Chapter 1

Introduction

...

Chapter 2

The SpiCE Corpus

2.1 Introduction

Most of our knowledge about spoken language and speech processing comes from monolingual individuals producing scripted speech in laboratory settings. Monolingual lab speech allows for researchers to exercise tight control over the linguistic backgrounds of the speakers and the linguistic material (e.g. reading or repeating sounds, words, or sentences). While highly informative, these controlled monolingual speech samples are not representative of spoken language in the real world. Multilingualism is the norm, not the exception, and individuals regularly make creative linguistic choices.

Crucially, corpus-based research with conversational or spontaneous speech is important in the fields of phonetics and psycholinguistics, as the research conclusions drawn from corpus and lab-based experiments do not always coincide (Gahl et al., 2012). Conversational speech allows for a more accurate empirical description of spoken language, as it represents more realistic and natural productions than scripted laboratory speech, even compared to scripted connected speech. Conversational speech also crucially permits for field testing of speech production theories (Bell et al., 2009; Gahl et al., 2012) in their natural habitats.

The discrepancies between results for conversational and lab speech have been found for monolingual (English) speech, but are likely be found with bilingual speech as well. Resources to query bilingual conversational speech are limited,

however, as the necessary resources permitting this type of inquiry are rare. As a step towards filling this gap, this paper introduces the **SpiCE** corpus of conversational bilingual **Sp**eech **i**n **C**antonese and **E**nglish. This open-access corpus was explicitly developed with phonetic bilingualism research in mind. The corpus design is based on key aspects of widely used existing corpora, such as the Buckeye corpus of conversational speech (Pitt et al., 2005). It crucially includes speech from the same individual in more than one language, as is the case in the Bangor corpora of Spanish-English, Welsh-English, and Welsh-Spanish bilingual speech (Deuchar et al., 2014), but with a more controlled recording setup, allowing for more nuanced acoustic-phonetic measurements.

The primary motivation for collecting this corpus is to have comparable high-quality recordings of conversational speech from early bilinguals in two languages, which in turn enables large scale phonetic analysis on a within-speaker basis.[1] To our knowledge, this type of resource does not yet exist for any pair of languages, much less for a typologically distinct pair like Cantonese (Sino-Tibetan) and English (Indo-European). Furthermore, Cantonese is a relatively understudied language, despite there being approximately 55 million native speakers around the world (Matthews et al., 2013), though this is changing with new corpora (Luke and Wong, 2015) and tools (Lee, 2018).

This paper provides a detailed overview of the corpus design and collection procedures, a description of the speakers, and the transcription and annotation pipeline. It concludes with descriptive statistics, and highlights key opportunities created by this corpus when it is made publicly available.

2.2 Corpus design and creation

This section provides detail about the speakers (Section 2.2.1), the procedures used to ensure high-quality recordings (Section 2.2.2), and the three tasks that each participant completed in both Cantonese and English (Section 2.2.3).

[1] SpiCE is a large speech corpus for the type of speech involved, and is comparable in size to the widely used Buckeye corpus (Pitt et al., 2005). Larger speech corpora tend to comprise existing recordings of broadcasters, phone conversations, or read speech. While these types of speech corpora are certainly useful, they are designed for different purposes.

No.	ID	Interview Order	Age	Gender	Age of Acquisition	
					English	Cantonese
1	VF19A	E → C	19	F	0	0
2	VF19B	E → C	19	F	0	0
3	VF19C	E → C	19	F	3	0
4	VF19D	C → E	19	F	2	0
5	VF20A	C → E	20	F	4	0
6	VF20B	C → E	20	F	5	0
7	VF21A	E → C	21	F	0	0
8	VF21B	C → E	21	F	3	0
9	VF21C	C → E	21	F	4	0
10	VF21D	E → C	21	F	0	0
11	VF22A	C → E	22	F	0	0
12	VF23B	E → C	23	F	2	0
13	VF23C	C → E	23	F	0	0
14	VF26A	C → E	26	F	0	0
15	VF27A	E → C	27	F	0	0
16	VF32A	C → E	32	F	3	0
17	VF33B	C → E	33	F	0	0

Table 2.1: Summary of basic language background information for participants in the SpiCE corpus, including age, gender, age of acquisition for both languages, and the order in which the interviews occurred.

2.2.1 Participants

The recordings in this corpus will comprise the speech of 34 early Cantonese-English bilinguals, 17 of which are female. At the time of submission, ten of 17 male participants had been recorded. All participants were between the ages of 19 and 35 (inclusive), reported normal speech and hearing, and resided in Vancouver, Canada at the time of recording. All participants in the study completed an extensive language background questionnaire, which included questions about language background, proficiency, use, and general demographics. A summary of language background information for all recorded speakers in the corpus is provided in Table 2.1. A more comprehensive language background summary will be released along with the corpus audio and transcripts (see Section 2.5 for more details about

releases).

Definitions of bilingualism are highly variable in the literature, as there are many different types of bilinguals (Amengual, 2017). For the purposes of this corpus, an early bilingual is someone who acquired both Cantonese and English before starting primary school (approximately age 5), and reports consistent use of both languages since that time. It is important to highlight that the Cantonese-English bilingual community in Vancouver (and Canada more generally) is incredibly diverse, both in terms of dialects spoken, as well as the regions from which families originally emigrated (Yu, 2013). Furthermore, given the prevalence of Cantonese in Vancouver (Statistics Canada, 2017), and longevity of the community (Yu, 2013), immigration from other Cantonese-speaking areas continues today.

This corpus reflects the diverse nature of Cantonese-English bilingualism in Vancouver, as it includes Canadian-born heritage speakers, recent immigrants from Hong Kong, as well as Cantonese speakers from other parts of the Cantonese diaspora. As a result, while all speakers are early bilinguals, various dialects are represented. The most well-represented dialect is Hong Kong Cantonese, as 20 of 27 participants report having at least one caretaker from Hong Kong (14 report only Hong Kong born caretakers).

2.2.2 Recording Setup

Recording took place in a quiet room in the linguistics laboratory building at the University of British Columbia in Vancouver, Canada. Two Cantonese-English bilingual research assistants (the third and fourth authors) and the participant were seated around a table. The interviewer was a female Cantonese-English bilingual from Metro Vancouver. The recording process was monitored by a male Cantonese-English bilingual from Hong Kong, who moved to Vancouver to attend university. The interviewer and participant were outfitted with AKG C520 head-mounted microphones positioned approximately 3 cm from the corner of the mouth. The microphones were connected to separate channels on a Sound Devices USBPre2 Portable Audio Interface. Stereo recordings were made with Audacity 2.2.2 (Audacity Team, 2018) on a PC laptop, and saved according to best archival practices, with a 44.1 kHz sampling rate, and 24-bit resolution.

