural representations like that posited in in Articulatory Phonology (Browman & Goldstein, 1992). For example, the phonetic realization of English aspirated stops has been assumed throughout to arise from two mappings: one from underlying segments such as /p/ to surface segments such as [p^h]; and another from the surface feature [+spread glottis] to a laryngeal spreading and timing target in the phonetics. However, both of these mappings may operate on the same natural class of consonants (the stops) in the same environment (here, and the beginning of word-initial stressed syllables). This raises the possibility that a single mapping from underlying segments to phonetics, with a uniform target supplied to all voiceless stops in the relevant context, would suffice. More generally, because both allophonic rules and target uniformity induce similarity among members of a natural class, it may be desirable to eliminate redundancies between them. In the same vein, uniformity could be stated as a 'static' condition on gestural scores, requiring identity of specification and relative timing within certain contexts.

The dissertation largely focused on structured variation among segments and the mapping from distinctive features to phonetic targets. Another research extension would be to investigate whether uniformity also applies to prosodic structure and its influence on phonetic targets associated with prosodic boundaries, stress, phrasal accents, tones, etc. As discussed in Chapter 2, covariation of VOT may be related to a prosodic parameter for domain-initial strengthening, which would make the prediction that strong covariation should be observed among all phonetic correlates affected by this prosodic parameter . Alternative models of prosodic structure could then be evaluated in future

research through the 'reverse engineering' discussed earlier. This approach would assess whether the posited structure gives rise to replicable phonetic patterns across talkers (e.g., covariation of talker-specific boundary or accent effects).

The empirical findings of phonetic covariation among speech sounds also has important implications for perceptual adaptation to novel talkers. First, it should be investigated whether listeners generalize talker-specific phonetic characteristics across speech sounds in accordance with the observed covariation patterns for other segments and phonetic dimensions. Second, it remains to be seen whether listeners employ knowledge of uniformity principles, or instead directly use empirical covariation relations, for the purpose of phonetic adaptation and generalization. Cases in which natural or artificial languages violate target uniformity would provide a critical testing ground for distinguishing these hypotheses. In addition, formalizations of prior perceptual knowledge of covariation should be incorporated into computational models of the cognitive processes underlying talker adaptation.

5.8 Conclusion

Variation in the phonetic realization of surface segments is extensive and highly structured. Structured variation has implications for the theory of phonetic realization and models of perceptual adaptation. In particular, the proposed uniformity constraints restrict variation in the phonetic implementation of segments with a shared phonological feature value, and prior knowledge of phonetic covariation among speech sounds can allow for rapid adaptation to novel talkers. More generally, structured variation and uniformity contribute to our understanding of the grammar of phonetic realization, and should be

evaluated along further acoustic and articulatory dimensions for additional segments, phonological units, and languages.

6 Appendix

Table 6.1. Papers cited in cross-linguistic VOT meta-analysis in Chapter 2, section 2.5.

Abdelli-Beruh (2009), Antoniou et al. (2011), Bandeira & Zimmer (2012), Banov (2014), Beckman et al. (2011), Behlau et al. (1988), Bennett (2010), Bortolini et al. (1995), Byrd (1993), Caramazza et al. (1973), Chao & Chen (2008), Chao et al. (2006), Cheng (2011), Cheng (2014), Cho & Ladefoged (1999), Cubrovic (2011), de Carvalho (2011), Docherty (1992), Dubyné (2014), Flege & Eefting (1987), Flemming et al. (2008), Gallagher (2010), Ganenkov (2011), Gósy (2009), Hajek & Stevens (2005), Hawkins (1979), Homma (1981), Inglis (2013), Istre (1980), Johnson & Wilson (2002), Keating (1980), Keating (1981), Klatt (1975), Knoll (2015), Kollia (1993), Kopczynski (1977), Kozminska (2015), Lisker & Abramson (1964), Lousada et al. (2010), Lundeborg et al. (2012), Maddieson et al. (2001), Madhu et al. (2014), Mayr & Montanari (2015), McCarthy et al. (2013), Midtlyng (2011), Mikuteit & Reetz (2007), Monaka (2005), Morgan (2011), Mortensen & Tøndering (2013), Obler (1982), Öğüt et al. (2006), Pasquale (2005), Post (2007), Raphael et al. (1983), Raphael et al. (1995), Recasens (1985), Rochet & Yanmei (1991), Rosner et al. (2000), Shi & Liao (1986), Shimizu (1989), Silva (2006), Simental (2014), Simon (2010), Stevens & Hajek (2004), Stewart (2015), Stölten & Engstrand (2002), Suomi (1980), Vagges et al. (1978), van Alphen & Smits (2004), Vanlocke (2011), Vicenik (2008), Williams (1977), Wu & Lin (1989), Yao (2007)

Table 6.2. Acoustic measures of a) the initial fricative and b) the initial fricative-vowel portion in the [z]-initial stimuli reported in Chapter 4, section 4.3.1.2. Spectral measures included frequencies up to 10 kHz unless otherwise specified. Within each cell, the high-level value is on the left and the low-level value is on the right.

a)

Word	[z] COG	[z] COG 550 Hz	[z] $Freq_M$
zeet	4987 | 765	8528 | 6509	6740 | 6406
zit	7813 | 770	8389 | 6566	6212 | 6072
zet	7422 | 4379	8220 | 7065	6848 | 6438
zate	7331 | 939	7634 | 6450	6988 | 6094
zat	7837 | 873	7845 | 6435	6697 | 6471
zoot	4971 | 374	7515 | 4675	6277 | 4457
zoat	7822 | 559	8495 | 5219	5825 | 5168
zought	3437 | 535	7169 | 5429	6449 | 5190
zot	7884 | 5483	8515 | 5910	6848 | 5749
zut	4948 | 2063	8515 | 5919	6966 | 5254

b)

Meg	CV COG	CV COG 550 Hz	CV $Freq_M$	Word duration
zeet	1360 | 403	7953 | 4620	6613 | 3141	198
zit	1292 | 430	7273 | 2623	6460 | 3211	209
zet	616 | 544	1626 | 1407	6920 | 6565	234
zate	469 | 436	2883 | 1855	3243 | 3243	254
zat	864 | 587	1516 | 904	5351 | 6481	187
zoot	1011 | 376	6633 | 3071	6398 | 4307	212 | 226
zoat	668 | 425	2860 | 913	5273 | 5203	247
zought	671 | 446	1932 | 784	6998 | 5243	278
zot	567 | 570	819 | 850	5192 | 5214	286
zut	1371 | 590	2520 | 972	6794 | 5146	246
Kim	**CV COG**	**CV COG 550 Hz**	**CV $Freq_M$**	**Word duration**
zeet	925 | 397	7677 | 5003	6651 | 6212	225
zit	2209 | 481	7506 | 2918	6455 | 6164	183
zet	1055 | 677	2855 | 1991	6923 | 6562	195
zate	538 | 404	3022 | 1527	3133 | 3133	238
zat	562 | 532	6875 | 2662	5198 | 3033	249
zoot	814 | 335	6875 | 2662	6395 | 4708	237
zoat	865 | 398	3161 | 1031	5278 | 5192	233 | 223
zought	1110 | 504	2316 | 844	6608 | 5338	251
zot	927 | 1038	2244 | 2371	5370 | 6188	217 | 213
zut	2074 | 577	5548 | 1917	6791 | 5469	216

Table 6.3. Acoustic measures of a) the initial fricative and b) the initial fricative-vowel portion of the [v]-initial stimuli reported in Chapter 4, section 4.4.1.2. Spectral measures included frequencies up to 10 kHz unless otherwise specified. Within each cell, the high-level value is on the left and the low-level value is on the right.

a)

