# Chapter 4

# The Structure of Voice Onset Time Variation in Bilingual Long-lag Stops

## 4.1 Introduction

One of the primary goals in this chapter is to investigate what languages can share in the mental representation of similar speech sound categories.[1] The idea of representation is intended here in the manner typically meant by psycholinguists (e.g., Llompart & Reinisch, 2018), exemplar theory proponents (e.g., Amengual, 2018), and Flege & Bohn (2021) in their revised Speech Learning Model (SLM-r). These groups use similar language to describe representation, emphasizing distributions of sensory experiences rather than theoretical linguistic descriptions. For example, Flege & 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 conversations" (2021, p. 32). The SLM-r also posits that the speech sounds

[1] Note that "similar" is an often ill-defined concept that will be grappled with in Section 4.1.1.

of a bilingual's languages exist in a shared phonetic space, regardless of what they share in their representations. In starting with the SLM-r—where distributions of exemplars for different categories cohabit the same phonetic space—this chapter addresses the extent to which languages share representation(s). Much like Chapter 3, the approach here is one of leveraging the structure of variation to understand the system. Additionally, as with the preceding chapters, this chapter focuses on the speech of early bilinguals, as discussed in Section 1.1. This is an important reminder, as the conceptual model referenced throughout this chapter—Flege & Bohn's (2021) SLM-r—focuses more on the speech of second language (L2) learners and late bilinguals, as opposed to the early bilinguals in this dissertation.

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 for bilinguals in the first place and addresses key terminology, like "similarity." Section 4.1.2 builds on this by briefly summarizing the relevant crosslinguistic influence literature, addressing assimilation, dissimilation, and how they reflect on the idea of shared representation for sound categories. 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. After Section 4.1.4, the chapter moves onto methods and results in Sections 4.2 and 4.3.

### 4.1.1 Identifying "links" across bilinguals' 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 describes "links or connections of one sort or another

between the phonetic categories" (2016, 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. This definition is somewhat problematic, as it would fail to identify (real) links in cases where influence is less likely to occur (cf. Grosjean, 2011). Links behave dynamically, 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 (this body of literature will be reviewed in Section 4.1.2). While there may be alternative explanations (i.e., global influence), the concept of links is widely assumed in accounting for bilinguals' behavior.

Flege & 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 & Bohn (2021) rely on perceptual metrics. This is not to say that either kind of metric is inherently better or worse, but rather, that perception- and production-based metrics offer different types of insight into the linguistic system. The reason for the challenge of pairing up sound categories is that perception and production do not always line up neatly. Flege & Bohn assert that similarity "must be assessed perceptually rather than acoustically because acoustic measures sometimes diverge from what listeners perceive" (2021, p. 33). The disconnect between the two processes arises from both linguistic and physiological bases.

In an overview chapter on similarity in crosslinguistic influence, Chang (2015) suggests an alternative—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 phone-

mic 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). While "distributional facts" seems to be intended in a broad sense, the example Chang gives is co-occurrence restrictions. 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 & 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 in the mind 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 particular 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 & Bohn, 2021). Assimilation is a merging of phonetic properties and arguably occurs when bilinguals and learners do not "discern a phonetic difference (or differences) between the realizations" of sound categories (Flege & Bohn, 2021, p. 40). It is not entirely clear at what level discernment between sound categories occurs, though it is most likely intended to be a linguistic level rather than a purely auditory one. Dissimilation, then, is the reverse—a diverging of phonetic properties that occurs when a difference is

detected but is too small to maintain. These two processes are defined in rather absolute terms, which are tempered to some extent in the following paragraphs' continued discussion of the SLM-r.

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 produced differences across languages for voice onset time (VOT) and the standard deviation of burst frequency. These bilinguals did not differentiate based on other spectral moments of the burst that monolingual comparison populations did, and as a result, Sundara et al. (2006) illustrate divergence on some properties and convergence on others.

The motivation for the outcomes of assimilation and dissimilation arises from two simple constraints from the production and perception systems—effectively, do not get too close to each other in perception, and do not get too complicated in production (Guion, 2003; Lindblom & Maddieson, 1988; Flege & 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. Bilinguals can, after all, perceive subtle acoustic differences between similar sound categories (Ju & Luce, 2004). In SLM-r, the potential outcomes of instability are assimilation and dissimilation. Considering that bilinguals are fully capable of distinguishing highly 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. Yet, given SLM-r's primary focus on L2 learning, it is perhaps not a surprising approach. It may thus be more appropriate in the case of early bilinguals to also consider contrast maintenance alongside dissimilation. That is, early bilinguals may not need to assimilate or dissimilate a pair of similar sounds but rather simply maintain the subtle contrast as is.

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 &

Bohn term *composite categories*. Suppose sound categories from two languages are phonetically 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 & 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, the 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 outside of late bilingual and L2 speech. 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., 2016b; Goldrick et al., 2014; Bullock & Toribio, 2009; Olson, 2016). Casillas (2021) found that early bilinguals did not produce "compromise" stop categories. That is, early Spanish-English bilinguals did not, on the whole, produce VOT that was somehow intermediate to the canonical productions by monolinguals of either language. Instead, crosslinguistic influence can be fully attributed to what Casillas describes as "performance category mismatches that result from dynamic phonetic interactions associated with language activation" (2021, p. 16). In this sense, 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 at play. This finding echoes arguments made by Bullock & Toribio (2009) on the sophistication and control that bilinguals exert over their range of 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 for two sound categories to 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—each 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 & Amengual, 2019; Sancier & Fowler, 1997), or compare groups of individuals in different language mode conditions (Antoniou et al., 2010). At a more local level, some experimental designs leverage language switching across blocks (Sundara et al., 2006), across trials (Goldrick et al., 2014), or within trials (i.e., prompted code-switching; Bullock & 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., 2016b; Balukas & Koops, 2015) and cognate production (Brown & 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 & 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. This kind of linguistic behavior, arguably, reflects deep metalinguistic knowledge on behalf of bilinguals like those studied in Bullock & Toribio (2009).

In these approaches, the ability to examine crosslinguistic influence for any given pair of sounds hinges on the presence of an observable acoustic difference under some set of conditions. Arguably, for this reason, most prior work in crosslinguistic 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 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., 2016b; Goldrick et al., 2014; Bullock & 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 & 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. (2016b) 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., 2016b). An experimental example comes from Simonet & 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 degrees of 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 focusing on both is rare, it is not entirely absent from the literature. 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 by Tsui et al. (2019) compares English long-lag and Cantonese short-lag stop production in bilinguals. The study used a picture naming task in the context of a language switching design, where participants named pictures in both languages. The crucial comparison in the study is between trials that occurred immediately after a trial of the same language or the other language (i.e., whether there was a switch). While the design closely mirrored Goldrick et al. (2014), the results were murky—balanced bilinguals showed no evidence of crosslinguistic influence. The decision to use English long-lag and Cantonese short-lag stops as the stimuli could explain this outcome. This comparison reflects the need for stimuli to be acoustically distinct beforehand—as noted above—yet, 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. The long-lag 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 suggests that (Tsui et al., 2019) would have gotten more insightful results by comparing Cantonese *long-lag* stops to English long-lag stops. More importantly, however, it highlights the design constraints of the paradigms used in crosslinguistic influence research. Such methods are better suited for detecting links between acoustically distinct 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 mental representations—it is ephemeral, occurring "online." Transfer, on the other hand, occurs on a longer time scale and affects the representations themselves. While Grosjean (2011) argues that disentangling the two types of influence is difficult, the methods described above seem tailored more towards interference, given the way that they promote activation of both languages.

A good example of interference comes from Catalan-Spanish bilinguals' vowel production. Simonet & 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. However, 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 & 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 & Bohn, 2021), even though more of the literature reviewed in support of the model uses designs that center inter-

ference. 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 during the process of producing speech. 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. 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 only surface forms.

### 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. Recall that mental representations are defined in psycholinguistic terms, as outlined at the beginning of this chapter. Unlike prior work focusing on variable convergence and divergence, this chapter addresses whether a single category is used in both languages or whether each lan-

guage 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 & 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 & 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 may be valid and even related to one another. Given the focus of this chapter on phonetic and psycholinguistic accounts of category formation and representation (rather than phonological), the articulatory account—with its accompanying acoustic consequences—is likely more appropriate, even if this dissertation does not directly engage with articulatory phonetics.

In this light, if a set of segments share an attribute (i.e., share a description such as "long-lag" or belong to the same natural class), 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 & Wilson, in press), and stop consonant VOT (Chodroff & 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—using different gestures—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 & 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 & 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 & Baese-Berk (2019) found a greater degree of between-talker variability with non-native speakers compared to the prior monolingual work (Chodroff & Wilson, 2017), the within-talker structure was robust. However, the uniformity framework has not yet been extended to early bilingual speech to compare how bilinguals produce phonetically similar sounds in each language. 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.

There is also motivation for uniformity in perception. As outlined in Section 1.2, Orena et al. (2019) speculated that a bilingual advantage at generalizing across languages in talker identification might derive from sensitivity to crosslinguistic structural similarity. While this remains speculation on behalf of crosslinguistic generalization, there is evidence that within-language uniformity facilitates talker identification, above and beyond typical talker-indexical components of a voice (Ganugapati & Theodore, 2019). It follows that this boon would also extend to bilingual talker identification, provided that phonetic variation across languages also exhibits uniform structure.

