Chapter 4

The structure of voice onset time variation in bilingual long-lag stop categories

4.1 Introduction

A consequence of bilingualism is that individuals must navigate segment inventories that exist in a shared phonetic space, in which categories may or may not share aspects of their mental representations (Flege and Bohn, 2021). One of the primary goals in this chapter is to investigate what languages share in the mental representation of similar speech sound categories. The idea of representation is intended here in the manner typically meant by psycholinguists (e.g., Llompart and Reinisch, 2018), exemplar theory proponents (e.g., Amengual, 2018), and Flege and Bohn (2021) in their revised Speech Learning Model (SLM-r). These groups use similar language to describe representation, emphasizing distributions of sensory experiences over theoretical linguistic descriptions. For example, Flege and Bohn describe the units of a multilingual segment inventory as categories comprising input distributions of exemplars: "the sensory stimulation associated with...speech sounds that are heard and seen during production by others...in meaningful con-

versations" (Flege and Bohn, 2021, p. 32). So, if sound categories from different languages exist in the same phonetic space and are represented by distributions of exemplars, how, then, can the extent to which languages share representation(s) be assessed? Much like Chapter **??**, the approach here is one of leveraging the structure of variation to understand the system.

There are many pieces to this puzzle, and the literature has already addressed some of them. The introduction to this chapter proceeds as follows—section 4.1.1 addresses which sound categories are candidates for shared representation in the first place. Section 4.1.2 briefly summarizes the relevant crosslinguistic influence literature, addressing assimilation, dissimilation, and how they reflect on the idea of shared representation. Section 4.1.3 identifies a limitation of the existing paradigms in crosslinguistic influence and proposes adapting the uniformity framework as a way to fill the gap. This framework offers a way to interpret the structure of variation for a given acoustic dimension. Section 4.1.4 introduces the focus of this particular study—long-lag stops in Cantonese and English—and outlines the specific research questions and hypotheses.

4.1.1 Identifying "links" across languages

At first glance, the best candidates for shared representation are sound categories that are "linked" together. The definition of links, however, can be frustratingly vague in the multilingualism literature. In a handbook chapter on bilingual phonetics and phonology, Simonet (2016) describes "links or connections of one sort or another between the phonetic categories" (p. 10). Despite being vaguely defined, links nonetheless represent a crucial concept. In the most basic sense, links are defined by the behavior they account for—they exist between sound categories that exert influence on one another under some set of circumstances. Links behave dynamically, as as such, Simonet also notes that "these connections...are transiently strengthened in contexts that induce the activation of both languages and inhibited in contexts that favor the use of only one of the languages" (2016, p. 10). Arguably, such links must exist because crosslinguistic influence can be observed.

While there may be alternative explanations (i.e., global influence), the concept of links is widely assumed in accounting for bilinguals' behavior.

Flege and Bohn (2021) expand on the idea of links in the SLM-r by providing a framework for predicting which sound categories will be linked together. The proposal is simple—namely, sound categories will be linked to the closest category in the other language. Determining which categories pair up, however, remains an empirical challenge from the perspective of speech production, and as a result, Flege and Bohn (2021) rely on perceptual metrics. The reason for this challenge is because perception and production do not always line up neatly. Flege and Bohn assert that similarity "must be assessed perceptually rather than acoustically because acoustic measures sometimes diverge from what listeners perceive" (2021, p. 33). This assertion is echoed in an overview chapter on crosslinguistic segment similarity, where Chang (2015) argues that in accounting for behavior, similarity is best captured abstractly. Chang states that crosslinguistic influence at the segmental level tends to occur between sounds that share "(1) similar positions in the respective phonemic inventories (when considering the contrastive feature oppositions—or, more broadly, the 'relative phonetics'—of the sounds in relation to other sounds in the inventory), and (2) similar distributional facts" (2015, p. 201). This approach to similarity emphasizes a general role for abstraction but does not necessarily invite a formal phonological analysis. Developing such an analysis would likely constitute a dissertation in itself—Mielke (2012) highlights the challenges of applying phonological features across languages, given the sheer variety of phonetics-phonology mappings in the world's languages.

While Flege and Bohn (2021) and Chang (2015) take different approaches—perceptual ratings and relative phonetics—they ultimately accomplish a similar goal, by accounting for abstraction and nonlinearity in listeners' mental representations. In sum, abstract similarity seems to be a prerequisite for the emergence of a link between two sound categories, given how it does a better job of accounting for when and where crosslinguistic influence occurs. The presence of abstract similarity, however, does not address what happens next. It does not entail any par-

ticular outcome, and it does not directly address how representation is structured for the sound categories in question.

4.1.2 Crosslinguistic influence and representation

The next step in the puzzle is understanding what happens to linked sound categories. The SLM-r outlines two primary outcomes for sound categories in a shared system—assimilation and dissimilation (Flege and Bohn, 2021). Assimilation is a merging of phonetic properties and arguably occurs when bilinguals and learners do not perceive a difference between two categories. Dissimilation, then, is the reverse—a diverging of phonetic properties that occurs when a difference is perceived. Notably, these processes need not impact all phonetic properties in the same way. For example, in a study of coronal stops produced by simultaneous French-English bilinguals, Sundara et al. (2006) found that bilinguals differentiated languages on voice onset time (VOT) and the standard deviation of burst frequency. These bilinguals did not differentiate based on other spectral moments that monolingual comparison populations did. This study illustrates convergence on some but not all properties.

The motivation for these outcomes arises from two simple constraints from the productive and perceptual systems. Effectively, do not get too close to each other in perception, and do not get too complicated in production (Guion, 2003; Lindblom and Maddieson, 1988; Flege and Bohn, 2021). These constraints lead SLM-r to posit that proximity leads to instability, even if what counts as close for early bilinguals remains unclear. In SLM-r, the potential outcomes of instability are assimilation and dissimilation. Considering how bilinguals are fully capable of maintaining subtle distinctions for similar sound categories across languages (e.g., Sundara et al., 2006; Lein et al., 2016; Casillas, 2021), this is not a trivial point to make. It may thus be more appropriate in the case of early bilinguals to also consider contrast maintenance alongside dissimilation.

Following what SLM-r posits, a relatively simple account is that dissimilation for similar sound categories would lead to distinct representations of those

categories and that assimilation leads to a shared representation. The picture is complicated, however, by the idea of imperfect assimilation and what Flege and Bohn term *composite categories*. Suppose sound categories from two languages are phonetically too close to each other but do not fully assimilate. In that case, the SLM-r proposes that they will remain linked in a composite category "defined by the statistical regularities present in the combined distributions of the perceptually linked...sounds." (Flege and Bohn, 2021, p. 41). This scenario might be characterized as imperfect or partially shared representation, where certain dimensions are kept apart and others overlap. For example, the place of articulation may be shared across languages even if VOT differs—this scenario is observed by Sundara et al. (2006), described above. Alternatively, this lack of clear-cut examples of assimilation in the literature may instead indicate that assimilation and dissimilation might be better cast as ends of a spectrum for a gradient, context-sensitive phenomenon. This re-conceptualization makes room for things like contrast maintenance and composite categories.

There are a few potential reasons for the lack of clear-cut assimilation. First, true assimilation might just be rare in bilingual speech. This reason is supported by a recent meta-analysis of crosslinguistic influence for Spanish and English initial stop consonants (Casillas, 2021). In this environment, English long-lag stops and Spanish short-lag stops are linked to one another (Fricke et al., 2016; Goldrick et al., 2014; Bullock and Toribio, 2009; Olson, 2016). Casillas found that early bilinguals did not produce "compromise" stop categories. That is, early Spanish-English bilinguals did not produce VOT that was somehow intermediate to the canonical productions by monolinguals of either language. Instead, the production of each category was influenced by task demands and factors such as social context. So while some assimilation occurs, it is far from the only process. This finding echoes arguments made by Bullock and Toribio (2009) on the sophistication and control that bilinguals exert over their possible forms. So while there is clear evidence of a link between the two sounds—and perhaps even evidence for a composite category—"compromise" seems inappropriate for capturing behavior.

Instead, bilinguals produce a wide range of forms appropriate to and influenced by different contexts. Without considering task and context factors, it is perhaps not surprising that the two sound categories masquerade as a single composite category.

A second reason for the rarity of complete assimilation arises from the experimental and corpus-based approaches typically used to study crosslinguistic influence. Experimental approaches to crosslinguistic influence use paradigms such as sentence reading, isolated word production, or picture naming paired with various common manipulations. At a more global scale, some studies set the language mode of the full session and compare individuals across sessions (Grosjean, 2011; Simonet and Amengual, 2019; Sancier and Fowler, 1997), or compare groups of individuals in different language mode conditions (Antoniou et al., 2010). At a more local level, experimental designs leverage language switching across blocks (Sundara et al., 2006), trials (Goldrick et al., 2014), or within trials (i.e., prompted code-switching Bullock and Toribio, 2009; Antoniou et al., 2011; Olson, 2016). Both types of experimental studies often include both cognate and non-cognate items as a focus or manipulation (e.g., Goldrick et al., 2014). Corpus-based approaches similarly tend to focus on proximity to code-switching (Fricke et al., 2016; Balukas and Koops, 2015) and cognate production (Brown and Amengual, 2015). Across all study types, there are common findings. Typically, cognates, words occurring before a language switch, and words produced in more bilingual modes show increased convergence. Conversely, unilingual modes, non-cognates, and words occurring far from a code-switch tend to show a greater degree of contrast maintenance (or divergence). While there is a tendency for convergence, Bullock and Toribio (2009) demonstrate that proximity to a code-switch in a formal experimental setting leads some individuals to exaggerate the difference between English and Spanish VOT.