Word	[v] COG	[v] COG 550 Hz	[v] $Freq_M$
veet	294 | 243	4472 | 3891	4867 | 5017
vit	310 | 243	4759 | 2388	6395 | 3747
vet	277 | 225	3619 | 4759	4027 | 6858
vate	287 | 228	3706 | 1865	3898 | 6019
vat	640 | 263	5721 | 2686	5103 | 6912
voot	673 | 217	6546 | 1154	6761 | 3704
voat	452 | 230	6894 | 3761	6934 | 6998
vought	472 | 227	6651 | 1167	4802 | 4328
vot	705 | 259	4694 | 3397	4048 | 6998
vut	1197 | 228	5116 | 2390	3413 | 6740

b)

Meg	CV COG	CV COG 550 Hz	CV $Freq_M$	Word duration
veet	316 | 244	2866 | 3909	3260 | 5017	449 | 467
vit	418 | 389	2067 | 1503	4153 | 3147	395
vet	278 | 226	3722 | 3633	4027 | 6858	403
vate	524 | 460	2308 | 1747	4996 | 4062	440
vat	591 | 590	813 | 814	3066 | 3066	530
voot	380 | 371	2045 | 1638	3601 | 3599	408
voat	453 | 231	6914 | 3771	6944 | 6998	452
vought	488 | 496	706 | 700	3741 | 3782	430
vot	622 | 632	803 | 802	3023 | 3004	479
vut	541 | 501	893 | 742	3047 | 3052	391
Kim	**CV COG**	**CV COG 550 Hz**	**CV $Freq_M$**	**Word duration**
veet	319 | 306	3498 | 3307	3082 | 3090	371 | 385
vit	357 | 325	1450 | 1054	3133 | 3133	323
vet	423 | 416	820 | 761	3093 | 3106	308
vate	410 | 306	2940 | 1193	4996 | 3147	350
vat	397 | 392	1130 | 980	5195 | 3077	323
voot	333 | 312	2987 | 1211	5103 | 3836	348
voat	414 | 414	833 | 798	3779 | 3760	341
vought	394 | 394	838 | 800	3356 | 3074	429
vot	412 | 414	891 | 864	3160 | 3001	366
vut	394 | 522	1071 | 784	5133 | 3109	314

7 References

Abdelli-Beruh, N. B. (2009). Influence of place of articulation on some acoustic correlates of the stop voicing contrast in Parisian French. *Journal of Phonetics*, *37*(1), 66–78. http://doi.org/10.1016/j.wocn.2008.09.002

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant & M. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp.103–113). London: Academic Press.

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4), 2031–2039.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*(6), 3171–3183. http://doi.org/10.1121/1.1701898

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. http://doi.org/10.1121/1.1528172

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of Phonetics*, *39*(4), 558–570. http://doi.org/10.1016/j.wocn.2011.03.001

Assmann, P. F., Nearey, T. M., & Bharadwaj, S. (2008). Analysis of a vowel database. *Canadian Acoustics*, *36*(3), 148–149.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, *71*(4), 975–989. http://doi.org/10.1121/1.387579

Autesserre, D., Roubeau, R., Di Cristo, A., Chevrie-Muller, C., Hirst, D., Lacau, J., & Maton, B. (1987). Contribution du cricothyroidien et des muscles sous-hyoidiens aux variations de la fréquence fondamentale en français: Approche électromyographique. In *Proceedings 11th International Congress of Phonetic Sciences* (Vol. 3, pp. 35-38). Academy of Sciences of the Estonian SSR, Tallin, Estonia.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, *24*(4), 527–554. http://doi.org/10.1080/01690960802299378

Bandeira, M. T., & Zimmer, M. C. (2012). The dynamics of interlinguistic transfer of VOT patterns in multilingual children. *Linguagem & Ensino, Pelotas*, *15*(2), 341–364.

Bang, H.-Y., & Clayards, M. (2016). Structured variation across sound contrasts, talkers, and speech styles. Poster presented at *LabPhon15: Speech Dynamics and Phonological Representation*. Ithaca, NY.

Banov, I. K. (2014). *The Production of Voice Onset Time in Voiceless Stops by Spanish-English Natural Bilinguals*. Master's Thesis. Brigham Young University.

Barik, H. C. (1977). Cross-linguistic study of temporal characteristics of different types of speech materials. *Language and Speech*, *20*(2), 116–126.

Barton, D. (1976). *The Role of Perception in the Acquisition of Phonology*. PhD Dissertation. Stanford University.

Barton, D., & Macken, M. A. (1980). An instrumental analysis of the voicing contrast in word-initial stops in the speech of four-year-old English-speaking children. *Language and Speech*, *23*(2), 159–169.

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. (2011). Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics*, *39*(1), 39–49.

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259–284. http://doi.org/10.1017/S0022226712000424

Behlau, M. S., Pontes, P. A. D. L., Ganança, M. M., & Tosi, O. (1988). Análise espectrográfica de formantes das vogais do português brasileiro. *Acta Awho*, *7*(2), 74-85.

Behrens, S., & Blumstein, S. E. (1988). On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *The Journal of the Acoustical Society of America*, *84*(3), 861–867. http://doi.org/10.1121/1.396655

Benjamin, B. J. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research*, *11*(2), 159–167. http://doi.org/10.1007/BF01068218

Bennett, R. (2010). Contrast and laryngeal states in Tz'utujil: A preliminary investigation. *UCSC Linguistics Research Center*, 93–120.

Bertelson, P., Vroomen, J., & Gelder, B. De. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Blacklock, O.S. (2004). *Characteristics of Variation in Production of Normal and Disordered Fricatives, Using Reduced-Variance Spectral Methods*. PhD Dissertation. University of Southampton.

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.

Boersma, P. & Weenink, D. (2015). Praat: Doing Phonetics by Computer [Computer program]. Version 6.0.05, retrieved 06 November 2015 from http://www.praat.org/.

Bortolini, U., Zmarich, C., Fior, R., & Bonifacio, S. (1995). Word-initial voicing in the productions of stops in normal and preterm Italian infants. *International Journal of Pediatric Otorhinolaryngology*, *31*(2–3), 191–206. http://doi.org/10.1016/0165-5876(94)01091-B

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. http://doi.org/10.1016/j.cognition.2007.04.005

Brandschain, L., Graff, D., & Walker, K. (2013). *Mixer 6 Speech LDC2013S03*. Hard Drive. Philadelphia: Linguistic Data Consortium.

Brandschain, L., Graff, D., Cieri, C., Walker, K., & Caruso, C. (2010). The Mixer 6 corpus: Resources for Cross-Channel and Text Independent Speaker Recognition. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)* (pp. 2441–2444). Malta.

Browman, C. P., & Goldstein, L. M. (1992). Articulatory Phonology: An Overview. *Advances in Psychology*, *94*(C), 67–84. http://doi.org/10.1016/S0166-4115(08)62789-2

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The Journal of the Acoustical Society of America*, *130*(6), 3980–3991. http://doi.org/10.1121/1.3658386

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86. http://doi.org/10.1016/j.jml.2015.12.009

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *The Journal of the Acoustical Society of America*, *92*(1), 593–596. http://doi.org/10.1121/1.404271

Byrd, D. (1993). 54,000 American stops. *UCLA Working Papers in Phonetics*, *83*, 97–116.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*(2), 173–199. http://doi.org/10.1006/jpho.1998.0071

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech*, *86*(1), 52–68. http://doi.org/10.1215/00031283-1277510

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, *54*(2), 421–428. http://doi.org/10.1121/1.1913594

Caravolas, M., & Volín, J. (2001). Phonological spelling errors among dyslexic children learning a transparent orthography: The case of Czech. *Dyslexia*, *7*(4), 229–245. http://doi.org/10.1002/dys.206

Carignan, C. (2013). *When Nasal is More Than Nasal: The Oral Articulation of Nasal Vowels in Two Dialects of French*. PhD Dissertation. University of Illinois at Urbana-Champaign.

Carignan, C., Shosted, R., Shih, C., & Rong, P. (2011). Compensatory articulation in American English nasalized vowels. *Journal of Phonetics*, *39*(4), 668–682.

Chao, K., & Chen, L. M. (2008). A Cross-Linguistic Study of Voice Onset Time in Stop Consonant Productions. *Computational Linguistics and Chinese Language Processing*, *13*(2), 215–232.