2.2.3 Recording Procedure

Upon arrival, participants were provided with an overview of the recording session procedures, and informed of the corpus publication process. Subsequently, participants were asked to provide written consent. Upon consent, participants completed a session in English, and a session in Cantonese. The order of languages was counterbalanced across participants (see Table 2.1). Each session consisted of three tasks—sentence reading, storyboard narration, and a conversational interview—described in the following sections. Together, these three tasks took approximately 30 minutes in each language. Along with the consent process, and a break between interviews, participants spent approximately 90 minutes in the lab.

Sentence Reading

Participants first read the sentences listed in Table 2.3 and Table 2.2 aloud, pausing between sentences. Participants were not instructed to speak in a particular style. As participants had varying levels of Cantonese reading ability, they were simultaneously presented with both Cantonese characters and the Jyutping romanization.[2] If necessary, participants could make use of the phrase's English translation. The Cantonese sentences are well-known declarative phrases, typically associated with Chinese New Year. While a more explicitly balanced set of sentences could have been used, participants' familiarity was deemed more important, as many Cantonese-English bilinguals in Canada are not literate in Cantonese. The English sentences included the Harvard Sentences list number 60 (IEEE, 1969), as well as series of holiday-themed declarative sentences to better match the content of the Cantonese sentences. This task was relatively formal, and typically lasted less than one minute.

Sentence reading was included in the session to insure that different participants produced a set of identical items, considering the core of the session was unscripted conversational interview (described in Section 2.2.3). While these sentences do not exhaustively reflect the sound systems of Cantonese and English, they provide samples of identical items for all individuals, which is advantageous

[2]Jyutping is one of the primary Cantonese romanization systems (Matthews et al., 2013), and is widely used in Cantonese corpus research (Nagy, 2011; Tse, 2019)

for future analyses or projects that require matched utterances.

	English
1	Stop whistling and watch the boys march
2	Jerk the cord, and out tumbles the gold
3	Slide the tray across the glass top
4	The cloud moved in a stately way and was gone
5	Light maple makes for a swell room
6	Set the piece here and say nothing
7	Dull stories make her laugh
8	A stiff cord will do to fasten your shoe
9	Get the trust fund to the bank early
10	Choose between the high road and the low
11	Wish on every candle for your birthday
12	Deck the halls with boughs of holly
13	Ring in the new year with a kiss
14	Have a spooky Halloween
15	Enjoy the vacation with your loved ones
16	Be filled with joy and peace during this time
17	Relax on your holiday break

Table 2.2: Sentences 1–10 comprise the Harvard Sentences List 60. Sentences 11–17 are holiday-themed original imperatives, designed to thematically match the Cantonese sentences.

No.	Cantonese	Jyutping	English translation

Table 2.3: All Cantonese sentences are widely-known imperatives associated with Chinese New Year.

Storyboard Narration

For the second task, participants narrated a short story from a cartoon storyboard in detail (Littell, 2010). The storyboard followed a simple plot about receiving gifts and writing thank you notes to family members and friends—a topic that

Cantonese-English bilinguals in the corpus were expected to be familiar with in both languages. This task was less formal than the sentence reading task, and ensured that different participants produced some of the same words in a more spontaneous context. Similar to the sentences, these same words may be useful for future analyses or projects that require matched utterances. Participants narrated the same cartoon in each language, which ensured that some of the same content was conveyed in each language (e.g., productions of *mother* in both languages). It lasted 4–5 minutes, and allowed participants time to get used to the recording setup and helped them get into the right language mode before the interview. This is important, because language mode is known to affect the degree of crosslinguistic influence in speech production (Simonet and Amengual, 2019).

Conversational Interviews

The conversational interviews formed the bulk of the recording time for each participant, lasting around 25 minutes. Participants were informed of the general interview structure ahead of time. The casual interview format was inspired by the Buckeye corpus of conversational speech (Pitt et al., 2005), and included everyday topics such as family, school, culture, hobbies, and food. These topics were selected to be relevant, interesting, and encourage storytelling, but to not delve into the personal details typically elicited in a sociolinguistic interview (Nagy, 2011). A major goal was for participants—who knew they were being recorded for linguistic inquiry—to feel at ease and freely discuss the questions. Questions were loosely laid out under general topic headings, with optional follow-up questions. While the English and Cantonese interviews had the same structure and general topic areas, the particular questions differed. Furthermore, each interview took its own shape, and was guided by what the participant wanted to talk about, anywhere from three to six topic areas covered. As a result, the speech samples from each language are comparable, but the specific questions differ between interviews and across participants.

Participants were encouraged to code-switch between languages by the interviewer, who included code-switches in some of her questions, and asked about topics that encouraged switches (e.g., Chinese foods in English; university course

work in Cantonese). While code-switching was encouraged, it was not a primary focus for the session.

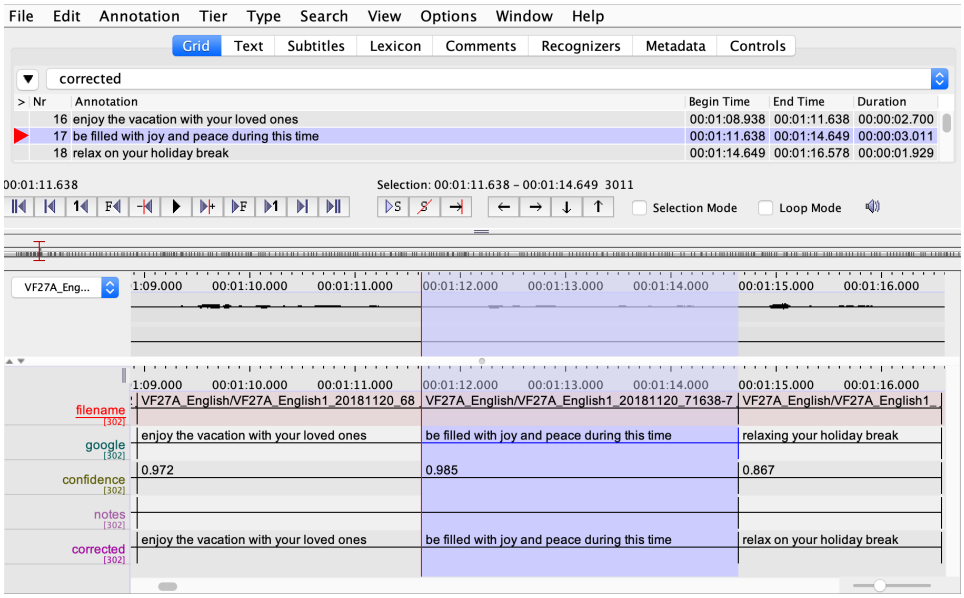

Figure 2.1: This screenshot from ELAN showcases a sample of hand-corrected English from the sentence reading task for participant VF27A. The audio waveform is displayed in two channels, with one for the participant (top) and the other for the interviewer (bottom). The annotation tiers include (1) the short audio chunk's filename, (2) the raw speech-to-text transcript, (3) the speech-to-text confidence rating, (4) space for transcriber notes, if any, and (5) the corrected transcript. Note that "relaxing" was corrected to "relax on" in the rightmost section displayed.