In these approaches, the ability to examine crosslinguistic influence for any given pair of sounds hinges on the presence of an observable difference under some set of conditions. Arguably, for this reason, most prior work in crosslinguis-

tic influence has focused on sounds that are phonologically similar (i.e., abstract, relative phonetics) yet phonetically distinct. A common example of this arises from languages that differ in their initial stop voicing contrasts. For example, as discussed at the beginning of this section, North American English contrasts long- and short-lag stops in the initial position. Conversely, Spanish contrasts short-lag and prevoiced initial stops. Despite the clear difference in how languages encode a laryngeal timing contrast, there is nonetheless strong evidence for a crosslinguistic link between English long-lag and Spanish short-lag stops (Casillas, 2021; Fricke et al., 2016; Goldrick et al., 2014; Bullock and Toribio, 2009; Olson, 2016).

These studies demonstrate phonetic convergence—or variable assimilation—in two ways. First, VOT is shorter for English initial stops produced by bilinguals when compared to monolingual control groups. This result is attributed to the influence on English long-lag stops from the short-lag category in the other language (Olson, 2016; Johnson and Babel, 2021). Similarly, French-English bilinguals are more likely to produce lead voicing in initial English voiced stops compared to English monolinguals (Sundara et al., 2006). Second, evidence of crosslinguistic influence can also come from comparing bilinguals to themselves across different circumstances. For example, Fricke et al. (2016) use a spontaneous speech corpus to demonstrate that Spanish-English bilinguals produce shorter, more Spanish-like VOT in the lead up to an English-to-Spanish code switch (Fricke et al., 2016). An experimental example comes from Simonet and Amengual (2019), where individuals participate in multiple sessions in which language mode is carefully controlled. While this body of work makes the presence of a link clear, it also highlights that there are distinct aspects of how these sound categories are represented in the bilingual mind (Casillas, 2021). In the SLM-r, these examples might be considered composite categories. Alternatively, they might be examples of contrasts being maintained in the face of proximity.

In any case, this focus presents a conundrum. By using methods where observing similarity hinges on the ability to detect a difference, researchers often preemptively exclude some of the best candidates for shared representation—those

that share both abstract *and* acoustic similarity. While rare, this is not always the case. One example comparing highly similar sound categories in the early bilingualism literature comes from a lab-based study of Mandarin-English bilingual children (Yang, 2019). The authors found that highly proficient bilingual 5 to 6-year-olds produced equivalent VOT for Mandarin and English long-lag stops, even though the monolingual comparison groups were consistently different. Yang's result suggests that the difference is either too small to maintain or that 5 to 6-year-old children have not yet mastered it. These claims should be tempered, however, as Yang (2019) did not control for language mode, and adult bilingual behavior was not considered.

Despite some inroads, there is nonetheless a distinct paucity of work examining highly phonetically similar speech sounds across languages, even when such a connection would make sense. A recent study of crosslinguistic influence in Cantonese-English bilinguals compares English long-lag and Cantonese short-lag stops in the context of a language switching paradigm (Tsui et al., 2019). While this comparison reflects the need for stimuli to be acoustically distinct beforehand—as noted above—it glosses over the fact that both languages contrast short-lag and long-lag VOT in initial position. The best candidates for links—and accompanying crosslinguistic influence—would be the long-lag stops in each language. These stops occupy the same relative position in their respective inventories and bear resemblance physically (e.g., see references in Section 4.1.4). Tsui et al.'s (2019) null result with balanced bilinguals is thus unsurprising.

This criticism is not intended to suggest that (Tsui et al., 2019) would have gotten more insightful results by comparing Cantonese long-lag stops to English long-lag stops. Rather, it highlights the design constraints of the paradigms used in crosslinguistic influence research. Such methods are better suited for detecting links between sound categories and documenting the circumstances that undergird parallel activation of languages. In Grosjean's (2011) terms, these methods are best suited to detecting *interference*, as opposed to *transfer*. Interference is the kind of crosslinguistic influence observed between simultaneously activated

representations, while transfer occurs on a longer time scale, affecting the representations themselves. While Grosjean (2011) argues that disentangling the two types of influence is difficult, these methods seem tailored more towards interference.

A good example of interference comes from Catalan-Spanish bilinguals' vowel production. Simonet and Amengual (2019) compare vowels on a within-talker basis from two separate sessions—unilingual Catalan and bilingual Spanish-Catalan—and found that Catalan /a/ was produced more like its Spanish counterpart in the bilingual session. While this result is straightforward, it is unique in that the authors show a dynamic within-talker process facilitated by language mode. In a monolingual setting, talkers maintain a contrast. In contrast, the same talkers show partial assimilation in the bilingual setting. When both languages are activated, Catalan interferes with Spanish, leading to the observed outcome of phonetic convergence. Simonet and Amengual (2019) argue that these sounds are linked and thus simultaneously activated but ultimately have separate representations in long-term memory (i.e., do not reflect transfer). In its discussion of category formation, SLM-r seems more concerned with assimilation and dissimilation at the level of long-term representations (i.e., transfer Flege and Bohn, 2021). In such cases, the methods should not have to rely on the presence of detectable acoustic differences. Notably, however, transfer and interference are difficult to disentangle (Grosjean, 2011).

To summarize, most work in crosslinguistic influence has focused on phonologically similar yet phonetically distinct pairs of segments and how they interfere with one another in online processing. These pairs are not strong candidates for transfer and shared mental representation (as defined at the beginning of this chapter). This widespread focus likely arises for several different reasons. The established paradigms—which greatly facilitate research—tend to require a detectable difference. It is also possible that assimilation in long-term mental representations is rare for early bilinguals, which would limit the options for studying such a phenomenon. Lastly, comparisons of categories that already exhibit both abstract

and phonetic similarity may be taken for granted and not considered an interesting problem to focus on, despite the nature of the mental representation of sound categories being a key focus in psycholinguistics (Samuel, 2020).

While many psycholinguists are indeed concerned with representation, processing seems to have taken center stage in the psycholinguistics of bilingualism. In a prominent example of this, Fricke et al. (2019) argue that "bilingualism has the potential to reveal the fundamental breadth and underlying nature of variation in language processing" (2019, p. 204). This chapter foregrounds the argument that bilingualism also offers a window into understanding the nature of mental representation. In the interest of understanding it, the best category candidates would be the hardest to distinguish using surface forms in the first place.

4.1.3 Adapting the uniformity framework

The study described in this chapter focuses on assessing whether phonetically similar sounds share a mental representation or not. Unlike prior work focusing on variable convergence and divergence, this chapter addresses whether a single category deploys to both languages or whether each language carries a separate representation of similar categories. Testing directly for shared structure in this way means that the set of methods that rely on detecting and modulating differences is not appropriate. To this end, this chapter extends the articulatory uniformity framework to the study of multilingual segment inventories.

Articulatory uniformity is conceptualized as a constraint on within-talker phonetic variation, in which articulatory gestures or phonological primitives are implemented systematically in speech production (Chodroff and Wilson, 2017; Faytak, 2018; Ménard et al., 2008). The core idea of the articulatory uniformity framework is that phonetic variation is highly structured. While Chodroff and Wilson (2017) draw tight connections between uniformity and phonological features, Faytak (2018) instead emphasizes how talkers learn and reuse articulatory gestures. This articulatory account builds on earlier work by Ménard et al. (2008), who argue that the stability of the first formant in French vowel production is best accounted

for by stability in the tongue height gesture (i.e., reuse of the gesture). While the specific theoretical accounts vary somewhat by author, there is nothing to suggest that such accounts are incompatible with one another. Both articulatory and phonological explanations are likely valid. Given the focus of this chapter on phonetic and psycholinguistic accounts of category formation and representation, the articulatory account—with its accompanying acoustic consequences—is perhaps more appropriate.

In this light, if a set of segments share an attribute, then talkers should implement the segments with the same phonetic target or articulatory gesture. This systematicity has been observed for vowel height (Ménard et al., 2008), tongue shape (Faytak, 2018), fricative peak frequency (Chodroff, 2017), and stop consonant VOT (Chodroff and Wilson, 2017). In the case of VOT in particular, the relationship between a laryngeal gesture and its acoustic consequence is clear. This allows for the extension of Ménard et al.'s (2008) argument regarding F1 and tongue height to VOT and its corresponding laryngeal gesture. Reusing the gesture across sounds that share the relevant attribute "may simplify the somatosensory feedback needed to control the speech task" (Ménard et al., 2008, p. 26). In simple terms, reusing gestures is easier than the alternative in the case of high vowels. The same argument could easily be extended to long-lag stops.

Findings for within-language stop consonant uniformity appear to be quite robust. Chodroff and Wilson (2017) report consistent results across a lab study based on reading a list of CVC words and a corpus study comprising connected read speech. Chodroff and Baese-Berk (2019) replicate the uniformity findings for stop consonants with connected read speech samples from 140 non-native English speakers with a wide range of native languages in the ALLSSTAR corpus (Bradlow et al., 2011). While Chodroff and Baese-Berk (2019) found a greater degree of between-talker variability with non-native speakers compared to the prior monolingual work (Chodroff and Wilson, 2017), the within-talker structure was robust. However, the uniformity framework has not yet been extended to early bilingual speech, in particular as a mechanism for comparing how bilinguals pro-

duce phonetically similar sounds in each of their languages. Extending the framework across languages follows the framing of uniformity as arising from articulatory reuse (Faytak, 2018), effectively asking whether or not reuse extends across languages.

4.1.4 Long-lag stops in Cantonese and English

English and Cantonese initial long-lag stops are strong candidates for shared mental representation because they exhibit both relative and physical phonetic similarity, akin to the difference for Mandarin and English in Yang (2019). Consider the initial stop [k^h]—in citation speech—with a mean VOT of 80 ms in American English (Lisker and Abramson, 1964) and 91 ms in Hong Kong Cantonese (Clumeck et al., 1981). While these values are objectively different—though based on small sample sizes—it seems that using the same laryngeal timing gesture would be advantageous given the small difference across monolingual populations (that may or may not be perceptible). There is ample work documenting VOT across different varieties of English and speaking styles, with values as low as the 30–50 ms in spontaneous speech (Stuart-Smith et al., 2015). There is far less work documenting Cantonese long-lag VOT; nonetheless, descriptive work casts it as having generic long-lag aspiration similar to English (Matthews et al., 2013; Bauer and Benedict, 1997; Chan and Li, 2000; Mielke and Nielsen, 2018). For example, Matthews et al. (2013) describe initial stops in both English and Cantonese as voiceless and aspirated, even though they differ in their phonological features.