Chao, K. Y., Khattab, G., & Chen, L. M. (2006). Comparison of VOT patterns in Mandarin Chinese and in English. In *Proceedings of the 4th Annual Hawaii International Conference on Arts and Humanities* (Vol. 840, p. 859).

Cheng, A. (2011). *Finding Remo: A Preliminary Phonetic Analysis of the Language*. Bachelor's Thesis. Swarthmore College.

Cheng, M. (2014). Exploration of the phonetic difference in stops between Hakka infant-directed speech and adult-directed speech. *Concentric: Studies in Linguistics*, *40*(1), 1–35. http://doi.org/10.6241/concentric.ling.40.1.01

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*(2), 155–190. http://doi.org/10.1006/jpho.2001.0131

Cho, T., & Keating, P. A. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485. http://doi.org/10.1016/j.wocn.2009.08.001

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. http://doi.org/10.1006/jpho.1999.0094

Chodroff, E., Maciejewski, M., Trmal, J., Khudanpur, S., & Godfrey, J. J. (2016). New release of Mixer-6: Improved validity for phonetic study of speaker variation and identification. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariana, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1323–1327). Portorož, Slovenia.

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America*, *136*(5), 2762–2772. http://doi.org/Doi 10.1121/1.4896470

Chomsky & Halle, M. (1968). *The Sound Patterns of English*. Cambridge, Mass.: MIT Press.

Chung, H., Kong, E. J., Edwards, J. R., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, *131*(1), 442–454. http://doi.org/10.1121/1.3651823

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. http://doi.org/10.1121/1.1815131

Clarke, C. M., & Luce, P. A. (2005). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. In V. Hazan & P. Iverson (Eds.), *Proceedings of ISCA Workshop on Plasticity in Speech Perception* (pp. 23–26). London, UK.

Clayards, M. A. (2018). Individual talker and token covariation in production of multiple cues to stop voicing. *Phonetica*, *75*(1), 1–23.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809. http://doi.org/10.1016/j.cognition.2008.04.004

Clements, G. N. (2003). Feature economy as a phonological universal. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 371–374). Barcelona, Spain. Retrieved from http://nickclements.free.fr/publications/2003e.PDF

Cohn, A. C. (1993). Nasalisation in English: Phonology or phonetics. *Phonology*, *10*(1), 43–81. http://doi.org/10.1017/S0952675700001731

Cole, J. S., Choi, H., Kim, H., & Hasegawa-Johnson, M. (2003). The effect of accent on the acoustic cues to stop voicing in Radio News speech. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 15–18). Barcelona, Spain.

Cole, J. S., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, *35*(2), 180–209. http://doi.org/10.1016/j.wocn.2006.03.004

Cole, R. A., Stern, R. M., Phillips, M. S., Brill, S. M., Pilant, A. P., & Specker, P. (1983). Feature-based speaker-independent recognition of isolated English letters. In *IEEE Conference on Acoustics, Speech and Signal Processing* (Vol. 8, pp. 731–733). Boston, MA.

Cooper, A. M. (1991). Laryngeal and oral gestures in English /p, t, k/. In *Proceedings of the 12th International Congress of Phonetic Sciences* (Vol. 2, p. 50). Aix-en-Provence, France.

Corrigan, K. P. (2010). *Irish English, Volume 1: Northern Ireland*. Edinburgh: Edinburgh University Press.

Cox, S. (1995). Predictive speaker adaptation in speech recognition. *Computer Speech and Language*, *9*(1), 1–17. http://doi.org/10.1006/csla.1995.0001

Cubrovic, B. (2011). Voice onset time in Serbian and Serbian English. *ELOPE: English Language Overseas Perspectives and Enquiries*, *(8)*1, 9–18. http://doi.org/10.4312/elope.8.1.9-18

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2006). How abstract phonemic categories are ncessary for coping with speaker-related variation. *Laboratory Phonology 10*, 91–111. http://doi.org/10.1017/CBO9781107415324.004

Dankovičová, J. (1997). Czech. *Journal of the International Phonetic Association*, *27*(1-2), 77-80.

Darlington, R. B. (1990). *Regression and Linear Models*. (J. D. Anker & B. Boylan, Eds.). New York: McGraw-Hill Publishing Company.

Das, S., & Hansen, J. H. L. (2004). Detection of voice onset time (VOT) for unvoiced stops (/p/, /t/, /k/) using the Teager energy operator (TEO) for automatic detection of accented English. In J. M. A. Tanskanen (Ed.), *Proceedings of the 6th Nordic Signal Processing Symposium* (pp. 344–347). Espoo, Finland.

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

De Cara, B., & Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 416–423. http://doi.org/10.3758/BF03195470

de Carvalho, F. O. (2011). Oral consonant acoustics in Tikuna (Yuri-Tikuna). In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong.

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

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Studia Linguistica*, *16*(1–2), 104–122.

DiCanio, C. T., Nam, H., Amith, J. D., García, R. C., & Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. *Journal of Phonetics*, *48*, 45–59. http://doi.org/10.1016/j.wocn.2014.10.003

Disner, S. F. (1983). *Vowel Quality: The Relation Between Universal and Language-Specific Factors*. PhD Dissertation. UCLA.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77–95. http://doi.org/http://dx.doi.org/10.1016/j.wocn.2014.12.005

Docherty, G. J. (1992). *The Timing of Voicing in British English Obstruents*. Berlin: Walter de Gruyter.

Dryer, Matthew S. & Haspelmath, Martin (eds.) (2013). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2017-05-15.)

Dubyné, L. E. (2014). *The Effect of Voice-Onset-Time on Dichotic Listening with Consonant-Vowel Syllables: A Replication Study*. Bachelor's Thesis. University of Pittsburgh.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Edwards, J. R., & Beckman, M. E. (2008). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language Learning and Development*, *4*(2), 122–156. Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. http://doi.org/10.2307/2289144

Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta oto-laryngologica. Supplementum*, 68(*257)*, 1-51.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. http://doi.org/10.3758/BF03206487

Elvin, J., & Escudero, P. (2014). Comparing acoustic analyses of Australian English vowels from Sydney: Cox (2006) versus AusTalk. In *Proceedings of the International Symposium on the Acquisition of Second Language Speech. Concordia Working Papers in Applied Linguistics* (pp. 145–156).

Ernestus, M., Kočková-Amortová, L., & Pollak, P. (2014). The Nijmegen Corpus of Casual Czech. In *Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation* (pp. 365–370). Retrieved from http://www.mirjamernestus.nl/Ernestus/NCCCz/Ernestus_Kočkova-Amortova_Pollak_2014_LREC.pdf

Evanini, K., Isard, S., & Liberman, M. Y. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proceedings of Interspeech* (pp. 1655–1658). Brighton, UK.

Evans, J.W. (1996). Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing.

Fisher-Jorgensen, E. (1954). Acoustic analysis of stop consonants. *Miscellanea Phonetica, 2,* 42-59.

Flege, J. E., & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication*, *6*(3), 185–202. http://doi.org/10.1016/0167-6393(87)90025-2

Flege, J. E., Frieda, E. M., Walley, A. C., & Randazza, L. A. (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, *20*(2), 155–187.

Flemming, E. S. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, *18*(1), 7–44.

Flemming, E. S. (2004). Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *The Phonetic Bases of Phonological Markedness* (pp. 232–276). Cambridge, MA: University Press. http://doi.org/10.1017/CBO9780511486401.008

Flemming, E., Ladefoged, P., & Thomason, S. (2008). Phonetic structures of Montana Salish. *Journal of Phonetics*, *36*, 465–491. http://doi.org/10.1016/j.wocn.2007.10.002

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research*, *42*(3), 663–677. http://doi.org/10.1016/j.psychsport.2013.04.005

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. http://doi.org/10.1121/1.396977

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. http://doi.org/10.1121/1.418332

Foulkes, P., Scobbie, J. M., & Watt, D. (2010). Sociophonetics. *The Handbook of Phonetic Sciences*, 703–754. http://doi.org/10.1002/9781444317251.ch19

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*, 268–294. http://doi.org/10.1016/j.wocn.2007.06.005

Fruehwald, J. (2013). *The Phonological Influence on Phonetic Change*. PhD Dissertation. University of Pennsylvania.