2.3 Annotation

All recordings have been (or will be) processed according to the pipeline outlined in this section. As much as possible, automatic tools have been leveraged to expedite hand correction.

2.3.1 Cloud Speech-to-Text

Google Cloud Speech-to-Text was used to produce an initial transcript of the interviews (Google, 2019). This was done using the Short Audio option, with the language variety set to Canadian English (en-CA) or Hong Kong Cantonese (yue-Hant-HK). In order to use this speech recognition product, the participant's speech from the recordings was first segmented into short chunks, typically under 15 seconds in duration.[3] No attention was paid to constituents at this point; rather, breaks were placed at breaths and other pauses. Short chunks were necessary for speech recognition and desirable for transcribers in the subsequent hand correction phase. With the audio files prepared in this way, speech recognition was completed using the Python client library for Google Cloud Speech-to-Text. The output included both a transcript and a confidence rating for each audio chunk. While the transcripts generated in this fashion were far from perfect, they serve the function of expediting the hand-correction process, and allow for a much earlier initial corpus release (see Section 2.5 for details).

2.3.2 Orthographic Transcription Hand-Correction

The automatically generated transcripts were converted into multi-tiered ELAN transcription files (Sloetjes and Wittenburg, 2008), with tiers for the automatically generated transcript, phrase transcription confidence, notes, and corrected transcript. During hand-correction, research assistants adjusted the transcript in the corrected tier, and took note of anything pertinent to the given audio chunk. Figure 2.1 depicts an example of corrected English transcriptions in ELAN (Sloetjes and Wittenburg, 2008). Direct identifiers (e.g., names) were marked during this phase, and will be silenced from the recordings prior to release. Transcriber guidelines were adapted from the multilingual Heritage Language Variation and Change corpus, which includes Cantonese (Nagy, 2011).

In both languages, the following conventions were used:

- The placeholder "xxx" denotes unintelligible speech.

[3] The interviewer's speech is included in the recordings for the purpose of context, but is not transcribed.

- Fragments are transcribed using "&" followed by the fragment produced (e.g., "&s").
- The "?" symbol marks questions; other punctuation is not used.

Cantonese-specific conventions include:

- Where possible, transcription is in characters.
- Words without a standard character are transcribed with Jyutping (e.g., *jyut6ping3*).
- Words produced in Mandarin Chinese are transcribed in Mandarin characters with "@m" appended to each.[4]

English-specific conventions include:

- Standard spelling is used.
- Proper nouns are capitalized and hyphenated if composed of multiple words (e.g., "British-Columbia").
- Filled pauses are transcribed with "um", "er", "uh", and other similar forms.
- Numbers are written out in word form (e.g., "one hundred").

2.3.3 Forced Alignment

Force-aligned transcripts were produced with the Montreal Forced Aligner (McAuliffe et al., 2017), using the hand-corrected orthographic transcripts and short audio chunks. Forced alignment was completed with the Train-and-Align option for each of the two languages, as a pretrained model is currently only available for English, and not necessarily representative of the dialects in this corpus. For English, the pronunciation dictionary provided with the Montreal Forced Aligner was used, which broadly reflects North American English varieties. For Cantonese, a pronunciation dictionary with segmental content only (no tones) was generated by mapping characters to the segments in the Jyutping romanization with the *PyJyutping* (Yau, 2019) and *PyCantonese* Python packages (Lee, 2018), as in prior work with Cantonese corpus research (Tse, 2019).

[4]Most participants report knowledge of Mandarin, though age of acquisition, use, and proficiency all vary drastically. In all cases, Mandarin was learned later than Cantonese and English.

2.4 Descriptive Statistics

As hand-correction is currently underway, the descriptive statistics reported in the section are based on the Google Cloud Speech-to-Text transcripts of all three tasks (sentence reading, storyboard narration, interview) for the 27 participants listed in Table 2.1.

2.4.1 Cantonese Interviews

The Cantonese interviews include approximately 9.43 hours of participant speech.[5] There were a total of 1,836 character types, and 98,401 character tokens. The number of characters varies somewhat drastically by participant, with a mean of 3,514 characters per interview (SD=734, min=2,171, max=5,410). The overall distribution of character frequency in the Cantonese interviews is depicted in Figure 2.2. As expected, there are a relatively small number of characters occurring frequently (e.g. pronouns, function words, etc.), while a majority are mid and low frequency.

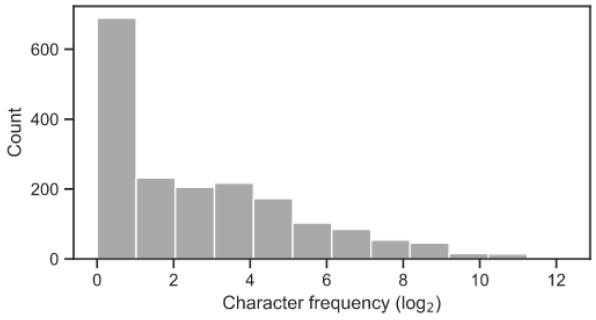

Figure 2.2: The distribution of log character frequency in the Cantonese interviews.

The decision to report descriptive statistics for Cantonese in characters rather than words arises from the difficulty of defining wordhood in Cantonese (Wong, 2006), a lack of tools for parsing written Cantonese,[6] and because written Can-

[5]Note that this excludes the duration of the interviewer questions, as well as pauses in the participant's speech. Furthermore, with the addition of the remaining seven male participants, this will increase to approximately 12 hours

[6]The *PyCantonese* package is a notable exception to this (Lee, 2018), though it does not include

tonese does not include spaces between words. An approximation of the word count can be devised by calculating the average word length from the Hong Kong Cantonese Corpus (Luke and Wong, 2015)—1.3 characters. By this estimation, the Cantonese interviews here include approximately 75,115 words.

While Google Cloud Speech-to-Text primarily transcribes Cantonese in characters, it also inserts English text.[7] As a result, it is possible to get a rough idea of how much code-switching there is in the Cantonese interviews. There were 1,321 English word types in the Cantonese interviews, and 2,858 word tokens. This does not necessarily indicate the number of switches, as participants may have produced more than one English word in a row for a given switch. Nonetheless, it demonstrates that there is a substantial amount of code-switching from Cantonese to English. The mean number of English words in the Cantonese interviews is 102 (s.d.=46, min=30, max=198).

2.4.2 English Interviews

The English interviews include a total of 5,494 word types and 91,828 word tokens in 9.89 hours of participant speech.[8] As in the Cantonese interviews, the number of words varies substantially by participant, with a mean word count of 3,729 (s.d.=701, min=2,113, max=4,518). The distribution of log word frequency in the English interviews is portrayed in Figure 2.3. Word frequency follows a similar pattern to Cantonese character frequency, with most words occurring infrequently, and a smaller proportion occurring very frequently.