### 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).[2] Consider the initial stop [$k^h$]—in citation speech—with a mean VOT of 80 ms in American English (Lisker & 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 long-lag 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 & Benedict, 1997; Chan & Li, 2000; Mielke & 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—English is typically analyzed with a ±voicing distinction and Cantonese with an ±aspiration distinction (Matthews et al., 2013).

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 & 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 & Wilson, 2017, 2018; Chodroff & Baese-Berk, 2019) allows for a new perspec-

[2]Please refer to Tables 3.1 and 3.2 in the previous chapter for a summary of the segmental inventories in both languages.

tive 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. The hypothesis in this chapter 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 2. 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 2 for additional information about the talkers.

### 4.2.2 Segmentation and 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. The identification of tokens was based on the phone tier of the transcripts, and as a result, derives from the forced aligner's identification of stops given a lexical item and its entry in the pronunciation dictionaries. For English, only words with initial stress were included in the initial sample (Lisker & Abramson, 1967).[3] 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 initial total count of /p t k/ across talkers and languages included 10,428 tokens.

[3] Chodroff & Wilson (2017) specifically excludes the extremely high-frequency English word "to." The initial stress requirement implicitly accomplishes this here—while "to" only has one syllable, the most commonly used pronunciation variant in the dictionary is unstressed.

While forced alignment performed reasonably well, anecdotally speaking, it was not perfect. Additionally, forced alignment includes both the closure and release of the stop in the marking of stop consonants. For these reasons, 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 & Blumstein, 1988). The window used with AutoVOT was defined as the force-aligned segment boundaries plus or minus 31 ms (as recommended by Chodroff & 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. This would occur, for example, in cases where a short vowel separates two stops, as may be the case in a phrase like "too tall" in running speech.

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), as in Chodroff & Wilson (2017).

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 in both the initial and final sample is likely due primarily to lexical distributional reasons. In addition to reporting on token frequency, Table 4.1 also summarizes the number of word types for each of the segments in each language.

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 token frequency of 97) than English (n = 158, max token frequency of 215).

**Table 4.1:** The number of stop tokens (overall and range across talkers) and word types for each language and sound category.

| **Language** | **Frequency** | **/p/** | **/t/** | **/k/** |
|---|---|---|---|---|
| Cantonese | Token (overall) | 374 | 1373 | 1688 |
| | Token (range) | 0–32 | 17-79 | 19–116 |
| | Type (overall) | 60 | 157 | 68 |
| English | Token (overall) | 1035 | 1336 | 3155 |
| | Range (tokens) | 4–96 | 15–150 | 52–294 |
| | Type (overall) | 158 | 143 | 208 |

## 4.3 Analysis and results

The articulatory uniformity framework offers solid theoretical grounds for interpreting the structure of VOT variation within and across talkers. This analysis provides a qualitative description 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 (i.e., how they are ordered by VOT). 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 where /t/ falls in the middle. The argument for this widely attested pattern is based on vocal tract aerodynamics and articulatory constraints (Cho & Ladefoged, 1999). One of the major contributions of Chodroff & Wilson (2017) is that these ordinal relationships are much more constrained than would be expected from a purely ordinal perspective. Ordinal relationships are a starting place, and 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 in English, as Chodroff & Wilson (2017) review literature indicating coronal behavior to be more variable across dialects of English, the relationship between /p/ and /k/ is inconsistent across talkers in the SpiCE corpus. Table 4.2 reports the proportion of talkers whose mean VOT values followed the expected /p/ $<$ /t/ $<$ /k/ relationships. Note that one talker (VM25A) did not have any instances of Cantonese /p/ in the final sample. The unexpected results were as follows. Cantonese /t/ is typically longer than Cantonese /p/—the opposite holds for English. Cantonese /k/ tends to be longer than Cantonese /t/, but this is almost never the case for English. The ordering of /p/ and /k/ is a toss-up in both languages.

Prior work with English connected speech reports rates of adherence in the 80-90% range for all pairwise combinations, with the exception of /t/ $<$ /k/ being drastically lower for native English speakers in Chodroff & Baese-Berk (2019). While the English /t/ $<$ /k/ comparison is remarkably low here at 6%, only the English /p/ $<$ /t/ ordering 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 in the standard errors indicates that strict ordering may not be appropriate here, as there is not a great deal of confidence in the means' ordering. 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. In fact, only three talkers in the corpus exhibit the same pattern of means across languages (VM21B, VM23A, and VM25A).

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

| **Language** | **/p/ $<$ /t/** | **/t/ $<$ /k/** | **/p/ $<$ /k/** | **/p/ $<$ /t/ $<$ /k/** | **n** |
|---|---|---|---|---|---|
| Cantonese | 0.24 | 0.61 | 0.39 | 0.15 | 33 |
| English | 0.82 | 0.06 | 0.47 | 0.00 | 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 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 & Wilson (2017), this correlation analysis aims to elucidate within-talker invariance and between-talker variability. Tight correlations for 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 the pattern of mean values for long-lag VOT (Chodroff & Wilson, 2017; Cho

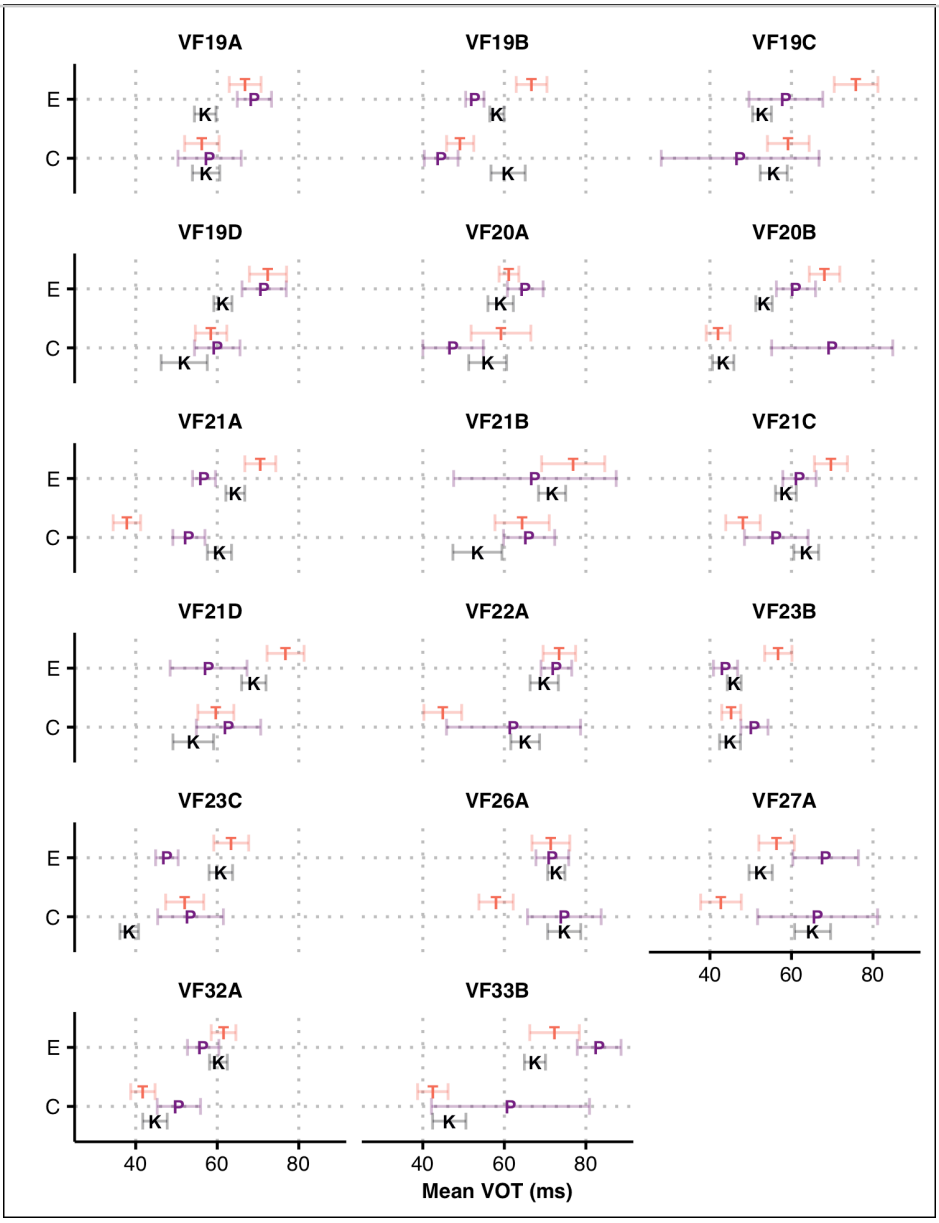


**Figure 4.1:** This figure depicts the ordinal relationships for the female talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

