Chapter 4

The structure of voice onset time variation in bilingual sound categories

4.1 Introduction

A consequence of bilingualism is that individuals must navigate overlapping segment inventories. This paper is concerned with what languages share, if anything, in the mental representation of speech sound categories. As representation means different things across linguistic disciplines, defining and situating the term is first necessary. The approach in this chapter largely falls out of the revised Speech Learning Model (SLM-r; Flege and Bohn, 2021) and its exemplar-flavored take on what phonetic categories look like in linguistic systems with more than one language.

SLM-r is a widely used and respected model used in second language acquisition and multilingualism research. Unlike some other models in the same space, SLM-r grapples with both perception and production. SLM-r assumes that speech sound categories from different languages exist in a shared phonetic space and are subject to constraints from the perceptual and productive 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, 1995). These constraints lead SLM-r to posit that proximity leads to instability, even if what counts as close remains unclear. Considering how bilinguals are fully capable of maintaining subtle distinctions for similar sound categories across languages (e.g., Sundara et al., 2006), this is not a trivial point to make.

So, what does representation look like in this system? SLM-r outlines a few potential outcomes for sound categories in a shared system—they can assimilate or dissimilate. A relatively simple take on this is that assimilation equals shared mental representation, while dissimilation equals separate. 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. This particular characterization is salient in a recent meta-analysis of crosslinguistic influence for Spanish and English initial stop consonants. In this study, Casillas (2021) 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. This finding echoes arguments made by Bullock and Toribio (2009) on the sophistication and control that bilingual exert over their possible forms. There is no compromise but rather a wide range of forms that bilinguals can deploy according to context.

This idea of composite categories is similar to other concepts in multilingualism literature, namely that of linked categories. While the idea is pervasive, it is somewhat vaguely defined. 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). Simonet then 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). Presumably, sound categories could be linked whether they surface in dissimilated or composite (assimilated) forms. As such, the idea behind composite categories is more useful than mere links in thinking about how representation works in the bilingual mind.

Most prior work has focused on sounds that are phonologically similar, yet phonetically distinct.

One example is the comparison between initial voiceless stops in English (long-lag) and Spanish (short-lag). Despite the substantial phonetic differences, these sounds are clearly linked in the bilingual mind (Fricke et al., 2016; Antoniou et al., 2010; Goldrick et al., 2014; Sundara et al., 2006). The studies cited here all examine initial voice-onset time (VOT) for bilinguals who speak English and a language with a different initial voicing contrast—Greek, Spanish, or French—and demonstrate convergence 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 influence on English long-lag stops from the short-lag category in the other language. Second, bilinguals appear more likely to produce lead voicing in initial English voiced stops compared to English monolinguals (Sundara et al., 2006). In both cases, evidence of crosslinguistic influence arises from comparing bilinguals to monolinguals. Corpus research demonstrates 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; Bullock and Toribio, 2009).

The studies mentioned so far focus on VOT, but represent a small subset of the crosslinguistic influence literature. There are many examples of contrasts that are maintained across languages, yet still subject to crosslinguistic influence—for example, with vowels (Guion, 2003), laterals (Amengual, 2018; Barlow, 2014), and fricatives (Peng, 1993)).

The ability to examine crosslinguistic influence between phonetically and/or phonologically similar sounds hinges on the presence of an observable difference

