Crosslinguistic (dis)similarity in Cantonese-English bilingual speech production

by

Khia Anne Johnson

B.A. Linguistics, University of Washington, 2013

a thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in

the faculty of arts
(Linguistics)

The University of British Columbia
(Vancouver)

December 2021

© Khia Anne Johnson, 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Crosslinguistic (dis)similarity in Cantonese-English bilingual speech production

submitted by **Khia Anne Johnson** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Linguistics**.

Examining Committee:

Molly Babel, Linguistics
Supervisor

Kathleen Currie Hall, Linguistics
Supervisory Committee Member

Márton Sóskuthy, Linguistics
Supervisory Committee Member

Abstract

...

Lay Summary

...

Preface

...

Table of Contents

List of Tables

List of Figures

Acknowledgments

...

Chapter 1

Introduction

...

Chapter 2

The SpiCE Corpus

2.1 Introduction

Much of our formal knowledge about the phonetics of spoken language and speech processing comes from monolingual individuals producing scripted speech in laboratory settings. While far from the only source of knowledge, especially as areas like sociophonetics and corpus phonetics continue to grow, laboratory speech perhaps retains an outsize role. Monolingual lab speech allows 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 represent a minority of the contexts in which spoken languages are used around the world. Bilingualism is the norm, not the exception, and individuals regularly make creative linguistic choices in their spontaneous speech.[1]

Conversational speech allows for a richer empirical description of spoken language compared to—or at the very least, in addition to—laboratory elicited speech. It provides a more realistic representation of how individuals produce language in everyday contexts that isolated word production and sentence reading do not faithfully capture. It enables and facilitates the study of non-formal speech styles, style-

[1] Throughout this chapter, and in the literature more broadly, multilingualism and bilingualism are used somewhat interchangeably. While there is a growing area focused on trilingualism and language acquisition beyond two languages, multilingualism research tends to focus on two languages.

shifting, and more. Conversational speech also crucially permits for field testing of speech production theories in their natural habitats. 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, given the differences in communicative contexts, attentional demands, and speaking rate variability (e.g., Gahl et al., 2012; Johnson and Babel, 2021).

Discrepancies between results for conversational and lab speech have been found for monolingual (English) speech, but are likely to be found with bilingual speech as well. Research on bilingual conversational speech is limited, however, as the resources needed for this type of inquiry are relatively rare. Table 2.1 provides a sample of prominent bilingual speech corpora, summarizing key information such as the title, balance of languages, speech style, and suitability for within-talker and phonetic research questions. While Table 2.1 is far from a comprehensive list, it nonetheless accurately reflects that most corpora of this type have focused on bilinguals of two European languages.

As a step towards filling this gap, this chapter introduces the **SpiCE** corpus of conversational bilingual **Sp**eech **i**n **C**antonese and **E**nglish (Johnson, 2021). In contributing to filling this gap, SpiCE will allow researchers to address a set of research questions that were previously not possible, using both conversational bilingual speech and sophisticated phonetic measurements, at scale.

As will become apparent later in this chapter, the SpiCE corpus focuses on early bilingualism. In light of this, Table 2.1 summarizes a selection speech corpora that involve similar populations, as opposed to corpora focused on late bilinguals and language learners (e.g., Bradlow et al., 2011).

To preview the end product before diving into the details—SpiCE is a corpus of bilingual speech in Cantonese and English, comprising high-quality recordings of 34 early Cantonese-English bilinguals. The participants were young adult members of the heterogeneous bilingual speech community in the Vancouver, Canada area. Each participant completed a few different tasks—reading sentences, narrating a cartoon storyboard, and conversing freely in a semi-structured interview with a bilingual peer as the interviewer. All of the recordings were manually transcribed at the word level and force-aligned at the phone level. This chapter describes each

Corpus	Language balance	Size	Style	Within-talker	Recording quality
Bangor Miami (Deuchar et al., 2014)	63% English 34% Spanish	CHECK	Conversational, code-switching	Yes, most talkers	CHECK
Bangor Patagonia (Deuchar et al., 2014)	78% Welsh 17% Spanish <0.5% English	CHECK	Conversational, code-switching	CHECK	CHECK
Bangor Siarad (Deuchar et al., 2014)	84% Welsh 4% English	CHECK	Conversational, code-switching	CHECK	CHECK
add more here!					

Table 2.1: A selection of prominent bilingual speech corpora, with summary information for the balance of langauges and speaking styles produced by the bilinguals in the corpus. [More will be added here! There are a number of Spanish-English bilingual corpora I could dig up info on! Also: https://biling.talkbank.org/access]

of these components in detail and offers justification for the decisions where warranted.

The SpiCE corpus design is based on key aspects of widely used existing spontaneous speech corpora, such as the Buckeye corpus of conversational speech (Pitt et al., 2005). In many ways, the Buckeye corpus is treated as a gold standard in the field of corpus phonetics. And while the SpiCE corpus does not copy its structure and level of detail exactly, the Buckeye corpus nonetheless serves as inspiration, particularly given its casual interview style and high recording quality. The goal, after all, is to facilitate phonetics research with spontaneous bilingual speech.

Given the bilingual design, SpiCE crucially includes speech from the same individual in more than one language. Inspiration in this regard is drawn from the Bangor corpora of Spanish-English, Welsh-English, and Welsh-Spanish bilingual speech (Deuchar et al., 2014). The Bangor corpora include speech from the same individual in more than one language but largely comprise field recordings, some

of which are noisy. For example, many of the recordings in the Spanish-English Bangor corpus were made with a lapel microphone worn on the participant's belt, and others with a radio microphone placed on a table. This variable—and often noisy—recording quality limits the scope of phonetics research using the corpora. Additionally, the Bangor corpora were designed for understanding code-switching in everyday situations. While this facilitates understanding broad patterns of language use, it also means that the corpora are not necessarily balanced for the languages involved—people do not necessarily use their languages in equal proportions. So while these corpora are incredibly valuable for linguistics research, there are nonetheless limitations. Compared to these corpora (and those listed in Table 2.1), SpiCE uses a more controlled and balanced recording setup, which allows for more nuanced acoustic-phonetic measurements. This is, however, at the expense of other criteria (e.g., naturalness), in which the Bangor corpora excel.