Unlike for the Cantonese interviews, it is not possible to get a sense of code-switching in the English interviews, as Google Cloud Speech-to-Text for Canadian English does not insert Cantonese characters. Anecdotally, the interviewers reported that while most of the English interviews contained some code-switches, there was overall less English-to-Cantonese code-switching. This remains to be quantified when the orthographic hand-correction is completed.

tools for splitting sentences into words.

[7] It is unclear at this point how accurate the English text in the Cantonese interviews is, though anecdotally speaking, it is accurate at least some of the time.

[8] As with the Cantonese interviews, this will increase to approximately 12 hours with the remaining participants.

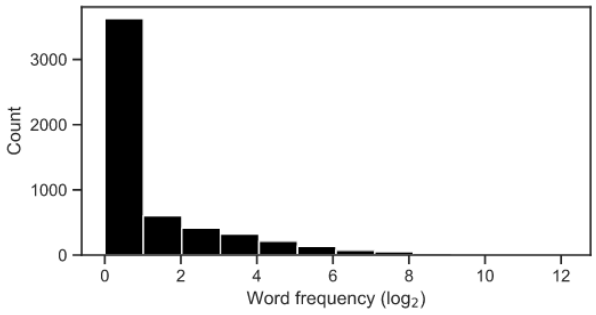

Figure 2.3: The distribution of log word frequency in the English interviews.

2.5 SpiCE Corpus Releases

The SpiCE corpus will be made publicly available in a series of releases with increasing transcription accuracy and breadth of annotation. The first release is planned for mid-2020, and will include all audio files with accompanying Google Cloud Speech-to-Text transcriptions and a detailed language background questionnaire summary. The second release will include hand-corrected orthographic transcriptions, with force-aligned phonetic transcriptions. These files will be released as they are completed. Further releases will depend on resources, but would likely include phone-level hand correction and/or corrections for a particular subset of phones. All full releases will be described in the online documentation,[9] shared through the LRE Map,[10] and announced on the first author's website.

2.6 Discussion & Conclusion

While various bilingual corpora exist, they lack in different ways. The SpiCE corpus described here enables within-speaker phonetic comparisons across languages. While this would be possible with some of the bilingual speakers in resources like the Bangor corpora (Deuchar et al., 2014), the recording quality limits the scope of phonetic queries. With the release of SpiCE and its high-quality recordings, scholars have the ability to ask and answer empirically and theoretically motivated re-

[9] https://spice-corpus.readthedocs.io/
[10] http://lremap.elra.info

search questions within the speech and language sciences using more sophisticated phonetic measurement techniques (e.g., spectral measures, in addition to temporal measures). This offers substantial potential for increasing our understanding of bilingual spoken language from both phonetic and psycholinguistic perspectives. While the recording quality of this corpus offers these particular advantages, SpiCE is also suitable for any other standard corpus-based inquiry with conversational speech, whether linguistic or paralinguistic in nature. The opportunities made available with SpiCE are especially important given the typological difference between the languages under consideration, and the fact that Cantonese is an understudied language.

Chapter 3

The structure of acoustic voice variation

3.1 Introduction

Voices can tell you a lot about the person who is talking, and have been discussed as an "auditory face" (Belin et al., 2004). They simultaneously provide information about the talker's current physical and emotional state, as well as cues to who they are (Belin et al., 2004). However, voices are highly variable, and while different voices share some dimensions of acoustic variability, the way that voices vary is argued to be largely idiosyncratic (Lee et al., 2019). Despite the high degree of variability and idiosyncrasies, listeners nonetheless use talker-specific information to recognize and discriminate voices. Listeners are good at identifying familiar voices, but perform poorly on the same tasks with unfamiliar voices (Nygaard and Pisoni, 1998). It was suggested by Lee et al. (2019) that familiarity with a voice largely comes from learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions.

Bilingualism brings an additional dimension of variability into the picture. Prior research in both perception and production suggests that while some aspects of voice variability differ for linguistic reasons, other talker-indexical features remain constant across languages, and still others can be influenced by both linguistic and non-linguistic factors. That bilingual listeners are sensitive to this information

signals its importance (Orena et al., 2019; Fricke et al., 2016).

In this study, we examine how voice varies across a bilingual's two languages. Some differences are expected. Languages differ in terms of their consonant and vowel inventories, which affect the spectral properties of a language. While all languages have consonants and vowels, they differ with respect to distribution, articulation, and acoustics (e.g., Munson et al., 2010). Suprasegmental and prosodic properties also vary across languages. Languages can differ in terms of whether a suprasegmental dimension is exploited at all, in addition to how a language might carve up a potential suprasegmental space (e.g., number of tonal contrasts). Within an individual bilingual, the acoustic variability within each language can also be related to the social identities a talker adopts within each language (see discussion in Cheng (2020)).

In an effort to understand what aspects of an individual's voice vary across languages and what are more or less fixed talker-specific attributes, researchers have compared spectral properties of bilingual speech. Results have been decidedly mixed (Cheng, 2020; Altenberg and Ferrand, 2006; Ryabov et al., 2016). For example, a small group of English-Cantonese bilinguals (n=9) in did not differ in mean fundamental frequency (F0), but exhibited greater variability in F0 (Altenberg and Ferrand, 2006). This was not the case in Ng et al. (2012), which examined voice differences with Cantonese-English bilinguals reading passages (n = 40). Based on Long-Term Average Spectral measures, females exhibited higher F0 in English than Cantonese, but males did not (Ng et al., 2012). In the same study, all participants had greater mean spectral energy values (mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese (Ng et al., 2012). Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English.

Together, these bodies of literature invite us to consider whether bilingual talkers have the "same" voice in each of their languages. Using a new corpus of conversational Cantonese-English bilingual speech—SpiCE (Johnson et al., 2020)—we look at spectral properties (Cheng, 2020; Altenberg and Ferrand, 2006; Ryabov et al., 2016; Ng et al., 2012), and also examine how acoustic variation is structured, following the work of Kreiman et al. (2014) and Lee et al. (2019).

3.2 Methods

3.2.1 Data

The 34 talkers (17 female, 17 male) studied here come from the SpiCE corpus of bilingual speech in Cantonese and English (Johnson et al., 2020). The corpus comprises two conversational interviews with early bilinguals—one in each language. While these interviews were conducted in either Cantonese or English, code-switching occurred regularly within each interview. Talkers were of a similar age (*Mdn* = 21, *M* = 22.4, *SD* = 4), learned both languages from an early age (Cantonese: *Mdn* = 0, *M* = 0.03, *SD* = 0.2; English: *Mdn* = 0, *M* = 1.32, *SD* = 1.8), and considered themselves to be proficient speakers (Cantonese: Good/Excellent = 29, Fair = 5; English: Good/Excellent = 33; Fair = 1). Talkers, however, had varied language background profiles (e.g., locations lived, dialects, knowledge of other languages), and should not be characterized as a homogeneous bilingual group. This heterogeneity is an accurate representation of the Cantonese speech community (Liang, 2015) and it is considered a boon in this study, as analyses are conducted on a within-talker basis.

