# Chapter 4

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

## 4.1 Introduction

A consequence of bilingualism is that individuals must navigate segment inventories that exist in a shared phonetic space, in which sound categories may or may not share aspects of their mental representations (Flege and Bohn, 2021). One of the primary goals in this chapter is to address what languages share, if anything, in the mental representation of speech sound categories. The idea of representation is intended here in the manner typically meant by psycholinguists (e.g., Llompart and Reinisch, 2018), exemplar theory proponents (e.g., Amengual, 2018), and Flege and Bohn (2021), who take a similar approach in the revised Speech Learning Model (SLM-r). They 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" (Flege and Bohn, 2021, p. 32). So, if sound categories from different languages exist in the same phonetic space and are repre-

sented by distributions of exemplars, how, then, can the extent to which languages share representation(s) be assessed?

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

### 4.1.1 Identifying "links" across languages

At first glance, the best candidates for shared representation are sound categories that are discussed as being "linked" together. The idea of links can be frustratingly vague in the multilingualism literature—in a handbook chapter on bilingual phonetics and phonology, Simonet (2016) describes "links or connections of one sort or another between the phonetic categories" (p. 10). While vague, links nonetheless represents a crucial concept. In the most basic sense, links exist between sound categories that exert influence on one another under some set of circumstances. Links behave dynamically, as as such, Simonet also notes that "these connections...are transiently strengthened in contexts that induce the activation of both languages and inhibited in contexts that favor the use of only one of the languages" (2016, p. 10). Arguably, the links must exist because crosslinguistic influence can be observed. While there may be alternative explanations, the concept of links is widely used in accounting for bilinguals' behavior.

Flege and Bohn (2021) expand on links in the SLM-r, providing a framework for predicting which sounds categories will be linked together. The proposal is simple—namely, sound categories will be linked to the closest category in the

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

While Flege and Bohn (2021) and Chang (2015) take different approaches—relative phonetics and perceptual ratings accomplish a similar goal—accounting for abstraction and nonlinearity in listeners' mental representations. In sum, abstract similarity seems to be a prerequisite for the emergence of a link between two sound categories, given how it does a better job of accounting for when and where crosslinguistic influence occurs. The presence of abstract similarity does not address what happens next. It does not entail any particular outcome, and it does not directly address how representation manifests 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. Assimilation is... Dissimilation is... The motivation for these outcomes arises from two simple constraints from the productive and perceptual systems. Effectively, don't get too close to each other in perception, and don't get too complicated in production (Guion, 2003; Lindblom and Maddieson, 1988; Flege and Bohn, 2021). These constraints lead SLM-r to posit that proximity leads to instability, even if what counts as close remains unclear, and that the outcome of instability is assimilation and dissimilation. Considering how bilinguals are fully capable of maintaining subtle distinctions for similar sound categories across languages (e.g., Sundara et al., 2006; Casillas, 2021), this is not a trivial point to make.

By following what SLM-r posits, a relatively simple account is that dissimilation for similar sound categories would lead to distinct representations of those categories. A similarly straightforward account is that assimilation leads to a shared representation. The picture is complicated, however, by the idea of imperfect assimilation and what Flege and Bohn term *composite categories*. In the SLM-r, if sounds from two languages are phonetically too close to each other, they will remain linked in a composite category "defined by the statistical regularities present in the combined distributions of the perceptually linked...sounds." (Flege and Bohn, 2021, p. 41). This scenario might be characterized as an imperfectly shared representation, where certain dimensions are kept apart, and others overlap. Alternatively, this lack of clear-cut examples of assimilation in the literature may instead indicate that assimilation and dissimilation might be better cast as ends of a spectrum for a gradient, context-sensitive phenomenon.

There are a few potential reasons for the lack of clear-cut assimilation. First, true assimilation might just be rare in bilingual speech. This reason is supported by a recent meta-analysis of crosslinguistic influence for Spanish and English initial stop consonants (Casillas, 2021). In this environment, English long-lag stops

and Spanish short-lag stops are linked to one another. Casillas found that early bilinguals did not produce "compromise" stop categories. That is, early Spanish-English bilinguals did not produce voice onset time that was somehow intermediate to canonical productions by monolinguals of either language. Instead, the production of each category was influenced by task demands and factors such as social context. This finding echoes arguments made by Bullock and Toribio (2009) on the sophistication and control that bilingual exert over their possible forms. So while there is clear evidence of a link between the two sounds, there is no compromise category. Instead, bilinguals produce a wide range of forms appropriate to and influenced by different contexts. Without considering task and context factors, it is perhaps not surprising that the two sound categories masquerade as a single composite category.

A second reason for the rarity of complete assimilation arises from the experimental and corpus-based approaches typically used to study crosslinguistic influence. In these approaches, the ability to examine crosslinguistic influence for any given pair of sounds hinges on the presence of an observable difference under some set of conditions. I posit that 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. 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 voice onset time, there is strong evidence for a crosslinguistic link between English long-lag and Spanish short-lag stops (Casillas, 2021; Fricke et al., 2016; Goldrick et al., 2014; Bullock and Toribio, 2009; Olson, 2016).