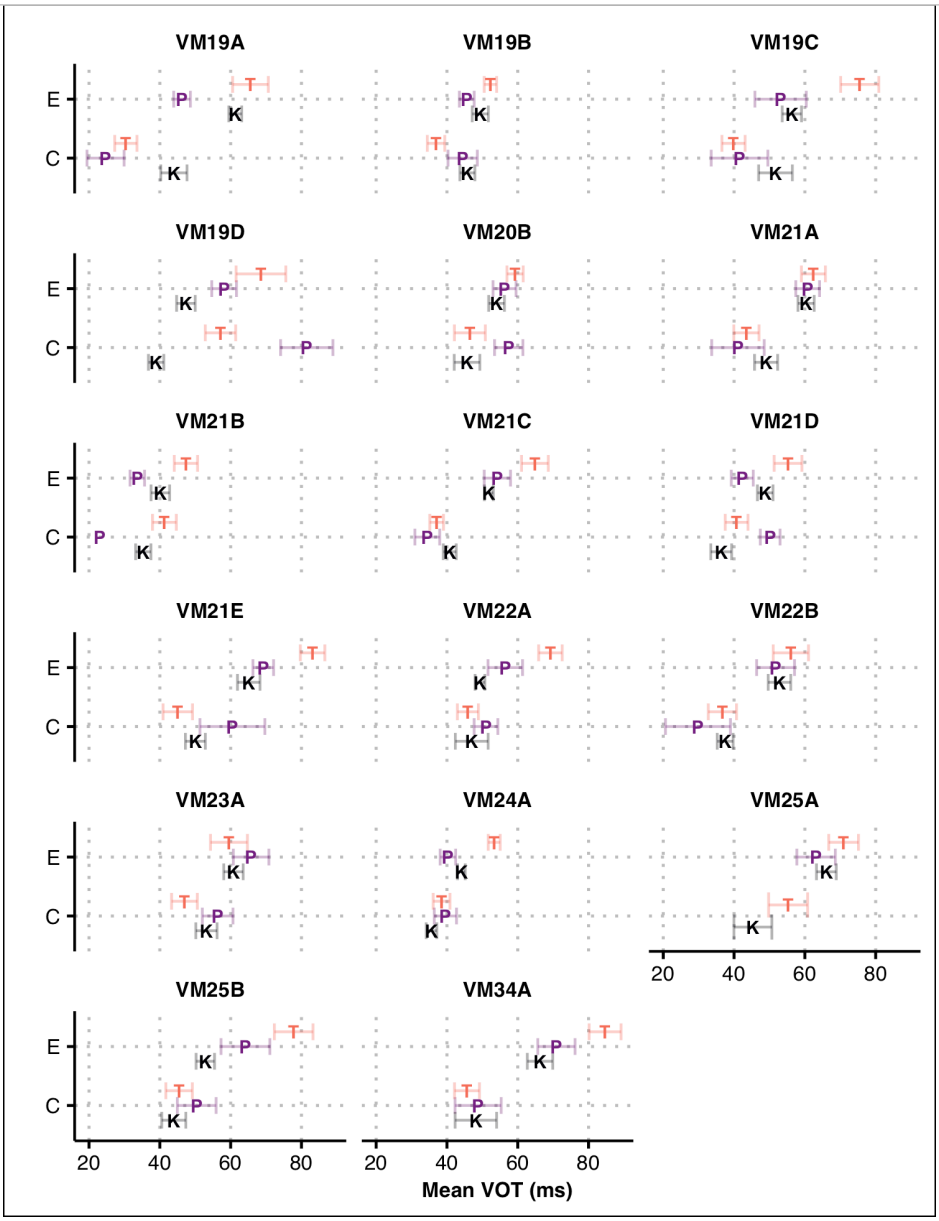


**Figure 4.2:** This figure depicts the ordinal relationships for the male talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows. VM25A had no /p/ tokens.

& 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 & Wilson, 2017; Chodroff & 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 ($0.5 < r < 0.7$; $p < 0.01$; n = 6) or weak and non-significant (n = 8). 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 ($r = 0.59$; $p = 0.003$). 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 ($r = 0.40$, $p = 0.11$). 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.

Chodroff & 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 & Wilson found that "the magnitudes 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" (2017, p. 34).

**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 |

A similar analysis was done 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.[4] 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 minor differences in significance and magnitude. Both versions of the analysis support a conclusion in which the patterns are weak overall.

[4]The canonical form was pulled from the pronunciation dictionaries used during forced alignment. In cases where there was more than one form in the dictionary, the entry with the higher number of phones was used.

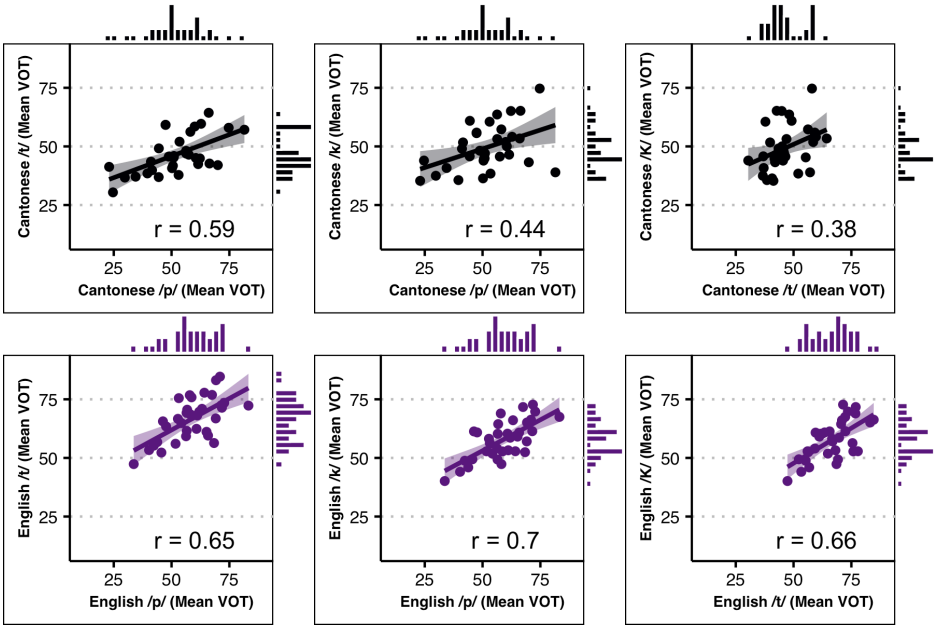


**Figure 4.3:** Correlations for within-language pairwise comparisons of raw mean VOT are depicted with points representing talker means for the segments on the x and y axes and superimposed regression lines. The margins display histograms for each of the axes. Within-Cantonese comparisons are depicted in black, and within English comparisons in purple. *Note that while some of the distributions in the margins appear different, they are not. This is an artifact of plotting the same distribution on different axes in different plots—they only appear mirrored.*

While these relationships indicate some degree of articulatory reuse, the overall picture is far from compelling, particularly when considered alongside the analysis of the ordinal relationships in Section 4.3.1. Compared to prior work, these correlations are less consistent and generally weaker, indicating that the uniformity constraint discussed in Section 4.1.3 may not be as robust as previously argued. This point will be returned to in this chapter's discussion (Section 4.4).

The next steps in Chodroff & Wilson's (2017) methods focus on validating the strength of the correlations. Their approach includes estimating confidence

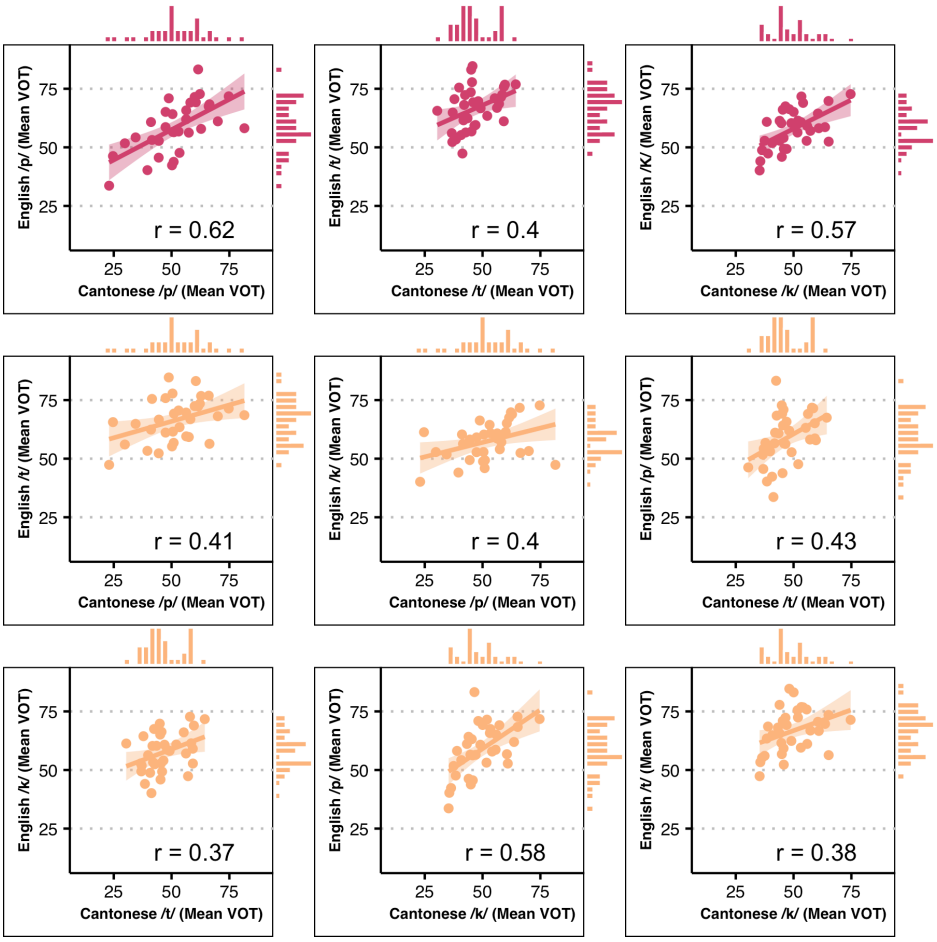


**Figure 4.4:** Correlations for the across-language comparisons of raw mean VOT are depcited in the same manner as Figure 4.3. Comparisons at the same place of articulation are depicted in pink, and comparisons at different places of articulation are in orange.