under some set of conditions. This observable difference could take any number of forms—acoustic, gestural, or cognitive (i.e., retrieval time). The sounds typically selected are not discussed as being the same—phonetic character choice notwithstanding. As such, links tend to be described as connecting similar and subject-to-influence sounds that ultimately have distinct representations in either the phonetics, the phonology, or both (Antoniou et al., 2010; Simonet, 2016; Bullock and Toribio, 2009). In the revised Speech Learning Model (SLM-r) (Flege and Bohn, 2021) introduced earlier, these examples would be considered composite categories—combined distributions of phonetic information from linked categories that presumably retain "peaks" for each language. While composite categories are widely attested, there are fewer good examples of full category convergence, at least in the early bilingualism literature. One example comes from a lab-based study of Mandarin-English bilingual children in which highly proficient 5–6 year olds did not differ in VOT across Mandarin and English long-lag stops, despite differences across the monolingual comparison groups (Yang, 2019). This suggests that the difference is either too small to maintain or that 5–6 year old children have not yet mastered it. The claims in (Yang, 2019) should be tempered, however, as language mode was not well-controlled for 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 clearly 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 linkages—and accompanying crosslinguistic influence—should be the long-lag stops in each language. The null result with balanced bilinguals is thus unsurprising. This is not to suggest that the (Tsui et al., 2019) would have gotten more insightful results by comparing long-lag

to long-lag, but rather to highlight that paradigms designed to modulate crosslinguistic influence tend to focus on *telling things apart*, as opposed to *telling things together*.

English and Cantonese initial long-lag stops are strong candidates for shared underlying representation, because they exhibit both phonetic and phonological *similarity* akin to the difference for Mandarin and English in (Yang, 2019). In an overview chapter on crosslinguistic segment similarity, (Chang, 2015) argues that the notion of similarity is best captured abstractly, by relative within-inventory position as opposed to physical characteristics. In an example from (Chang, 2015), English and Mandarin /u/ are considered to be linked—both occupy the highest, backest, rounded position—despite English /u/-fronting rendering it more physically similar to Mandarin /y/. This abstract "relative phonetics" elegantly accounts for various phenomena (Chang, 2015), while simultaneously shying away from making claims about whether or not segments share a mental representation or theoretical phonological specifications across languages.

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 representation at the phonetic/gestural/mental level. This common focus on telling things apart is likely an artifact of commonly-used paradigms requiring differences to detect influence. Alternatively, comparisons of categories that already show strong evidence of both phonological and phonetic similarity may be taken for granted and not considered an interesting problem to focus on, despite the nature of mental representation of sound categories being a key focus of psycholinguistics in general—especially in perception (Samuel, 2020). But also in production... (CITE) In the interest of understanding mental representation, the best candidates would be the hardest to distinguish using surface forms in the first place.

The present study is focused instead on *telling things together*, and in doing so 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 phonological primitives (e.g., features) are implemented systematically in speech production (Chodroff and Wilson, 2017; Faytak, 2018; Ménard et al., 2008). Put differently, if a set of segments share a phonological feature, that feature should be implemented 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, the relationship between laryngeal gesture and acoustic consequence is clear. While there are straightforward ties to theoretical phonology from articulatory uniformity, the selection of a particular framework is not a straightforward task in a bilingual context. English and Cantonese stops are typically analyzed with different distinctive features—[voice] and [spread glottis], respectively—despite surfacing with long-lag VOT in initial position and occupying the same relative position. The study reported here focuses only on the relative phonetics and sidesteps theoretical phonology for the time being. This is consistent with the argument that theoretical linguistic descriptions do not always neatly map onto psycholinguistic phenomena (Samuel, 2020).

Within-language uniformity has been observed for initial stops in non-native English, such that the relationship between stops for an individual is clear even if between-talker variability is larger than for native speakers (Chodroff and Baese-Berk, 2019). 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, however, follows the conceptualization of uniformity arising from articulatory reuse (Faytak, 2018). In the case of an early Cantonese-English bilinguals, 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 in this case would be advantageous given the small difference across

monolingual populations, that may or may not be perceptible. While this remains an empirical question, it follows 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, and 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 of speech in Cantonese and English (Johnson et al., 2020). The corpus includes recordings of 34 early Cantonese-English bilinguals (half female, half male) in both languages, with the order of languages counterbalanced. SpiCE also includes hand-corrected orthographic and force-aligned phone level transcripts. The design of the SpiCE corpus is well-suited to the present study, as it includes comparable samples of spontaneous speech from the same set of individuals in two languages, though it differs from prior studies that use larger read speech corpora (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019).