These studies demonstrate phonetic convergence—or variable assimilation—in two ways. First, VOT is shorter for English initial stops produced by bilinguals when compared to monolingual control groups. This result is attributed to the influence on English long-lag stops from the short-lag category in the other language (Olson, 2016). Second, bilinguals appear more likely to produce lead voicing in

initial English voiced stops compared to English monolinguals (Sundara et al., 2006). Evidence of crosslinguistic influence arises from comparing bilinguals to monolinguals comparing bilinguals to themselves across different circumstances. For example, Fricke et al. (2016) use a spontaneous speech corpus to demonstrate that Spanish-English bilinguals produce shorter, more Spanish-like VOT in the lead up to an English-to-Spanish code switch (Fricke et al., 2016). While this body of work makes the presence of a link clear, it also highlights that there are distinct aspects of how these sound categories are represented in the bilingual mind (Casillas, 2021). In the SLM-r, these examples might be considered composite categories. Alternatively, they might be examples of contrasts being maintained in the face of proximity.

In any case, this focus presents a conundrum. By using methods where observing similarity hinges on the ability to detect a difference, researchers preemptively exclude the best candidates for shared representation—those that share both abstract and acoustic similarity. One example comparing highly similar sound categories in the early bilingualism literature comes from a lab-based study of Mandarin-English bilingual children (Yang, 2019). The authors found that highly proficient bilingual 5 to 6-year-olds produced equivalent VOT for Mandarin and English long-lag stops, even though the monolingual comparison groups were consistently different. Yang's result suggests that the difference is either too small to maintain or that 5 to 6-year-old children have not yet mastered it. These claims should be tempered, however, as Yang (2019) did not control for language mode, and adult bilingual behavior was not considered.

Despite some inroads, there is nonetheless a distinct paucity of work examining highly phonetically similar speech sounds across languages, even when such a connection would make sense. A recent study of crosslinguistic influence in Cantonese-English bilinguals compares English long-lag and Cantonese short-lag stops in the context of a language switching paradigm (Tsui et al., 2019). While this comparison reflects the need for stimuli to be acoustically distinct beforehand, 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—should be the long-lag stops in each language. The null result with balanced bilinguals is thus unsurprising. I do not mean to suggest that the (Tsui et al., 2019) would have gotten more insightful results by comparing Cantonese long-lag stops to English long-lag stops. Instead, I aim to highlight the design constraints of the paradigms used in crosslinguistic influence research. Such methods are better suited for detecting dissimilation or contrasts that have been maintained. Conversely, methods for assessing assimilation must not rely on the presence of detectable acoustic differences.

To summarize, most work in crosslinguistic influence has focused on phonologically similar yet phonetically distinct pairs of segments, which are not strong candidates for shared mental representation (as defined at the beginning of this chapter). This focus of this work on telling sounds apart is likely a result of commonly-used paradigms requiring differences to detect influence and the possibility that assimilation may be rare in early bilingual speech. Additionally, 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). In this sense, studying interactions between systems in bilingual speech offers yet another window into linguistic processing (Fricke et al., 2019). In the interest of understanding mental representation, the best category candidates would be the hardest to distinguish using surface forms in the first place.

### 4.1.3 Adapting the uniformity framework

The study described in this chapter focuses on assessing whether phonetically similar sounds share representation or not. Unlike prior work that tells sound categories apart, this chapter aims to tell sound categories together—that is, does a single category deploy to each language, or does each language have its own version. 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, I extend the articulatory uniformity framework to the study of multilingual segment inventories.

Articulatory uniformity is conceptualized as a constraint on within-talker phonetic variation, in which articulatory gestures or phonological primitives are implemented systematically in speech production (Chodroff and Wilson, 2017; Faytak, 2018; Ménard et al., 2008). The core idea of the articulatory uniformity framework is that phonetic variation is highly structured. While Chodroff and Wilson (2017) draw tight connections between uniformity and phonological features, Faytak (2018) emphasizes how talkers learn and reuse articulatory gestures. The 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. While the theoretical account varies somewhat by author, there is nothing to suggest that such accounts are incompatible. That is, both articulatory and phonological explanations are likely valid. Given the focus of this chapter on phonetic and psycholinguistic accounts of category formation and representation, the articulatory account is perhaps more appropriate.

In this light, if a set of segments share an attribute, then talkers should implement the segments with the same phonetic target or articulatory gesture—which may or may not have an acoustic consequence. This systematicity has been observed for in vowel height (Ménard et al., 2008), tongue shape (Faytak, 2018), fricative peak frequency, and stop consonant VOT (Chodroff and Wilson, 2017). In the case of VOT in particular, the relationship between a laryngeal gesture and its acoustic consequence is clear. This allows for the extension of Ménard et al.'s (2008) argument regarding F1 and tongue height to VOT and its corresponding laryngeal gesture. Reusing the gesture across sounds that share the relevant attribute "may simplify the somatosensory feedback needed to control the speech task" (Ménard et al., 2008, p. 26). In simple terms, reusing gestures is easier than the alternative in the case of high vowels. I posit that the same argument could easily be extended for long-lag stops.

