INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Leveraging the uniformity framework to examine crosslinguistic similarity for long-lag stops in spontaneous Cantonese-English bilingual speech

Khia A. Johnson

Department of Linguistics, University of British Columbia, Canada
khia.johnson@ubc.ca

Abstract

While crosslinguistic influence is widespread in bilingual speech production, it is less clear which aspects of representation are shared across languages, if any. Most prior work examines phonetically distinct yet phonologically similar sounds, for which phonetic convergence suggests a cross-language link within individuals [1]. Convergence is harder to assess when sounds are already similar, as with English and Cantonese initial long-lag stops. Here, the articulatory uniformity framework [2, 3, 4] is leveraged to assess whether bilinguals share an underlying laryngeal feature across languages, and describe the nature of cross-language links. Using the SpiCE corpus of spontaneous Cantonese-English bilingual speech [5], this paper asks whether Cantonese-English bilinguals exhibit uniform voice-onset time for long-lag stops within and across languages. Results indicate moderate patterns of uniformity within-language—replicating prior work [2, 6]—and weaker patterns across languages. The analysis, however, raises many questions, as correlations were generally lower compared to prior work, and talkers did not adhere to expected ordinal relationships by place of articulation. Talkers also retained clear differences for /t/ and /k/, despite expectations of similarity. Yet at the same time, more of the overall variation seems to derive from individual-specific differences. While many questions remain, the uniformity framework shows promise.

Index Terms: bilingual speech, articulatory uniformity, crosslinguistic influence, voice-onset time, corpus linguistics

1. Introduction

A consequence of bilingualism is that individuals must navigate overlapping segment inventories [7]. This paper is concerned with what languages share, if anything, in the mental representation of speech sounds. Most prior work has focused on phonologically similar yet phonetically distinct sounds, as with the comparison between initial voiceless stops in English (long-lag) and Spanish (short-lag). Despite substantial phonetic differences, these sounds are clearly linked [8, 9, 10, 11]. 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 (e.g., Spanish) 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 the influence on English long-lag stops from short-lag stops in the other language. Second, bilinguals are more likely to produce lead voicing in initial English voiced stops than English monolinguals [11]. In both cases, evidence of crosslinguistic influence (CLI) 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 [8, 12]. The studies mentioned here focus on VOT—a small subset of the CLI literature. There are many examples of contrasts maintained across languages yet still subject to CLI (e.g., vowels [13], laterals [14, 15], and fricatives [16]).

The ability to examine CLI between similar sounds hinges on the presence of an observable difference under some set of conditions. The sounds typically selected are not discussed as being the same—phonetic character choice notwithstanding. As such, links are described as connecting similar, subject-to-influence sounds that presumably retain distinct representations [9, 17, 12]. In the revised Speech Learning Model (SLM-r) [7], these examples could be considered composite categories: combined distributions of phonetic information from linked categories that retain "peaks" for each language. While composite categories can be readily identified, there are fewer good examples of full category convergence, at least in the early bilingualism literature. One example comes from a lab study of Mandarin-English bilingual children: highly proficient 5 to 6-year-olds did not differ in VOT across Mandarin and English long-lag stops, despite differences across the monolingual comparison groups [18]. This result suggests the difference is either too small to maintain or that 5 to 6-year-old children have not yet mastered it. However, the claims in [18] should be tempered as language mode was not well-controlled for and adult bilingual behavior was not considered.

Despite some inroads, there is a distinct paucity of work examining highly similar speech sounds across languages, even when such a comparison would make sense. A recent CLI study of Cantonese-English bilinguals compares English long-lag and Cantonese short-lag stops in a language switching paradigm [19]. This comparison reflects the need for acoustically distinguishable stimuli but glosses over the fact that both languages contrast short-lag and long-lag VOT in initial position. The best candidates for CLI and links are the long-lag stops in each language. The null result in [19] is thus unsurprising, though it does not necessarily suggest that comparing long-lag with long-lag would have led to more insightful results. Instead, it highlights that paradigms designed to modulate CLI emphasize *telling things apart*, as opposed to *telling things together*.

The idea of telling things apart or together fits within SLM-r [7], where categories from different languages exist in a shared phonetic space. These categories are subject to constraints from the perceptual and productive systems: don't get too close to each other in perception and don't get too complicated in production [13, 20, 21]. SLM-r assumes that close proximity leads to instability but fails to define what counts as close. Considering the proximity that bilinguals are capable of maintaining, this is not a trivial point to make. Assuming that convergence is one outcome of proximity, the original SLM would argue that if two segments sound the same, they must share a representation [21]. Note, however, that this does not necessarily apply to composite categories where differences persist.