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 (i.e., a system where the only requirement was the relative VOT ordering of segments) 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 this 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 talkers' sound categories compare to one another while simultaneously accounting for factors known to influence long-lag VOT, such as speaking rate and prosodic position. This section builds on the frequentist mixed-effects model analysis in Chodroff & Wilson (2017). Also, 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 analyzed 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. First, Bayesian modeling and Bayesian inference are described in broad terms. References are provided that point the reader to further reading on the topic. Second, the structure of the model used in this chapter is described and motivated. Lastly, the results of the model are reported. All code used in this analysis is available on GitHub, at https://github.com/khiajohnson/dissertation.[5]

[5]Note that this repository is currently private.

### Bayesian inference

The corpus sample was analyzed with a Bayesian linear mixed-effects 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, as well as derived parameters (e.g., some combination of existing 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, while also not biasing the model toward any specific outcome. 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. HDI stands for highest density interval, and it is 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. This decision criterion provides a useful scaffolding for interpreting the magnitudes of standardized effects when presented alongside the posterior distributions.

### Modeling multilingual VOT

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

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, which in turn allows both to vary by group. While Bayesians tend to refer to the random effects structure in terms of partial pooling, it is important to note that partial pooling and random effects refer to the same thing, regardless of whether the analysis is frequentist of Bayesian.

The model was specified as follows. First, Equation 4.1 gives the formula used in *brms*. Immediately afterward is a description of the model's parameters and how they were specified.

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

**VOT** was the dependent variable—it was standardized (i.e., centered and scaled) 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 & 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 & Alday, 2021). Specifically, weighted effect coding ensures that a simple effect is equivalent to what the main effect would be in a model without the interaction. Coding was implemented using the *wec* R 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 (weights: 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 (see Footnote 4). 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 to calculate the speaking rate. Average phone duration was standardized (i.e., centered and scaled).

**(Preceding) Pause** is a binary variable that indicates whether or not the token occurred after a pause or not. Pauses were identified using the force-aligned transcripts and include instances where the preceding phone was "sil" (silence) or "sp" (silent pause). Pause was also implemented with weighted effect coding (weights: False $= -0.33$, True $= 1$).

**Word** indicates the word that the VOT measurement comes from.

**Talker** indicates which of the 34 talkers produced the item.

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 as 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 5,000 iterations (2,500 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 $\hat{R}$ 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. 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 indicates that English long-lag stops were produced with longer VOT than Cantonese ($\beta = 0.16$, 98.9% HDI outside ROPE). The effect of Place is not consistently modulated by Language, as both of the Place $\times$ Language interaction terms overlapped substantially with the ROPE. This interpretation is not, however, supported by the model predictions summarized in Figure 4.6, which shows the conditional effects for Place and Language—that is, the predicted means for each of the six combinations. The predictions look exactly as would be expected in the case of an interaction—the distance between means is absent for /p/, larger for /t/, and still larger for /k/. That this doesn't emerge in the parameter summary in Figure 4.5 and Table 4.4 may be due to how the categorical variables were coded. In cases such as these, McElreath (2020) argues that parameter summaries are often less informative (and useful) than model predictions.

To dig into this interaction and justify its inclusion in the model, a second model was fit without the interaction term. All other aspects were identical to the model described in Equation 4.1. The models with and without the interaction were then compared using the expected log pointwise predictive density (ELPD Vehtari et al., 2017), as implemented in the *loo* R package (Vehtari et al., 2020). The result of this comparison demonstrates the importance of the interaction term—it substantially improves the model's predictive accuracy (ELPD difference: $-13.4$, SE difference: 5.7). This result suggests that the interaction visible in the model's predictions in Figure 4.6 is valid, even if the parameterization of the population-level variables does not show such an outcome.

Returning to a summary of the original model, 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-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. The intercept can be interpreted as the model estimate when all parameters are at their zero value—in the case of continuous parameters, this is zero, while in the case of the categorical variables, it is the weighted grand mean. The intercepts in 4.8 thus reflect individuals' deviations from this overall intercept. 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 & Wilson (2017), this result highlights between-talker variability

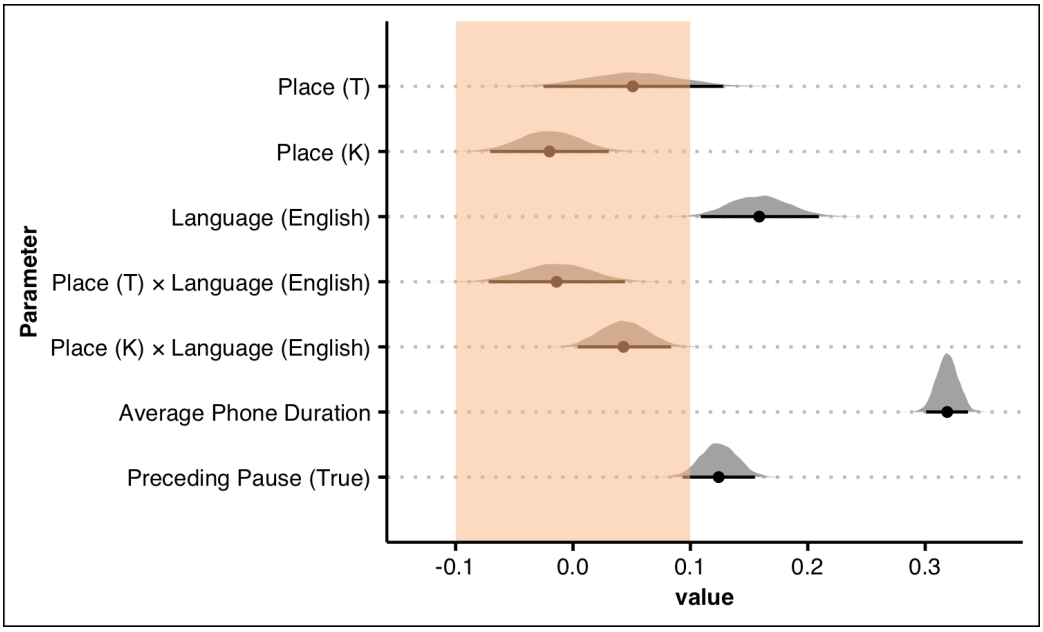


**Figure 4.5:** This figure depicts the 95% HDI posterior distributions for each of the population-level parameters, with the posterior mean indicated by the dot. The orange shaded section represents the ROPE. Recall how to interpret ROPEs—accept the null if posterior is fully within bounds and reject it if the posterior is fully outside ROPE; otherwise, withhold a decision.

and within-talker stability.

At first glance, the results of the mixed-effects model are puzzling in how they seem to contradict what is apparent in the raw data described and analyzed in Sections 4.3.1 and 4.3.2. While there seems to be a cross-language difference in VOT for /t/ and /k/, the model here suggests there is more uniformity than those analyses would support. Why? Given the uncontrolled and spontaneous nature of this speech data, and the large amount of variability captured in partial pooling for words, a simple answer to this question is that talkers simply use different words. To test this, a third model was fit without Word intercepts (but otherwise identical to Equation 4.1). Qualitatively, the exclusion of Word intercepts drastically

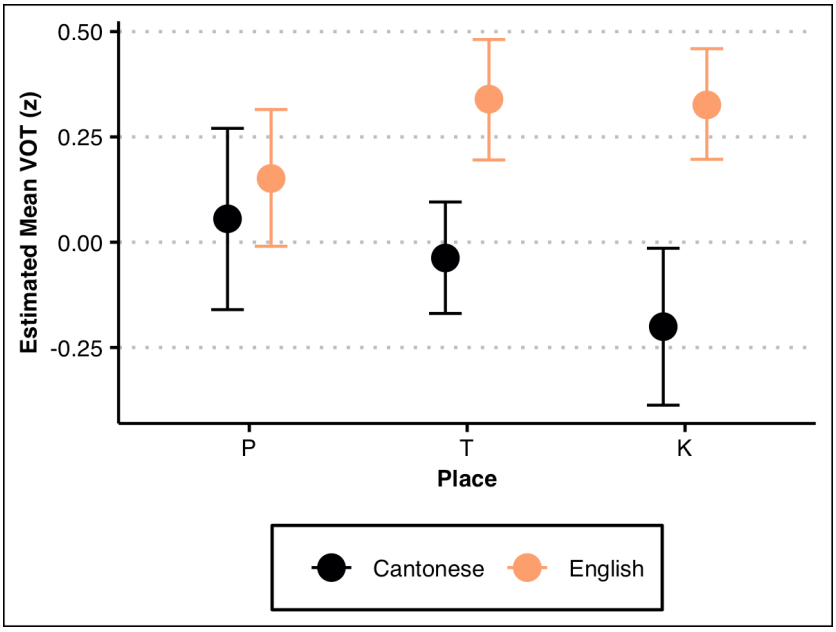


**Figure 4.6:** This figure depicts the model's predicted value and standard error of the predicted value for each of the places of articulation by language, using the fitted method in *brms*' conditional effects function. Notably, the error overlaps almost completely for /p/, but not at all for /t/ and /k/.