Fruehwald, J. (2017). The role of phonology in phonetic change. *Annual Review of Linguistics*, *3*, 25–42. http://doi.org/10.1146/annurev-linguistics-011516-034101

Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? *Turbulent Sounds. An Interdisciplinary Guide*, (June), 281–302. http://doi.org/10.1515/9783110226584.281

Furui, S. (1980). A training procedure for isolated word recognition systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(2), 129–136.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, *12*(2), 75–98. http://doi.org/10.1006/csla.1998.0043

Gallagher, G. E. S. (2010). *The Perceptual Basis of Long-Distance Laryngeal Restrictions*. PhD Dissertation. Massachusetts Institute of Technology.

Ganenkov, D. (2011). Acoustic characteristics of ejective / unaspirated stops in Udi. Paper presented at the *Conference on Caucasian Languages*. Leipzig.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*(7), 2332–2352. http://doi.org/10.3758/s13414-011-0167-x

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. D., & Dahlgren, N. L. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia: Linguistic Data Consortium.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gendrot, C., & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech* (pp. 2453–2456). Lisbon, Portugal.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 78–89. http://doi.org/10.1109/TAU.1968.1161953

Gick, B., Campbell, F., Oh, S., & Tamburri-Watt, L. (2006). Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics*, *34*(1), 49–72. http://doi.org/10.1016/j.wocn.2005.03.005

Gilbert, J. H. (1977). A voice onset time analysis of apical stop production in 3-year-olds. *Journal of Child Language*, *4*(1), 103-110.

Glasberg, B. R., & Moore, B. C. . (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. http://doi.org/10.1016/0378-5955(90)90170-T

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)* (Vol. 1, pp. 517–520). http://doi.org/10.1109/ICASSP.1992.225858

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*(02), 141–174. http://doi.org/10.1017/S0025100302001020

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193. Retrieved from http://jcaa.caa-aca.ca/index.php/jcaa/article/view/2476

Gósy, M., & Ringen, C. (2009). Everything you always wanted to know about VOT in Hungarian. In *IXth International Conference on the Structure of Hungarian*. Debrecen, Hungary.

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables ? *Journal of Pragmatics*, *52*, 63–71. http://doi.org/10.1016/j.pragma.2012.12.019

Guy, G. R., & Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, *172–173*, 1–9. http://doi.org/10.1016/j.lingua.2016.01.001

Haggard, M. P., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of Acoustical Society of America*, *47*(2B), 613–617. http://doi.org/10.3758/BF03197617

Hagiwara, R. (1995). *Acoustic Realizations of American /r/ as Produced by Women and Men*. PhD Dissertation. UCLA.

Hajek, J., & Stevens, M. (2005). On the acoustic characterization of ejective stops in Waima'a. In *Proceedings of Interspeech* (pp. 2889–2893). Lissabon.

Hale, M., & Reiss, C. (2000). "Substance abuse" and "dysfunctionalism": Current trends in phonology. *Linguistic Inquiry*, *31*(1), 157–169.

Hamann, S. (2002). Postalveolar fricatives in slavic languages as retroflexes. *OTS Yearbook*, 105–127.

Hamann, S. (2004). Retroflex fricatives in Slavic languages. *Journal of the International Phonetic Association*, *34*(01), 53–67. http://doi.org/10.1017/S0025100304001604

Hardcastle, W. J. (1973). Some observations on the tense-lax distinction in initial stops in Korean. *Journal of Phonetics*, *1*(3), 263-272.

Harnsberger, J. D. (2000). A cross-language study of the identification of non-native nasal consonants varying in place of articulation. *The Journal of the Acoustical Society of America*, *108*(2), 764–783. http://doi.org/10.1121/1.429610

Hasegawa-Johnson, M., Chen, K., Cole, J. S., Borys, S., Kim, S. S., Cohen, A., Zhang, T., Choi, J.-Y., Kim, H., Yoon, T., & Chavarria, S. (2005). Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. *Speech Communication*, *46*(3–4), 418–439. http://doi.org/10.1016/j.specom.2005.01.009

Hawkins, S. (1979). Temporal coordination in the speech of children: Further data. *Journal of Phonetics*, *7*(3), 235-267.

Hayden, R. E. (1950). The Relative frequency of phonemes in General-American English. *Word*, *6*(3), 217–223. http://doi.org/10.1080/00437956.1950.11659381

Heffernan, K. (2004). Evidence from HNR that /s/ is a social marker of gender. *Toronto Working Papers in Linguistics*, *23*, 71–84. Retrieved from https://twpl.library.utoronto.ca/index.php/twpl/article/view/6208

Hickey, R. (1989). R-coloured vowels in Irish English. *Journal of the International Phonetic Alphabet*, *15*(02), 44–58.

Hickey, R. (2007). Southern Irish English. In D. Britain (Ed.), *Language in the British Isles* (Vol. 1, pp. 135–151). Cambridge: University Press. http://doi.org/10.1515/9783110197181

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. http://doi.org/10.1121/1.411872

Hoit, J. D., Solomon, N. P., & Hixon, T. J. (1993). Effect of lung volume on voice onset time (VOT). *Journal of Speech and Hearing Research*, *36*(3), 516–521. http://doi.org/10.1044/jshr.3603.516

Holliday, J. J., Reidy, P. F., Beckman, M. E., & Edwards, J. R. (2015). Quantifying the robustness of the Engilsh sibilant fricative contrast in children. *Journal of Speech, Language, and Hearing Research*, *58*(3), 622–637. http://doi.org/10.1044/2015

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312. http://doi.org/10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5), 2801–2817. http://doi.org/10.1121/1.2354071

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1), 156–169.

Homma, Y. (1981). Durational relationship between Japanese stops and vowels. *Journal of Phonetics*, *9*, 273-281.

Honda, K., & Fujimura, O. (1991). Intrinsic vowel F0 and phrase-final F0 lowering: phonological vs. biological explanations. In J. Gauffin, B. Hammarbergs (Eds.), *Vocal fold physiology: acoustic, perceptual, and physiological aspects of voice mechanisms* (pp. 149-157), San Diego: Singular Publishing Group.

Hoole, P. (1997). Techniques for investigating laryngeal articulation and the voice-source. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* (Vol. 35). Retrieved from http://www.phonetik.uni-muenchen.de/~hoole/pdf/larymeth_fipkm.pdf

Hoole, P. (2006). *Experimental Studies of Laryngeal Articulation*. PhD Dissertation. LMU Munich.

Hoole, P., & Pouplier, M. (2015). Interarticulatory Coordination. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 133–157). Somerset, MA: Wiley. http://doi.org/10.1002/9781118584156.ch3

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *The Journal of the Acoustical Society of America*, *28*(2), 303–310.

Hunnicutt, L., & Morris, P. (2016). Pre-voicing and aspiration in Southern American English. In *University of Pennsylvania Working Papers in Linguistics* (Vol. 22, pp. 215–224). http://doi.org/10.1017/CBO9781107415324.004

Inglis, D. (2013). Oral stop consonants in Tai Khamti : An acoustic study in voice onset time establishing manner distinctions of articulation. In *46th International Conference on Sino-Tibetan Languages and Linguistics*. Dartmouth College.

Istre, G. L. (1980). Fonologia transformacional e natural: uma introdução crítica. Unpublished manuscript. Florianópolis: UFSC.

Jacewicz, E., Fox, R. A., & Lyle, S. (2009). Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association*, *39*(3), 313–334. http://doi.org/10.1017/S0025100309990156

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Oxford: Blackwell Publishers. http://doi.org/10.1002/9780470757024.ch15

Johnson, C. E., & Wilson, I. L. (2002). Phonetic evidence for early language differentiation: Research issues and some preliminary data. *The International Journal of Bilingualism*, *6*(3), 271–289.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. http://doi.org/10.1121/1.1288413

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419–454.