The audio recordings are high quality, with a 44.1 kHz sampling rate, 24-bit resolution, and minimal background noise. Both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. The participant channel was extracted, including any code-switches they made during the interview, as this process was done without orthographic transcripts. We then identified all voiced segments with the *Point Process (periodic, cc)*[1] and *To TextGrid (vuv)* Praat algorithms (Boersma and Weenink, 2021), implemented with the Parselmouth Python package (Jadoul et al., 2018). While speech from the interviewer can occasionally be heard in the participant channel, it is quiet enough to have been ignored by the Praat algorithms. This method captures vowels and approximants, as well as some voiced obstruents, differing slightly from (Lee et al., 2019).

[1]The pitch range settings differed by gender (female: 100–500, male: 75–300).

3.2.2 Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by Lee et al. (2019)[2] in VoiceSauce (Shue et al., 2011). The suite of selected measurements are drawn from a psychoacoustic voice quality model (Kreiman et al., 2014). Measurements were made every 5 ms during voiced segments, as in Lee et al. (2019).

F0 is a correlate of pitch, associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics. **F1**, **F2**, and **F3** are the three formant frequencies typically discussed in relation to linguistic contrasts (e.g., for vowels and sonorant consonants). The fourth formant frequency, **F4**, is not typically discussed in linguistic contexts, and is instead associated with talker characteristics.

The corrected amplitude difference between the first two harmonics, **H1*–H2***, is a measure of harmonic spectral slope associated with phonation type. The asterisk indicates that the value has been corrected (Iseli et al., 2007). The corrected amplitude difference between the second and fourth harmonics, **H2*–H4***, is a measure of harmonic spectral slope in a slightly higher frequency band. **H4*–H2kHz*** is the corrected amplitude difference between the fourth harmonic and the harmonic closest to 2000 Hz; it is a measure of harmonic spectral slope in a higher frequency band. The corrected amplitude difference between the harmonics closest to 2000 Hz and 5000 Hz, **H2kHz*–H5kHz***, is a measure of harmonic spectral slope that does not depend on F0. Together, these harmonic-based measures characterize source spectral shape (Kreiman et al., 2014).

Cepstral Peak Prominence (**CPP**) is a measure of the ratio between harmonic energy and spectral noise, and is associated with non-modal phonation types. Root Mean Square (RMS) **Energy** is a measure of overall amplitude. The subharmonics-harmonics amplitude ratio (**SHR**) is a measure of spectral noise associated with period doubling or irregularities in phonation. Together, these three measures characterize key aspects of spectral noise.

[2]The exception was formant dispersion, which was excluded because it was almost perfectly correlated with the measured value of F4 and led to problems with the principal components analyses.

3.2.3 Exclusionary criteria and post-processing

Observations were removed if they included impossible or erroneous values, including instances where F0, F1–F4, CPP, or H5kHz was equal to zero. Filtering was done with just one of the uncorrected harmonic amplitude measures, as erroneous values tended to co-occur on the same observation, and the distribution of H5kHz did not span zero, with the exception of (erroneous) values equal to zero. This minimizes the removal of correctly measured zero values.

Moving standard deviations were calculated for each of the 12 measures using a centered 50 ms window ($\approx$ 10 observations). This captures dynamic changes for each of the voice quality measures, which is important as they may better reflect what listeners attend to in talker identification and discrimination tasks (Lee et al., 2019).[3] Observations missing a moving standard deviation value (i.e., those near a voicing boundary) were removed. Including both the values measured in VoiceSauce and their moving standard deviations, a total of 24 measures were used in the analysis described in the next section. Across the 34 talkers, there were 3,126,267 observations after winnowing the data.

3.2.4 Principal components analysis

Principal components analysis (PCA) is a dimensionality reduction technique appropriate for data that include a large number of (potentially) correlated variables. The distillation into components helps identify and facilitate describing the internal structure, in this case, of a voice. We adapt methods from work on voices (Lee et al., 2019) and faces (Burton et al., 2016; Turk and Pentland, 1991). The goal is to the capture similarities or differences for each talker's voice across languages. As such, we conducted PCAs separately for each talker-language pair, and compared the results of each talker's English and Cantonese PCAs. All 24 measures were normalized (z-scored) on by-PCA basis for the analysis. PCAs were implemented with the *parameters* package (Makowski et al., 2019) in R (R Core Team, 2020), using an oblique *promax* rotation to simplify the factor structure, as the measurements reported in the previous section were expected to be somewhat correlated

[3] We used *SD*, as opposed to the coefficients of variation used by Lee et al. (2019). Regardless, all variables were scaled prior to inclusion the PCAs.

(Lee et al., 2019).

Each PCA included the number of components for which all resulting eigenvalues were greater than 0.7 times the mean eigenvalue, following Jolliffe's (Jolliffe, 2002) recommended adjustment to the Kaiser-Guttman rule. We used this rule, rather than a more sophisticated test (e.g., broken sticks), as it is not detrimental to our exploratory analysis to err on the side of including marginal components. Additionally, across each of the components, only loadings with an absolute value of 0.32 or higher were interpreted (Lee et al., 2019; Tabachnick and Fidell, 2013); however, all loadings were retained for the canonical redundancy analysis described in the next section.

3.2.5 Canonical redundancy analysis

In order to assess whether variation in a talker's voice is structurally similar across both languages, we compare the PCA output from English and Cantonese by calculating redundancy indices from a canonical correlation analysis (CCA) (Stewart and Love, 1968; Jolliffe, 2002). CCA is a statistical method used to explore how groups of variables are related to one another. The two sets of variables are transformed such that the correlation between the rotated versions is maximized. This is useful here, as a talker may have similar components in their English PCA and Cantonese PCA, but these components might not necessarily be in the same order, even if they account for comparable amounts of variance.

Redundancy is a relatively simple way to characterize the relationship between the loadings matrices of two PCAs—the two sets of variables under consideration here. The two indices represent the amount of variation in a talker's Cantonese PCA output that can be accounted for via canonical variates by their English PCA output, and vice versa. Notably, the two redundancy indices are not symmetrical (Stewart and Love, 1968).

We computed redundancy indices for all pairwise combinations, including cases where similar values were expected (same talker, different language), and cases where we expected dissimilarity (different talker and language). Considering that the PCA analyses retain the lower-dimensional structure within each language, these redundancy indices effectively reflect the degree to which the lower-dimensional

structure of the voice variability is retained across a talker's two languages.

3.3 Results

3.3.1 Crosslinguistic comparison of acoustic measurements

For each acoustic measurement and talker, we conducted a Student's *t*-test and calculated Cohen's *d*, in order to give a high-level assessment of whether variable means differed across the two languages. These comparisons have no bearing on how a given variable *varies*. Table 3.1 reports counts of talkers by effect size. Notably, across all talkers and variables, only 21.1% yielded non-trivial Cohen's *d* values. Most talkers (32/34) had at least one non-trivial comparison. The distribution of these counts is depicted in Figure 3.1.