**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] |

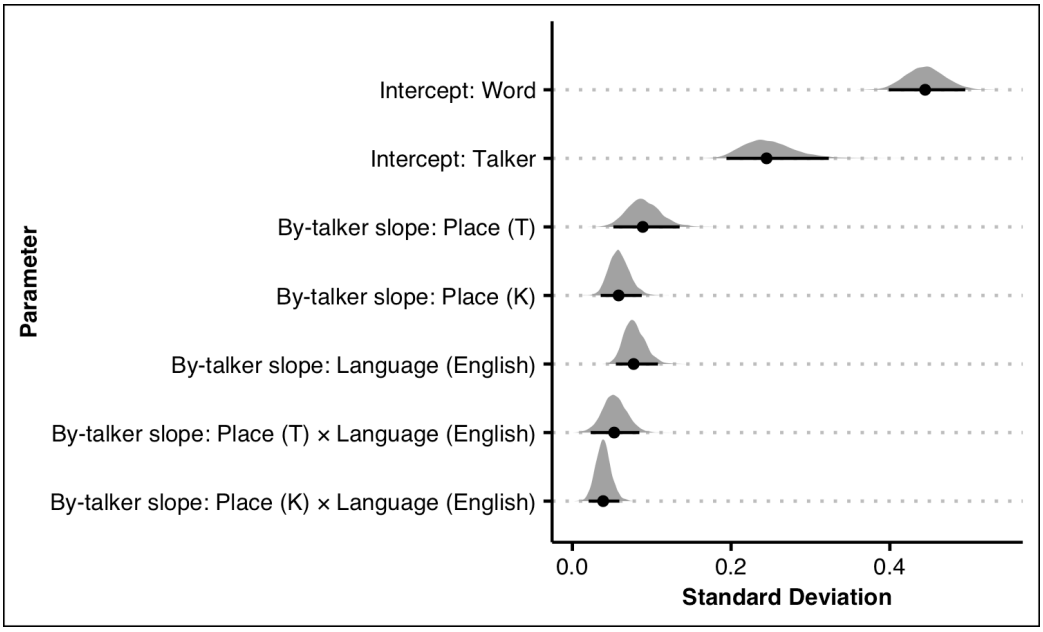


**Figure 4.7:** This figure depicts the posterior distributions for the standard deviation of each of the grouping parameters, both intercepts and slopes.

changes the model output. All of the remaining standard deviation parameters increased—the standard deviations for Talker intercepts and slopes for Place, Language, and Place × Language interaction. Additionally, the interaction between Place and Language predicted by the original model disappears. Without partial pooling for words, only the difference between English and Cantonese /t/ remains. These differences indicate that the apparent discrepancy between the mixed-effects model and the ordinal and correlational analyses can be explained via differences in word distributions across talkers. Essentially, talkers vary in the words they use, and this variation may render the results of Sections 4.3.1 and 4.3.2 somewhat misleading.

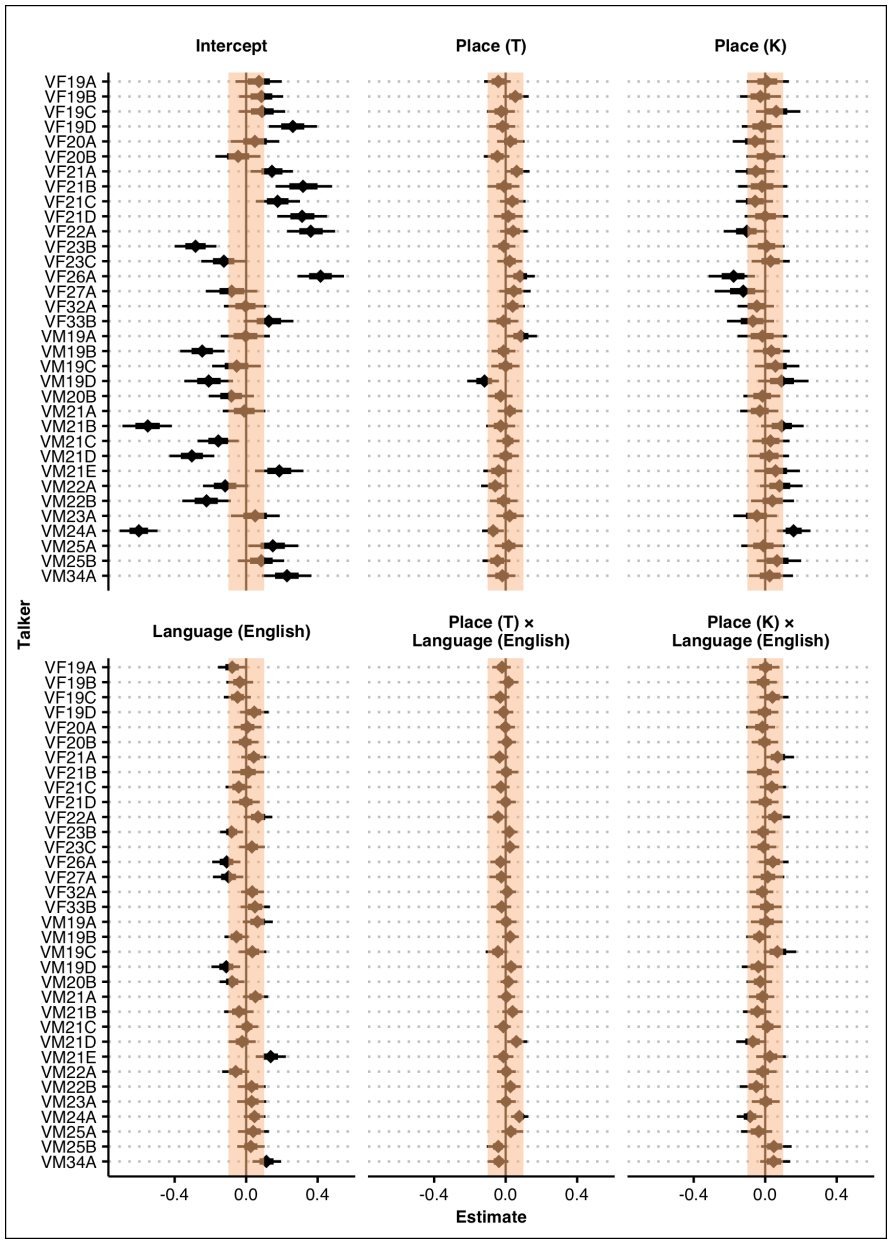


**Figure 4.8:** This figure depicts the 95% HDI for each talker across the talker intercepts and by-talker slope terms. The shaded orange interval represents the ROPE.

## 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 2. It leverages the uniformity framework to assess VOT similarity within and across languages from a few different angles—via ordinal relationships, pairwise correlations, and a Bayesian linear mixed-effects model. In broad strokes, the evidence for uniformity both within and across languages was somewhat mixed. Yet, uniformity was apparent in the mixed-effects model—along with a clear crosslinguistic difference—which is arguably the most robust and reliable of the methods used here (Haines et al., 2020).

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 attributable 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 & Wilson, 2017; Cho & Ladefoged, 1999; Lisker & Abramson, 1964). In this body of work, Chodroff & Wilson (2017) found weaker correlations in reading compared to isolated word production, indicating that speech style plays a role. The yet weaker correlations in this chapter could be interpreted as extending that pattern to a more casual style. Another possible (complementary) account of the results in this chapter is one based on distributions of words produced by the talkers. Both isolated word production and reading offer tight control over the words produced, which means that the distribution of words produced is constant across talkers. The same cannot be said of spontaneous speech, in which talkers produce different words and utterances over the course of conversational interviews. The mixed-effects model in Section 4.3.3 highlights how this is a crucial difference across talkers in this chapter. In this sense, speaking style is an important factor, as it impacts both form and content.

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 & Wilson, 2017; Chodroff & 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 structured variation is far from compelling. This result was surprising, given the typological survey of VOT relationships in Chodroff et al. (2019), which found uniform structure at the level of languages. Recall that the prediction for this chapter included both within-language and across-language uniformity.

While the murky outcome of the ordinal relationships and correlational analyses was largely unexpected, observing a different outcome for a different speech style is not without precedent. For example, the correlation magnitudes that Chodroff & 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 even weaker relationships—likely because of an increase in variability. 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 & Wilson, 2017); and, similarly, for corpus studies of monolingual and bilingual speech (Johnson, 2019).

The Bayesian mixed-effects model offers insight into sources of variation, including the specific role that language and place of articulation play in accounting for VOT variation. The model showed 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 & 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, the model showed a consistent difference across languages for VOT, and very few talkers

deviated from this overall pattern in a meaningful way. This result suggests that the population-level account provides an appropriate generalization across talkers. While there is precedent for languages having different long-lag VOT settings (Chodroff et al., 2019), it is not immediately clear *why* English and Cantonese, in particular, would show different patterns on a within-speaker basis. It could be due to lexical distributional reasons, broad language timing differences, or underlying representational differences—such explanations would be mere speculation at this point. A much greater amount of variation in the model is captured by variable intercepts for word and talker. The model, then, supports the argument for uniformity—between-talker differences vary drastically, while talkers tend to be more internally consistent. Internal consistency, in this case, seems to be more on a "macro" level. That is, talkers with longer /p/ VOT tend to also have longer /t/ and /k/ VOT in both languages (as with speech rate in Bradlow et al., 2017). This kind of macro internal consistency says nothing about the specific ordinal relationships across languages or sound categories, just the general ballpark that VOT production falls into. This macro level, however, may ultimately be what listeners have access to for talker identification—this idea will be expanded upon in the next paragraph.