Keating, P. A. (1980). *A Phonetic Study of a Voicing Contrast in Polish*. PhD Dissertation. Brown University.

Keating, P. A. (1981). A cross-language study of range of voice onset time in the perception of initial stop voicing. *The Journal of the Acoustical Society of America*, *70*(5), 1261. http://doi.org/10.1121/1.387139

Keating, P. A. (1984a). Phonetic and phonological representation of stop consonant voicing. *Language*, *60*(2), 286–319.

Keating, P. A. (1984b). Universal phonetics and the organization of grammars. *UCLA Working Papers in Phonetics*, (59), 35–49.

Keating, P. A. (1985). Universal Phonetics and the Organization of Grammars. In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 115–132).

Keating, P. A. (1990). Phonetic representations in a generative grammar. *Journal of Phonetics*, *18*(3), 321–334.

Keating, P. A. (2003). Phonetic and other influences on voicing contrasts. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 20–23). Barcelona, Spain. Retrieved from http://www.linguistics.ucla.edu/people/keating/ICPhS_phon_pk.pdf

Keating, P. A., Byrd, D., Flemming, E. S., & Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, *14*(2), 131–142.

Kendall, T. (2009). *Speech Rate, Pause and Linguistic Variation: An Examination through the Sociolinguistic Archive and Analysis Project*. PhD Dissertation. Duke University.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168. http://doi.org/10.1006/jpho.1996.0039

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics*, *26*(2), 117–128. http://doi.org/10.1006/jpho.1997.0069

Kirby, J. P., & Ladd, D. R. (2015). Stop voicing and f0 perturbations: Evidence from French and Italian. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper number 0740. Glasgow, UK.

Kirby, J. P., & Ladd, D. R. (2016). Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. The *Journal of Acoustical Society of America*, *140*(4), 2400–2411. http://doi.org/10.1121/1.4962445

Kirov, C., & Wilson, C. (2012). The specificity of online variation in speech production. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 587–592). Sapporo, Japan.

Klatt, D. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review*, *122*(2), 148–203.

Kleinschmidt, D. F., & Jaeger, T. F. (submitted). Inferring listeners' beliefs about unfamiliar talkers. Unpublished manuscript.

Knoll, K. (2015). The perception of English, Mandarin and Polish word-initial stops by Polish schoolchildren and adults, *Theoretical and Applied Linguistics, 1*(3), 71–91.

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language and Hearing Research*, *43*(5), 1211–1228.

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1175–1189.

Kollia, H. B. (1993). Segmental duration changes due to variations in stress, vowel, place of articulation, and voicing of stop consonants in Greek. *The Journal of the Acoustical Society of America*, *93*(4), 2298–2298. http://doi.org/10.1121/1.406483

Kong, E. J. (2009). *The Development of Phonation-type Contrasts in Plosives: Cross-linguistic Perspectives*. PhD Dissertation. The Ohio State University.

Kong, E. J., & Edwards, J. R. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57. http://doi.org/10.1016/j.wocn.2016.08.006

Kopczynski, Andrzej. (1977). *Polish and American English Consonant Phonemes: A Contrastive Study.* Warsaw: Panstwowe Wydawnictwo Naukowe.

Kozminska, K. (2015). Preliminary results of a sociophonetic study of VOT and Polish transnational identities in the UK. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper number 0827. Glasgow, UK.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. http://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. http://doi.org/10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. http://doi.org/10.1016/j.jml.2006.07.010

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics 4(4), 303-320*.

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *The Journal of the Acoustical Society of America*, *70*(2), 340–349. http://doi.org/10.1121/1.386782

Kurath, H., & McDavid, R. I. (1961). *The Pronunciation of English in the Atlantic States: Based upon the Collections of the Linguistic Atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English*. New York: Mouton de Gruyter.

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65.

Ladefoged, P. (1988). The many interfaces between phonetics and phonology. *UCLA Working Papers in Phonetics*, (70), 13–23.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. http://doi.org/10.1121/1.1908694

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker Normalization via general auditory processes. *Frontiers in Psychology*, *3*, 1–9. http://doi.org/10.3389/fpsyg.2012.00203

Lasry, M. J., & Stern, R. M. (1984). A posteriori estimation of correlated jointly Gaussian mean vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(PAMI-6), 530–535.

Leggetter, C. J., & Woodland, P. C. (1994). Speaker adaptation of continuous density HMMs using multivariate linear regression. In *ICSLP (Vol. 94)* (pp. 451–454).

Leggetter, C. J., & Woodland, P. C. (1995). Flexible speaker adaptation using maximum likelihood linear regression. In *Proceedings ARPA Spoken Language Technology Workshop (Vol. 9)*. Austin, Texas.

Levon, E., & Holmes-Elliott, S. (2013). East end boys and west end girls: /s/-fronting in Southeast England. *University of Pennsylvania Working Papers in Linguistics*, *19*(2), 111–120. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1309&context=pwpl

Li, F., Edwards, J. R., & Beckman, M. E. (2007). Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 917–920).

Liberman, M. (2017a). Towards progress in theories of language sound structure. In Brentari, D. & Lee, J. (Eds.), *Shaping Phonology* (festschrift in honor of John Goldsmith). Chicago: University of Chicago Press.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Lindau, M., & Wood, P. (1977). Vowel features. *UCLA Working Papers in Phonetics*, *38*, 41–48.

Lindblom, B. (1967). Vowel duration and a model of lip-mandible coordination. *Transmission Laboratory Quarterly Progress and Status Report (Royal Institute of Technology, Stockholm)*, 1–29

Lindblom, B. (1983). Economy of speech gestures. In *The Production of Speech* (pp. 217–245). New York: Springer.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. Jaeger (Eds.), *Experimental Phonology* (pp. 13–44). Orlando: Academic Press.

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). The Netherlands: Springer Netherlands.

Linker, W. (1982). *Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study*. PhD Dissertation. UCLA.

Linville, S. E. (1998). Acoustic correlates of perceived versus actual sexual orientation in men's speech. *Folia Phoniatrica et Logopaedica*, *50*(1), 35–48. http://doi.org/10.1159/000021447

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608.

Löfqvist, A., & McGarr, N. (1987). Laryngeal dynamics in voiceless consonant production. In Baer T., Sasaki C., Harris K. (Eds.), *Laryngeal function in phonation and respiration* (pp. 391–402). Boston: College Hill.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183. http://doi.org/10.3758/BF03193667

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–619. http://doi.org/10.3758/BF03206049

Lotto, A. J., & Knill, D. C. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–619. http://doi.org/10.3758/BF03206049

Lotto, A. J., & Sullivan, S. C. (2008). Speech as a sound source. In *Auditory Perception of Sound Sources* (pp. 281–305). http://doi.org/10.1007/978-0-387-71305-2_10

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *The Journal of the Acoustical Society of America*, *113*(1), 54–56. http://doi.org/10.1121/1.1527959

Lousada, M., Jesus, L. M. T., & Hall, A. (2010). Temporal acoustic correlates of the voicing contrast in European Portuguese stops. *Journal of the International Phonetic Association*, *40*(3), 261–275. http://doi.org/10.1017/S0025100310000186

Lundeborg, I., Larsson, M., Wiman, S., & Mcallister, A. M. (2012). Voice onset time in Swedish children and adults. *Logopedics Phoniatrics Vocology*, *37*(3), 117–122. http://doi.org/10.3109/14015439.2012.664654

Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice-onset time in word-initial stop consonants. *Journal of Child Language*, *7*(1), 41–74.

Maddieson, I. (1997). Phonetic universals. In J. Laver & W. J. Hardcastle (Eds.), *Handbook of Phonetic Sciences* (pp. 619–639). Oxford: Blackwells Publishers.

Maddieson, I., Smith, C. L., & Bessell, N. (2001). Aspects of the phonetics of Tlingit. *Anthropological Linguistics*, *43*(2), 153–176.