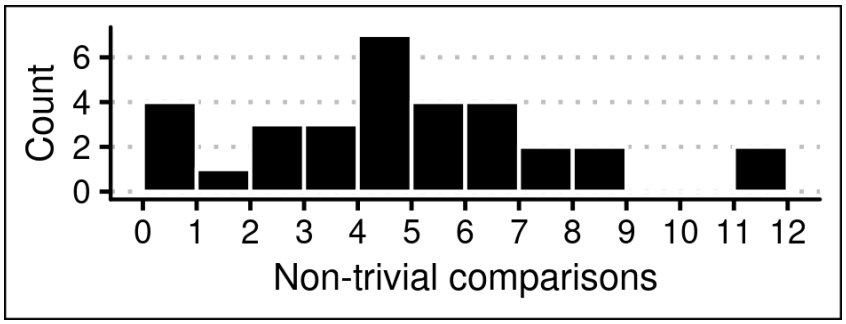

Figure 3.1: A summary of the number of non-trivial comparisons from Table 3.1 across the 34 talkers.

For the non-trivial comparisons, there were consistent patterns across languages for H1*–H2* and F0. For the remaining variables, while some talkers exhibited a difference in mean values, the direction of the difference varied, or relatively few talkers exhibited the difference.

H1*–H2* was significantly higher in Cantonese for a relatively large subset of the talkers (13/34), lower for a small number (3/34), but trivial for most (18/34). While based on a different measure than (Ng et al., 2012), this is consistent with the finding that Cantonese tends to be breathier, or English creakier—the current analysis does not distinguish between these interpretations.

Table 3.1: This table reports counts of Cohen's *d* for crosslinguistic comparisons of each of the acoustic measurements by talker. Degrees of freedom ranged between 49,274–136,644 across t-tests. For most talkers and variables, the difference in means was trivial, which is reflected in that column's high counts.

	Cohen's *d*		
	Trivial	**Small**	**Medium**
Variable	***0.0–0.2***	***0.2–0.5***	***0.5–0.8***
F0	21	10	3
F0 s.d.	34	0	0
F1	24	9	1
F1 s.d.	29	5	0
F2	26	8	0
F2 s.d.	32	2	0
F3	24	9	1
F3 s.d.	29	5	0
F4	30	3	1
F4 s.d.	28	6	0
H1*–H2*	18	15	1
H1*–H2* s.d.	32	2	0
H2*–H4*	25	9	0
H2*–H4* s.d.	31	3	0
H4*–2kHz*	25	8	1
H4*–2kHz* s.d.	34	0	0
H2kHz*–5kHz*	23	10	1
H2kHz*–5kHz* s.d.	31	3	0
CPP	21	10	3
CPP s.d.	32	2	0
Energy	17	14	3
Energy s.d.	18	16	0
SHR	31	3	0
SHR s.d.	29	5	0

If there was a non-trivial difference in F0 across languages, then Cantonese had a lower mean F0 than English (13/34; Female = 7), though most talkers did not exhibit a difference (21/34). This is consistent with prior findings that when a difference between English and Cantonese was found, Cantonese had a lower mean F0 for females (Ng et al., 2012; Altenberg and Ferrand, 2006). We also observe this difference for a small number of males.

3.3.2 PCA results

The PCAs across both languages for all 34 talkers resulted in 10–15 components and accounted for 74.6–85.8% of the total variation. To assess whether talkers exhibit the same structure in voice variability across their languages, we first consider the patterns present across the different PCAs, as this provides context for understating what unique structural characteristics in talkers' voices looks like. To this end, we briefly summarize common patterns across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 3.2 shows the first four components of a single talker's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages.

Broadly speaking, there were a lot of similarities in component composition across both talkers and languages, with the eight most commonly occurring components summarized in Table 3.2. For context, recall that PCAs had anywhere from 10–15 components total. These eight components consisted of source spectral shape, spectral noise, as well as formant variables. On the other hand, F0 co-occurred with a wide variety of variables (often Energy), but in a manner that was less consistent across talkers. There were additional components (not reported here) that were shared by less than half of talkers. In summary, despite the greater amount of shared structure across PCAs than found in Lee et al. (2019), there is still ample room for idiosyncratic variation, both in terms of which variables co-occur, as well as in how much variance different components account for.

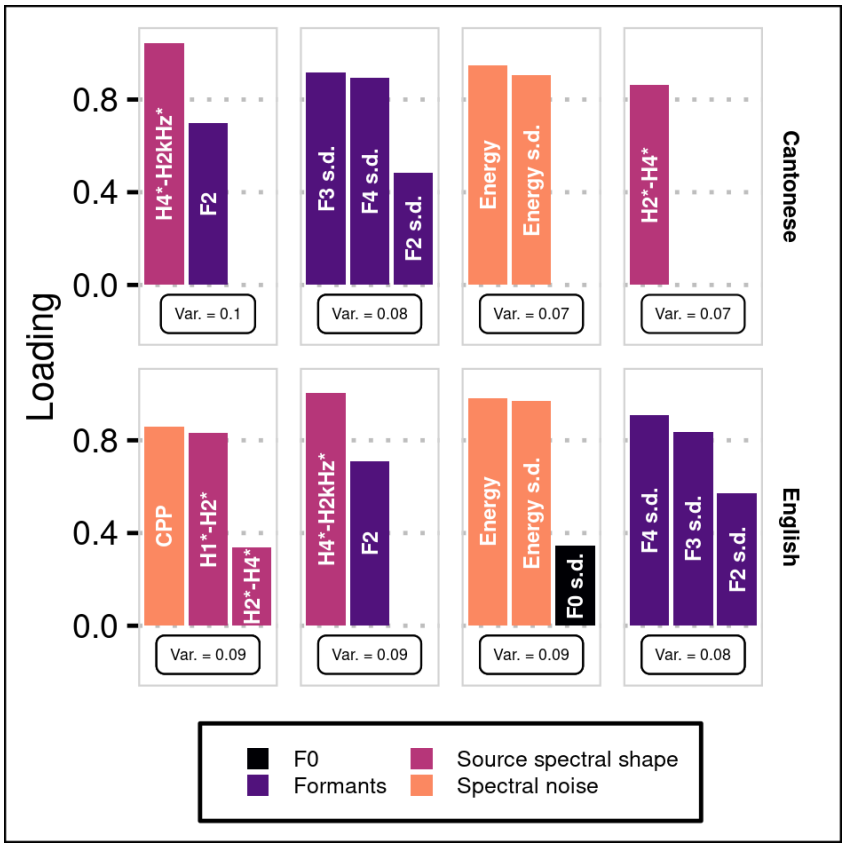

Figure 3.2: In the first four components of a talker's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

3.3.3 Within-talker analysis

A slight majority of talkers had the same number of components for each of their languages (18/34). Of the remainder, most talkers had a difference of one in the number components (14/34), and far fewer differed by two (2/34). Redundancy indices for within-talker comparisons ranged from 0.82 to 0.99, (Mdn = 0.93, M = 0.92, SD = 0.04), and are displayed in Figure 3.3, with the two redundancy indices for a given pair plotted against one another. Comparisons across talkers within-

Table 3.2: A summary of the most commonly occurring components across all PCAs. Variables are only included if |Loading| > 0.32. Italics indicate additional variables that were present on a component for a subset of talkers (i.e., an alternative but related configuration). *N* indicates the number of times a component occurred (out of 34), and *Var. %* gives the range of percent variance accounted for by the component.