In light of the evidence for uniformity, 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. This is a relatively small difference, yet, it is worth flagging that this kind of difference is often smallest in spontaneous speech compared to lab speech. Regardless, a difference of this magnitude is not likely to be perceptible in categorization (and similar) paradigms. Work by McMurray et al. (2002), however, demonstrates a gradient and fine-grained effect of processing VOT in increments as small as 5 ms. In this study, McMurray et al. monitor participants' eye moments in an experiment using the visual world paradigm to ascertain how small VOT differences impact lexical access. The crucial result is participants access the correct word, but that modulating within-category VOT

impacted processing difficulty. Further, as research on mergers in sound change has demonstrated, individuals do not always perceive differences that they produce (Yu & Zellou, 2019; Cheng et al., 2021). 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. This conclusion essentially questions whether full assimilation actually occurs in the speech of early bilinguals, and as a result, questions its utility for this kind of population. Partial and context-dependent assimilation (i.e., due to interference) seem to be more fruitful directions.

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. The study in Bradlow et al. (2017) offers an example of this dual influence in the case of speech rate, using native and non-native speech from the ALLSSTAR corpus. In this study, Bradlow et al. demonstrate that talkers who speak faster in their first language (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 demonstrate differences across languages, with some L1s tending to be slower and others faster. This interpretation could be applied fairly transparently to the study in this chapter: talkers with long VOT in one language would also have long VOT in the other language, even if they maintain a difference between languages.

Yet another possibility is that the VOT specification may be even more distinct underlyingly (i.e., $>$ 4 ms) but ultimately brought closer together by the bilingual language mode of the SpiCE interviews. While the interviews were set up such that sentence reading and storyboard narration tasks preceded the conversational interviews (see Section 2.2), and thus helped talkers get into the language mode of the

interview, the context is nonetheless bilingual—it promotes a bilingual language mode (see Grosjean, 2011). In this context, observing a meaningful difference across languages for VOT suggests that under different circumstances, such a difference might be more pronounced. While this account seems to contradict the one offered in the preceding paragraph, it does not specify where the difference arises from—it could be in the representation of VOT and/or in timing factors. The accounts are thus not necessarily contradictory.

The results presented in this chapter provide some support for a crosslinguistic uniformity constraint, in addition to providing an empirical description of bilingual long-lag stops. The weaker (or merely "macro") constraint on within-talker variability compared to prior work has implications for representation and perception. 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). Further, uniformity in structure is useful for talker identification within a single language (Ganugapati & Theodore, 2019). While these accounts are straightforward to interpret in the context of clear and strong relationships, the chapter raises questions about how useful fine-grained structure is in the case of spontaneous speech. It would be worth exploring in future work whether the "macro" structure discussed above is sufficient to confer a benefit in talker identification or if the tight structure of uniformity is necessary. Given the presence of macro structure in the results—as well as the importance of salient factors like pitch (over more subtle factors; Perrachione et al., 2019)—it seems that macro structure might be sufficient. If this interpretation stands in perception, it would lend insight and nuance into the utility of uniformity as a perceptual strategy in real communicative contexts.

Overall, this chapter highlights the need to study spontaneous speech and demonstrates the utility—and some limitations—of the uniformity framework for better understanding crosslinguistic similarity. This chapter also provides evidence for the speculations about what drives multilingual talker identification (Orena et al., 2019) outlined in Section 1.2.

# Bibliography


Afouras, T., Chung, J. S., & Zisserman, A. (2020). Now you're speaking my language: Visual language identification. In *Proceedings of Interspeech 2020*, (pp. 2402–2406). https://doi.org/10.21437/Interspeech.2020-2921 → pages 46, 82

Alderete, J., Chan, Q., & Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, *191*, 103952. https://doi.org/10.1016/j.cognition.2019.04.021 → page 11

Altenberg, E. P., & Ferrand, C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, *20*(1), 89–96. https://doi.org/10.1016/j.jvoice.2005.01.005 → pages 49, 52, 61, 63, 83

Amengual, M. (2017). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, *23*(5), 954–970. https://doi.org/10.1177/1367006917741364 → pages 3, 13

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. https://doi.org/10.1016/j.wocn.2018.04.002 → page 88

Antoniou, M., Best, C. T., Tyler, M. D., & 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. https://doi.org/10.1016/j.wocn.2010.09.005 → page 94

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. https://doi.org/10.1016/j.wocn.2011.03.001 → page 94

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 4218–4222). Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.520 → page 11

Audacity Team (2018). Audacity (R): Free audio editor and recorder. https://www.audacityteam.org/ → page 19

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443. https://doi.org/10.1177/1367006913516035 → page 94

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001 → page 115

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. *ArXiv Preprints*, (pp. 1–21). http://arxiv.org/abs/1506.04967 → page 115

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110823707 → page 101

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008 → page 37

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. Version 6.1.38. http://www.praat.org/ → page 55

Bolton, K., Bacon-Shone, J., & Lee, S.-L. (2020). Societal multilingualism in Hong Kong. In *Multilingual Global Cities*, (pp. 160–184). Routledge. https://doi.org/10.4324/9780429463860-12 → page 15

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., & Mok, K. (2011). Language- and talker-dependent variation in global features

of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, (pp. 356–359). Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Bradlow/Bradlow.pdf → pages 10, 100

Bradlow, A. R., Kim, M., & 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. https://doi.org/10.1121/1.4976044 → pages 53, 109, 126, 127, 135

Brehm, L., & Alday, P. M. (2021). A decade of mixed models: It's past time to set your contrasts. *OSF Preprints*. https://osf.io/3tgq6/ → page 116

Brown, E. L., & 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. https://doi.org/10.1515/shll-2015-0003 → page 94

Brown, E. L., & Harper, D. (2009). Phonological evidence of interlingual exemplar connections. *Studies in Hispanic and Lusophone Linguistics*, *2*(2), 257–274. https://doi.org/10.1515/shll-2009-1052 → page 3

Bruggeman, L., & Cutler, A. (2019). No L1 privilege in talker adaptation. *Bilingualism: Language and Cognition*, (pp. 1–13). https://doi.org/10.1017/S1366728919000646 → page 137

Bullock, B. E., & Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In L. Isurin, D. Winford, & K. deBot (Eds.) *Studies in Bilingualism*, vol. 41, (pp. 189–206). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/sibil.41.12bul → pages 4, 47, 53, 93, 94, 95, 135, 136

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01 → page 114

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances.

*Cognitive Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231 → pages 41, 43, 69

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. https://doi.org/10.3390/languages6010009 → pages 92, 93, 95, 139

Ćavar, M., Ćavar, D., & Cruz, H. (2016). Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, (pp. 4004–4011). Portorož, Slovenia. https://aclanthology.org/L16-1632/ → page 29

Chan, A. Y. W., & 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. https://doi.org/10.1080/07908310008666590 → page 101

Chan, L., Johnson, K. A., & Babel, M. (2020). Lexically-guided perceptual learning in early Cantonese-English bilinguals. *The Journal of the Acoustical Society of America*, *147*(3), EL277–EL282. https://doi.org/10.1121/10.0000942 → pages 3, 137

Chang, C. B. (2015). Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity. In E. Raimy, & C. E. Cairns (Eds.) *The Segment in Phonetics and Phonology*, (pp. 199–217). Chichester, UK: John Wiley & Sons, Inc., 1 ed. https://doi.org/10.1002/9781118555491.ch9 → pages 90, 91

Cheng, A. (2020). Cross-linguistic F0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, *147*(2), EL67–EL73. https://doi.org/10.1121/10.0000498 → pages 51, 52

Cheng, L. S. P., Babel, M., & Yao, Y. (2021). Production and perception across three Hong Kong Cantonese consonant mergers: Community- and individual-level perspectives. Manuscript under review. → page 127

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. https://doi.org/10.1006/jpho.1999.0094 → pages 105, 106, 124, 131

Chodroff, E., & Baese-Berk, M. (2019). Constraints on variability in the voice onset time of L2 English stop consonants. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 661–665). Melbourne, Austrailia. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_710.pdf → pages 100, 101, 105, 109, 125, 131, 136

Chodroff, E., Golden, A., & 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. https://doi.org/10.1121/1.5088035 → pages 113, 125, 126

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 → pages 39, 99, 100, 101, 102, 103, 105, 106, 109, 111, 113, 116, 119, 124, 125, 131, 139

Chodroff, E., & 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). https://doi.org/10.1515/lingvan_2017_0047 → page 101

Chodroff, E., & Wilson, C. (in press). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. *Language*. Expected publication in June 2022. https://eleanorchodroff.com/articles/ChodroffWilson_UniformitySibilants_Language_Accepted_2022.pdf → page 99

Clumeck, H., Barton, D., Macken, M. A., & 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. https://www.jstor.org/stable/23753507 → pages 101, 125

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., & Carter, D. (2014). Building bilingual corpora. In E. M. Thomas, & I. Mennen (Eds.) *Advances in the Study of Bilingualism*, (pp. 93–110). Multilingual Matters. https://doi.org/10.21832/9781783091713_008 → pages 9, 34