English and Cantonese initial long-lag stops are strong

http://dx.doi.org/10.21437/Interspeech.2021-1780

candidates for shared underlying representation. They exhibit both phonetic and phonological similarity, like the Mandarin-English comparison in [18]. Similarity is arguably best captured abstractly by relative within-inventory position as opposed to physical characteristics [1]. In an example from [1], English and Mandarin /u/ are linked—both occupy the highest, most back, rounded position—despite English /u/-fronting rendering it more physically similar to Mandarin /y/. "Relative phonetics" elegantly accounts for various phenomena [1] but refrains from making claims about whether or not segments share representation or theoretical phonological specifications.

To summarize, most work in CLI has focused on phonologically similar yet phonetically distinct pairs of segments, which are not strong candidates for shared representation. This focus on telling things apart likely derives from commonly-used paradigms requiring differences to detect CLI. Alternatively, comparisons of highly similar categories may not be considered an interesting problem for researchers, even if the nature of representation is a key focus in psycholinguistics—especially in perception [22].

The present study is focused instead on telling things together, and to do so, extends the *articulatory uniformity framework* to the study of bilingual 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 [2, 3, 4]. Put differently, if segments that share a phonological feature should implement it with the same phonetic target or articulatory gesture (which may or may not have an acoustic consequence). This systematicity has been observed for vowel height [4], tongue shape [3], fricative peak frequency, and stop consonant VOT [2]. In the case of VOT, the relationship between a laryngeal gesture and acoustic consequence is clear. While there are clear ties between theoretical phonology and articulatory uniformity, choosing a particular phonological framework is not straightforward 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. This study focuses only on relative phonetics of [1]—sidestepping theoretical phonology—and is consistent with the observation that theoretical linguistic descriptions do not always neatly map onto psycholinguistic phenomena [22].

Within-language uniformity has been observed for initial stops in non-native English [6]. However, the uniformity framework has not yet been extended to bilingual speech, particularly as a mechanism for comparing how bilinguals produce similar sounds across languages. Leveraging the framework in this way follows the conceptualization of uniformity as arising from articulatory reuse [3]. In the case of early Cantonese-English bilinguals, consider the initial stop [k^h] with a mean VOT of 80 ms in American English [23] and 91 ms in Hong Kong Cantonese [24]. 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 and possibly imperceptible difference across monolingual populations. While this remains an empirical question, it follows the finding that bilingual Mandarin-English children did not distinguish between languages in VOT [18]. Following the predictions of SLM-r [7], long-lag items with minimally distinct VOT are more likely to assimilate or dissimilate than coexist in such close proximity.

This study asks: Do Cantonese-English bilinguals uniformly produce long-lag stops within and across each of their languages? Given their proximity [7] and evidence of uniformity within non-native English speech [6], we predicted that early Cantonese-English bilinguals would uniformly implement long-lag stops within and across languages. Regardless of the results, the methodology from [2, 6] allows for a new perspective on the structure of variation and nature of representation in bilinguals and facilitates the study of already similar speech sounds in ways that other paradigms do not.

2. Methods

2.1. Corpus

This study uses conversational interviews from the SpiCE corpus of speech in Cantonese and English [5, 25]. The corpus includes recordings of 34 early Cantonese-English bilinguals in both languages (half female, half male; order of languages counterbalanced). SpiCE also includes hand-corrected orthographic and force-aligned phone level transcripts [26]. 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. However, it differs from prior studies that use larger read speech corpora [2, 6].

2.2. Segmentation & measurement

All instances of prevocalic word-initial /p t k/ were identified from the SpiCE corpus' force-aligned TextGrid transcripts ($n = 10,428$). For English, only items with initial stress were included in the initial sample [27].[1] While forced alignment performed reasonably well—anecdotally—VOT estimates were refined using AutoVOT [28], 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 segment 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 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, with Cantonese /p/: $n = 374$, /t/: $n = 1373$, and /k/: $n = 1688$; and English /p/: $n = 1035$, /t/: $n = 1336$, and /k/: $n = 3155$. 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, as the SpiCE corpus has a comparable amount of recorded speech in each language. 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).