Findings for stop consonant within-language uniformity appear to be quite robust. Chodroff and Wilson (2017) report consistent results across a lab study based on reading a list of CVC and a corpus study using connected read speech. Chodroff and Baese-Berk (2019) replicate the uniformity findings for stop consonants with read speech samples from 140 non-native English speakers in the ALLSSTAR corpus (wide range of native languages Bradlow et al., 2011). While Chodroff and Baese-Berk (2019) found a greater degree of between-talker variability with non-native speakers compared to the prior monolingual work, the within-talker structure was robust. However, the uniformity framework has not yet been extended to early bilingual speech, in particular as a mechanism for comparing how bilinguals produce phonetically similar sounds in each of their languages. Extending the framework in this way follows the framing of uniformity as arising from articulatory reuse (Faytak, 2018).

### 4.1.4 Long-lag stops in Cantonese and English

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

While the presence of articulatory reuse within Cantonese and across lan-

guages remains an empirical question, it follows from the finding that bilingual Mandarin-English children did not distinguish between languages in VOT (Yang, 2019). Following the predictions of the SLM-r (Flege and Bohn, 2021), this work suggests that long-lag items of minimally distinct VOT would assimilate or dissimilate but not be stable in such close proximity.

Thus, the present study asks: Do Cantonese-English bilinguals uniformly produce long-lag stops within and across each of their languages? Leveraging the methodology from Chodroff and colleagues (Chodroff and Wilson, 2017, 2018; Chodroff and Baese-Berk, 2019) allows for a new perspective on the structure of variation and nature of representation in bilinguals. It also facilitates the study of phonetically similar speech sounds in ways that other paradigms do not. As may be clear from the framing of the introduction, the hypothesis was that bilinguals would indeed exhibit crosslinguistic uniformity.

## 4.2 Methods

### 4.2.1 Corpus

This study uses 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 **??** for additional information about the talkers.

### 4.2.2 Segmentation & measurement

All instances of prevocalic word-initial /p t k/ were identified from the SpiCE corpus' force-aligned Praat TextGrid transcripts. For English, only items with initial stress were included in the initial sample (Lisker and Abramson, 1967)—this means the extremely high-frequency English word "to" was excluded, as was the case in Chodroff and Wilson (2017). 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 Cantonese interviews, and likewise for English. The initial total count across talkers and languages included 10,428 tokens.

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

After running AutoVOT, instances of /p t k/ were subjected to exclusionary criteria to catch errors. Items 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 ($n = 567$), 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 items had a VOT more than 2.5 standard deviations above the grand mean ($> 129.5$ ms; $n = 191$).

Of the initial sample, 14.1% was excluded, resulting in 8,961 stops, summarized in Table 4.1. Talkers had a median of 97 Cantonese stops (range: 54-194) and 150.5 English stops (range: 73-540). While Cantonese stops were culled at a slightly higher rate (43% of initial sample, 38% of final sample), the higher number of English stops is likely due primarily to lexical distributional reasons. Additionally, English has a greater number of highly frequent /k/-initial word types, while

Cantonese /p/ occurs in fewer, less frequent word types in the final sample ($n = 60$, max frequency of 97) than English ($n = 158$, max frequency of 215).

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

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

## 4.3 Analysis & Results

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

### 4.3.1 Ordinal relationships

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

The results 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, the relationship between /p/ and /k/ is inconsistent across talkers. Table 4.2 reports the proportion of talkers whose mean VOT values followed the expected /p/ $<$ /t/ $<$ /k/ relationships. Prior work with connected speech reports rates of adherence in the 80-90% range, except for English /t/ $<$ /k/ being drastically lower for native English speakers (Chodroff and Baese-Berk, 2019). While the English /t/ $<$ /k/ comparison is remarkably low here (6%), only English /p/ $<$ /t/ (82%) falls in the range that prior work suggests. 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. In many cases, the standard errors for the different segments in a given talker's panel overlap, indicating that strict ordering may not be appropriate here. Additionally, talkers do not appear to be consistent across languages. For example, talker VF19B in Figure 4.1 exhibits a clear /p/ $<$ /t/ $<$ /k/ relationship in Cantonese, but a clear /p/ $<$ /k/ $<$ /t/ relationship in English.

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

| **Language** | **p<t** | **t<k** | **p<k** | **n** |
|---|---|---|---|---|
| Cantonese | 0.24 | 0.61 | 0.39 | 33 |
| English | 0.82 | 0.06 | 0.47 | 34 |

### 4.3.2 Pairwise correlations

To examine the relationship between stops within and across languages, 15 pairwise Pearson's *r* correlations were calculated across talker means. Each correlation compares talkers means for two different segments. The full pairwise correlations include 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 and Wilson (2017),