Madhu, S. R., Kumar, M., & Sreedevi, N. (2014). Voice onset time across gender and different vowel contexts in Telugu. *Language in India*, *14*(2), 252–263.

Major, R. (1976). *Phonological Differentatiation of a Bilingual Child*. PhD Dissertation. The Ohio State University.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962–3973. http://doi.org/10.1121/1.2990715

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412. http://doi.org/10.3758/BF03204884

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [sh]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228. http://doi.org/10.3758/BF03204377

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. http://doi.org/10.1080/03640210802035357

Mayr, R., & Montanari, S. (2015). Cross-linguistic interaction in trilingual phonological development: The role of the input in the acquisition of the voicing contrast. *Journal of Child Language, 42*(05), 1006–1035.http://doi.org/10.1017/S0305000914000592

McCarthy, K. M., Evans, B. G., & Mahon, M. (2013). Acquiring a second language in an immigrant community: The production of Sylheti and English stops and vowels by London-Bengali speakers. *Journal of Phonetics*, *41*(5), 344–358. http://doi.org/10.1016/j.wocn.2013.03.006

McDonough, J. & Ladefoged, P. (1993). Navajo stops. *UCLA Working Papers in Phonetics, 84,* 151-164.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246. http://doi.org/10.1037/a0022325

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126. http://doi.org/10.1207/s15516709cog0000

Menyuk, P., & Klatt, M. (1975). Voice onset time in consonant cluster production by children and adults. *Journal of Child Language*, *2*(02), 223-231.

Midtlyng, P. J. (2011). The effects of speech rate on VOT for initial plosives and click accompaniments in Zulu. In E. G. Bokamba (Ed.), *Proceedings of the 40th Annual Conference on African Linguistics* (pp. 105–118). Somerville, MA: Cascadilla Proceedings Project.

Mikuteit, S., & Reetz, H. (2007). Caught in the ACT: The timing of aspiration and voicing in East Bengali. *Language and Speech*, *50*(2), 247–277.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, *50*(1), 271–285.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1-3), 106–115.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and Speech*, *21*(3), 221–241.

Möbius, B. (2004). Corpus-based investigations on the phonetics of consonant voicing. *Folia Linguistica*, *38*(1–2), 5–26. http://doi.org/10.1515/flin.2004.38.1-2.5

Monaka, K. C. (2005). Shekgalagari laryngeal contrasts: The plosives. *South African Journal of African Languages*, *25*(4), 243–257. http://doi.org/10.1080/02572117.2005.10587262

Morgan, D. W. (2011). *Resetting VOT in a Bilingual Region*. Master's Thesis. The University of Texas at El Paso.

Morris, R. J., & Brown, W. S. (1994). Age-related differences in speech variability among women. *Journal of Communication Disorders*, *27*(1), 49–64. http://doi.org/10.1016/0021-9924(94)90010-8

Mortensen, J., & Tøndering, J. (2013). The effect of vowel height on voice onset time in stop consonants in CV sequences in spontaneous Danish. In R. Eklund (Ed.), *Proceedings of Fonetik* (pp. 49–52). Linköping, Sweden.

Morton, J. R., Sommers, M. S., & Lulich, S. M. (2015). The effect of exposure to a single vowel on talker normalization for vowels. *The Journal of the Acoustical Society of America*, *137*(3), 1443–1451. http://doi.org/10.1121/1.4913456

Munhall, K. G., Ostry, D. J., & Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(4), 457–474.

Munson, B., Jefferson, S. V, & McDonald, E. C. (2006). The influence of perceived sexual orientation on fricative identification. *The Journal of the Acoustical Society of America*, *119*(4), 2427–2437. http://doi.org/10.1121/1.2173521

Najafian, M., Safavi, S., Weber, P., & Russell, M. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey 2016* (pp. 132–139). Bilbao, Spain.

Nartey, J. N. A. (1982). *On Fricative Phones and Phonemes*. PhD Dissertation. UCLA.

Nearey, T. M. (1978). *Phonetic Feature System for Vowels*. PhD Dissertation. University of Alberta.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. http://doi.org/10.1121/1.397861

Nearey, T. M. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, *101*(6), 3241–3254.

Nearey, T. M., & Assmann, P. F. (2007). Probabilistic "sliding template" models for indirect vowel normalization. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 246–270). New York: Oxford University Press.

Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, *24*(1), 1–18. http://doi.org/10.1017/S0025100300004965

Newlin-Łukowicz, L. (2013). TH-stopping in New York City: Substrate effect turned ethnic marker? *Penn Working Papers in Linguistics*, *19*(2), 151–160.

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *The Journal of the Acoustical Society of America*, *113*(5), 2850–2860. http://doi.org/Doi 10.1121/1.1567280

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196. http://doi.org/10.1121/1.1348009

Nielsen, K. Y. (2007). Implicit phonetic imitation is constrained by phonemic contrast. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1961–1964). Saarbrücken, Germany.

Nielsen, K. Y. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142. http://doi.org/10.1016/j.wocn.2010.12.007

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, *32*(1), 120–132.

Nordström, P. E. (1976). Female and infant vocal tracts simulated from male area functions, *Journal of Phonetics*, *5*(1), 81–92.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. http://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. http://doi.org/10.3758/BF03206860

Obler, L. K. (1982). The parsimonious bilingual. In *Exceptional language and linguistics* (pp. 339–346). New York: Academic Press.

Öğüt, F., Kiliç, M. A., Engin, E. Z., & Midilli, R. (2006). Voice onset times for Turkish stop consonants. *Speech Communication*, *48*, 1094–1099. http://doi.org/10.1016/j.specom.2006.02.003

Ohala, J. J. (1979). The contribution of acoustic phonetics to phonology. In *Frontiers of Speech Communication Research* (pp. 355–363).

Ohala, J. J. (1980). Moderator's summary of symposium on "Phonetic universals in phonological systems and their explanation." In *Proceedings of the 9th International Congress of Phonetic Sciences* (Vol. 3, pp. 181–194). Copenhagen.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of Acoustical Society of America*, *75*(1), 224–230.

Ohde, R. N. (1985). Fundamental frequency correlates of stop consonant voicing and vowel quality in the speech of preadolescent children. *The Journal of the Acoustical Society of America*, *78*(5), 1554–1561. http://doi.org/10.1121/1.392791

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. http://doi.org/10.1121/1.389856

Ostendorf, M., Shafran, I., Shattuck-Hufnagel, S., Carmichael, L., & Byrne, W. (2001). A prosodically labeled database of spontaneous speech. In *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding* (pp. 5–7). Red Bank, New Jersey. Retrieved from http://www.isca-speech.org/archive_open/archive_papers/prosody_2001/prsr_022.pdf

Ostry, D. J., Feltham, R. F., & Munhall, K. G. (1984). Characteristics of speech motor development in children. *Developmental Psychology*, *20*(5), 859–871.

Pajak, B., Bicknell, K., & Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. In V. Demberg & R. Levy (Eds.), *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 11–20). Sofia, Bulgaria.

Palková, Z. (1994). *Fonetika a fonologie češtiny: s obecným úvodem do problematiky oboru*. Prague: Karolinum.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). Brisbane, Queensland, Australia: IEEE.

Pasquale, M. D. (2005). Variation of Voice Onset Time in Quechua-Spanish Bilinguals. *Contactos y contextos linguísticos*, 227-235.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR Corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 357–362).

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. http://doi.org/10.1016/j.jneumeth.2006.11.017

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.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pierrehumbert, J. B., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Docherty & D. R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 90–117). Cambridge: Cambridge University Press.

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. http://doi.org/10.1016/j.specom.2004.09.001

Podesva, R. J., & Van Hofwegen, J. (2014). How conservatism and normative gender constrain variation in inland California: The case of /s/. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 128–137. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1820&context=pwpl

Polland, B., & Hála, B. (1926). Artikulace českých zvuků v roentgenových obrazech (skiagramech). In *Mimo rozpravy* (Vol. 41). CAVU.

Port, D. K., & Preston, M. S. (1972). Early apical stop production: A voice onset time analysis. *Haskins Laboratories Status Report on Speech Research, SR-29/30,* 125–149.