While the presence of articulatory reuse within Cantonese and across languages remains an empirical question, it aligns with the finding that bilingual Mandarin-English children did not distinguish between languages in VOT (Yang, 2019). Additionally, the predictions of the SLM-r (Flege and Bohn, 2021) suggest that long-lag items of minimally distinct VOT would assimilate or dissimilate but not be stable in such proximity. Thus, the present study asks: Do Cantonese-English bilinguals uniformly produce long-lag stops within and across each of their languages? Leveraging the methodology from Chodroff and colleagues (Chodroff

and Wilson, 2017, 2018; Chodroff and Baese-Berk, 2019) allows for a new perspective on the structure of variation and nature of mental representation in bilinguals' segment inventories. It also facilitates the study of phonetically similar speech sounds in ways that other paradigms do not. As may be clear from the framing of the introduction, the hypothesis was that bilinguals would indeed exhibit crosslinguistic uniformity and leverage articulatory reuse across Cantonese and English.

4.2 Methods

4.2.1 Corpus

This study uses the conversational interview recordings from the SpiCE corpus described in Chapter **??**. As a reminder, the corpus comprises recordings of 34 early Cantonese-English bilinguals in both languages. The analysis in this chapter builds on the force-aligned phone transcripts. Please refer to Chapter **??** for additional information about the talkers.

4.2.2 Segmentation & measurement

All instances of prevocalic word-initial /p t k/ were identified from the conversational interview portion of the SpiCE corpus' force-aligned Praat TextGrid transcripts. For English, only words with initial stress were included in the initial sample (Lisker and Abramson, 1967)—this means the extremely high-frequency English word "to" was excluded, as was the case in Chodroff and Wilson (2017).[1] Code-switches out of the interview's primary language were not aligned, and as a result, they do not appear in the phone tier of the TextGrids. This limitation of forced alignment means that Cantonese /p t k/ were only considered if they occurred in the predominantly Cantonese interviews, and likewise for English. The

[1] While "to" only has one syllable, the most commonly used pronunciation variant in the dictionary is unstressed.

initial total count of /p t k/ across talkers and languages included 10,428 tokens.

While forced alignment performed reasonably well, anecdotally speaking, VOT estimates were refined using AutoVOT (Keshet et al., 2014)—a command-line software tool that facilitates automated measurement of positive VOT. AutoVOT identifies the onset and offset of positive VOT within a specified window and with a minimum duration. Here, the minimum allowed VOT was set to 15 ms. This value was selected as the stops under consideration are all long-lag stops, and aspiration values under 15 ms are typical of short-lag stops (Lieberman and Blumstein, 1988). The window used with AutoVOT was defined as the force-aligned segment boundaries plus or minus 31 ms (as in Chodroff and Wilson, 2017). If stops were too close for a 31 ms buffer, the onset of the second stop's window was set as the offset of the preceding window, as TextGrids do not permit overlapping intervals and AutoVOT uses the full TextGrid.

After running AutoVOT, instances of /p t k/ were subjected to exclusionary criteria to catch errors and exclude tokens immediately after a code switch. Tokens were excluded if there was substantial enough misalignment such that the AutoVOT offset did not fall within the original force-aligned boundaries of the word ($n = 567$). Tokens were also excluded if the previous word was unknown (i.e., unintelligible or in a different language; $n = 263$), if VOT was equal to the minimum value of 15 ms ($n = 446$), or if tokens had a VOT more than 2.5 standard deviations above the grand mean (> 129.5 ms; $n = 191$).

Of the initial sample, 14.1% was excluded, resulting in 8,961 stop tokens, summarized in Table 4.1. Talkers had a median of 97 Cantonese stops (range: 54-194) and 150.5 English stops (range: 73-540). Cantonese stops were culled at a slightly higher rate—they represent 43% of the initial sample, but only 38% of the final, post-exclusions sample. As there were comparable amounts of recorded speech in each language, the higher number of English stops is likely due primarily to lexical distributional reasons. Additionally, English has a greater number of highly frequent /k/-initial word types, while Cantonese /p/ occurs in fewer, less frequent word types in the final sample ($n = 60$, max frequency of 97) than English ($n = 158$,

max frequency of 215).

Table 4.1: The number of stop tokens for each language and sound category.

Language	/p/	/t/	/k/
Cantonese	374	1373	1688
English	1035	1336	3155

4.3 Analysis & Results

The articulatory uniformity framework offers solid theoretical grounds for interpreting the structure of VOT variation within and across talkers. This analysis qualifies and quantifies that structure from a few different angles. Section 4.3.1 describes the ordinal relationship between each of the segments across talkers and languages. Section 4.3.2 reports on a series of pairwise correlations of talker means for each of the three segments in each language. Lastly, Section 4.3.3 comprises the results of a Bayesian mixed-effects model aimed at elucidating the role of language while accounting for variables known to impact VOT.

4.3.1 Ordinal relationships

Prior work with lab and read speech strongly suggests an expected ordinal relationship for VOT across places of articulation, in which /p/ is consistently shorter than /k/ and /t/ tends to fall in the middle. The argument for this widely attested pattern is based on vocal tract aerodynamics and articulatory constraints (Cho and Ladefoged, 1999). One of the major contributions of Chodroff and Wilson (2017) is that these relationships are tighter than would be expected from a purely ordinal perspective. While ordinal relationships are a starting place, they represent just one piece of the puzzle.

The results presented in this section suggest that *puzzle* is an appropriate characterization, as talkers largely did not adhere to the expected order. While there is

some reason to expect coronals not to pattern accordingly—particularly in English—the relationship between /p/ and /k/ is inconsistent across talkers. Table 4.2 reports the proportion of talkers whose mean VOT values followed the expected /p/ $<$ /t/ $<$ /k/ relationships. Prior work with connected speech reports rates of adherence in the 80-90% range, with the exception of /t/ $<$ /k/ being drastically lower for native English speakers (Chodroff and Baese-Berk, 2019). While the English /t/ $<$ /k/ comparison is remarkably low here at 6%, only English /p/ $<$ /t/ falls in the range that prior work suggests, at 82%. This lack of adherence is apparent in the relative ordering or markers in Figures 4.1 and 4.2, which depict the mean and standard error of VOT for each segment, language, and talker. The goal of Figures 4.1 and 4.2 is to showcase the variety of patterns across individuals and to highlight that a single summary plot of means only would be inappropriate. In many cases, the standard errors for the different segments in a given talker's panel overlap, as is the case for VM21B in Figure 4.2. Such overlap indicates that strict ordering may not be appropriate here. Additionally, talkers do not appear to be consistent across languages. For example, talker VF19B in Figure 4.1 exhibits a clear /p/ $<$ /t/ $<$ /k/ relationship in Cantonese, but a clear /p/ $<$ /k/ $<$ /t/ relationship in English.

Table 4.2: Proportion of talker means that adhered to expected ordinal relationship for VOT: /p/ $<$ /t/ $<$ /k/ VOT durations. Note that talker VM25A has no instances of Cantonese /p/ in the final sample.

Language	**/p/<t/**	**/t/<k/**	**/p/<k/**	**N**
Cantonese	0.24	0.61	0.39	33
English	0.82	0.06	0.47	34

4.3.2 Pairwise correlations

To examine the relationship between stops within and across languages, 15 pairwise Pearson's *r* correlations were calculated using talker means. Each correlation compares talkers means for two different segments. The full set of pairwise correlations includes three within English, three within Cantonese, and nine comparing