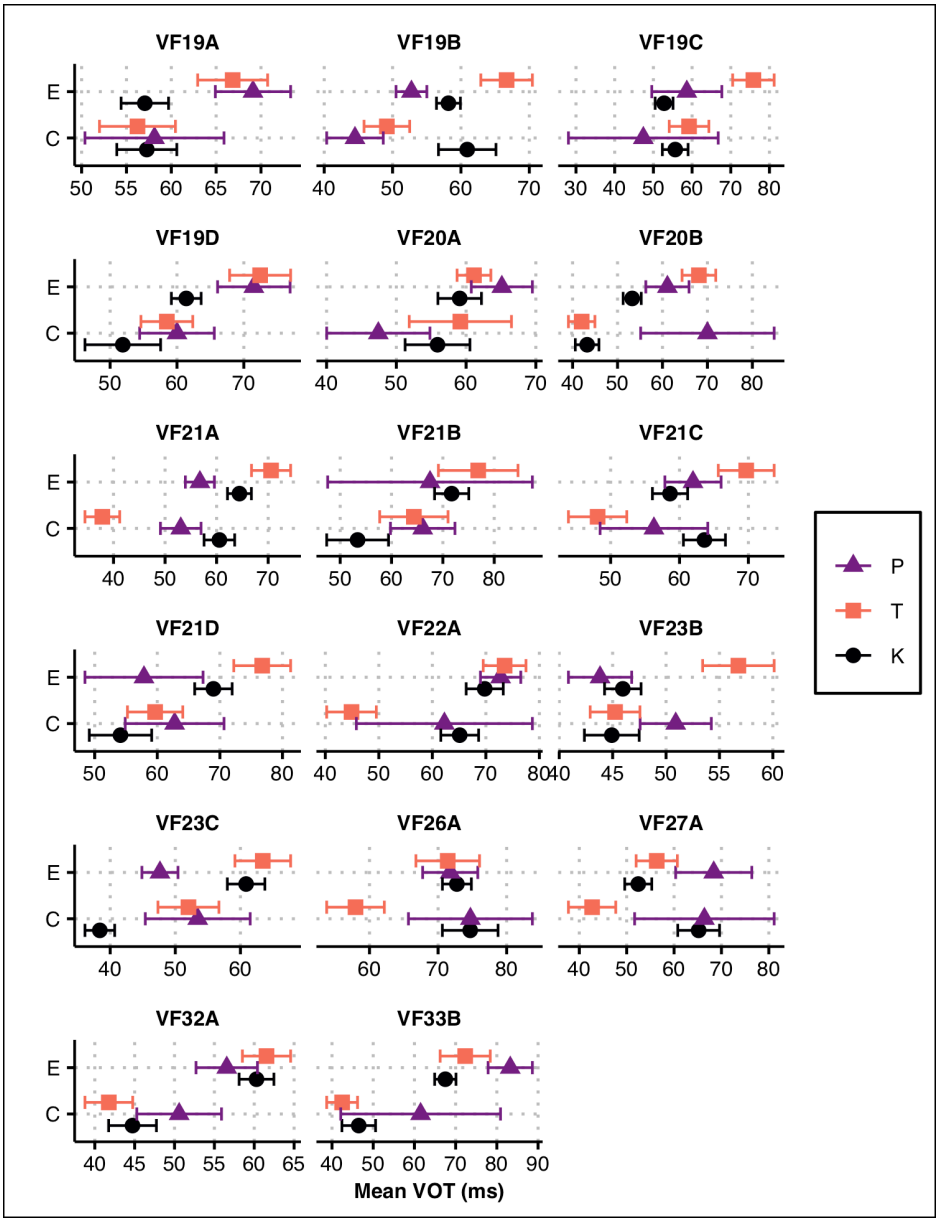


**Figure 4.1:** This figure depicts the ordinal relationships for the female talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