SpiCE is also unique in the population it represents. Many of the resources available to researchers on sites like BilingBank, ELRA, and elsewhere feature late bilinguals and second language learners and vary widely in task and recording quality. One example of a Cantonese-English resource that fits this description is the ShefCE corpus (Ng et al., 2017). ShefCE is a parallel corpus featuring L1 Hong Kong Cantonese and L2 English read speech, where participants read lectures in each language one sentence at a time. While there are similarities with what SpiCE aims to accomplish (e.g., promoting research with Cantonese-English bilingual speech), ShefCE occupies a different niche in the speech sciences—it was designed for L2 pronunciation assessment and training speech recognition models.

The primary motivation for collecting this corpus was to have comparable high-quality recordings of conversational speech from early bilinguals in two languages, which enables large-scale phonetic analysis on a within-speaker basis. It is worth noting that corpus size is a subjective measure, as different fields have different standards in this respect. For the type of corpus, SpiCE is relatively large (32.8 total recording hours and approximately 219,000 words), being slightly smaller in size than the Buckeye corpus (approximately 40 total recording hours and 307,000 words Pitt et al., 2005). Both of these are purpose-built corpora recorded in person. Truly large corpora tend to be collected from existing recordings (radio, YouTube, audiobooks, etc.; e.g., Librispeech, 1000 hours: Panayotov et al., 2015), crowd-

sourced online (e.g. Mozilla Common Voice, 2500 hours: Ardila et al., 2020), via phone (e.g., SWITCHBOARD, 260 hours: Godfrey et al., 1992), and other similar more scalable methods. The reason? High-quality, purpose-built corpora are expensive and time-consuming to create.

To my 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 85 million speakers around the world (Ethnologue, 2021), though this is changing with new Cantonese language corpora (Luke and Wong, 2015; Leung and Law, 2001; Winterstein et al., 2020; Alderete et al., 2019), natural language processing tools (Lee, 2018; Yau, 2019), and support in speech technology applications (Google, 2019).

While some of the design choices have been touched upon already, the remainder of this chapter provides a detailed overview of the corpus. Sections 2.2 covers the design and collection procedures and includes a detailed description of the participants. Section 2.3 describes the transcription and annotation pipeline. Section 2.4 concludes with descriptive statistics summarizing the corpus.

2.2 Corpus design and creation

This section provides detail about the speakers (Section 2.2.2), the procedures used to ensure high-quality recordings (Section 2.2.3), and the three tasks that each participant completed in both Cantonese and English (Section 2.2.4).

Data collection took place between November 2018 and March 2020. Orthographic transcription began shortly after the first interview was recorded and was completed in April 2021. The corpus was made available to the public in May 2021 via Scholars Portal Dataverse at https://doi.org/10.5683/SP2/MJOXP3. Additionally, detailed documentation for the corpus is available both with the corpus download and at https://spice-corpus.readthedocs.io/.

2.2.1 Recruitment

Participants were recruited for the SpiCE corpus through a variety of methods at the University of British Columbia. This included word of mouth, the Linguistics

Human Subject Pool, the Psychology Paid Studies list, advertisements in department email lists, advertisements in linguistics courses, printed flyers, and posts on various club forums.

The recruitment process focused on fluent speakers of Cantonese and English, between the ages of 19 and 35, with normal speech and hearing, who began learning both languages from early childhood (age 5 or earlier). One goal of recruitment was to maintain a balance of male- and female-identifying speakers, and as a result, once 17 females had participated, the recruitment language was adjusted to focus on male- or nonbinary-identifying participants.

Before scheduling a session, participants first completed a language background survey. If an individual signed up to participate but did not meet the criteria for participation, their session was canceled and they were contacted with an explanation.

All participants who came into the lab were compensated for their time with partial course credit or $15 CAD.

2.2.2 Participants

The recordings in SpiCE comprise the speech of 34 early Cantonese-English bilinguals. Throughout this chapter and the corpus, participants are identified by participant IDs. The IDs are designed to provide basic information about the participant. For example, VF19A indicates that the participant was recorded in **V**ancouver, identified as **F**emale, and was 19 years old at the time of recording. The letter at the end distinguishes participants of the same age and gender. There were 17 participants who self-identified as female and 17 as male. Participants ranged in age from 19 to 34 years old at the time of recording. Apart from one talker who reported mild high-frequency hearing loss (VM25A), all participants reported normal speech and hearing. Additionally, all participants resided in the Metro Vancouver, Canada area at the time of recording. The SpiCE corpus also includes a detailed summary extracted from an extensive language background survey administered before the recording session (without the researchers present), as well as a copy of the survey itself. Basic summary information is included in Table 2.2, and in visualizations throughout this chapter. All participant information is based on self-reported participant data from the survey.

There were a handful of additional individuals who participated in the study but were ultimately excluded from the published SpiCE corpus due to missing language background questionnaire information (n=1), recording issues (n=2), or not starting learning Cantonese until age eight (n=1).