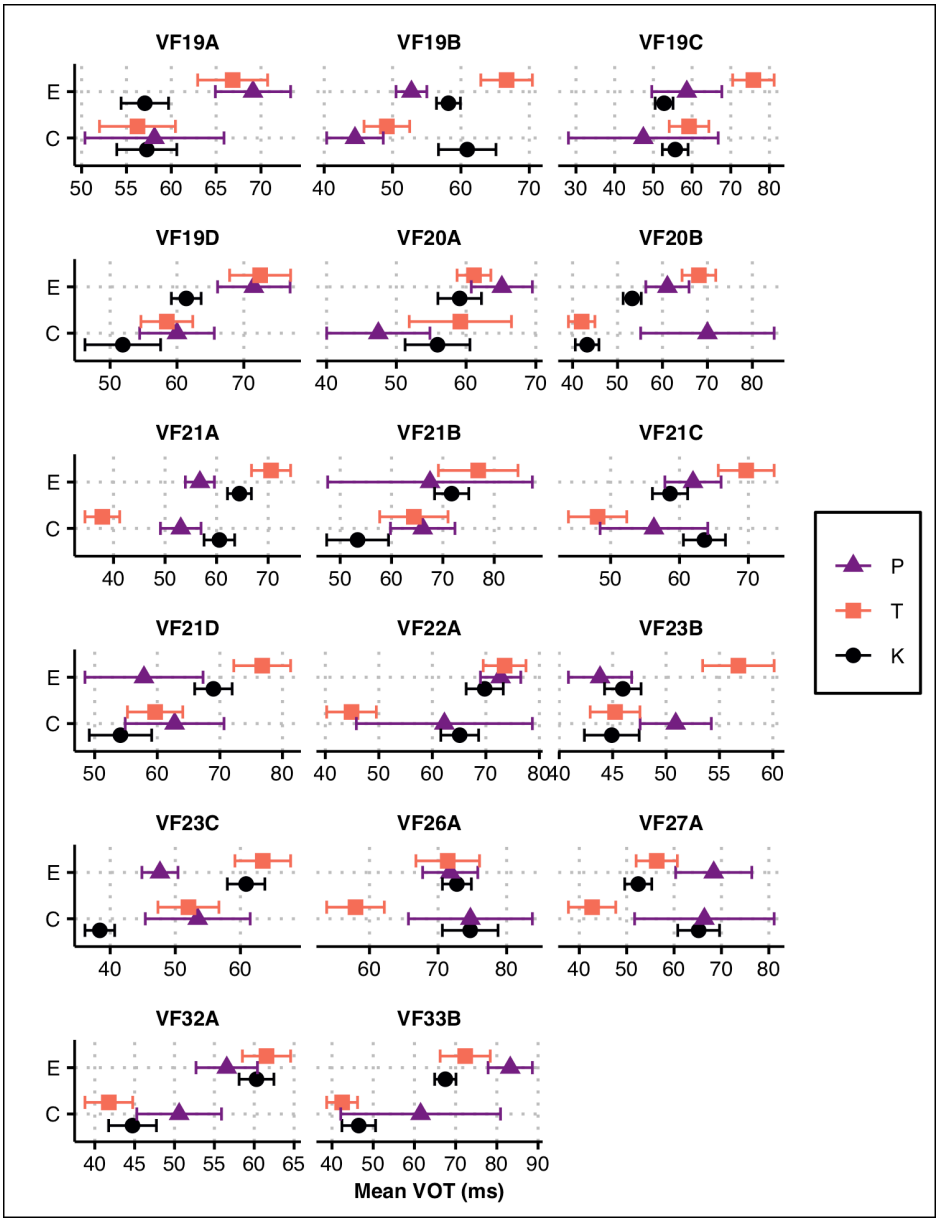

Figure 4.1: This figure depicts the ordinal relationships for the female talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

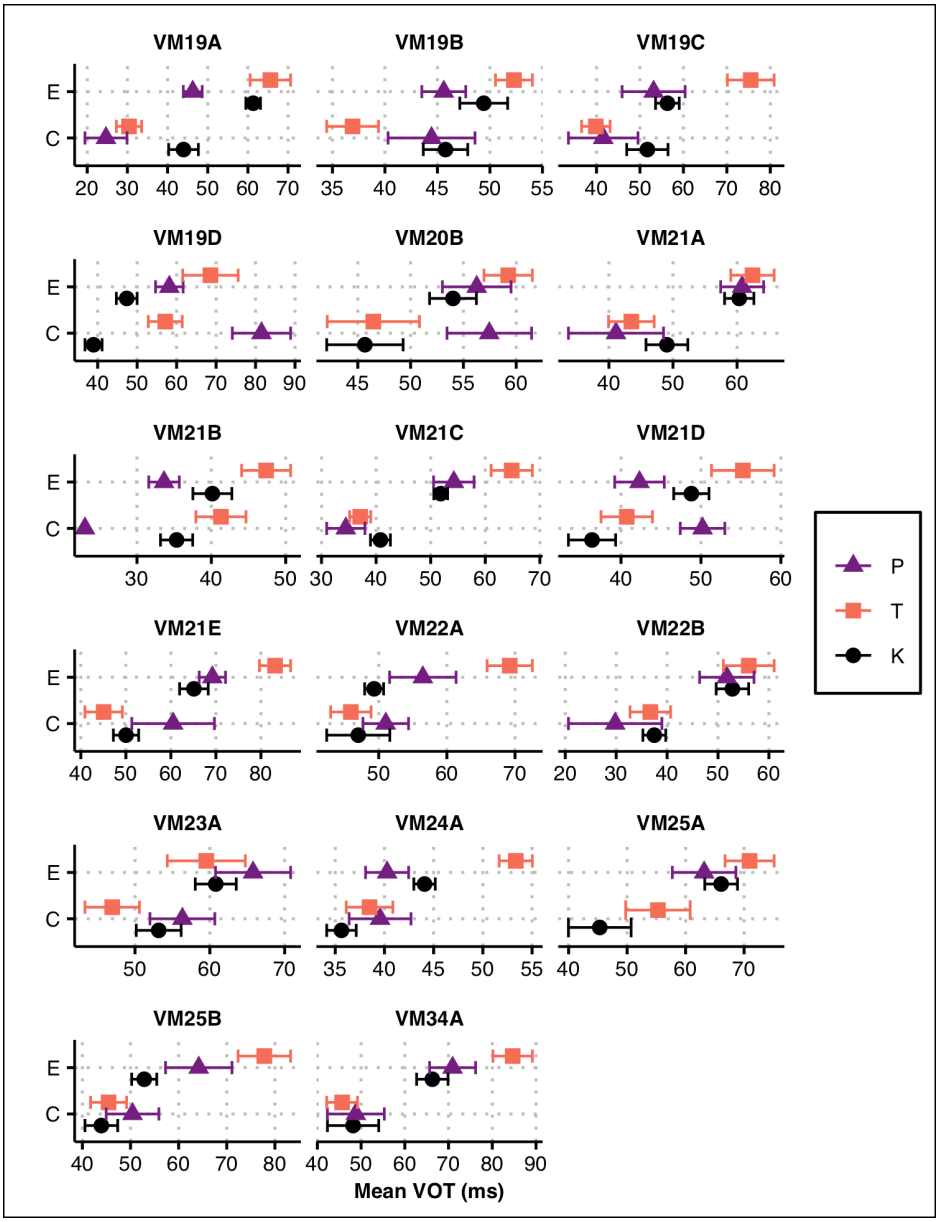

Figure 4.2: This figure depicts the ordinal relationships for the male talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

English to Cantonese. These correlations are reported along with Holm-adjusted *p*-values to account for multiple comparisons. This analysis uses the *psych* (Revelle, 2021) package in R (R Core Team, 2020). As in Chodroff and Wilson (2017), this correlation analysis aims to elucidate within-talker invariance and between-talker variability. Tight correlations between-talker means signals within-talker invariance, while a wide-spread between points signals between-talker variability. While using means ignores information about within-category variability—a major shortcoming of this approach—prior work sets up strong, clear expectations about how the mean values for long-lag VOT pattern (Chodroff and Wilson, 2017; Cho and Ladefoged, 1999). The mixed-effects analysis in the following section takes this variation into account.

Table 4.3 summarizes the output of all 15 correlations in text form. Figure 4.3 depicts the six within-language correlations and Figure 4.4 depicts the across language correlations. While there is some evidence for both within- and across-language structured variation, the correlations reported here are considerably lower than prior work on English read speech, where within-language comparisons had $r > 0.7$ (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019). With the exception of the English /p/ $\sim$ /k/ ($r = 0.70$, $p < 0.001$), all of the correlations here were either moderate ($0.5 < r < 0.7$; $p < 0.01$) or weak and non-significant. Within-English correlations were the most consistent—all three had *r* at or above 0.65 ($p < 0.001$). Of the within-Cantonese correlations only /p/ $\sim$ /t/ was significant ($r = 0.59$; $p = 0.003$), though the correlation for /p/ $\sim$ /k/ was marginal ($r = 0.44$; $p = 0.08$). This disparity across English and Cantonese for the same set of talkers highlights the need to study a variety of typologically distinct languages to understand how the structure of variation *varies*.

Two of three across-language correlations at the same place of articulation were significant, with moderate *r* values (/p/ $\sim$ /p/: $r = 0.62$, $p = 0.001$; /k/ $\sim$ /k/: $r = 0.57$, $p = 0.004$). Notably, the correlation for /t/ $\sim$ /t/ was not significant ($r = 0.40$, $p = 0.11$), which may reflect the variable behavior of English coronals more broadly, even if the particular sample here only includes cases where aspiration was produced. Of

the across-language comparisons that do not share a place of articulation, only one was significant—Cantonese /k/ $\sim$ English /p/ (r=0.58, p=0.003). Again, /t/ is absent here.

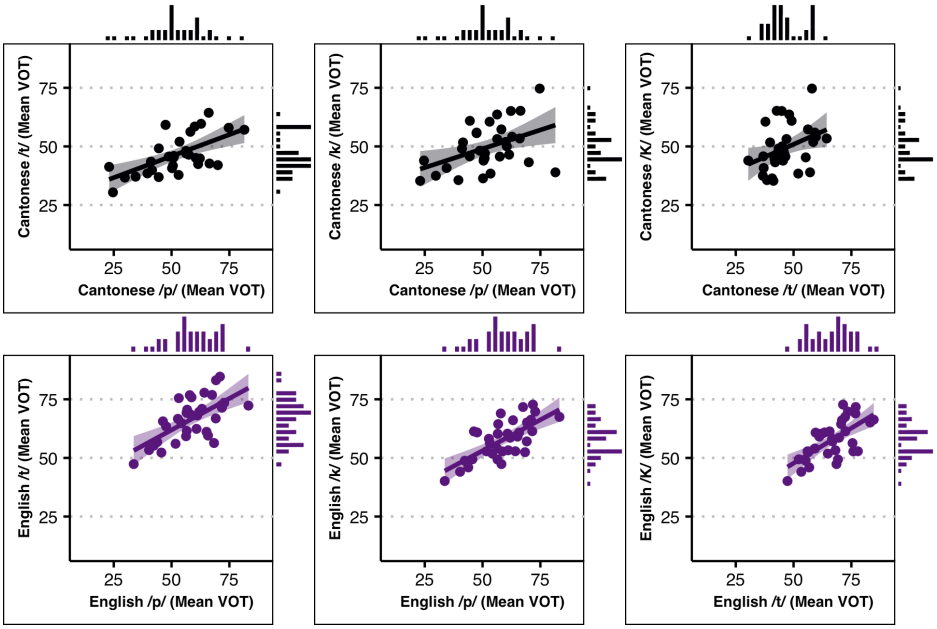

Figure 4.3: Correlations for within-language pairwise comparisons of raw mean VOT are depicted with points representing talker means for the segments on the x and y axes, and superimposed linear smooths. The margins display histograms for each of the axes. Within-Cantonese comparisons are depicted in black, and within English comparisons in purple.