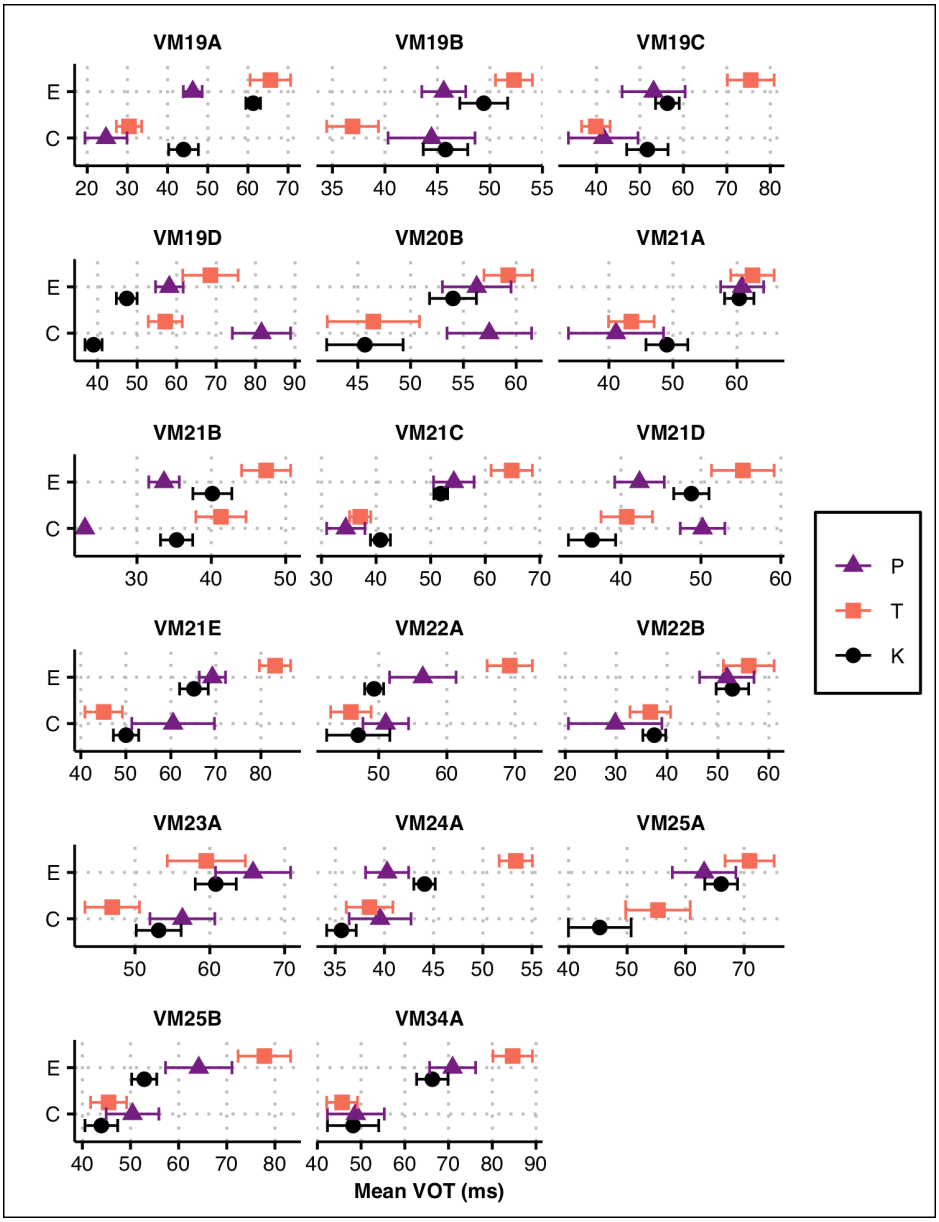


**Figure 4.2:** This figure depicts the ordinal relationships for the male talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

this correlation analysis aims to elucidate within-talker invariance and between-talker variability. While using means ignores information about within-category variability, prior work sets up strong, clear expectations about how the mean values for long-lag VOT pattern (Chodroff and Wilson, 2017; Cho and Ladefoged, 1999). Additionally, Section 4.3.3 digs into individual differences and variability, thus addressing such concerns (see Haines et al., 2020).

Table 4.3 summarizes the output of all 15 correlations in text form. Figures 4.3 and 4.4 depict each of the 15 correlations in the fashion of Chodroff and Wilson (2017). 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. Similar within-language comparisons had $r > 0.7$ (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019). With the exception of the English /p/ $\sim$ /k/ ($r = 0.70$, $p < 0.001$), all of the correlations were either moderate ($0.5 < r < 0.7$; $p < 0.01$) or not significant. Within-English correlations were the most consistent—all three had *r* at or above 0.65 ($p < 0.001$). Of the within-Cantonese correlations only /p/ $\sim$ /t/ was significant ($r$= 0.59; $p$=0.003), though the correlation for /p/ $\sim$ /k/ was marginal ($r$= 0.44; $p$=0.08). 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). 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).

Chodroff and Wilson (2017) also repeat the correlation analysis in a way the coarsely accounts for speaking rate. This consideration is important, as the local speaking rate is known to influence long-lag VOT in spontaneous speech (Stuart-Smith et al., 2015) and because prior work demonstrates both talker and language effects on speech rate (Bradlow et al., 2017). In comparing the two versions of the correlation analysis, Chodroff and Wilson found that "the magnitudes of the correlations among voiceless stops did not deviate from the original magnitudes, demonstrating that differences among talkers in the realization of these sounds cannot be reduced to talker-specific speaking rates" (2017, p. 34).