4.2.2 Segmentation & measurement

All instances of prevocalic word-initial /p t k/ were identified from the SpiCE corpus' force-aligned TextGrid transcripts ($n = 13,488$). VOT estimates were refined using AutoVOT (Keshet et al., 2014), with the minimum allowed VOT value set to 15 ms. AutoVOT identifies the onset and offset of positive VOT within a specified window (here, force-aligned boundaries +/- 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. 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 that the AutoVOT offset did not fall within the original force-aligned boundaries of the word ($n = 600$), if the previous word was unknown (i.e., unintelligible or in a different language; $n = 268$), if VOT was equal to the minimum value of 15 ms ($n = 618$), or if items had a VOT more than 2.5 s.d. above the grand mean (> 127.8 ms; $n = 249$). Lastly, following (Chodroff and Wilson, 2017), instances of the English word "to" were excluded from the analysis given its propensity for reduction and extremely high frequency ($n = 2295$).

Of the initial sample, 29.9% was excluded, resulting in 9,458 long-lag stops, with Cantonese /p/: $n = 374$, /t/: $n = 1376$, and /k/ $n = 1687$; and English /p/ $n = 1129$, /t/ $n = 1497$, and /k/ $n = 3395$. Talkers had a median of 97 Cantonese stops (range: 59-194) and 166 English stops (range: 69-574). The higher number of English stops is likely due to lexical distributional reasons. The SpiCE corpus has a similar amount of recorded speech in each language, and while Cantonese stops were culled at a slightly higher rate in the exclusions specified above, they made up a smaller proportion to begin with (33% of initial sample vs. 29% of sample before excluding "to"). English also seems to have more highly frequent /k/-initial word types. Conversely, Cantonese /p/ occurs in fewer, less frequent word types in the final sample ($n = 60$, max frequency of 97) than English ($n = 185$, max frequency of 214).

4.3 Analysis & Results

The articulatory uniformity framework offers strong theoretical grounds for interpreting the structure of VOT variation within and across talkers. The analysis qualifies and quantifies that structure from a few different perspectives. In all cases, the pattern of results is depicted by Figure XX, which plots individuals' mean and standard errors for each of the three stops by language—showcasing both variability and commonalities.

4.3.1 Ordinal relationships

Prior work with lab and read speech strongly suggests an expected ordinal relationship for VOT across places of articulation: /p/ $<$ /t/ $<$ /k/. One of the major contributions of (Chodroff and Wilson, 2017) is that these relationships are tighter that 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 for the SpiCE corpus suggest that *puzzle* is an appropriate characterization, as talkers largely did not adhere the expected order. Table 4.1 reports the proportion of talkers whose mean VOT values followed the expected /p/ $<$ /t/ $<$ /k/ relationships. Prior work on connected speech reports rates of adherence in the 80-90% range (Chodroff and Baese-Berk, 2019), with the exception of English /t/ $<$ /k/ being drastically lower for native English speakers. While the /t/ $<$ /k/ comparison is also low here (18%), only the English /p/ $<$ /t/ proportion (0.74) is at all close to previous work. This lack of adherence is apparent in... many cases the standard errors overlap, suggesting that a strict ordering by means may not be appropriate. Additionally, many talkers are not internally consistent across languages... this is depicted by... LOTS TO ADD HERE

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 and are reported

Table 4.1: 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 sample.

Language	**p<t**	**t<k**	**p<k**	n
Cantonese	0.27	0.61	0.40	33
English	0.74	0.18	0.41	34

along with Holm-adjusted p-values where significant. In each case, means were calculated over *residual* VOT values from a simple linear regression in which VOT was predicted by average phone duration within the word—a proxy for speech rate calculated as the 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. Using residual VOT means mitigates the impact of talker- and language-specific speech rate for these comparisons. This is important, as speech rate is known to influence VOT (Chodroff and Wilson, 2017), and because prior work demonstrate talker and language effects on speech rate (Bradlow et al., 2017).