Definitions of bilingualism are highly variable in the literature, as there are many different types of bilinguals (Amengual, 2017). For this corpus, an early bilingual is someone who began learning both Cantonese and English before starting primary school (approximately age 5), reports consistent use of both languages since that time, and self-selected to participate in a research study involving an interview in each language. 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 or varieties spoken, as well as in 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's presence in Vancouver (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, Cantonese speakers from other parts of the Cantonese diaspora, and individuals who do not neatly fit into these particular categories. As a result, while all speakers are early bilinguals, various dialects are represented. Figure 2.1 depicts where SpiCE participants reported living during different age intervals. These intervals were selected after reviewing freeform participant responses comprising when they lived in different places. Specifically, Figure 2.1 reports the number of participants who indicated that they lived in a given country during the age ranged for the panel. For example, if a participant moved from Hong Kong to Canada at age 7, they would be counted in both bars in that panel.

Soliciting Cantonese dialect information directly would have been challenging, as many of the participants in the corpus would not have straightforward dialect classifications. This is especially true for individual who were born and/or raised in the Cantonese diaspora, but to Hong Kongers as well, given the extent of globalization in Hong Kong (cite). In light of this, it is useful to summarize where the SpiCE participants' caretakers were primarily raised. Figure 2.2 does exactly this.

No.	ID	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
18	VM19A	E → C	19	M	0	0
19	VM19B	C → E	19	M	2	0
20	VM19C	E → C	19	M	0	0
21	VM19D	C → E	18	M	1	1
22	VM20B	E → C	20	M	0	0
23	VM21A	E → C	21	M	0	0
24	VM21B	E → C	21	M	0	0
25	VM21C	C → E	21	M	0	0
26	VM21D	C → E	21	M	0	0
27	VM21E	C → E	21	M	5	0
28	VM22A	C → E	22	M	4	0
29	VM22B	E → C	22	M	0	0
30	VM23A	E → C	23	M	0	0
31	VM24A	E → C	24	M	3	0
32	VM25A	E → C	25	M	4	0
33	VM25B	E → C	25	M	0	0
34	VM34A	C → E	34	M	0	0

Table 2.2: Basic participant information from the language background survey, including age, gender (M for male and F for female), age of acquisition (phrased as "age began learning"), and the order the interviews occurred (E for English and C for Cantonese). See Section 2.2.4 for information about interview order.

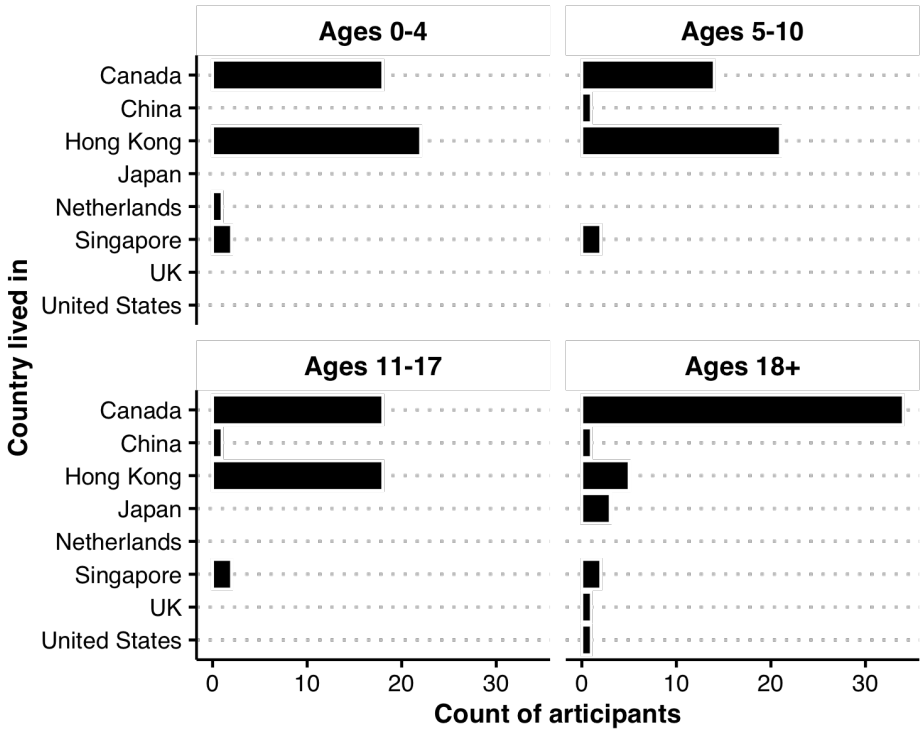

Figure 2.1: This four panel bar chart summarizes where the SpiCE participants lived during different portions of their lives.

The most well-represented group is Hong Kong, as 29 of 34 participants report having at least one caretaker who was primarily raised in Hong Kong. Of these, 20 report only having caretakers raised in Hong Kong. If caretakers birth location is considered instead, the numbers are 27 and 18, respectively.

Additionally, calling an individual a bilingual does not preclude knowledge of additional languages. All but one of the individuals represented in the SpiCE corpus report some degree of proficiency in a language other than Cantonese or English. The most common by far is Mandarin. The age SpiCE talkers began learning other languages varies widely, but is consistently later than (or simultaneous with) Cantonese and English. This information is depicted in Figures 2.3 and 2.4, with a panel for each participant.

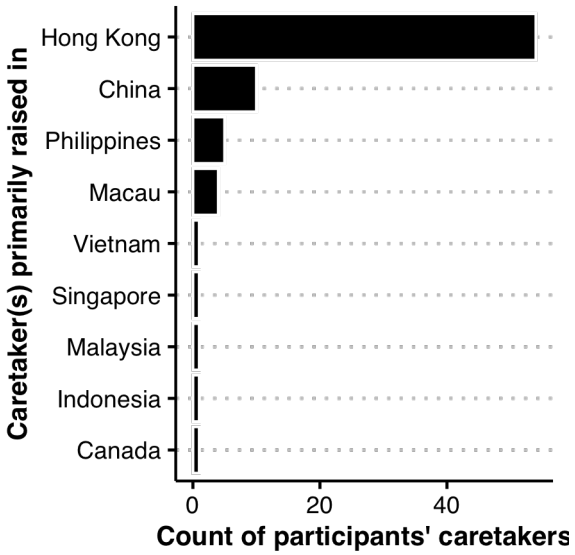

Figure 2.2: This bar chart summarizes the number of caretakers who were raised in various locations. Note that the number of caretakers reported by individual participants varies.