[1] Note that this means the extremely high-frequency word "to" was excluded, as in [2].

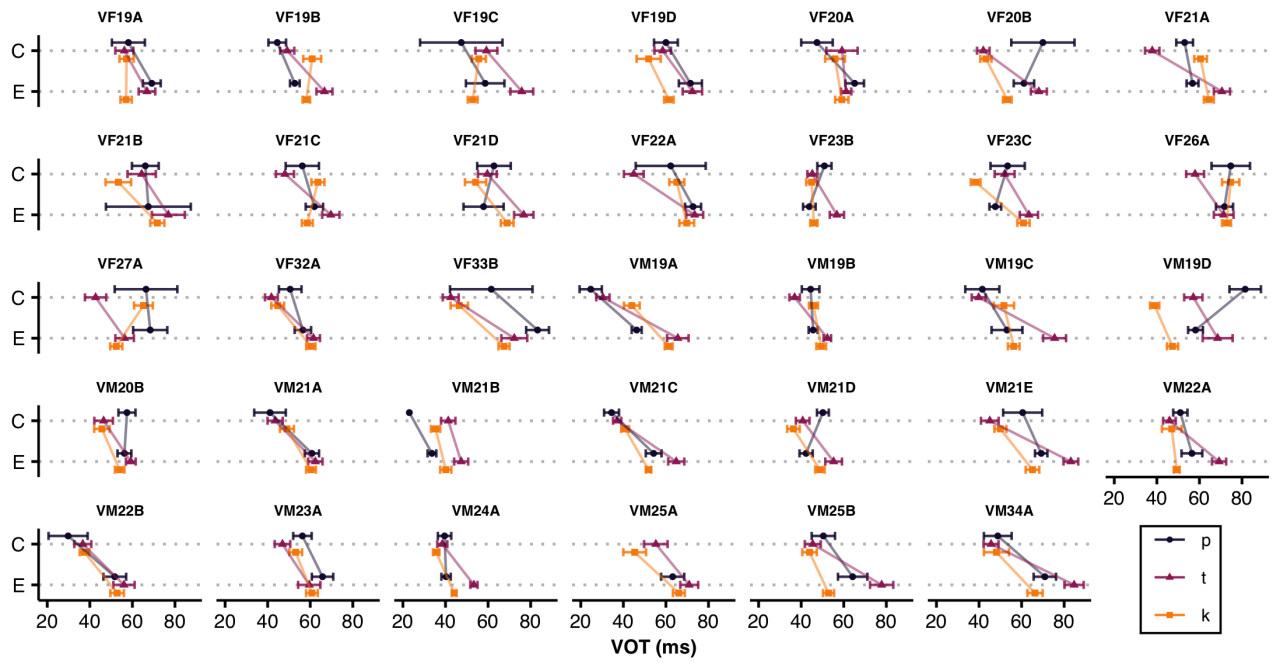

Figure 1: *Mean and SE for VOT across place of articulation, language, and individuals for C(antonese) and E(nglish). The vertical offset is included to facilitate visualizing overlapping segments.*

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. In all cases, the pattern of results is depicted by Figure 1, which plots individuals' mean and standard errors for each of the three stops by language—showcasing both variability and commonalities.

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 [2] 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. Table 1 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 [6]. While the English /t/ < /k/ comparison is remarkably low here (6%), only English /p/ < /t/ (82%) falls in the range prior work suggests. This lack of adherence is apparent in the relative ordering or markers in Figure 1, though in many cases, the standard errors overlap, suggesting that a strict ordering by means may be inappropriate. Additionally, crossed lines in Figure 1 indicated that many talkers are not internally consistent across languages.

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 along with Holm-adjusted p-values, using the *psych* [29] package in R [30]. 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 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 consideration is important, as speech rate is known to influence VOT [2] and because prior work demonstrates both talker and language effects on speech rate [31].

Table 2 summarizes the output of all 15 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 connected speech. Similar within-language comparisons had $r > 0.7$ [2, 6]. With the exception of the English /p/ ∼ /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-language correlations occurred more consistently occurred (5 of 6 significant) than across-language comparisons (3 of 9). While these relationships indicate some degree of articulatory reuse, the overall picture is not particularly compelling.