Chodroff and Wilson (2017) also repeat the correlation analysis in a way the coarsely accounts for speaking rate. This consideration is important, as the local speaking rate is known to influence long-lag VOT in spontaneous speech (Stuart-Smith et al., 2015) and because prior work demonstrates both talker and language effects on speech rate (Bradlow et al., 2017). In comparing the two versions of the correlation analysis, Chodroff and Wilson found that "the magnitudes of the correlations among voiceless stops did not deviate from the original magnitudes,

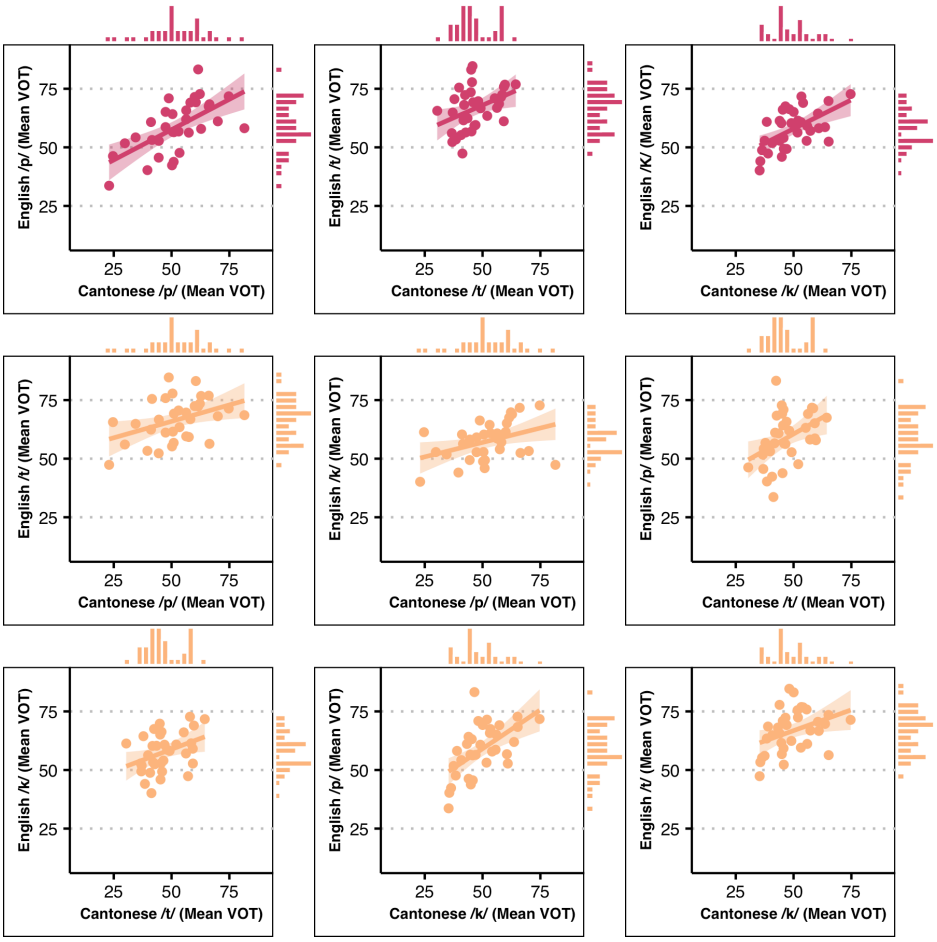

Figure 4.4: Correlations for the across-language comparisons of raw mean VOT are depcited in the same manner as Figure 4.3. Comparisons at the same place of articulation are depicted in pink, and comparisons at different places of articulation are in orange.

Table 4.3: All 15 correlations are based on raw mean VOT—and separately, residual VOT after accounting for speaking rate—for each talker, language, and segment. Each row indicates the comparison, Pearson's *r*, and the Holm-adjusted *p*-value given 15 comparisons.

		Raw		Residualized	
Type	**Comparison**	*r*	*p*	*r*	*p*
Within-Cantonese	Cantonese /p/ ∼ Cantonese /t/	0.59	0.003	0.59	0.003
Within-Cantonese	Cantonese /p/ ∼ Cantonese /k/	0.44	0.08	0.55	0.01
Within-Cantonese	Cantonese /t/ ∼ Cantonese /k/	0.38	0.11	0.34	0.21
Within-English	English /p/ ∼ English /t/	0.65	<0.001	0.63	0.001
Within-English	English /p/ ∼ English /k/	0.70	<0.001	0.70	<0.001
Within-English	English /t/ ∼ English /k/	0.66	<0.001	0.60	0.002
Across-language	Cantonese /p/ ∼ English /p/	0.62	0.001	0.57	0.01
Across-language	Cantonese /t/ ∼ English /t/	0.40	0.11	0.35	0.21
Across-language	Cantonese /k/ ∼ English /k/	0.57	0.004	0.54	0.01
Across-language	Cantonese /p/ ∼ English /t/	0.41	0.11	0.29	0.31
Across-language	Cantonese /p/ ∼ English /k/	0.40	0.11	0.29	0.31
Across-language	Cantonese /t/ ∼ English /p/	0.43	0.08	0.37	0.20
Across-language	Cantonese /t/ ∼ English /k/	0.37	0.11	0.27	0.31
Across-language	Cantonese /k/ ∼ English /p/	0.58	0.003	0.59	0.003
Across-language	Cantonese /k/ ∼ English /t/	0.38	0.11	0.37	0.20

demonstrating that differences among talkers in the realization of these sounds cannot be reduced to talker-specific speaking rates" (2017, p. 34).

I conducted a similar analysis here, using means calculated over *residual* VOT values from a simple linear regression in which VOT was predicted by average phone duration within the word. Average phone duration is a proxy for speech rate. It was calculated as the difference between the word's AutoVOT-estimated onset and force-aligned offset, divided by the number of segments in the canonical form of the word. The results—Pearson's *r* and Holm-adjusted *p* values—are reported in the rightmost columns of Table 4.3. Qualitatively, the results mostly mirror the correlations based on raw VOT, though there are some differences in significance and magnitude. This difference can likely be attributed to the generally weak correlations found.

While these relationships indicate some degree of articulatory reuse, the overall picture is far from compelling, particularly when considered alongside the results of the analysis of the ordinal relationships in Section 4.3.1. Compared to prior work, these correlations are less consistent and generally weaker.

The next steps in Chodroff and Wilson's (2017) methods focus on validating the strength of the correlations. Their approach includes estimating confidence intervals for the correlations using a bootstrap procedure. In a later paper, Chodroff et al. (2019) simulate what would emerge from a purely ordinal relationship between stops and demonstrate that the observed correlations are much stronger—ultimately arguing for a uniformity constraint on phonetic variation. Given that the correlations found in this chapter are drastically lower and largely do not adhere to the expected ordinal relationships, the remainder of this analysis takes a different approach.

4.3.3 Linear mixed-effects model

The analysis in the section leverages a Bayesian multilevel linear model to elucidate the sources of variation within and across talkers. As Bayesian modeling emphasizes the estimation of effect magnitudes, the model can be used to assess how talker's sound categories compare to one another while simultaneously accounting for factors known to influence long-lag VOT. This modeling approach is more in line with the generative modeling approach advocated for by Haines et al. (2020)—in a way that the correlation and ordinal relationships analyses in the preceding sections are not. Specifically, this approach retains the variation lost when working with means and also uses a response variable distribution that aligns with the constraints of the variable in question.

This section proceeds as follows. Section 4.3.3 describes Bayesian modeling and inference in broad terms and points the reader towards sources for further reading on the topic. Section 4.3.3 motivates and describes the structure of the model used in this chapter. Lastly, Section 4.3.3 reports the results of the model. All code used in this analysis is available at a GitHub repository that will be made

public later on.

Bayesian inference

The corpus sample was analyzed with a Bayesian multilevel generalized linear model using the brms package in R (Burkner, 2017; R Core Team, 2020). The brms package provides a simple, formula-based interface to Stan—a widely used probabilistic programming language for estimating Bayesian statistical models via Hamiltonian Monte Carlo and No-U-Turn Sampling (Stan Development Team, 2021). Bayesian models are desirable in the case of modeling multilingual VOT for both practical and theoretical reasons. Practically, they are not subject to the convergence problems that plague comparable frequentist models. Theoretically, they allow for graded statements regarding the strength of evidence for all parameters, both population-level (i.e., fixed effects) and group-level (i.e., random effects) parameters. While there are many other benefits, readers are referred to Vasishth et al.'s (2018) recent in-depth tutorial paper on Bayesian modeling in the phonetic sciences for further argumentation.

Inference in Bayesian models is based on the posterior distributions of parameters in the model, which reflect the range and probability of credible values for parameters. The posterior combines information from prior knowledge and the likelihood of observing the data given the specified model. While some Bayesian models use detailed and specific prior knowledge, it is perhaps more common to use weakly informative, regularizing priors (Gelman et al., 2017), which constrain the parameter space to possible values and down weight extreme or unlikely values. The model described in the next section uses regularizing priors.

While Bayesian modeling typically emphasizes parameter estimation in a probabilistic framework, there are decision criteria that facilitate hypothesis testing. One such technique is to use Kruschke's (2011) ROPE+HDI method. The ROPE is a "region of practical equivalence" surrounding the null value. The HDI is the highest density interval typically used to describe Bayesian posterior distributions. Kruschke's (2011) decision criterion is simple: if the HDI falls entirely within the

ROPE, then the null value can be accepted; if the HDI falls entirely outside the ROPE, then it can be rejected; if there is overlap, then a decision should be withheld. In the case of standardized data, Kruschke (2011) recommends the convention of setting a ROPE to be $[-0.1, 0.1]$—half the size of a small Cohen's *d* effect. While Bayesians tend to shy away from using decision criteria, it nonetheless provides a useful scaffolding for interpreting the magnitudes of standardized effects, when presented alongside the full posterior distributions.