Port, R. F., & Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *The Journal of the Acoustical Society of America*, *66*(3), 654–682. http://doi.org/10.1121/1.383692

Post, M. W. (2007). *A Grammar of Galo*. PhD Dissertation. La Trobe University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwartz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on ASRU*.

Raphael, L. J., Tobin, Y., Faber, A., Most, T., Kollia, H. B., & Milstein, D. (1995). Intermediate values of voice onset time. In Bell-Berti, F. Raphael, L.J. (Eds)., *Producing Speech: Contemporary Issues: for Katherine Safford Harris*, 117–127.

Raphael, L. J., Tobin, Y., & Most, T. (1983). Atypical VOT categories in Hebrew and Spanish. *The Journal of the Acoustical Society of America*, *74*(S1), S89. http://doi.org/10.1121/1.2021206

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech*, *28*(2), 97–114. http://doi.org/10.1177/002383098502800201

Reidy, P. F. (2016). Spectral dynamics of sibilant fricatives are contrastive and language specific. *The Journal of the Acoustical Society of America*, *140*(4), 2518–2529. http://doi.org/10.1121/1.4964510

Reidy, P. F., & Beckman, M. E. (2012). The effect of spectral estimator on common spectral measures for sibilant fricatives. In *Proceedings of Interspeech*. Portland, Oregon.

Repp, B. H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, *22*(2), 173–189.

Rochet, B., & Yanmei, F. (1991). Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception. *Canadian Acoustics*, *19*(4), 105–106.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. http://fave.ling.upenn.edu

Rosner, B. S., Lopez-Bascuas, L. E., Garcia-Albea, J. E., & Fahey, R. P. (2000). Voice-onset times for Castilian Spanish initial stops. *Journal of Phonetics*, *28*, 217–224. http://doi.org/10.006/jpho.2000.0113

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn, J. T. (1979). Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review*, *86*(5), 415–451. http://doi.org/10.1037/0033-295X.86.5.415

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition*, *48*(1), 31–51. http://doi.org/10.1006/brcg.2001.1302

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*, 96–109. http://doi.org/10.1016/j.wocn.2010.11.006

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, *43*(5), 1178–1179.

Scobbie, J. M. (2006). Flexibility in the face of incompatible English VOT systems. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory Phonology 8 Varieties of Phonological Competence. Phonology and Phonetics 4-2*. (pp. 367–392). New Haven, Conn.

Seidl-Friedman, A., Kobayashi, M., & Cieri, C. (1999). American English Spoken Lexicon LDC99L23. DVD. Philadelphia: Linguistic Data Consortium.

Shadle, C. H. (1985). *The Acoustics of Fricative Consonants*. PhD Dissertation. Massachusetts Institute of Technology.

Shadle, C. H., Chen, W., & Whalen, D. H. (2016). Stability of the main resonance frequency of fricatives despite changes in the first spectral moment, *The Journal of the Acoustical Society of America*, *140*(4), 3219–3220.

Shadle, C. H., Koenig, L. L., & Preston, J. L. (2014). Acoustic characterization of /s/spectra of adolescents: Moving beyond moments. *Proceedings of Meetings on Acoustics*, *12*(060006), 1–20. http://doi.org/10.1121/1.4862854

Shadle, C. H., & Scully, C. (1995). An articulatory-acoustic-aerodynamic analysisof [s] in VCV sequences. *Journal of Phonetics*, *23*(1–2), 53–66. http://doi.org/10.1016/S0095-4470(95)80032-8

Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: Evidence from patterns of temporal stability. *Phonology*, *26*(1), 187–215. http://doi.org/10.1017/S0952675709001754

Shi, F., & Liao, R. R. (1986). Zhongmei xuesheng hanyu seyin shizhi duibi fenxi (Contrastive analysis of the length of stops in Mandarin between Chinese and American students)*, Language Teaching and Linguistic Studies*, *4*, 67-83.

Shimizu, K. (1989). A cross-language study of voicing contrasts of stops. *Studia Phonologica*, *23*, 1012.

Shosted, R., Carignan, C., & Rong, P. (2012). Managing the distinctiveness of phonemic nasal vowels : Articulatory evidence from Hindi. *The Journal of the Acoustical Society of America*, *131*(1), 455–465. http://doi.org/10.1121/1.3665998

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95-EL101. http://doi.org/10.1121/1.4736711

Silbert, N., & de Jong, K. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. *The Journal of the Acoustical Society of America*, *123*(5), 2769–2779. http://doi.org/10.1121/1.2890736

Silva, D. J. (2006). Variation in voice onset time for Korean stops: A case for recent sound change. *Korean Linguistics*, *13*, 1–16.

Šimáčková, Š., Podlipský, V. J., & Chládková, K. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association*, *42*(2), 225–232. http://doi.org/10.1017/S0025100312000102

Simental, G. (2014). *Phonetic Realization of /p, t, k/ in Spanish-English Code-Switching*. Master's Thesis. The University of Texas at El Paso.

Simon, E. (2010). Phonological transfer of voicing and devoicing rules: evidence from L1 Dutch and L2 English conversational speech. *Language Sciences*, *32*, 63–86. http://doi.org/10.1016/j.langsci.2008.10.001

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception & Psychophysics*, *73*, 1195–1215. http://doi.org/10.3758/s13414-011-0096-8

Skaličková, A. (1974). *Srovnávací Fonetika Angličtiny a Češtiny (Comparative Phonetics of English and Czech)*. Academia: Prague.

Smiljanić, R., & Bradlow, A. R. (2008). Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics*, *36*, 91–113. http://doi.org/10.1016/j.wocn.2007.02.002

Smith, B. L. (1978). Effects of place of articulation and vowel environment on "voiced" stop consonant production. *Glossa*, *12*(2), 163–173.

Smith, N. V. (1973). *The Acquisition of Phonology: A Case Study*. Cambridge: Cambridge University Press.

Solé, M.-J. (2007). Controlled and mechanical properties in speech. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 302–321). Oxford: Oxford University Press.

Solé, M.-J., & Estebas, E. (2000). Phonetic and phonological phenomena: VOT. A cross-language comparison. In *Proceedings of the XVII AEDEAN Conference* (pp. 437–444). Vigo, Spain.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 976–984. http://doi.org/10.1121/1.387032

Sonderegger, M. (2015). Trajectories of voice onset time in spontaneous speech on reality TV. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Science*. Paper number 0903. Glasgow, UK.

Sonderegger, M., Bane, M., & Graff, P. (2017). The medium-term dynamics of accents on reality television. Manuscript to appear in *Language*.

Sonderegger, M., & Keshet, J. (2010). Automatic discriminative measurement of voice onset time. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of Interspeech* (pp. 2242–2245). Makuhari, Japan.

Sonderegger, M., & Keshet, J. (2012). Automatic discriminative measurement of voice onset time. *The Journal of the Acoustical Society of America*, *132*(6), 3965–3979. http://doi.org/10.1121/1.4763995

Stelmachowicz, P. G., Pittman, A. L., Hoover, B. M., & Lewis, D. E. (2001). Effect of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults. *The Journal of the Acoustical Society of America*, *110*(4), 2183–2190. http://doi.org/10.1121/1.1400757

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Stevens, M., & Hajek, J. (2004). A preliminary investigation of some acoustic characteristics of ejectives in Waima'a: VOT and closure duration. In *Proceedings of the 10th Australian International Conference on Speech Science & Technology* (pp. 277–282).

Stewart, J. (2015). *Production and Perception of Stop Consonants in Spanish, Quichua, and Media Lengua*. PhD Dissertation. University of Manitoba.

Stölten, K., & Engstrand, O. (2002). Effects of sex and age in the Arjeplog dialect: A listening test and measurements of preaspiration and VOT. *Proceedings of Fonetik, THM-QPSR*, *44*(1), 29–32.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. http://doi.org/10.1177/0261927X99018001006

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference*. Berlin: Mouton de Gruyter.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *The Journal of the Acoustical Society of America*, *60*(1), 213–224.