Table 4.2 summarizes the output of the significant correlations. While there is some evidence for both within- and across-language structured variation, the correlations reported here are considerably lower compared to prior work on English connected speech, where 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/ ($r = 0.75$, $p < 0.001$), all of the correlations were either moderate ($0.5 < r < 0.7$; $p < 0.01$) or not significant. Within-language correlations more consistently occurred (5 of 6 significant), compared to the across-language comparisons (3 of 9). Notably, most of the comparisons involving /t/ in either language, were not significant. While these relationships seem to indicate some degree of articulatory reuse, the overall picture is not particularly compelling.

Table 4.2: Correlations based on mean residual VOT by talker and language. Each row indicates the comparison, Pearson's *r*, and Holm-adjusted p-value.

Comparison	***r***	***p***
Cantonese /p/ ∼ Cantonese /t/	0.59	0.004
Cantonese /p/ ∼ Cantonese /k/	0.54	0.009
Cantonese /t/ ∼ Cantonese /k/	0.33	0.28
English /p/ ∼ English /t/	0.58	0.004
English /p/ ∼ English /k/	0.75	<0.001
English /t/ ∼ English /k/	0.57	0.005
Cantonese /p/ ∼ English /p/	0.57	0.006
Cantonese /t/ ∼ English /t/	0.31	0.29
Cantonese /k/ ∼ English /k/	0.55	0.006
Cantonese /p/ ∼ English /t/	0.23	0.33
Cantonese /p/ ∼ English /k/	0.35	0.29
Cantonese /t/ ∼ English /p/	0.43	0.08
Cantonese /t/ ∼ English /k/	0.31	0.29
Cantonese /k/ ∼ English /p/	0.56	0.006
Cantonese /k/ ∼ English /t/	0.24	0.33

4.3.3 Linear mixed effect model

In an effort to better account for variation due to known factors such as speech rate and the presence of a preceding pause, a linear mixed effect model was fit with the *lme4* R package (Bates et al., 2015). The aims of the model were two-fold: estimating the effect of language by segment, and elucidating the sources of variation in the random effect structure. The dependent variable, VOT (centered) was predicted by Average Phone Duration (standardized), Preceding Pause (False$= -0.32$, True$= 1$), Language (Cantonese$= -1.75$, English$= 1$), Place of Articulation (Place T: /p/$= -1.91$, /t/$= 1$, /k/$= 0$; Place K: /p/$= -3.38$, /t/$= 10$, /k/$= 1$), and the Language $\times$ Place interaction. As likely apparent from the parenthetical values, all categorical fixed effects were weighted effect coded (following Chodroff and Wilson, 2017). Random intercepts for Talker and Word were in-

cluded, as were by-Talker slopes for Language, Place, and their interaction.[1]

The model returned a significant intercept ($\beta = 3.62$, $SE = 1.22$, $p = 0.004$), significant main effects for Average Phone Duration ($\beta = 7.75$, $SE = 0.23$, $p < 0.001$) and Preceding Pause (True; $\beta = 2.96$, $SE = 0.38$, $p < 0.001$) as well as significant simple effect for Language (English; $\beta = 2.81$, $SE = 0.59$, $p < 0.001$), indicating that VOT was longer at slower speech rates, as well as after pauses and in English, compared to the weighted mean. Neither Place nor its interaction with Language was significant. As one of the mixed effect model analysis goals was to assess the effect of Language across places of articulation, pairwise post-hoc comparisons were computed for Language by Place of Articulation using emmeans, with a confidence level of 0.95, and the Kenward-Roger degrees-of-freedom method. The contrast between languages was significant for /t/ ($\beta = -7.96$, $SE = 2.25$, $p < 0.001$) and /k/ ($\beta = -9.66$, $SE = 2.43$, $p < 0.001$), but not for /p/ ($\beta = -0.81$, $SE = 2.28$, $p = 0.78$). This suggests that 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 ($SD = 11.45$) and Talker ($SD = 6.11$) accounted for the most variation, followed by the by-Talker slope standard deviations for Language ($SD = 1.76$), Place T ($SD = 2.76$), Place T $\times$ Language ($SD = 1.53$), Place K ($SD = 1.80$) and Place K $\times$ Language ($SD = 1.03$). This indicates that talkers and words differ substantially in mean VOT, and that the slopes for Place and Language effects are 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 (Johnson et al., 2020), and uses the uniformity framework to assess VOT similarity within and across languages. In broad strokes, the ev-