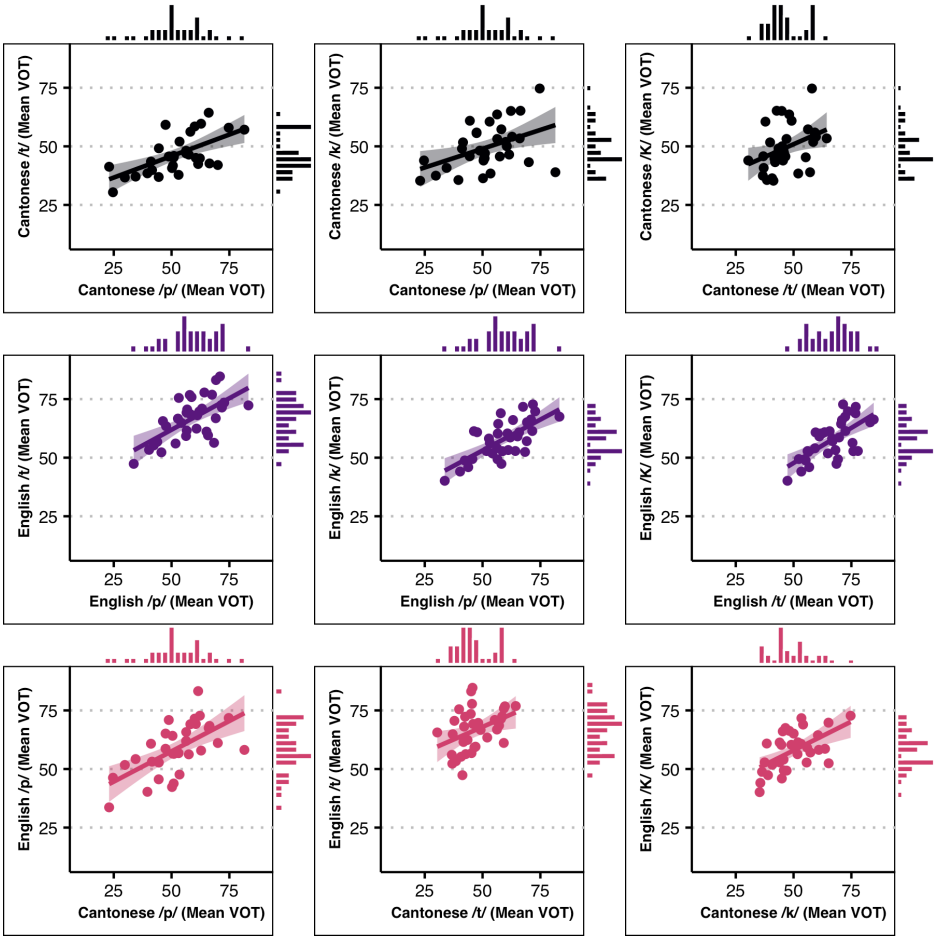


**Figure 4.3:** Correlations!

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

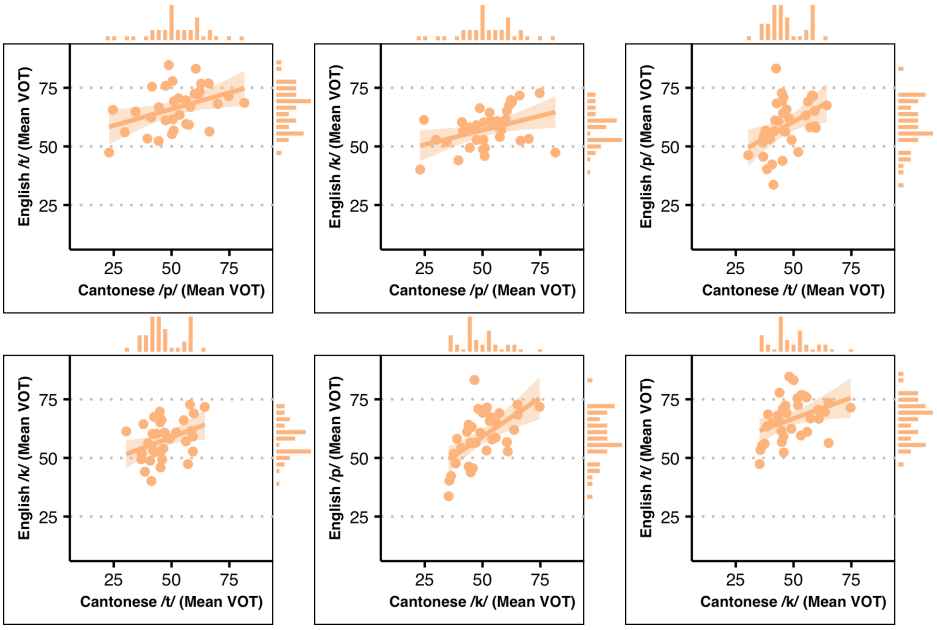


**Figure 4.4:** Correlations!

difference can likely be attributed to the generally weak correlations found.

While these relationships indicate some degree of articulatory reuse, the overall picture is far from compelling, particularly when considered alongside the results of the analysis of the ordinal relationships in Section 4.3.1. Compared to prior work, these correlations are less consistent and generally weaker. Section 4.3.3 digs into why this might be the case.

### 4.3.3 Linear mixed effect model

To better account for variation due to known factors such as speech rate and the presence of a preceding pause, a linear mixed-effects model was fit with the *lme4* package (Bates et al., 2015) in R (R Core Team, 2020). The model's aims were two-fold: estimating the effect of language by segment and elucidating the sources of variation in the random effects structure. Analyzing the data in this way mitigates the issues with conducting statistical tests on talker means, which ignores much of

**Table 4.3:** All 15 correlations based on raw mean VOT—and separately, VOT residuals 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 |

the variation in the data (Haines et al., 2020). The model's lmer formula was: VOT $\sim 1 + \text{Place} \times \text{Language} + \text{Average Phone Duration} + \text{Preceding Pause} + (\text{Place} \times \text{Language} \mid \text{Talker}) + (1 \mid \text{Word})$. The parameters were specified as follows I would have like to do everything Bayesian in my thesis but do not feel like I have the time.

**VOT** was the dependent variable—it was standardized.

**Average Phone Duration** represents the average duration of phones within the word. It was calculated as the difference between the word's AutoVOT-estimated onset and the force-aligned word offset, divided by the number of segments in the canonical form of the word. As noted earlier, average phone duration serves as a proxy for local speaking rate. A word-internal measure

is desirable, as many tokens were preceded by a pause and thus lack the preceding context from which to calculate it. Average phone duration was also standardized.