Stuart-Smith, J. (2016). Social dynamics and phonological representations: Observations from speech and society in Scotland. Invited talk presented at LabPhon15: Speech Dynamics and Phonological Representation. Ithaca, NY.

Stuart-Smith, J., Rathcke, T., Sonderegger, M., & Macdonald, R. (2015). A real-time study of plosives in Glaswegian using an automatic measurement algorithm. *Language Variation-European Perspectives V: Selected Papers from the Seventh International Conference on Language Variation in Europe (ICLaVE 7)*, *17*, 225–237.

Stuart-Smith, J., Timmins, C., & Wrench, A. (2003). Sex and gender differences in Glaswegian /s/. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*, 1851–1854. Retrieved from http://eresearch.qmu.ac.uk/2252/

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095. http://doi.org/10.1037/0096-1523.7.5.1074

Suomi, K. (1980). *Voicing in English and Finnish stops: A typological comparison with an interlanguage study of the two languages in contact*. PhD Dissertation. University of Turku.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309–1325.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, *75*(3), 983–992.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. http://doi.org/10.1121/1.393381

Tabain, M. (2001). Variability in fricative production and spectra: Implications for the hyper- and hypo- and quantal theories of speech production. *Language and Speech*, *44*(3), 57–94. http://doi.org/10.1177/00238309010440010301

Tatman, R. (2016). Speaker dialect is a necessary feature to model perceptual accent adaptation in humans. *4th Pacific Northwest Regional NLP Workshop: NW-NLP 2016*.

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2016). When size matters: Advantages of weighted effect coding in observational studies. *International Journal of Public Health*, 1–5. http://doi.org/10.1007/s00038-016-0901-1

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099. http://doi.org/10.1121/1.3467771

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982. http://doi.org/10.1121/1.3106131

Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, *42*(5), 324–333. http://doi.org/10.1016/j.jcomdis.2009.03.001

Torreira, F., & Ernestus, M. (2012). Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica*, *69*(3), 124–148. http://doi.org/10.1159/000

Turk, A., & Shattuck-Hufnagel, S. (2014). Timing in talking: What is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*, 20130395. http://doi.org/10.1098/rstb.2013.0395

Vagges, K., Ferrero, F. E., Magno-Caldognetto, E., & Lavagnoli, C. (1978). Some acoustic characteristics of Italian consonants. *Journal of Italian Linguistics Amsterdam*, *3*(1), 69-84.

van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics*, *32*(4), 455–491. http://doi.org/10.1016/j.wocn.2004.05.001

Vanlocke, J. (2011). *On the Production of Aspiration and Prevoicing: The Effect of Training on Native Speakers of Belgian Dutch*. Master's Thesis. Universiteit Gent.

Velten, H. V. (1943). The growth of phonemic and lexical patterns in infant language. *Language*, *19*(4), 281–292.

Vicenik, C. (2008). *An Acoustic Analysis of Georgian Stops*. PhD Dissertation. UCLA.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723–735. http://doi.org/10.1121/1.403997

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation and for spectral-envelope for speaker differences distortion, *96*(3), 1263–1282.

Watkins, A J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *99*(6), 3749–3757. http://doi.org/10.1121/1.414981

Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics*, *7*, 197–204.

Westbury, J. R., Hashi, M., & J. Lindstrom, M. (1998). Differences among speakers in lingual articulation for American English /r/. *Speech Communication*, *26*, 203–226. http://doi.org/10.1016/S0167-6393(98)00058-2

Westbury, J. R., & Keating, P. A. (1980). Central representation of vowel duration. *The Journal of the Acoustical Society of America*, *67*(S1), S37-S37.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *The Journal of the Acoustical Society of America*, *69*(1), 275–282. http://doi.org/10.1121/1.385348

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica*, *47*(1–2), 36–49.

Whalen, D. ., Gick, B., Kumada, M., & Honda, K. (1999). Cricothyroid activity in high and low vowels: exploring the automaticity of intrinsic F0. *Journal of Phonetics*, *27*(2), 125–142. http://doi.org/10.1006/jpho.1999.0091

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics*, *23*, 349–366.

Whalen, D. H., Levitt, A. G., & Goldstein, L. M. (2007). VOT in the babbling of French- and English-learning infants. *Journal of Phonetics*, *35*, 341–352. http://doi.org/10.1016/j.wocn.2006.10.001

Whiteside, S. P., & Irving, C. J. (1998). Speakers' sex differences in voice onset time: A study of isolated word production. *Perceptual and Motor Skills*, *86*(2), 651–654.

Wickens, T.D. (2002). *Elementary Signal Detection Theory*. New York: Oxford University Press.

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 469–481. http://doi.org/10.1109/89.326607

Williams, L. (1977). The voicing contrast in Spanish. *Journal of Phonetics*, *5*(2), 169-184.

Winn, M. (2014). *Praat Scripts: Make Fricative Continuum*. Retrieved 16 May 2017 from http://www.mattwinn.com/praat.html.

Wu, Z., & Lin, M. C. (1989). *Shiyan Yuyinxue Gaiyao (An Outline of Experimental Phonetics)*. Beijing: Higher Education.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*(1), 37–48. http://doi.org/10.1016/j.jml.2013.12.003

Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report*, 183–225.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. In M. Pak (Ed.), *Current Numbers in Unity and Diversity of Languages* (pp. 1122–1137). Seoul: Linguistic Society of Korea.

Yao, Y., Tilsen, S., Sprouse, R. L., & Johnson, K. (2010). Automated measurement of vowel formants in the Buckeye corpus. *UC Berkeley Phonology Lab Annual Reports*, (1994), 80–94.

Yeni-Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 966–975. http://doi.org/10.1121/1.387031

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837–1851. http://doi.org/10.1016/j.visres.2008.05.008

Yoon, T., & Kang, Y. (2013). *The Korean Phonetic Aligner Program Suite*. http://korean.utsc.utoronto.ca/kpa/

Yu, A. C. L., Abrego-Collier, C., Phillips, J., Pillion, B., & Chen, D. (2015). Investigating variation in English vowel-to-vowel coarticulation in a longitudinal phonetic corpus. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper number 0519. Glasgow, UK.

Yuan, J., & Liberman, M. Y. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08* (pp. 5687–5790). http://doi.org/10.1121/1.2935783

Yuan, J., & Liberman, M. Y. (2011). Automatic measurement and comparison of vowel nasalization across languages. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2244–2247). Hong Kong.

Zavaliagkos, G., Schwartz, R., & Makhoul, J. (1995). Batch, incremental and instantaneous adaptation techniques for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 676–679). Detroit, MI.

Zimman, L. (2017). Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society*, 1–32. http://doi.org/https://doi.org/10.1017/S004740451700007

Zlatin, M. A. (1974). Voicing contrast: Perceptual and productive voice onset time characteristics of adults. *The Journal of the Acoustical Society of America*, *56*(3), 981–994. http://doi.org/10.1121/1.1903359

Zlatin, M. A., & Koenigsknecht, R. A. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research*, *19*(1), 93–111.

Zue, V. W. (1976). *Acoustic Characteristics of Stop Consonants: A Controlled Study*. PhD Dissertation. Massachusetts Institute of Technology.

Zygis, M. (2003). Phonetic and phonological aspects of Slavic sibilant fricatives. *ZAS Papers in Linguistics*, *3*, 175–213.

Vita

Eleanor Chodroff was born on July 27, 1989 in York, Pennsylvania. She received a B.A. in German and Linguistics with *summa cum laude* and Phi Beta Kappa honors from New York University in May 2012. The following September, she joined the Ph.D. program in Cognitive Science at Johns Hopkins University. During her time at Johns Hopkins, Eleanor worked under the guidance of Dr. Colin Wilson and with researchers in the Center for Language and Speech Processing, Drs. Sanjeev Khudanpur and Jack Godfrey. In July 2017, Eleanor will continue research as a postdoctoral researcher in the Department of Linguistics at Northwestern University with Dr. Jennifer Cole.