[1]Formula: VOT $\sim$ 1 + Place $\times$ Language + Average Phone Duration + Preceding Pause + (Place $\times$ Language | Talker) + (1 | Word).

idence for uniformity both within and across languages was limited. A correlation analysis provides evidence for within-language uniformity and some across-language structure. The magnitudes were mostly moderate, and most did not involve coronal stops. 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 to be weaker than the evidence in prior work, where strong within-language patterns were observed (Chodroff and Wilson, 2017; Chodroff and Baese-Berk, 2019).

The far more interesting outcomes relate to unexpected results. The ordinal relationships should be interpreted with a grain of salt, as there are a number of potential explanations not immediately relevant to the research question. For example, means were based off of fewer tokens than in prior work (especially for /p/), which may render those proportions less reliable; and, the speech in SpiCE differs in style (conversational vs. read). Lastly, the error often overlaps, potentially making the ordinal relationships unreliable or less 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, and that lab and corpus results often differ (Gahl et al., 2012), as do corpus studies of monolingual and bilingual speech (Johnson, 2019).

While the results here do not necessarily provide evidence for a crosslinguistic uniformity constraint, they offer insight into what makes bilingual speech unique, as well as empirical descriptions of bilingual long-lag stop. In terms of describing the relationship between the long-lag stops in each language, talkers seem to maintain a crosslinguistic contrast despite the close proximity of the stops—for many talkers—in the long-lag space. This makes a composite category in SLM-r terms seem plausible (Flege and Bohn, 2021), and merits further investigation.

A lack of strong cross-language uniformity has implications for speech perception, in which tracking a uniformity-like constraint has been proposed as mecha-

nism 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 persist, then such a constraint may have limited use in real communicative contexts, whether or not listeners use it in a lab setting. On the whole, 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 phonetic 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 3

Antoniou, M., Best, C. T., Tyler, M. D., and Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, 38(4):640–653. → pages 3, 4

Barlow, J. A. (2014). Age of acquisition and allophony in Spanish-English bilinguals. *Frontiers in Psychology*, 5. → page 3

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 11

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts l2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899. → page 10

Bullock, B. E. and Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, L., Winford, D., and deBot, K., editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam. → pages 2, 3, 4

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. → page 2

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 5

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 6, 7, 9, 10, 13

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 6, 7, 8, 9, 10, 11, 13

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 7

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 6, 13

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. → page 6

Flege, J. E. (1995). Second-language speech learning: theory, findings, and problems. In Strange, W., editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 233–277. York Press, Timonium, MD. → page 2

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, 4, 7, 13

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. → page 3

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 13

Goldrick, M., Runnqvist, E., and Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, 25(4):1031–1036. → page 3

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, 60(2):98–128. → pages 2, 3

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 13

Johnson, K. A., Babel, M., Fong, I., and Yiu, N. (2020). SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4089–4095, Marseille, France. European Language Resources Association. → pages 7, 12

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 8

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 2

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 6, 13

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 6

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 14

Peng, S.-h. (1993). Cross-language influence on the production of Mandarin /f/ and /x/ and Taiwanese /h/ by native speakers of taiwanese amoy. *Phonetica*, 50(4):245–260. → page 3

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 14

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070. → pages 5, 6

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. → pages 2, 3, 4

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 2, 3

Tsui, R. K.-Y., Tong, X., and Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, 40(1):29–58. → page 4

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820. → pages 4, 5, 7