**Preceding Pause** indicates whether or not the token occurred after a pause or not. Pauses were defined as X. Preceding pause is a binary variable with weighted effect coding (False $= -0.33$, True $= 1$).

**Language** is a binary variable indicating whether a token occurred in an English or Cantonese word. It was weighted effect coded (Cantonese $= -1.61$, English $= 1$).

**Place** is a categorical variable indicating whether a token was produced with a labial, coronal, or dorsal place of articulation. Place had three levels and was weighted effect coded (Place T: /p/ $= -1.92$, /t/ $= 1$, /k/ $= 0$ ; Place K: /p/ $= -3.44$, /t/ $= 10$, /k/ $= 1$).

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. As apparent in the above list, the categorical fixed effects all employed weighted effect coding, using the *wec* R package (Nieuwenhuis et al., 2017). This decision follows Chodroff and Wilson (2017) and facilitates the interpretation of the simple effects in light of the interaction term (Brehm and Alday, 2021). Additionally, random intercepts for Talker and Word were included, as were by-Talker slopes for Language, Place, and their interaction. While this does not represent the maximal model (Barr et al., 2013), it instead follows guidelines for parsimonious model building (Bates et al., 2018), in which the parameters of direct interest are included as random slopes, and the controlling parameters are not.

Models had convergence issues mehhh

At an $\alpha = 0.01$ threshold, the model returned a significant intercept ($\beta = 0.18$, $SE = 0.49$, $p < 0.001$), significant main effects for Average Phone Duration

**Table 4.4:** Fixed-effects results from the linear mixed effects model.

| **Fixed Effect** | $\beta$ | ***SE*** | ***p*** |
|---|---|---|---|
| Intercept | 0.19 | 0.05 | $< 0.001$ |
| Place (K) | -0.02 | 0.03 | 0.41 |
| Place (T) | 0.05 | 0.04 | 0.22 |
| Language (English) | 0.16 | 0.03 | $< 0.001$ |
| Average Phone Duration | 0.32 | 0.01 | $< 0.001$ |
| Preceding Pause (True) | 0.12 | 0.02 | $< 0.001$ |
| Place (K) $\times$ Language (English) | 0.04 | 0.02 | 0.04 |
| Place (T) $\times$ Language (English) | -0.01 | 0.03 | 0.69 |

**Table 4.5:** Random effects results from the linear mixed effects model.

| **Group** | **Random effect** | ***SD*** |
|---|---|---|
| Talker | Intercept | 0.24 |
| | Place (K) | 0.07 |
| | Place (T) | 0.10 |
| | Language (English) | 0.07 |
| | Place (K) $\times$ Language (English) | 0.05 |
| | Place (T) $\times$ Language (English) | 0.07 |
| Word | Intercept | 0.44 |
| Residual | | 0.77 |

($\beta = 0.32$, $SE = 0.01$, $p < 0.001$) and Preceding Pause (True; $\beta = 0.12$, $SE = 0.02$, $p < 0.001$), as well as significant simple effect for Language (English; $\beta = 0.15$, $SE = 0.03$, $p < 0.001$), indicating that VOT was longer at slower speech rates, after pauses, and in English, compared to the weighted mean. Neither Place nor its interaction with Language was significant. As one of the linear mixed-effects model goals was to assess the effect of Language across places of articulation, pairwise post-hoc comparisons were computed for Language $\times$ Place using the *emmeans* package (Lenth, 2021), with a confidence level of 0.95, and the Kenward-Roger degrees-of-freedom method. The contrast between languages

was significant for /t/ ($\beta = -0.38$, $SE = 0.09$, $p < 0.001$) and /k/ ($\beta = -0.53$, $SE = 0.10$, $p < 0.001$), but not for /p/ ($\beta = -0.09$, $SE = 0.11$, $p = 0.39$): VOT is consistently longer in English for /t/ and /k/.

The second goal of the mixed-effects analysis was to gain insight into the sources of variation through the random effects structure. Of the random effects, the intercepts for Word ($SD = 0.20$) and Talker ($SD = 0.06$) accounted for the most variation, followed by the by-Talker slope standard deviations for Language ($SD = 0.005$), Place T ($SD = 0.01$), Place T $\times$ Language ($SD = 0.005$), Place K ($SD = 0.004$) and Place K $\times$ Language ($SD = 0.002$). Talkers and words differ substantially in mean VOT, while the slopes for Place and Language effects appear more consistent across talkers.

## 4.4 Discussion

This paper reports a study of long-lag stops in Cantonese-English bilingual speech from the SpiCE corpus described in Chapter **??**. It uses the uniformity framework to assess VOT similarity within and across languages. In broad strokes, the evidence for uniformity both within and across languages was limited. The correlation analysis provides evidence for within-language uniformity and some across-language structure. The magnitudes were largely weak and moderate. These results are corroborated by the random effects structure of the linear mixed-effects model, as more of the variation is attributable to Talker intercepts than to the Language and Place slope effects. In this sense, while there is some degree of structure in VOT variation, it seems weaker than the relationships described in prior work, where strong and clear within-language patterns were observed (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019).