Table 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/.*

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

Table 2: *All 15 correlations based on mean residual VOT by talker and language. Each row indicates the comparison, Pearson's r, and Holm-adjusted p-value. The header row specifies the languages in each pairwise comparison in order: C(antonese) or E(nglish).*

C ∼ C	***r***	***p***	**E ∼ E**	***r***	***p***	**C ∼ E**	***r***	***p***	**C ∼ E**	***r***	***p***	**C ∼ E**	***r***	***p***
p ∼ t	0.59	0.003	p ∼ t	0.63	<0.001	p ∼ p	0.57	0.006	t ∼ p	0.37	0.20	k ∼ p	0.59	0.003
p ∼ k	0.55	0.009	p ∼ k	0.70	<0.001	p ∼ t	0.29	0.31	t ∼ t	0.35	0.21	k ∼ t	0.37	0.20
t ∼ k	0.34	0.21	t ∼ k	0.60	0.002	p ∼ k	0.29	0.31	t ∼ k	0.27	0.31	k ∼ k	0.54	0.009

3.3. Linear mixed-effects 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-effects model was fit with the *lme4* package [32] in R [30]. The aims of the model were two-fold: estimating the effect of language by segment, and elucidating the sources of variation in the random effects structure. The dependent variable, VOT (standardized) was predicted by Average Phone Duration (standardized), Preceding Pause (False $= -0.33$, True $= 1$), Language (Cantonese $= -1.61$, English $= 1$), Place of Articulation (Place T: /p/ $= -1.92$, /t/ $= 1$, /k/ $= 0$; Place K: /p/ $= -3.44$, /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 [2], using the *wec* package [33]. Random intercepts for Talker and Word were included, as were by-Talker slopes for Language, Place, and their interaction.[2]

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 ($\beta = 0.32$, $SE = 0.01$, $p < 0.001$) 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 [34], 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. Discussion

This paper reports a study of long-lag stops in Cantonese-English bilingual speech from the SpiCE corpus [25]. 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 [2, 6].

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 differs in style (conversational vs. read). This outcome is perhaps not surprising, as [2] reported magnitude differences between isolated citation form and connected read speech. Lastly, the error often overlaps in Figure 1, 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 [24, 23]. No explanation is offered here other than to reiterate the casual speech style under examination. Additionally, lab and corpus results often differ [35, 2], as do corpus studies of monolingual and bilingual speech [36].

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 [7] 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 [37] and in multilingual talker identification [38]. 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.

5. Acknowledgements

I would like to thank Molly Babel, Márton Sóskuthy, Kathleen Currie Hall, and members of the Speech-in-Context lab for their comments and work on the SpiCE corpus. This research was supported by a UBC Public Scholar's initiative award.

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

6. References

[1] C. B. Chang, "Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity," in *The Segment in Phonetics and Phonology*, 1st ed., E. Raimy and C. E. Cairns, Eds. John Wiley & Sons, 2015, pp. 199–217.

[2] E. Chodroff and C. Wilson, "Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English," *Journal of Phonetics*, vol. 61, pp. 30–47, 2017.

[3] M. D. Faytak, "Articulatory uniformity through articulatory reuse: Insights from an ultrasound study of Sūzhōu Chinese," Doctoral dissertation, University of California, Berkeley, 2018.

[4] L. Ménard, J.-L. Schwartz, and J. Aubin, "Invariance and variability in the production of the height feature in French vowels," *Speech Communication*, vol. 50, no. 1, pp. 14–28, 2008.

[5] K. A. Johnson, "SpiCE: Speech in Cantonese and English [V1]," *Scholars Portal Dataverse*, 2021. [Online]. Available: https://doi.org/10.5683/SP2/MJOXP3

[6] E. Chodroff and M. Baese-Berk, "Constraints on variability in the voice onset time of L2 English stop consonants," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 661–665.

[7] J. E. Flege and O.-S. Bohn, "The revised speech learning model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*, R. Wayland, Ed. Cambridge University Press, 2021, pp. 3–83.

[8] M. Fricke, J. F. Kroll, and P. E. Dussias, "Phonetic variation in bilingual speech: A lens for studying the production-comprehension link," *Journal of Memory and Language*, vol. 89, pp. 110–137, 2016.

[9] M. Antoniou, C. T. Best, M. D. Tyler, and C. Kroos, "Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2," *Journal of Phonetics*, vol. 38, no. 4, pp. 640–653, 2010.

[10] M. Goldrick, E. Runnqvist, and A. Costa, "Language switching makes pronunciation less nativelike," *Psychological Science*, vol. 25, no. 4, pp. 1031–1036, 2014.