Modeling multilingual VOT

VOT was modeled using a Bayesian multilevel linear mixed-effects model. The model used in this section is provided in 4.1. While the model is not the maximal model (Barr et al., 2013), it instead follows guidelines for parsimonious model building (Bates et al., 2018), in which the parameters of direct interest are included as random slopes, and the controlling parameters are not.

The model used in this section is provided in 4.1. While the model is not the maximal model (Barr et al., 2013), it instead follows guidelines for parsimonious model building (Bates et al., 2018), in which the parameters of direct interest are included as random slopes, and the controlling parameters are not.

$$\begin{aligned}\text{VOT} \sim 1 + \text{Place} \times \text{Language} + \text{Average Phone Duration} + \text{Pause} + \\ (1 + \text{Place} \times \text{Language} \mid \text{Talker}) + (1 \mid \text{Word})\end{aligned} \tag{4.1}$$

One of the main benefits of multilevel modeling is partial pooling, where information for different levels of a variable is shared across those levels. McElreath argues that "any batch of parameters with exchangeable index values can and probably should be pooled [where exchangeable] just means the index values have no

true ordering" (2020, p. 435). Pooling can be done for both intercepts and slopes, much in the way that frequentist models allow for slopes and intercepts to vary in the random effects structure. As apparent in the formula in 4.1, this model includes partial pooling for all of the parameters of interest. Details about how the parameters were specified are as follows.

VOT was the dependent variable—it was standardized in order to facilite the specification of priors and a ROPE.

Place encodes place of articulation for the stops and has three levels. Following Chodroff and Wilson (2017), Place was weighted effect coded in order to account for unequal sample sizes across the three levels and to faciliate the interpretation of the simple effects in light of the interaction term (Brehm and Alday, 2021). Coding was implemented using the *wec* package (Nieuwenhuis et al., 2017), and leads to reporting Place effects for T (weights: /p/ $= -1.92$, /t/ $= 1$, /k/ $= 0$) and K (weights: /p/ $= -3.44$, /t/ $= 0$, /k/ $= 1$).

Language is a binary variable that encodes whether the VOT measurement comes from an English or Cantonese word. As with Place, Language was also weighted effect coded (Cantonese $= -1.61$, English $= 1$).

Average Phone Duration represents the average duration of phones within the word. It was calculated as the difference between the word's AutoVOT-estimated onset and the force-aligned word offset, divided by the number of segments in the canonical form of the word. As noted in Section 4.3.2, average phone duration serves as a proxy for local speaking rate. A word-internal measure is desirable here, as many tokens were preceded by a pause and thus lack the necessary preceding context from which to calculate it. Average phone duration was standardized—that is, both centered and scaled.

(Preceding) Pause indicates whether or not the token occurred after a pause or not. Pauses were identified using the force-aligned transcripts and include

any instance where the preceding phone was "sil" (silence) or "sp" (silent/short pause?). The presence of a preceding pause was also implemented with weighted effect coding (weights: False $= -0.33$, True $= 1$).

Word indicates the word that the VOT measurement comes from.

The interaction for Language $\times$ Place was included in the model—it directly addresses the research question relating to whether or not bilinguals maintain a difference across languages for these sounds. Additionally, the model includes partial pooling (i.e., random intercepts) for Word and Talker, as well as for the Language, Place, and Language $\times$ Place terms (i.e., random slopes).

As noted above, the priors were set to be weakly informative and regularizing, motivated by the discussions in Gelman et al. (2017) and McElreath (2020). Specifically, the priors were set at follows:

Intercept Student's *t* distribution with $\nu = 3$, $\mu = 0$, and $\sigma = 2.5$

Population-level parameters Normal distribution with $\mu = 0$ and $\sigma = 1$

Group-level standard deviations Half Student's *t* distribution with $\nu = 3$, $\mu = 0$, and $\sigma = 2.5$

Group-level correlations LKJ distribution with $\eta = 2$

The model was fit using four chains with 5000 iterations (2500 warmup) for a total of 10,000 post-warmup samples. The chains were well-mixed, based on a visual inspection of trace plots, a lack of divergent transitions, and r-hat values below 1.05. Additionally, the effective sample size was sufficiently large for all parameters (for discussion, see Vasishth et al., 2018).

Results

A summary of the model's population-level parameters is provided numerically in Table 4.4 and visually in Figure 4.5. Including all of these is probably not necessary, thoughts? The population parameters indicate that VOT is modulated by

language, local speaking rate, and the presence of a preceding pause. Recall that the categorical population-level parameters were weighted effect coded—this facilitates the interpretation of simple effects in the presence of interaction terms.

The overall effect of Language indicating that English long-lag stops were produced with longer VOT than Cantonese ($\beta = 0.16$, 98.9% HDI outside ROPE). The effect of Language did not seem to be modulated by place of articulation, as the HDIs for Place (T), Place (K), and both Place $\times$ Language interaction terms overlapped substantially with the ROPE. At most, 11% of the posterior distribution fell outside the ROPE for these four parameters. Figure 4.6 shows the conditional effects for Place and Language—that is, the expected means for each of the six combinations. The effect of Language is readily apparent for /t/ and /k/.

The control parameters behaved as expected. VOT was longer when the local speaking rate was slower. This effect is captured by the relatively high posterior mean for Average Phone duration ($\beta = 0.32$, 100.0% HDI outside ROPE). VOT was also longer after a pause, though the effect size was considerably smaller than for speaking rate ($\beta = 0.12$, 94.0% HDI outside ROPE).

Table 4.4: Population parameter summary.

Parameter	**Est.**	**95% HDI**	**% Outside ROPE**
Intercept	0.18	[0.09, 0.28]	95.1
Place (T)	0.05	[-0.03, 0.13]	11.1
Place (K)	-0.02	[-0.07, 0.03]	0.1
Language (English)	0.16	[0.11, 0.21]	98.9
Average Phone Duration	0.32	[0.30, 0.34]	100.0
Preceding Pause (True)	0.12	[0.09, 0.16]	94.0
Place (T) $\times$ Language (English)	-0.01	[-0.07, 0.04]	0.2
Place (K) $\times$ Language (English)	0.04	[0.00, 0.08]	0.3

While the main takeaway from the population parameters is the difference in VOT across languages, the model also offers insight into the sources of variation in this population. A summary of the variability in the model's grouping parameters is provided numerically in Table 4.5 and visually in Figure 4.7. The largest source of variability in the model is in the Word intercepts ($\beta = 0.44$). The second-

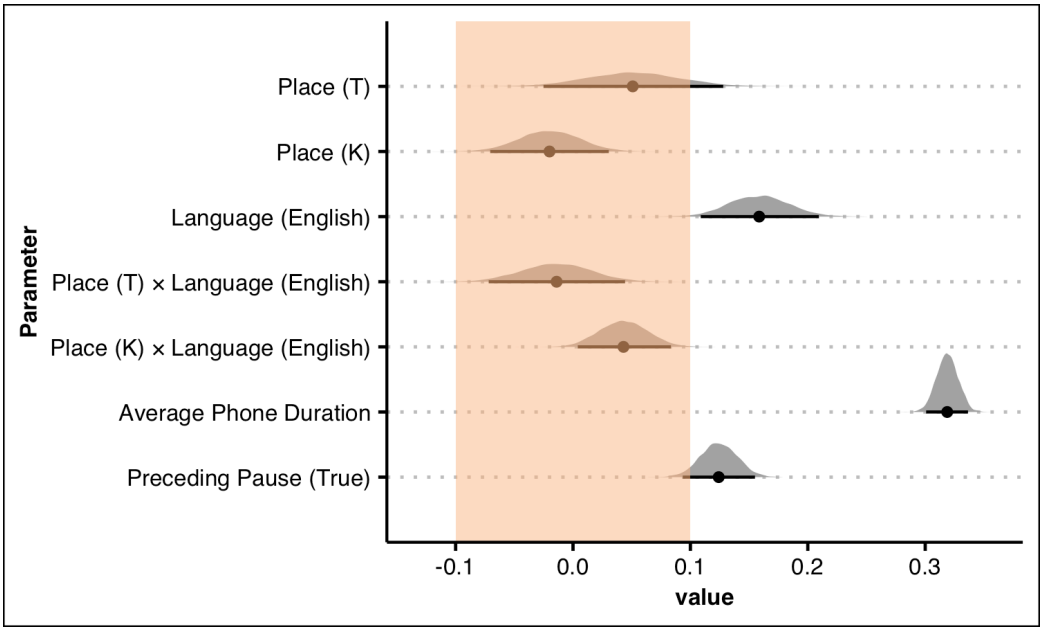

Figure 4.5: This figure depicts the posterior distributions for each of the population-level parameters, with the posterior mean indicated by the dot. The orange shaded section represents the ROPE.

largest source of variability is in the Talker intercepts ($\beta = 0.25$). While there is variability across talkers in the random slopes, few are meaningfully different from the corresponding population-level parameters—this is evident in Figure 4.8, which depicts the by-Talker intercept and slope deviations from the model. A sizable plurality of talker intercept posterior distributions falls outside of the ROPE, while the vast majority of the by-talker slopes overlap substantially or fall entirely within the ROPE. In line with Chodroff and Wilson (2017), this result highlights between-talker variability and within-talker stability.