The far more interesting outcomes here relate to unexpected results. The ordinal relationships should be interpreted with a grain of salt, as there are several potential explanations not immediately relevant to the research question. For example, means were based on fewer tokens than in prior work (especially for /p/), which may render those proportions less reliable; and, the speech in SpiCE dif-

fers in style (conversational vs. read). This outcome is perhaps not surprising, as Chodroff and Wilson (2017) reported magnitude differences between isolated citation form and connected read speech. Lastly, the error often overlaps in Figure XX, potentially making the ordinal relationships less reliable or meaningful. Another unexpected outcome is that English VOT seems to be consistently longer than in Cantonese—the opposite of what prior work suggested (Clumeck et al., 1981; Lisker and Abramson, 1964). No explanation is offered here other than to reiterate the casual speech style under examination. Additionally, lab and corpus results often differ (Gahl et al., 2012; Chodroff and Wilson, 2017), as do corpus studies of monolingual and bilingual speech Johnson (2019).

While the results here do not necessarily provide evidence for a bilingual crosslinguistic uniformity constraint, they offer insight into what makes bilingual speech unique and provide an empirical description of bilingual long-lag stops. In terms of describing the relationship between the long-lag stops in each language, talkers maintain a crosslinguistic contrast despite stops' proximity—for many talkers—in the long-lag space. The contrast is a strong candidate for a composite category in SLM-r (Flege and Bohn, 2021) and merits further investigation.

A lack of strong crosslinguistic uniformity has implications for speech 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). If the results of this study stand, then such a perceptual strategy may have limited use in real communicative contexts, whether or not listeners use it in a lab setting. Overall, this study highlights the need to study spontaneous speech and offers a first pass at leveraging the methods of the uniformity framework to better understand crosslinguistic similarity.

# Bibliography


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

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. → page 20

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2018). Parsimonious mixed models. *arXiv:1506.04967 [stat]*, pages 1–21. ArXiv preprint: 1506.04967. → page 20

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. → page 18

Bauer, R. S. and Benedict, P. K. (1997). *Modern Cantonese Phonology*. De Gruyter Mouton, Berlin. → page 9

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., and Mok, K. (2011). Language- and talker-dependent variation in global features of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 356–359, Hong Kong. → page 9

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts l2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899. → page 16

Brehm, L. and Alday, P. M. (2021). A decade of mixed models: It's past time to set your contrasts. Technical report, Open Science Framework. → page 20

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

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

Chan, A. Y. W. and Li, D. C. S. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' english pronunciation problems. *Language, Culture and Curriculum*, 13(1):67–85. → page 9

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

Cho, T. and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 27(2):207–229. → pages 12, 16

Chodroff, E. and Baese-Berk, M. (2019). Constraints on variability in the voice onset time of L2 English stop consonants. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 661–665, Melbourne, Austrailia. Australasian Speech Science and Technology Association Inc. → pages 9, 10, 13, 16, 22

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47. → pages 8, 9, 10, 12, 13, 16, 20, 22, 23

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4(s2). → page 10

Clumeck, H., Barton, D., Macken, M. A., and Huntington, D. A. (1981). The aspiration contrast in Cantonese word-initial stops: Data from children and adults. *Journal of Chinese Linguistics*, 9(2):210–225. → pages 9, 23

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. → pages 8, 9

Flege, J. E. and Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In Wayland, R., editor, *Second Language Speech Learning: Theoretical and Empirical Progress*, pages 3–83. Cambridge University Press. → pages 1, 2, 3, 4, 10, 23

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

Fricke, M., Zirnstein, M., Navarro-Torres, C., and Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, 22(1):200–207. → page 7

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806. → page 23

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

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

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., and Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. Technical report, PsyArXiv. → pages 16, 19

Johnson, K. A. (2019). Probabilistic reduction in Spanish-English bilingual speech. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1263–1267, Melbourne, Australia. Australasian Speech Science and Technology Association Inc. → page 23

Keshet, J., Sonderegger, M., and Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction (0.91) [Computer Software]. → page 11

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.6.0. → page 21

Lieberman, P. and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge. → page 11

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

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 9, 23

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

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

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → page 9

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

Mielke, J. and Nielsen, K. (2018). Voice onset time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America*, 144(4):2166–2177. → page 9

Ménard, L., Schwartz, J.-L., and Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50(1):14–28. → page 8

Nieuwenhuis, R., Grotenhuis, M. t., and Pelzer, B. (2017). Weighted Effect Coding for Observational Data with wec. *The R Journal*, 9(1):477–485. → page 20

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

Orena, A. J., Polka, L., and Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. *The Journal of the Acoustical Society of America*, 145(4):EL303–EL309. → page 23

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → pages 13, 18

Reinisch, E., Weber, A., and Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1):75–86. → page 23

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.1.3. → page 13

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070. → page 7

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

Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6(3-4):505–549. → page 16

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

Tsui, R. K.-Y., Tong, X., and Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, 40(1):29–58. → pages 6, 7

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820. → pages 6, 9, 10