2.2.3 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 undergraduate bilingual research assistants 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 with a 44.1 kHz sampling rate, and 16-bit resolution.[2]

[2]Many files were originally recorded with 24-bit or 32-bit depth, but were converted to 16-bit depth before the publication of the SpiCE corpus, for the purpose of consistency and maintaining a reasonable file size while still providing high-quality audio.

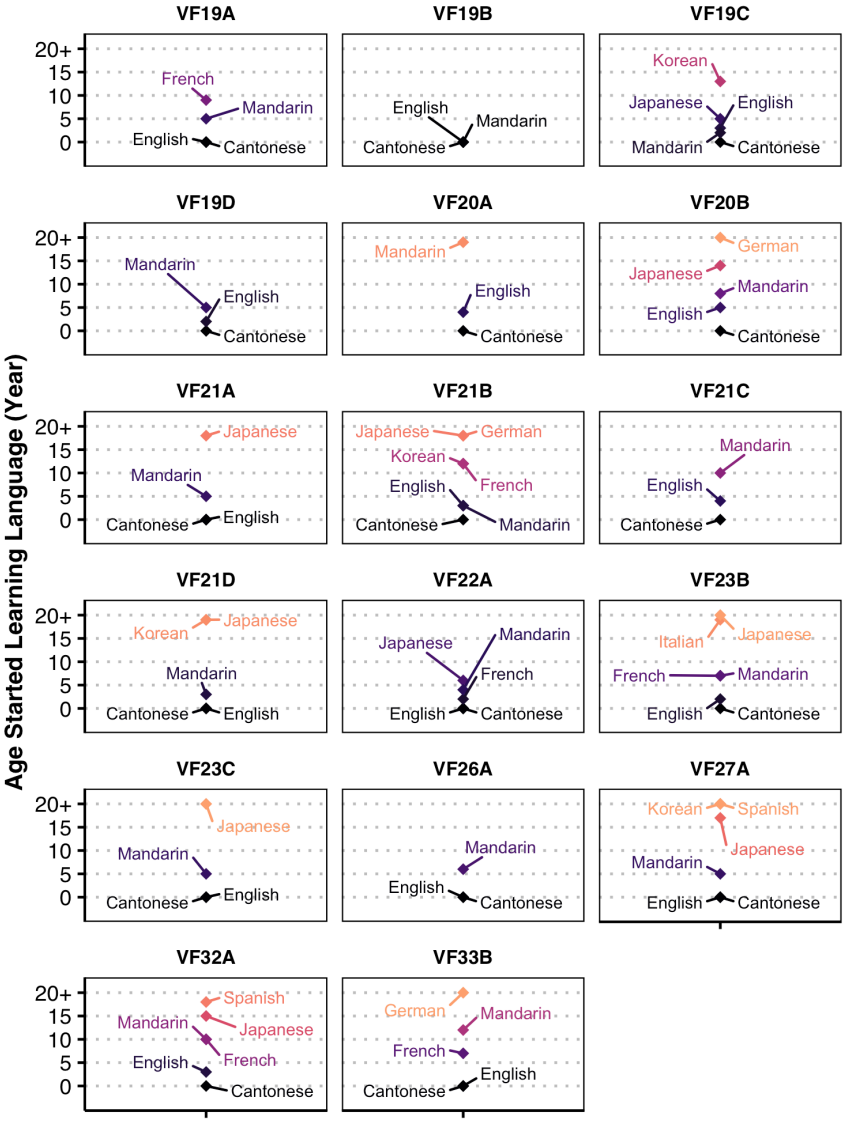

Figure 2.3: Multilingualism for the female participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