	Cantonese		English	
Variables	**N**	**Var. %**	**N**	**Var. %**
H4*–H2kHz*, H2kHz*–H5kHz*, F2, *F3*, *F4*	34	9.3–15.5	32	9.2–16.7
H4*–H2kHz* s.d., H2kHz*–H5kHz* s.d.	32	6.3–8.3	34	4.1–5.0
Energy, Energy s.d, *F0*	31	5.8–9.4	33	6.3–9.1
CPP s.d.	29	4.1–5.0	31	4.1–4.9
SHR, SHR s.d.	30	3.8–7.5	29	5.4–7.3
F3, F4, *F2*	26	6.0–8.5	29	5.8–8.5
F3 s.d., F4 s.d., *F2 s.d.*	26	5.3–8.6	29	4.7–8.6
H2*–H4* s.d., H1*–H2* s.d.	26	4.2–6.5	28	4.2–6.8

language (range: 0.63–0.98, $Mdn = 0.84$, $M = 0.84$, $SD = 0.6$) and across-language (range: 0.66–0.98, $Mdn = 0.83$, $M = 0.84$, $SD = 0.6$) are generally lower, but still relatively high. Within-talker values were confirmed to be higher than across-talker comparisons [*Welch's* $t(71.36) = -17.83$, $p < 0.001$, d = 1.76].

The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded components almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.3.2, there were a substantial number of commonly occurring patterns across talkers and languages.

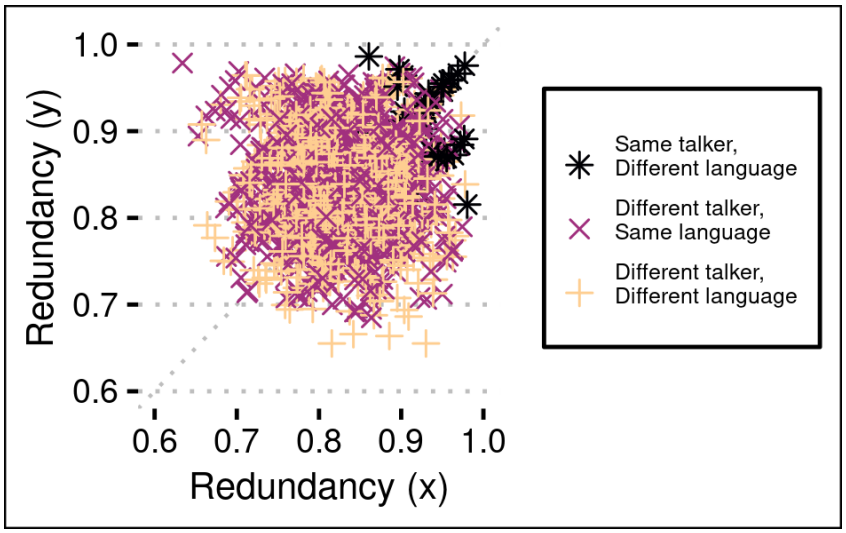

Figure 3.3: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are clustered at the top right.

3.4 Discussion and conclusion

This study examines spectral properties and structural similarities in an individual's voice in two languages. A clear result is that most of the bilinguals studied here exhibit similar spectral properties, and similar lower-dimensional structure in voice variation, despite substantial segmental and suprasegmental differences across English and Cantonese (Matthews et al., 2013). In this sense, a majority appear to have the same "voice" across languages, which renders voice-as-an-auditory-face an apt comparison.

The comparison of these 34 Cantonese-English bilinguals' voices across languages suggest more similarity for an individual across languages than found within a more tightly controlled group of monolingual English speakers (Lee et al., 2019)—several analysis decisions may have contributed to this. We compared similar components independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that any comparisons made are among like items. Any downside to this methodological decision is mitigated

by the fact that most components made relatively small contributions, accounting for 4.2–10.3% (95% highest density interval) of the PCA's total variance.

While statistical choices may have affected these results, the data differences between the current and previous studies are also important to note. This study uses substantially longer passages than the short samples in Lee et al. (2019). The larger speech sample may allow for a more stable underlying structure to showcase itself, as opposed to the potential for ephemeral variation in a shorter sample. This possibility is easily testable by manipulating the length of the speech sample in the analysis.

Ultimately, the goal is to understand how the acoustic variability and structure of talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination. Turning to listener and behavioural data will help in deciphering what is meaningful variation within a voice from low level noise that cannot be attributed to a particular vocal signature. Verification from listener performance will help adjudicate which statistical choices present an acoustic voice space that matches listener organization.

3.5 Acknowledgments

This project draws on research supported by the Social Sciences and Humanities Research Council of Canada (SSHRC), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the University of British Columbia Public Scholars Initiative. KAJ led conceptualization, design, analysis, interpretation, and writing. MB contributed to each of these areas. RAF contributed to analysis and interpretation.

Chapter 4

Croslinguistic uniformity for VOT

4.1 Introduction

A consequence of bilingualism is that individuals must navigate overlapping segment inventories (Flege and Bohn, 2021). This paper is concerned with the question of what languages share, if anything, in the representation of speech sounds. Most prior work has focused on sounds that are phonologically similar, yet phonetically distinct, as with 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 similar sounds hinges on the presence of an observable difference, at least under some set of conditions. 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 (Antoniou et al., 2010; Simonet, 2016; Bullock and Toribio, 2009). In the revised Speech Learning Model (SLM-r) (Flege and Bohn, 2021), 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 similar speech sounds across languages, even when such a comparison 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 crosslinguis-

tic 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*.

The idea of telling things apart or together fits within the SLM-r framework (Flege and Bohn, 2021), where categories from different languages exist in a shared phonetic space and are subject to constraints from the perceptual an productive systems: 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). This framework 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 an outcome of proximity at least sometimes, the original SLM would argue that if two segments sound like the same duck, then they must in fact be the same duck (i.e., share an underlying representation). Note, however, that this conclusion does not necessarily apply to composite categories where differences persist despite crosslinguistic influence.

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 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. 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 similarity may be taken for granted and not considered an interesting problem to focus on, despite the nature of representation being a key focus of psycholinguistics in general—especially in perception (Samuel, 2020). In the interest of understanding representation, the best candidates would be the hardest to distinguish 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 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 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 already 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 ($n = 2295$).