[11] M. Sundara, L. Polka, and S. Baum, "Production of coronal stops by simultaneous bilingual adults," *Bilingualism: Language and Cognition*, vol. 9, no. 1, pp. 97–114, 2006.

[12] B. E. Bullock and A. J. Toribio, "Trying to hit a moving target: On the sociophonetics of code-switching," in *Studies in Bilingualism*, L. Isurin, D. Winford, and K. deBot, Eds. John Benjamins Publishing Company, 2009, vol. 41, pp. 189–206.

[13] S. G. Guion, "The Vowel Systems of Quichua-Spanish Bilinguals," *Phonetica*, vol. 60, no. 2, pp. 98–128, 2003.

[14] M. Amengual, "Asymmetrical interlingual influence in the production of Spanish and English laterals as a result of competing activation in bilingual language processing," *Journal of Phonetics*, vol. 69, pp. 12–28, 2018.

[15] J. A. Barlow, "Age of acquisition and allophony in Spanish-English bilinguals," *Frontiers in Psychology*, vol. 5, 2014.

[16] S.-h. Peng, "Cross-language influence on the production of Mandarin /f/ and /x/ and Taiwanese /h/ by native speakers of taiwanese amoy," *Phonetica*, vol. 50, no. 4, pp. 245–260, 1993.

[17] M. Simonet, "The phonetics and phonology of bilingualism," in *Oxford Handbooks Online*, 2016.

[18] J. Yang, "Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults," *Second Language Research*, p. 0267658319851820, 2019.

[19] R. K.-Y. Tsui, X. Tong, and C. S. K. Chan, "Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching," *Applied Psycholinguistics*, vol. 40, no. 1, pp. 29–58, 2019.

[20] B. Lindblom and I. Maddieson, "Phonetic universals in consonant systems," in *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, L. M. Hyman and C. N. Li, Eds. Routledge, 1988, pp. 62–78.

[21] J. E. Flege, "Second-language speech learning: theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, W. Strange, Ed. York Press, 1995, pp. 233–277.

[22] A. G. Samuel, "Psycholinguists should resist the allure of linguistic units as perceptual units," *Journal of Memory and Language*, vol. 111, p. 104070, 2020.

[23] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[24] H. Clumeck, D. Barton, M. A. Macken, and D. A. Huntington, "The aspiration contrast in Cantonese word-initial stops: Data from children and adults," *Journal of Chinese Linguistics*, vol. 9, no. 2, pp. 210–225, 1981.

[25] K. A. Johnson, M. Babel, I. Fong, and N. Yiu, "SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4089–4095.

[26] M. McAuliffe, M. Socolof, E. Stengel-Eskin, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner (1.0.1) [Computer Software]," 2019. [Online]. Available: https://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[27] L. Lisker and A. S. Abramson, "Some effects of context on voice onset time in english stops," *Language and Speech*, vol. 10, no. 1, pp. 1–28, 1967.

[28] J. Keshet, M. Sonderegger, and T. Knowles, "AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction (0.91) [Computer Software]," 2014. [Online]. Available: https://github.com/mlml/autovot/

[29] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, 2021, r package version 2.1.3. [Online]. Available: https://CRAN.R-project.org/package=psych

[30] R Core Team, *R: A Language and Environment for Statistical Computing*, 2021. [Online]. Available: https://www.R-project.org/

[31] A. R. Bradlow, M. Kim, and M. Blasingame, "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*, vol. 141, no. 2, pp. 886–899, Feb. 2017.

[32] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[33] R. Nieuwenhuis, M. t. Grotenhuis, and B. Pelzer, "Weighted Effect Coding for Observational Data with wec," *The R Journal*, vol. 9, no. 1, pp. 477–485, 2017.

[34] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2021, R package version 1.6.0. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[35] S. Gahl, Y. Yao, and K. Johnson, "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," *Journal of Memory and Language*, vol. 66, no. 4, pp. 789–806, 2012.

[36] K. A. Johnson, "Probabilistic reduction in Spanish-English bilingual speech," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 1263–1267.

[37] E. Reinisch, A. Weber, and H. Mitterer, "Listeners retune phoneme categories across languages," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 1, pp. 75–86, 2013.

[38] A. J. Orena, L. Polka, and R. M. Theodore, "Identifying bilingual talkers after a language switch: Language experience matters," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. EL303–EL309, Apr. 2019.