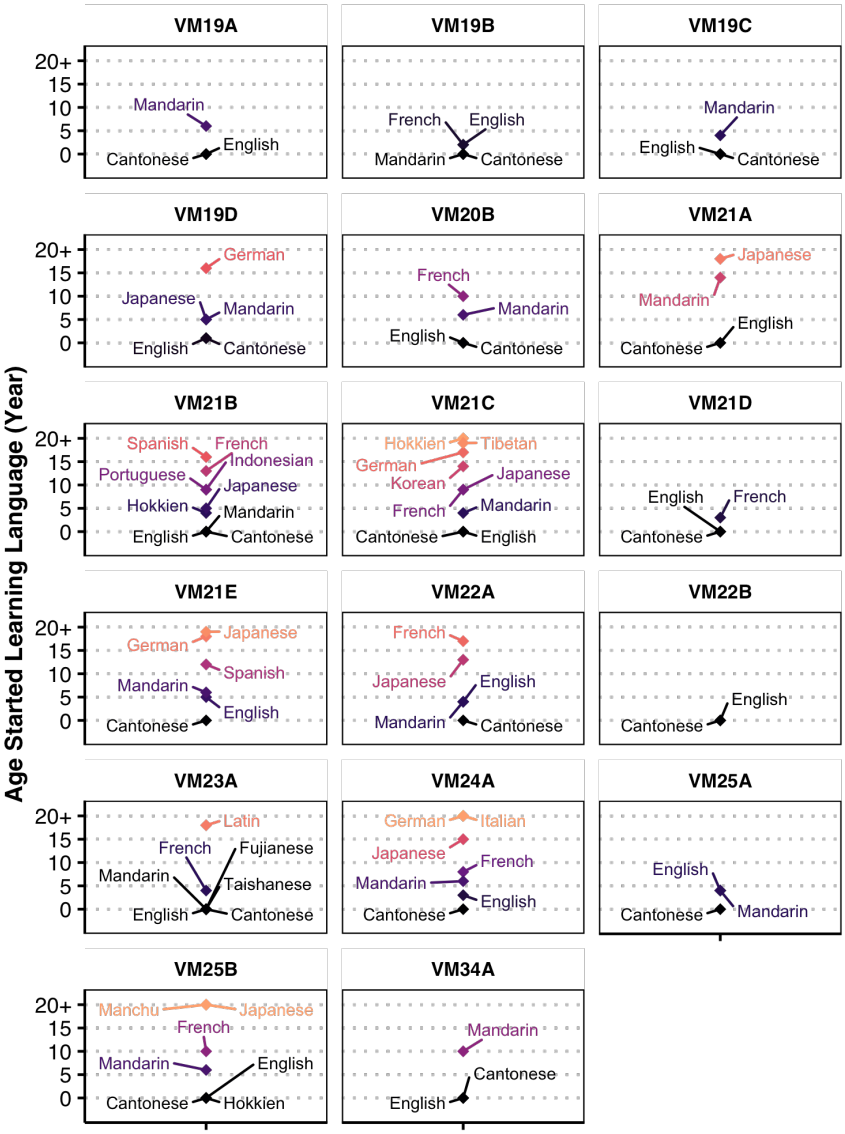

Figure 2.4: Multilingualism for the male participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

2.2.4 Recording Procedure

Upon arrival, participants were provided with an overview of the recording session procedures, and informed of the corpus publication process. This included informing participants that they would able to withdraw their data up until the SpiCE corpus' public release, and that they would receive notice at least 30 days before publication.[3] Subsequently, participants were asked to provide written consent. Upon consent, participants completed a set of tasks in English and the same set of tasks in Cantonese—all within the same session. The order of languages was counterbalanced across participants (see Table 2.2). This counterbalancing did not extend to other participant charactersitics, and as a result a higher proportion of the female participants completed the Cantonese part of the session before the English part, and vice versa for the male participants.

Each half of the session consisted of three tasks—sentence reading, storyboard narration, and a conversational interview—described in the following sections. While the primary focus of the recording session was the interview in each language, the sentence reading and storyboard narration tasks serve a practical purposes and add to the overall utility of the corpus. Rationale for each is described in the following sections. Each of these three tasks were recorded in the same audio file, though there are separate recordings for each half of the overall session. That is, each participant has a Cantonese recording and an English recording, each comprising the three tasks in that language. Together, each recording lasted approximately 30 minutes in each language. Along with the consent process, recording setup, and a break between interviews, participants spent up to 90 minutes in the lab.

Sentence Reading

Sentence reading was included in the session to ensure that different participants produced a set of identical items, considering the core of the session was an unscripted conversational interview (described in Section 2.2.4). 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

[3] No participants withdrew their data. At that time, participants were also encouraged to let the research team know if there were any portions of the interview they would like silenced from the published version.

for future analyses or projects that require matched utterances.

Participants first read the sentences listed in Table 2.4 and Table 2.3 aloud, pausing between sentences. Participants completed a single repetition and were not instructed to speak in a particular style. As participants had varying levels of Cantonese reading ability, they were simultaneously presented with the Cantonese characters, Jyutping romanization, and English translation.[4] The Cantonese sentences were well-known declarative phrases, typically associated with Chinese New Year.[5] 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 a 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.

In practice, the utility of these sentences may be somewhat limited, as sentences with speech errors were not necessarily repeated, and some Cantonese sentences were skipped altogether. In any case, the sentence reading task also served the purpose of getting participants into the appropriate Cantonese or English language mode before the upcoming interview. As such, they can be considered a warmup task.

Storyboard Narration

For the second task, participants narrated a short story from a cartoon storyboard originally developed for linguistic fieldwork (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. A reproduction of the storyboard is available with the corpus download. 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. Participants varied in how they approached

[4] 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).

[5] It is possible that familiarity and high frequency of some of these phrases led to them being produced with reduction patterns not present in typical reading. This is a limitation of the sentences.

No.	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 2.3: Sentences 1–10 comprise the Harvard Sentences List 60. Sentences 11–17 are holiday-themed imperatives created for this corpus to match the Cantonese sentences thematically.

No.	Cantonese	Jyutping	English translation
1	新年快樂	*san1 lin4 faai3 lok6*	Happy New Year
2	恭喜發財	*gung1 hei2 faat3 choi4*	Congratulations on happiness and prosperity
3	身體健康	*san1 tai2 gin6 hong1*	May your health be well
4	快高長大	*faai3 gou1 zoeng2 dai6*	Grow quickly
5	龍馬精神	*lung4 ma5 zing1 san4*	Have the spirit of the horse and dragon
6	學業進步	*hok6 yip6 zeon3 bou6*	Progress in your education
7	年年有餘	*lin4 lin4 yau5 yue4*	Excess in each year
8	出入平安	*cut1 yap6 ping4 on1*	Leave and enter in safety
9	心想事成	*sam1 soeng2 si6 sing4*	Accomplish that which is in your heart
10	生意興隆	*saang1 yi3 hing1 lung4*	Have a prosperous business
11	萬事如意	*maan6 si6 yu4 yi3*	A thousand things according to your will
12	天天向上	*tin1 tin1 hoeng3 soeng6*	Upwards and onwards every day
13	笑口常開	*siu3 hau2 soeng4 hoi1*	Laugh with an open mouth frequently
14	大吉大利	*daai6 gat1 daai6 lei6*	Much luck and much prosperity
15	五福臨門	*mm5 fuk1 lam4 mun4*	Five blessings for your household
16	招財進寶	*ziu1 coi4 zeon3 bou2*	Seek wealth welcome in the precious
17	盤滿缽滿	*pun4 mun5 but3 mun5*	Basins full of wealth