Ethnologue (2021). Chinese, Yue. In D. M. Eberhard, G. F. Simons, & C. D. Fennig (Eds.) *Ethnologue: Languages of the World*. Dallas, TX: SIL International, 24 ed. Online version. http://www.ethnologue.com → page 11

Faytak, M. D. (2018). *Articulatory Uniformity Through Articulatory Reuse: Insights from an Ultrasound Study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. https://escholarship.org/uc/item/0jr0010h → pages 99, 100

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.) *Second Language Speech Learning: Theoretical and Empirical Progress*, (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002 → pages 1, 3, 88, 89, 90, 91, 92, 93, 97, 101

Fricke, M., Baese-Berk, M. M., & Goldrick, M. (2016a). Dimensions of similarity in the mental lexicon. *Language, Cognition and Neuroscience*, *31*(5), 639–645. https://doi.org/10.1080/23273798.2015.1130234 → page 3

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016b). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, *89*, 110–137. https://doi.org/10.1016/j.jml.2015.10.001 → pages 55, 93, 94, 95, 138, 139

Fricke, M., Zirnstein, M., Navarro-Torres, C., & Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, *22*(1), 200–207. https://doi.org/10.1017/S1366728918000482 → pages 3, 98

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006 → pages 8, 125

Ganugapati, D., & Theodore, R. M. (2019). Structured phonetic variation facilitates talker identification. *The Journal of the Acoustical Society of America*, *145*(6), EL469–EL475. https://doi.org/10.1121/1.5100166 → pages 100, 128

Garellek, M. (2019). The phonetics of voice. In W. F. Katz, & P. F. Assmann (Eds.) *The Routledge Handbook of Phonetics*. Routledge. https://doi.org/10.4324/9780429056253_5 → pages 39, 40, 42, 57, 84

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555 → pages 114, 117

Gertken, L. M., Amengual, M., & Birdsong, D. (2014). Assessing language dominance with the Bilingual Language Profile. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.) *Measuring L2 proficiency: Perspectives from SLA*, (pp. 208–225). Bristol, UK: Multilingual Matters. https://doi.org/10.21832/9781783092291-014 → page 2

Godfrey, J., Holliman, E., & 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*. https://doi.org/10.1109/icassp.1992.225858 → page 11

Goldrick, M., Runnqvist, E., & Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, *25*(4), 1031–1036. https://doi.org/10.1177/0956797613520014 → pages 93, 94, 95, 96

Google (2019). Cloud speech-to-text. V1. https://cloud.google.com/speech-to-text/ → pages 11, 25

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics*, *0*. https://doi.org/10.1515/ling-2021-0094 → page 10

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. https://doi.org/10.1016/0093_934X(89)90048_5 → pages 1, 2, 3

Grosjean, F. (2011). An attempt to isolate, and then differentiate, transfer and interference. *International Journal of Bilingualism*, *16*(1), 11–21. https://doi.org/10.1177/1367006911403210 → pages 90, 94, 97, 98, 128

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, *60*(2), 98–128. https://doi.org/10.1159/000071449 → page 92

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/xr7y3 → pages 113, 124

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of speech and hearing research*, *37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769 → page 58

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058 → page 21

Ingvalson, E. M., Ettlinger, M., & Wong, P. C. M. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism*, *18*(1), 35–47. https://doi.org/10.1177/1367006912456586 → page 4

Iseli, M., Shue, Y.-L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America*, *121*(4), 2283–2295. https://doi.org/10.1121/1.2697522 → page 57

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001 → page 55

Järvinen, K., Laukkanen, A.-M., & Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, *38*(2), 47–51. https://doi.org/10.3109/14015439.2012.687764 → pages 50, 52, 55, 67

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama, & K. Maekawa (Eds.) *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, (pp. 29–54). Toyko, Japan: The National International Institute for Japanese Language. https://linguistics.berkeley.edu/~kjohnson/papers/Massive.pdf → page 5

Johnson, K. A. (2019). Probabilistic reduction in Spanish-English bilingual speech. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 1263–1267). Melbourne, Australia. → page 125

Johnson, K. A. (2021a). Leveraging the uniformity framework to examine crosslinguistic similarity for long-lag stops in spontaneous Cantonese-English

bilingual speech. In *Proceedings of Interspeech 2021*, (pp. 2671–2675). https://doi.org/10.21437/Interspeech.2021-1780 → page vi

Johnson, K. A. (2021b). SpiCE: Speech in Cantonese and English. V1. https://doi.org/10.5683/SP2/MJOXP3 → pages 6, 8, 130

Johnson, K. A., & Babel, M. (2021). Language contact within the speaker: Phonetic variation and crosslinguistic influence. *OSF Preprints*. https://doi.org/10.31219/osf.io/jhsfc → pages 8, 95, 135

Johnson, K. A., Babel, M., Fong, I., & Yiu, N. (2020a). SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 4089–4095). Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.503 → page vi

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020b). Bilingual acoustic voice variation is similarly structured across languages. In *Proceedings of Interspeech 2020*, (pp. 2387–2391). https://doi.org/10.21437/Interspeech.2020-3095 → page vi

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer-Verlag, 2 ed. https://doi.org/10.1007/b98835 → pages 69, 77

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science*, *15*(5), 314–318. https://doi.org/10.1111/j.0956-7976.2004.00675.x → pages 4, 92, 136

Kawahara, H., Agiomyrgiannakis, Y., & Zen, H. (2016). Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, (pp. 221–228). https://doi.org/10.21437/SSW.2016-36 → page 56

Keating, P., Kreiman, J., & Alwan, A. (2019). A new speech database for within- and between-speaker variability. In *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 736–739). Melbourne, Australia. https://www.assta.org/proceedings/ICPhS2019/papers/ICPhS_785.pdf → pages 41, 86

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*,

*132*(2), 1050–1060. https://doi.org/10.1121/1.4730893 → pages 45, 49, 50, 132

Keshet, J., Sonderegger, M., & Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction. Version 0.94. https://github.com/mlml/autovot/ → page 103

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. https://doi.org/https://doi.org/10.1111/tops.12331 → pages 5, 38

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1), e009. https://doi.org/10.3989/loquens.2014.009 → pages 39, 40, 41, 56, 57, 69

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, *149*(1), 457–465. https://doi.org/10.1121/10.0003331 → pages 40, 69

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925 → pages 114, 115

Labov, W., Ash, S., & Boberg, C. (2008). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton. https://doi.org/10.1515/9783110167467 → page 67

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 175. https://doi.org/10.3389/fpsyg.2011.00175 → pages 43, 83

Laver, J. (1980). *The phonetic description of voice quality*, vol. 31 of *Cambridge Studies in Linguistics*. New York: Cambridge University Press. https://www.cambridge.org/9780521108898 → page 39

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383 → pages 43, 83

Lee, B., & Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, *20*(3), 174–185. https://doi.org/10.1080/2050571x.2016.1273572 → pages 50, 51, 52, 55, 86, 132

Lee, J. L. (2018). PyCantonese. Version 2.2.0. https://pycantonese.org/ → pages 11, 28

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, *146*(3), 1568–1579. https://doi.org/10.1121/1.5125134 → pages 38, 39, 41, 42, 43, 45, 53, 54, 56, 60, 61, 69, 70, 72, 74, 76, 81, 82, 83, 84, 86, 130

Lee, Y., & Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. In *The 178th Meeting of the Acoustical Society of America*. San Diego, CA. Poster. https://doi.org/10.1121/1.5137431 → pages 41, 42, 76, 83, 86, 130

Lee, Y., & Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. In *The 179th Meeting of the Acoustical Society of America*. Acoustics Virtually Everywhere. Poster. https://doi.org/10.1121/1.5146847 → pages 41, 42, 46, 69, 76, 83, 130

Lein, T., Kupisch, T., & 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. https://doi.org/10.1177/1367006915589424 → page 92

Leung, M.-T., & Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, *6*(2), 305–325. https://doi.org/10.1075/ijcl.6.2.06leu → page 11

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science*, *10*(2), e1483. https://doi.org/10.1002/wcs.1483 → page 44

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. https://doi.org/10.1007/978-3-319-12619_7 → page 54

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279 → pages 5, 38

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, *5*(1), 91–107. https://doi.org/10.1146/annurev-linguistics-011516-033830 → page 10

Lieberman, P., & Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139165952 → page 103

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman, & C. N. Li (Eds.) *Language, Speech, and Mind: Studies in Honour of Victoria A. Fromkin*, (pp. 62–78). London: Routledge. → page 92

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830 → pages 101, 124, 125, 131

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28. https://doi.org/10.1177/002383096701000101 → page 102

Littell, P. (2010). Thank-you notes [Version 1.0: Agent focus]. http://totemfieldstoryboards.org/stories/thank_you_notes/ → page 21

Llompart, M., & 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*. https://doi.org/10.1016/j.wocn.2017.12.001 → page 88

Lloy, A., Johnson, K., & Babel, M. (2021). Examining the roles of language familiarity and bilingualism in talker recognition. In *The 13th International Symposium on Bilingualism*. Virtual. Poster. → pages 87, 138

Lloy, A., Johnson, K. A., & Babel, M. (2020). Bilingual talker identification with spontaneous speech in Cantonese and English: The role of language-specific knowledge. In *The 179th Meeting of the Acoustical Society of America*. Virtual. Poster. https://doi.org/10.1121/1.5147685 → pages 87, 138