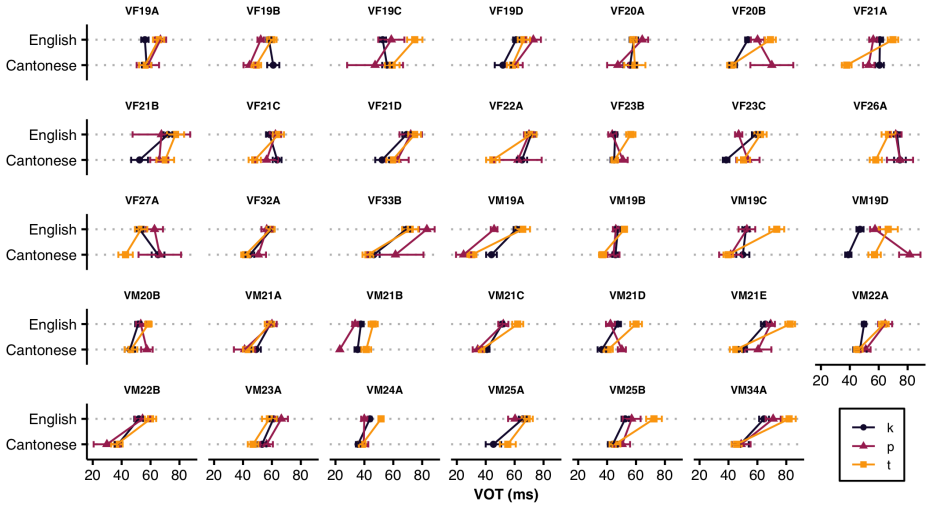

Figure 4.1: Mean and SE for VOT across place of articulation, language, and individuals in the SpiCE corpus.

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 4.1, 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 the relative ordering or markers in Figure 4.1, though in many cases the standard errors overlap, suggesting that a strict ordering by means may not be appropriate. Additionally, crossed lines in Figure 4.1 indicated that many talkers are not internally consistent across languages.

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

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 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/ $\sim$ Cantonese /t/	0.59	0.004
Cantonese /p/ $\sim$ Cantonese /k/	0.54	0.009
Cantonese /t/ $\sim$ Cantonese /k/	0.33	0.28
English /p/ $\sim$ English /t/	0.58	0.004
English /p/ $\sim$ English /k/	0.75	$<$0.001
English /t/ $\sim$ English /k/	0.57	0.005
Cantonese /p/ $\sim$ English /p/	0.57	0.006
Cantonese /t/ $\sim$ English /t/	0.31	0.29
Cantonese /k/ $\sim$ English /k/	0.55	0.006
Cantonese /p/ $\sim$ English /t/	0.23	0.33
Cantonese /p/ $\sim$ English /k/	0.35	0.29
Cantonese /t/ $\sim$ English /p/	0.43	0.08
Cantonese /t/ $\sim$ English /k/	0.31	0.29
Cantonese /k/ $\sim$ English /p/	0.56	0.006
Cantonese /k/ $\sim$ 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 (cen-

tered) 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 included, 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.

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

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 evidence 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 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 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 similarity.

Chapter 5

Discussion

...

Chapter 6

Conclusion

...

Bibliography

Altenberg, E. P. and Ferrand, C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, 20(1):89–96. → pages 17, 24

Amengual, M. (2017). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, 23(5):954–970. → page 5

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 30

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 29, 30

Audacity Team (2018). Audacity (R): Free audio editor and recorder. → page 5

Barlow, J. A. (2014). Age of acquisition and allophony in Spanish-English bilinguals. *Frontiers in Psychology*, 5. → page 30

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 37

Belin, P., Fecteau, S., and Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135. → page 16

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111. → page 2

Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Version 6.1.38. → page 18

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 36

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 30

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., and Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1):202–223. → page 20

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 31

Cheng, A. (2020). Cross-linguistic F0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, 147(2):EL67–EL73. → page 17

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 32, 33, 35, 37, 39

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 32, 33, 34, 35, 36, 37, 38, 39

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 33

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 33, 39

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., and Carter, D. (2014). Building bilingual corpora. In Thomas, E. M. and Mennen, I., editors, *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters. → pages 3, 14

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. → pages 32, 33

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 31

Flege, J. E. and Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In *Second Language Speech Learning*, pages 3–83. Cambridge University Press. → pages 29, 30, 31, 33, 39

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 17, 29, 30

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. → pages 2, 39

Goldrick, M., Runnqvist, E., and Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, 25(4):1031–1036. → page 29

Google (2019). Cloud speech-to-text. → page 10

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, 60(2):98–128. → pages 30, 31

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246. → page 6

Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America*, 121(4):2283–2295. → page 19

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 18

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 39

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 17, 18, 33, 39

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → page 21

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 34

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, 1(1):e009. → pages 17, 19

Lee, J. L. (2018). PyCantonese [Version 2.2.0]. → pages 3, 11, 12

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3):1568–1579. → pages 16, 17, 18, 19, 20, 21, 24, 27, 28

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. → page 18

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 31

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 33, 39

Littell, P. (2010). Thank-you notes [Version 1.0: Agent focus]. → page 7

Luke, K. K. and Wong, M. L. Y. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics*, 25(2015):309–330. → pages 3, 13

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 20

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → pages 3, 6, 27

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner [Version 1.0.1]. → page 11

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of vox humana. *Clinical linguistics & phonetics*, 24(4-5):245–260. → page 17

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 32

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, 43(1-2):65–84. → pages 6, 8, 10

Ng, M. L., Chen, Y., and Chan, E. Y. (2012). Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—a long-term average spectral analysis. *Journal of Voice*, 26(4):e171–e176. → pages 17, 22, 24

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376. → page 16

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. → pages 17, 40

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 30

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95. → pages 3, 8

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → page 20

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 40

Ryabov, R., Malakh, M., Trachtenberg, M., Wohl, S., and Oliveira, G. (2016). Self-perceived and acoustic voice characteristics of Russian-English bilinguals. *Journal of Voice*, 30(6):772.e1 – 772.e8. → page 17

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070. → page 32

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences*, volume 3, pages 1846–1849, Hong Kong. → page 19

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. → page 30

Simonet, M. and Amengual, M. (2019). Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, 24(2):208–221. → page 8

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco. European Language Resources Association (ELRA). → page 10

Statistics Canada (2017). Proportion of mother tongue responses for various regions in Canada, 2016 Census. → page 5

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163. → page 21

Sundara, M., Polka, L., and Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, 9(1):97–114. → page 29

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 21

Tse, H. (2019). *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA. → pages 6, 11

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 30, 31

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Sco. Press. → page 20

Wong, W. Y. P. (2006). *Syllable fusion in Hong Kong Cantonese connected speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. → page 12

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820. → pages 30, 31, 33

Yau, M. (2019). PyJyutping. → page 11

Yu, H. (2013). Mountains of gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, pages 108–121. Routledge. → page 5