Table 2.4: All Cantonese sentences are widely-known imperatives associated with Chinese New Year.

this task, with some treating it as a series of picture description tasks, and others taking a more narrative approach. Despite this difference, this task may be useful for future analyses or projects that require utterances in a matched semantic space, as participants narrated the same cartoon in each language. This ensured that some of the same content was conveyed in each language (e.g., productions of *mother* in both languages). The storyboard narration lasted 4–5 minutes in each session and allowed participants time to continue getting used to the recording setup. As with the sentences, the storyboard narration also facilitated participants getting into the language mode of the session before the conversational 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. While within each language, the possible sequence of questions was the same, each interview took its own course, guided by what the participant wanted to talk about. This means that that the total number of general topics covered ranged from three to six. The interview materials are included with the corpus download. As a result, the speech samples from each language are comparable, but the specific questions differ between interviews and across participants.

Participants were informed explicitly that code-switching was acceptable. Ad-

ditionally, participants were implicitly 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. As will become apparent later in this chapter, there was substantially more code-switching in the Cantonese part of the session.

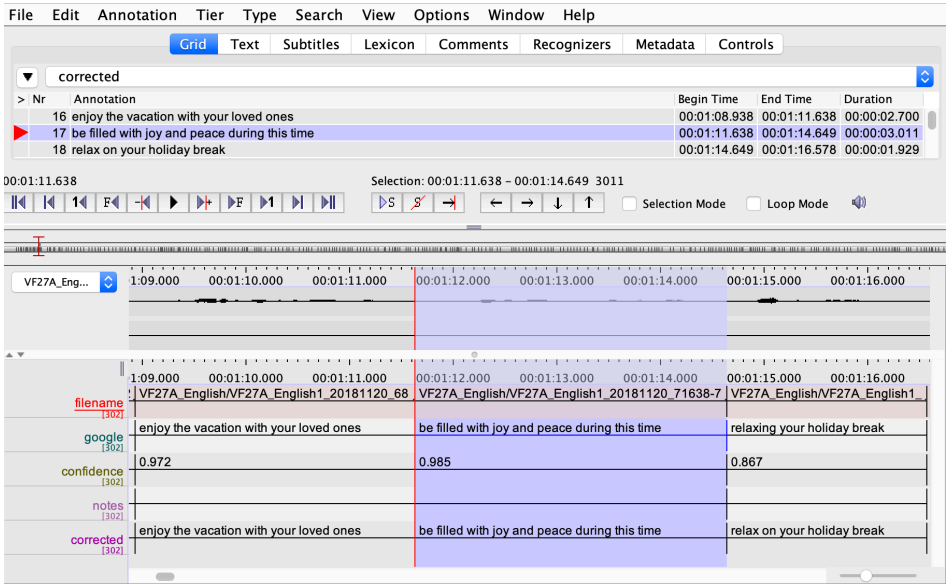

Figure 2.5: This screenshot from ELAN shows a sample of hand-corrected English from the sentence reading task for participant VF27A. The audio waveform is displayed in two channels, with one for the participant (top) and the other for the interviewer (bottom). The annotation tiers include (1) the short 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 were processed according to the pipeline outlined in this section. As much as possible, automatic tools were leveraged to expedite manual 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). To use this speech recognition product, the participant's speech was extracted from the participant's channel and segmented into short chunks, typically under 15 seconds in duration.[6] No attention was paid to constituents at this point; rather, breaks were placed at breaths and other pauses. Short chunks were necessary to use the speech recognition product with locally stored files, which was important for data privacy reasons. The short chunks would also prove useful 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 served the function of expediting the hand-correction process.

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 2.5 depicts an example of corrected English transcriptions in ELAN (Sloetjes and Wittenburg, 2008). Direct identifiers (e.g., names) were marked during this phase and silenced from the recordings prior to release. Transcriber guidelines were adapted from the multilingual Heritage Language Variation and Change cor-

[6] The interviewer's speech is included in the SpiCE corpus recordings for context but is not transcribed.

pus, which includes Cantonese (Nagy, 2011). Guidelines for Cantonese were developed in collaboration with the bilingual research assistant team.

In both languages, the following conventions were used:

- The placeholder "xxx" denotes unintelligible speech.
- Fragments are transcribed using "&" followed by the fragment produced (e.g., "&s").
- The "?" symbol marks questions but is not used consistently; other punctuation is not used.
- Words produced in a language other then English or Cantonese are transcribed in the language with, for example, "@m" appended to the end of each form for Mandarin (simplified characters), "@j" for Japanese, "@ml" for Malaysian, and "@i" for Indonesian.

Cantonese-specific conventions include:

- Where possible, transcription is in characters.
- Words without a standard character are transcribed in the Jyutping romanization system (e.g., *jyut6ping3*).
- Fully lexicalized syllable fusion[7] is transcribed with the smallest number of characters representing what was produced by the talker. For example, when fully fused, 乜嘢 (*mat1 ye5*, "what") is transcribed as 咩 (*me1*). In some instances, an intermediate form is produced. For this lexical item, the intermediate form would be transcribed as 咩嘢 (*me1 ye5*). Cases of fully lexicalized syllable fusion tend to be relatively clear to identify.
- Non-lexicalized (or ambiguous) cases of syllable fusion are transcribed with the full number of characters present in the un-fused form, but with brackets identifying which syllables are fused. For example, 朝頭早 ("morning") is pronounced *ziu1 tau4 zou2* in its full form but can be fused to *ziau14 zou2*—this fused form would be transcribed as 【朝頭】早.

[7] Syllable fusion is a phenomenon in which adjacent syllables in Cantonese are blended together. It ranges from assimilation at the syllable boundary to segment deletion and re-syllabification (Wong, 2006). Syllable fusion is common in Cantonese, though its frequency of occurrence and degree varies.