4.4 Discussion

This chapter reports on a study of long-lag stops in Cantonese-English bilingual speech from the SpiCE corpus described in Chapter **??**. It leverages the uniformity

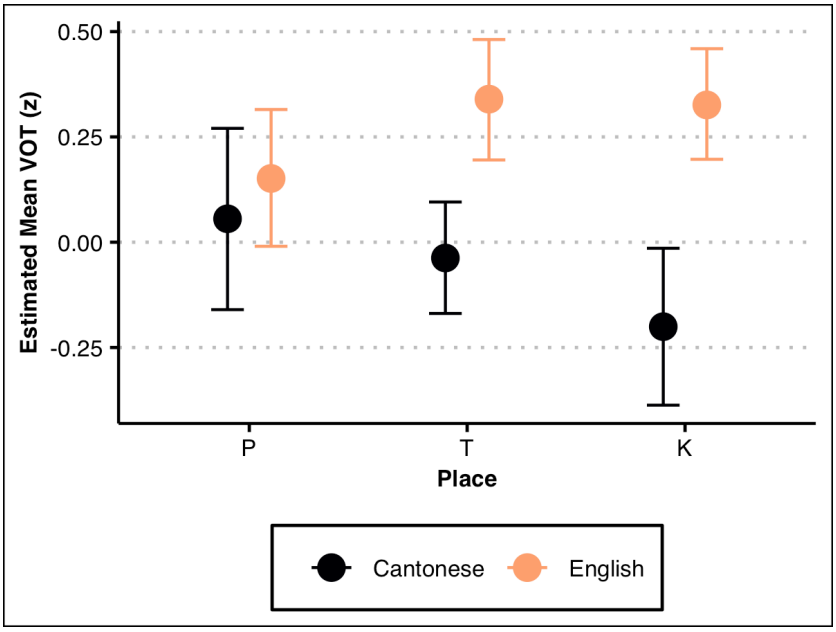

Figure 4.6: This figure depicts the model estimated mean for each of the places of articulation by language, using the fitted method of brms conditional effects function. (This caption needs revising, but this comment for keeping track.)

Table 4.5: Group parameter variability summary.

Group	Parameter S.D.	Est.	95% HDI
Word	Intercept	0.44	[0.40, 0.50]
Talker	Intercept	0.25	[0.19, 0.32]
Talker	Place (T)	0.09	[0.05, 0.14]
Talker	Place (K)	0.06	[0.04, 0.09]
Talker	Language (English)	0.08	[0.05, 0.11]
Talker	Place (T) × Language (English)	0.05	[0.02, 0.08]
Talker	Place (K) × Language (English)	0.04	[0.02, 0.06]

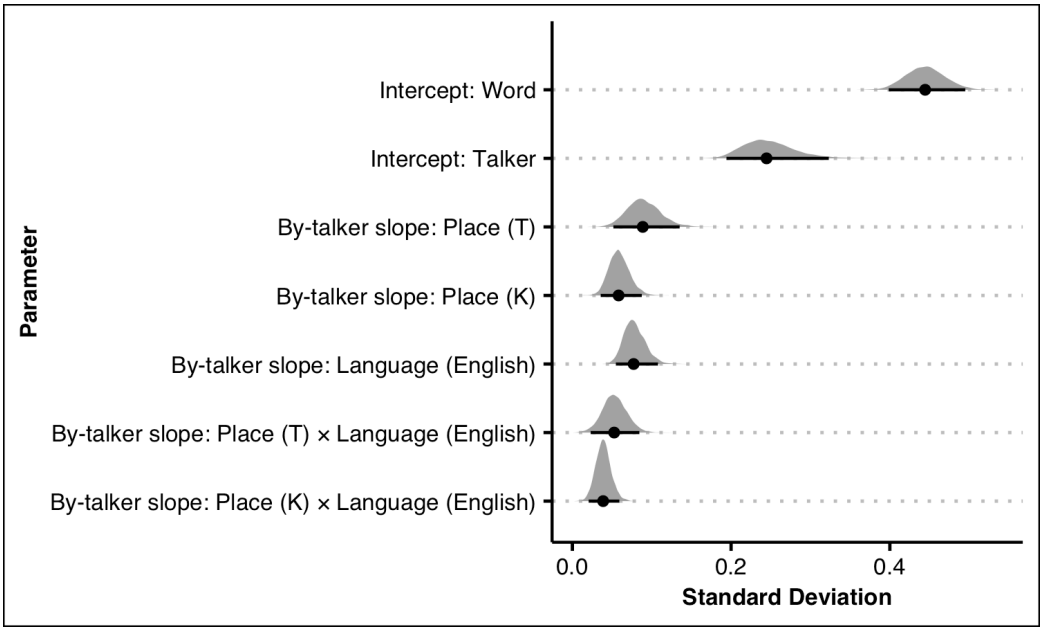

Figure 4.7: This figure depicts the posterior distributions for the standard deviation of each of the grouping parameters, both intercepts and slopes.

framework to assess VOT similarity within and across languages from a few different angles. In broad strokes, the evidence for uniformity both within and across languages was somewhat mixed.

An analysis of ordinal relationships between the duration of mean VOT for talkers in each language was inconclusive. Talkers largely did not adhere to the expected order, and further, talkers were not internally consistent across languages. This counters prior work establishing strong rates of adherence. However, the difference may be due entirely to speaking style. Most of the work documenting ordinal relationships among mean stop VOT is based on isolated word production and read speech (e.g., Chodroff and Wilson, 2017; Cho and Ladefoged, 1999; Lisker and Abramson, 1964). Conversely, the SpiCE corpus interviews comprise casual conversational speech. In any case, the analysis of ordinal relationship offers an incredibly coarse picture of how phonetic variation is structured.

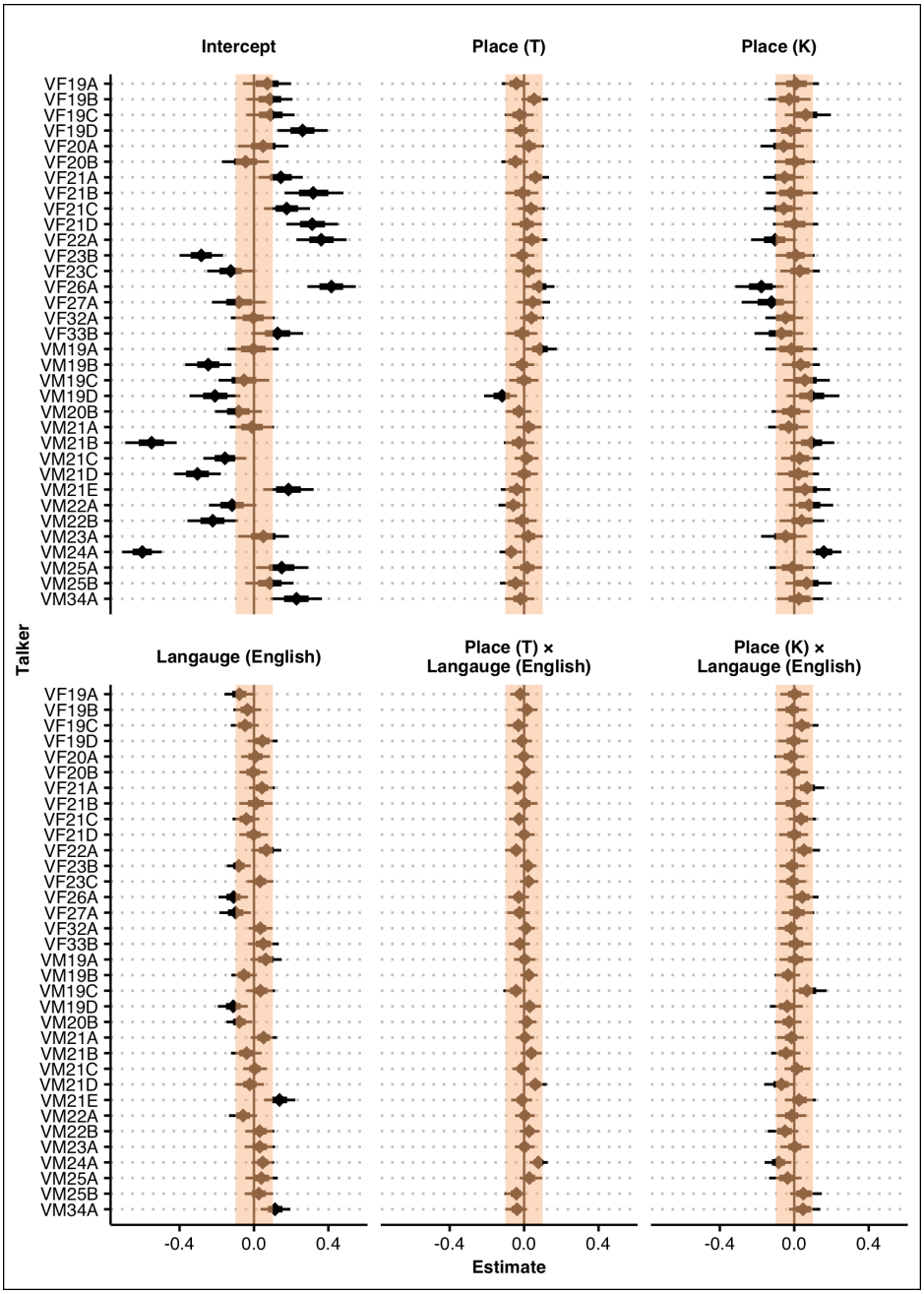

Figure 4.8: This figure depicts the 95% HDI for each talker across the talker intercepts and by-talker slope terms. The shaded orange interval represents the ROPE.

The correlation analysis offers a slightly more nuanced take on structure, and in doing so, provides evidence for within-language and, to a lesser extent, across-language uniformity. Across the board, the correlation magnitudes were weak or moderate, which differs from the strong and clear within-English patterns observed in prior studies (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019). The within-English comparisons were consistently moderate and significant, which replicates prior work. The within-Cantonese and across-language comparisons do not offer nearly as clear a picture. While there is some evidence of structure, particularly for /p/ and /k/ across languages, the evidence for tightly structure variation is far from compelling. While the murky outcome of the ordinal relationship and correlation analysis was largely unexpected, observing a different outcome for a different speech style is not without precedent. For example, the correlation magnitudes that Chodroff and Wilson (2017) found for connected, read speech were not as strong as those for isolated word production. It follows that an even less formal connected speech style would lead to still weaker relationships. Attending to speaking style is likely one of the main factors accounting for why lab and corpus results often differ (Gahl et al., 2012; Chodroff and Wilson, 2017); and, similarly, for corpus studies of monolingual and bilingual speech Johnson (2019).