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, *24*(1), 71–89. https://doi.org/10.1177/002383098102400105 → pages 51, 133

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445 → page 69

Luke, K. K., & Wong, M. L. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series*, *25*, 312–333. https://www.jstor.org/stable/26455290 → page 11

Matthews, S., Yip, V., & Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. https://doi.org/10.4324/9780203835012 → pages x, 20, 48, 82, 101, 134

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner. Version 1.0.1. https://montrealcorpustools.github.io/Montreal-Forced-Aligner/ → page 28

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman and Hall/CRC, 2 ed. https://doi.org/10.1201/9780429029608 → pages 115, 117, 118

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9 → page 126

Ménard, L., Schwartz, J.-L., & Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, *50*(1), 14–28. https://doi.org/10.1016/j.specom.2007.06.004 → pages 99, 100

Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second Language Research*, *26*(1), 13–41. → page 40

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006 → page 91

Mielke, J., & 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. https://doi.org/10.1121/1.5059493 → page 101

Munson, B., & Babel, M. (2019). The phonetics of sex and gender. In W. F. Katz, & P. F. Assmann (Eds.) *The Routledge Handbook of Phonetics*. Routledge. https://doi.org/10.4324/9780429056253_19 → page 57

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of vox humana. *Clinical linguistics & phonetics*, *24*(4-5), 245–260. → page 47

Myers-Scotton, C. (2011). The matrix language frame model: Developments and responses. In *Codeswitching Worldwide*, vol. 126 of *Trends in Linguistics. Studies and Monographs*. De Gruyter Mouton. https://doi.org/10.1515/9783110808742.23 → page 55

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, *43*(1-2), 65–84. http://digital.casalini.it/10.1400/190440 → pages 20, 23, 26

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. University of Adelaide, Adelaide, Australia. Version 0.5. http://ua.edu.au/ccs/teaching/lsr → page 61

Ng, M. L., Chen, Y., & Chan, E. Y. (2012). Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—a long-term average spectral analysis. *Journal of Voice*, *26*(4), e171–e176. https://doi.org/10.1016/j.jvoice.2011.07.013 → pages 49, 50, 53, 61, 63, 67, 83, 130, 132

Ng, M. L., Hsueh, G., & Sam Leung, C.-S. (2010). Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, *12*(3), 230–236. https://doi.org/10.3109/17549501003721080 → pages 49, 61, 83

Ng, R. W. M., Kwan, A. C., Lee, T., & Hain, T. (2017). ShefCE: A Cantonese-English bilingual speech corpus for pronunciation assessment. In

*Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 5825–5829). https://doi.org/10.1109/ICASSP.2017.7953273 → page 10

Nieuwenhuis, R., Manfred, t. G., & Pelzer, B. (2017). Weighted effect coding for observational data with wec. *The R Journal*, *9*(1), 477. https://doi.org/10.32614/rj-2017-017 → page 116

Olson, D. J. (2016). The role of code-switching and language context in bilingual phonetic transfer. *Journal of the International Phonetic Association*, *46*(3), 263–285. https://doi.org/10.1017/S0025100315000468 → pages 93, 94, 95

Ordin, M., & Mennen, I. (2017). Cross-linguistic differences in bilinguals' fundamental frequency ranges. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1493–1506. https://doi.org/10.1044/2016_JSLHR-S-16-0315 → page 52

Orena, A. J., Polka, L., & 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. https://doi.org/10.1121/1.5097735 → pages 4, 5, 44, 45, 100, 128, 134, 138, 139

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/icassp.2015.7178964 → page 11

Perrachione, T. K. (2018). Recognizing speakers across languages. In S. Frühholz, & P. Belin (Eds.) *The Oxford Handbook of Voice Perception*, (pp. 514–538). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.23 → page 44

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, *146*(5), 3384–3399. https://doi.org/10.1121/1.5126697 → pages 44, 46, 53, 76, 128

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. https://doi.org/10.1016/j.specom.2004.09.001 → pages 9, 10, 22

Pittam, J. (1987). The long-term spectral measurement of voice quality as a social and personality marker: A review. *Language and Speech*, *30*(1), 1–12. https://doi.org/10.1177/002383098703000101 → pages 39, 40

Podesva, R. J., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270 → pages 37, 38, 40

Polinsky, M. (2018). *Heritage Languages and their Speakers*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107252349 → page 135

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ → pages 61, 69, 106, 114

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75–86. https://doi.org/10.1037/a0027979 → page 128

Revelle, W. (2021). psych: Procedures for psychological, psychometric, and personality research. R package version 2.1.3. https://CRAN.R-project.org/package=psych → page 106

Ryabov, R., Malakh, M., Trachtenberg, M., Wohl, S., & Oliveira, G. (2016). Self-perceived and acoustic voice characteristics of Russian-English bilinguals. *Journal of Voice*, *30*(6), 772.e1–772.e8. https://doi.org/10.1016/j.jvoice.2015.11.009 → page 51

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. https://doi.org/10.1016/j.jml.2019.104070 → page 98

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. https://doi.org/10.1006/jpho.1997.0051 → pages 3, 94

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences*, vol. 3, (pp. 1846–1849). Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011 → page 56

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935345.013.72 → pages 89, 90

Simonet, M., & Amengual, M. (2019). Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, *24*(2), 208–221. https://doi.org/10.1177/1367006919826388 → pages 3, 22, 94, 95, 97

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, *3*(2), 621–640. https://doi.org/10.1111/j.1749-818X.2009.00125.x → page 55

Sjölander, K. (2004). The Snack Sound Toolkit. https://www.speech.kth.se/snack/ → page 57

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf → pages 25, 26

Sóskuthy, M., & Stuart-Smith, J. (2020). Voice quality and coda /r/ in Glasgow English in the early 20th century. *Language Variation and Change*, *32*(2), 133–157. https://doi.org/10.1017/S0954394520000071 → page 74

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*(2), 218–231. https://doi.org/10.3758/BF03193744 → pages 46, 82

Stan Development Team (2021). *Stan Modeling Language Users Guide and Reference Manual*. https://mc-stan.org → page 114

Statistics Canada (2017). Proportion of mother tongue responses for various regions in Canada, 2016 Census. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm → pages 13, 137

Stewart, D., & Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, *70*(3, pt.1), 160–163. https://doi.org/10.1037/h0026143 → page 77

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3-4), 505–549. https://doi.org/10.1515/lp-2015-0015 → pages 101, 109, 125

Sun, J. (2020). jieba. Version 0.42.1. https://github.com/fxsjy/jieba → page 28

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (pp. I–333–I–336). https://doi.org/10.1109/ICASSP.2002.5743722 → page 59

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, *9*(1), 97–114. https://doi.org/10.1017/S1366728905002403 → pages 92, 93, 94, 95

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 ed. → page 70

Tanner, J., Sonderegger, M., Stuart-Smith, J., & Fruehwald, J. (2020). Toward "English" phonetics: Variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00038 → page 6

Tse, H. (2019). *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA. http://d-scholarship.pitt.edu/35721/ → pages 20, 29

Tsui, R. K.-Y., Tong, X., & Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, *40*(1), 29–58. https://doi.org/10.1017/S0142716418000449 → pages 96, 97

Turk, M., & Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1991.139758 → page 69

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008 → pages 114, 117

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1. https://mc-stan.org/loo/ → page 118

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4 → page 118

Voigt, R., Jurafsky, D., & Sumner, M. (2016). Between- and within-speaker effects of bilingualism on F0 variation. In *Proceedings of Interspeech 2016*, (pp. 1122–1126). San Francisco, CA. https://doi.org/10.21437/Interspeech.2016-1506 → pages 51, 133

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9–30. https://doi.org/10.1093/applin/amx039 → page 47

Wilson, C., & Mihalicek, V. (2011). *Language Files: Materials for an Introduction to Language and Linguistics*. Columbus, UH: Ohio State University Press. https://linguistics.osu.edu/research/pubs/lang-files → pages x, 48, 82

Winterstein, G., Tang, C., & Lai, R. (2020). CantoMap: A Hong Kong Cantonese MapTask corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 2906–2913). Marseille, France. https://aclanthology.org/2020.lrec-1.355 → page 11

Wong, W. Y. P. (2006). *Syllable Fusion in Hong Kong Cantonese Connected Speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. http://rave.ohiolink.edu/etdc/view?acc_num=osu1143227948 → pages 26, 28

Xue, S. A., Hagstrom, F., & Hao, J. (2002). Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: A functional system approach. *Asia Pacific Journal of Speech, Language and Hearing*, *7*(1), 55–62. https://doi.org/10.1179/136132802805576544 → page 50

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, (p. 0267658319851820). https://doi.org/10.1177/0267658319851820 → pages 96, 101

Yang, Y., Chen, S., & Chen, X. (2020). F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Proceedings of Interspeech 2020*, (pp. 4163–4167). https://doi.org/10.21437/Interspeech.2020-2549 → page 51

Yau, M. (2019). PyJyutping. https://github.com/MacroYau/PyJyutping → page 11

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150. https://doi.org/10.1146/annurev-linguistics-011516-033815 → page 127

Yu, H. (2013). Mountains of gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, (pp. 108–121). Routledge. https://doi.org/10.4324/9780203100387.ch7 → page 13

Yuan, J., Ryant, N., & Liberman, M. (2014). Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 2539–2543). https://doi.org/10.1109/ICASSP.2014.6854058 → page 29

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 11475. https://doi.org/10.1038/srep11475 → page 45