- Filled pauses are transcribed with the character 啞 (*e6*), or using Jyutping if different (e.g., *m6*).
- Transcribers followed a shared set of guidelines for transcribing sentence final particles. This includes the following common particles:
  - 呀 is the sentence-final particle used at the end of lists, and for exclamations and questions.
  - 呢 was used for both *ne1* and *le1* in marking questions.
  - 囉 was used as a sentence-final particle for marking emphasis.
  - 嚟 was the final particle used to expresses something being done or completed.
  - 咗 was the particle used after verbs to mark past tense.
  - While not a *final* particle, 咁 was consistently used as a filler in the words 咁嗎 "obviously" and 咁嘛 "isn't it'.'

English-specific conventions include:

- Standard spelling is used.
- Proper nouns are capitalized (e.g., "British Columbia").
- Filled pauses are transcribed with "um", "er", "uh", and other similar, non-elongated 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. The output of the forced alignment process was phone-level annotations for each audio file.

In Cantonese, forced alignment was completed with the Train-and-Align option, as there was no pre-trained model available for Cantonese. As Cantonese orthography does not separate words with spaces, words segmentation was done

using the *jieba* Python library (Sun, 2020), along with a Cantonese word segmentation dictionary designed for use with *jieba*.[8] While using an automated tool such as this is likely an imperfect solution, it has the benefit of reproducibility and consistency. This is important, as it can be difficult to define wordhood in Cantonese (e.g., see Wong, 2006).

The Cantonese pronunciation dictionary was generated using the *PyCantonese* Python library (Lee, 2018). Pronunciations were identified by getting the Jyutping romanization from each character or when transcription was done in Jyutping, using that existing Jyutping transcription. Next, the Jyutping was separated into segments, and the tone number was appended to the syllable nucleus (i.e., vowel or syllabic nasal). Research assistants supplemented the dictionary with alternative pronunciations for words that participated in syllable fusion. This approach bears some similarity to that of Tse (2019) but differs in that it also includes tonal information—which has been shown to improve forced alignment as long as there are not too many tone-nucleus combinations (Ćavar et al., 2016; Yuan et al., 2014).

Forced alignment in English took advantage of the Montreal Forced Aligner's pre-trained English model and pronunciation dictionary, which uses the ARPABET phone set. This dictionary broadly reflects North American English varieties. The dictionary was supplemented with manual additions, to minimize the number of out-of-vocabulary items.

The word and phone output of the forced alignment process were included in a Praat TextGrid for each audio recording, along with annotation tiers for the task (sentences, storyboard, and interview) and utterance (the short chunks). In both sessions, any material not in the main language of the session was not force-aligned and appears as "<unk>" in the word tier and "spn" in the phone tier, representing unknown words and spoken noise, respectively.[9] The force-aligned transcripts were not manually corrected or checked. This means that any short chunk with code-switching or unintelligible speech will likely have poorer alignment because the model does not have a representation for that span of speech, either in the phone

[8]The Cantonese Word Segmentation GitHub page: https://github.com/wchan757/Cantonese_Word_Segmentation.

[9]The Montreal Forced Aligner uses Kaldi conventions, and "spn" is short for "spoken noise." While in some models, it can be used to represent specific kinds of spoken noise, it is used here as a catchall unknown phones.

set or the pronunciation dictionary. As a result, it is advisable to use stringent exclusionary criteria or perform checks before analyzing data from the corpus.

A sample output from one of the Cantonese interviews of the final corrected and force-aligned transcript is provided in Figure 2.6.

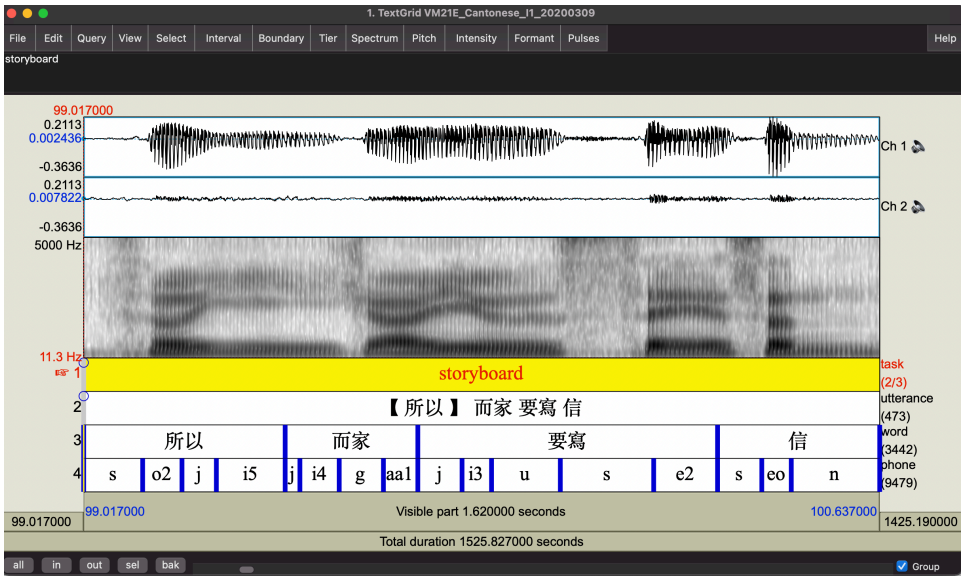

Figure 2.6: This screenshot from Praat shows what the final transcript looks like for a small portion of a Cantonese interview.

2.4 Descriptive Statistics

The descriptive statistics in this section are intended to give a general sense of the quantity and quality of the data in the corpus. They are based on the transcript data as described in the previous section, specifically the hand-corrected utterance tier and the force-aligned phone tier. Additionally, this section only reports on participant speech, though the interviewer's speech is included in its own channel in the stereo audio files.