The Bayesian mixed-effects model offers some insight into the primary sources of variation, including the specific role that language and place of articulation play in accounting for VOT variation. The model resulted in a clear difference in VOT by language, with English VOT being consistently longer for /t/ and /k/. While the introduction to this chapter reported on prior work indicating Cantonese might be longer (Clumeck et al., 1981; Lisker and Abramson, 1964), those numbers were not necessarily appropriate for the speaking style of the SpiCE corpus, particularly given the differences in English VOT across styles (Stuart-Smith et al., 2015). Interestingly, while the model shows a consistent difference across languages for VOT, very few talkers deviate from this overall pattern in a meaningful way. A much greater amount of variation is captured by variable intercepts for word and talker. The model's outcome, then, supports the argument for

uniformity—between-talker differences vary drastically, while talkers tend to be more internally consistent.

In light of the evidence for uniformity—albeit not particularly strong evidence—the small but consistent difference across languages is worth some attention. While Section 4.3.3 models standardized VOT, back-transforming the value into the original units suggests a difference across languages of approximately 4 ms. A difference of this magnitude is not likely to be perceptible in categorization (and similar) paradigms; work by McMurray et al. (2002) demonstrates a gradient and fine-grained effect of processing VOT in increments as small as 5 ms. Further, as research on mergers in sound change has demonstrated, individuals do not always perceive differences that they produce (Citation idea for this Molly??). As such, perception may not be the best indicator of whether or not this size difference is meaningful in practice. If this difference is indeed meaningful and bears out in future work, it carries implications for how similar sound categories are represented and discussed in the literature. If talkers can maintain such small distinctions across languages, it would reiterate the rarity of assimilation for early bilinguals and necessitate a broader version of models like SLM-r, that account for a wider variety of multilingual backgrounds.

Another possibility is that the underlying laryngeal gesture is "the same" but subject to global language timing factors. That is, talker-internal and language-internal factors both influence how VOT manifests. Bradlow et al.'s (2017) offers an example of this dual influence for speech rate multiple languages for learners, using native and non-native speech from ALLSSTAR corpus. In this study, Bradlow et al. demonstrates that talkers who speak faster in their L1 tend to also speak similarly fast in their L2. As this study examined a wide variety of L1s (the L2 was always English), Bradlow et al. also demonstrates differences across languages, with some L1s tending slower and others faster. This interpretation could be applied fairly transparently to the study in this chapter.

The results presented in this chapter provide limited support for a crosslinguistic uniformity constraint. Nonetheless, they offer insight into what makes bilingual

speech unique by providing an empirical description of bilingual long-lag stops. The weaker constraint on variability compared to prior work has implications for representation—as noted above—and processing. Tracking a uniformity-like pattern has been proposed as a mechanism for rapidly adapting to speech across languages (Reinisch et al., 2013), and in multilingual talker identification Orena et al. (2019). If the results of this study stand, then such a perceptual strategy may have limited use in real communicative contexts, whether or not listeners use it in a lab setting. Overall, this study highlights the need to study spontaneous speech and demonstrates the utility—and some limitations—of the uniformity framework for better understanding crosslinguistic similarity.

Bibliography

Amengual, M. (2018). Asymmetrical interlingual influence in the production of Spanish and English laterals as a result of competing activation in bilingual language processing. *Journal of Phonetics*, 69:12–28. → page 1

Antoniou, M., Best, C. T., Tyler, M. D., and Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, 38(4):640–653. → page 6

Antoniou, M., Best, C. T., Tyler, M. D., and 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. → page 6

Balukas, C. and Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, 19(4):423–443. → page 6

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. → page 25

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2018). Parsimonious mixed models. *arXiv:1506.04967 [stat]*, pages 1–21. ArXiv preprint: 1506.04967. → page 25

Bauer, R. S. and Benedict, P. K. (1997). *Modern Cantonese Phonology*. De Gruyter Mouton, Berlin. → page 12

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., and Mok, K. (2011). Language- and talker-dependent variation in global features of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 356–359, Hong Kong. → page 11

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts l2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899. → pages 20, 34

Brehm, L. and Alday, P. M. (2021). A decade of mixed models: It's past time to set your contrasts. Technical report, Open Science Framework. → page 26

Brown, E. L. and Amengual, M. (2015). Fine-grained and probabilistic cross-linguistic influence in the pronunciation of cognates: Evidence from corpus-based spontaneous conversation and experimentally elicited data. *Studies in Hispanic and Lusophone Linguistics*, 8(1):59–83. → page 6

Bullock, B. E. and Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, L., Winford, D., and deBot, K., editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam. → pages 5, 6, 7

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28. → page 24

Casillas, J. V. (2021). Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages*, 6(1):9. → pages 4, 5, 7

Chan, A. Y. W. and Li, D. C. S. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' english pronunciation problems. *Language, Culture and Curriculum*, 13(1):67–85. → page 12

Chang, C. B. (2015). Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity. In Raimy, E. and Cairns, C. E., editors, *The Segment in Phonetics and Phonology*, pages 199–217. John Wiley & Sons, Inc., Chichester, UK, 1 edition. → page 3

Cho, T. and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 27(2):207–229. → pages 15, 19, 31

Chodroff, E. (2017). *Structured variation in obstruent production and perception*. PhD dissertation, Johns Hopkins University, Baltimore, MD. → page 11

Chodroff, E. and Baese-Berk, M. (2019). Constraints on variability in the voice onset time of L2 English stop consonants. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 661–665, Melbourne, Austrailia. Australasian Speech Science and Technology Association Inc. → pages 11, 13, 16, 19, 33

Chodroff, E., Golden, A., and Wilson, C. (2019). Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, 145(1):EL109–EL115. → page 23

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47. → pages 10, 11, 12, 13, 14, 15, 19, 20, 22, 23, 26, 29, 31, 33

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4(s2). → page 13

Clumeck, H., Barton, D., Macken, M. A., and Huntington, D. A. (1981). The aspiration contrast in Cantonese word-initial stops: Data from children and adults. *Journal of Chinese Linguistics*, 9(2):210–225. → pages 12, 33

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. → pages 10, 11, 12

Flege, J. E. and Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In Wayland, R., editor, *Second Language Speech Learning: Theoretical and Empirical Progress*, pages 3–83. Cambridge University Press. → pages 1, 2, 3, 4, 5, 9, 12

Fricke, M., Kroll, J. F., and Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, 89:110–137. → pages 5, 6, 7

Fricke, M., Zirnstein, M., Navarro-Torres, C., and Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, 22(1):200–207. → page 10

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806. → page 33

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555. → pages 24, 27

Goldrick, M., Runnqvist, E., and Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, 25(4):1031–1036. → pages 5, 6, 7

Grosjean, F. (2011). An attempt to isolate, and then differentiate, transfer and interference. *International Journal of Bilingualism*, 16(1):11–21. → pages 6, 8, 9

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, 60(2):98–128. → page 4

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., and Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. Technical report, PsyArXiv. → page 23

Johnson, K. A. (2019). Probabilistic reduction in Spanish-English bilingual speech. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1263–1267, Melbourne, Australia. Australasian Speech Science and Technology Association Inc. → page 33

Johnson, K. A. and Babel, M. (2021). Language contact within the speaker: Phonetic variation and crosslinguistic influence. Technical report, OSF Preprints. → page 7

Keshet, J., Sonderegger, M., and Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction (0.91) [Computer Software]. → page 14

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3):299–312. → pages 24, 25

Lein, T., Kupisch, T., and van de Weijer, J. (2016). Voice onset time and global foreign accent in German–French simultaneous bilinguals during adulthood. *International Journal of Bilingualism*, 20(6):732–749. → page 4

Lieberman, P. and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge. → page 14

Lindblom, B. and Maddieson, I. (1988). Phonetic universals in consonant systems. In Hyman, L. M. and Li, C. N., editors, *Language, speech, and mind: studies in honour of Victoria A. Fromkin*, pages 62–78. Routledge, London. → page 4

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 12, 31, 33

Lisker, L. and Abramson, A. S. (1967). Some effects of context on voice onset time in english stops. *Language and Speech*, 10(1):1–28. → page 13

Llompart, M. and Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in German vowels. *Journal of Phonetics*, 67. → page 1

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → page 12

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, 2 edition. → pages 25, 26, 27

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2):B33–B42. → page 34

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, 122(2):145–163. → page 3

Mielke, J. and Nielsen, K. (2018). Voice onset time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America*, 144(4):2166–2177. → page 12

Ménard, L., Schwartz, J.-L., and Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50(1):14–28. → pages 10, 11

Nieuwenhuis, R., Grotenhuis, M. t., and Pelzer, B. (2017). Weighted Effect Coding for Observational Data with wec. *The R Journal*, 9(1):477–485. → page 26

Olson, D. J. (2016). The role of code-switching and language context in bilingual phonetic transfer. *International Phonetic Association. Journal of the International Phonetic Association; Cambridge*, 46(3):263–285. → pages 5, 6, 7

Orena, A. J., Polka, L., and Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. *The Journal of the Acoustical Society of America*, 145(4):EL303–EL309. → page 35

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → pages 19, 24

Reinisch, E., Weber, A., and Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1):75–86. → page 35

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.1.3. → page 19

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070. → page 10

Sancier, M. L. and Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25(4):421–436. → page 6

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. → page 2

Simonet, M. and Amengual, M. (2019). Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, 24(2):208–221. → pages 6, 7, 9

Stan Development Team (2021). *Stan Modeling Language Users Guide and Reference Manual*. Version. → page 24

Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6(3-4):505–549. → pages 12, 20, 33

Sundara, M., Polka, L., and Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, 9(1):97–114. → pages 4, 5, 6, 7

Tsui, R. K.-Y., Tong, X., and Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, 40(1):29–58. → page 8

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., and Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71:147–161. → pages 24, 27

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820. → pages 8, 12