2.4.1 Cantonese Interviews

The Cantonese recordings include 8.3 hours of speech: 13.6 minutes of sentences, 44.0 minutes of storyboard narration, and 7.4 hours of conversational interview data. These estimates are calculated from the summed duration of all non-silent intervals in the phone tier of the transcripts, and as such, do not include interviewer questions or any pauses in the participant's speech.

In the Cantonese interview sessions, there were a total of 8,112 word types and 90,512 word tokens. The number of words varies substantially across participants, with a mean of 749 word types (SD=157, minimum=483, maximum=1081) and 2,662 word tokens per interview (SD=637, minimum=1,654, maximum=4,212). The numbers reported here include all types of "words"—Cantonese words, English words, words in other languages, phonological fragments, and unintelligible stretches of speech. Figure 2.7 shows the split of these categories on a by-participant basis within the Cantonese interview sessions. Figure 2.7 indicates that all participants switch to English during the Cantonese interview sessions. The amount of switching varies across participants, with VF19D producing an especially large number of English words. While the other three categories also vary, they are comparatively small in number.

The overall distribution of word frequency in the Cantonese interviews is depicted in Figure 2.8. As expected, there are a relatively small number of words occurring frequently (e.g. pronouns, function words, etc.), while a majority are mid and low frequency. This pattern follows what is expected in a word frequency distribution, and is reassuring given the automated method of segmenting the Cantonese transcripts into words.

2.4.2 English Interviews

Using the same estimation technique as used for Cantonese, the English recordings include 8.9 hours of speech: 21.9 minutes of sentences, 45.7 minutes of storyboard narration, and 7.7 hours of conversational interview speech.

The English interviews include a total of 4,972 word types and 104,618 word tokens. As in the Cantonese interviews, the number of words varies substantially by participant, with a mean word type count of 609 (SD=119, minimum=434, maxi-

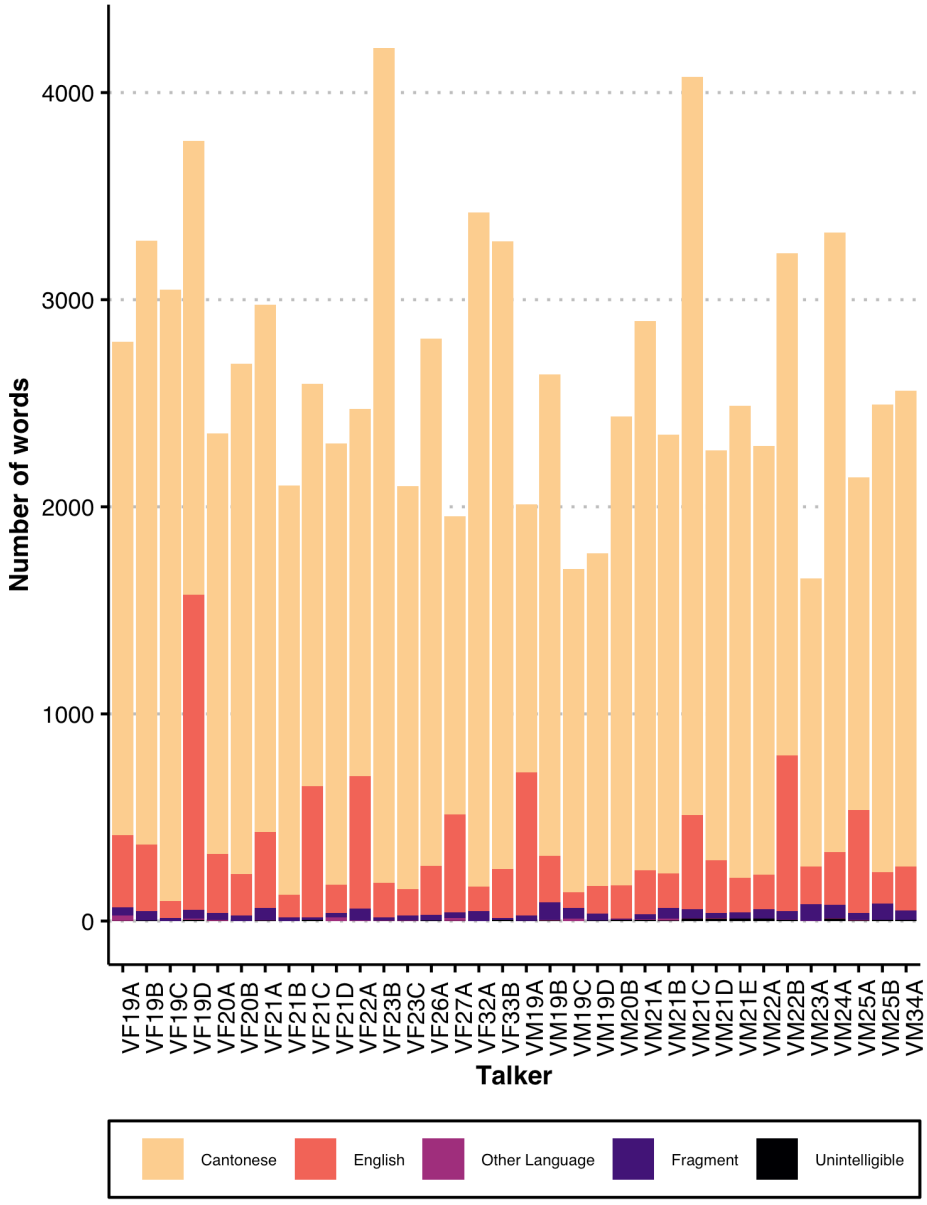

Figure 2.7: The total word count for each participant's Cantonese interview task is represented by bar height. Color indicates the kind of item counted.

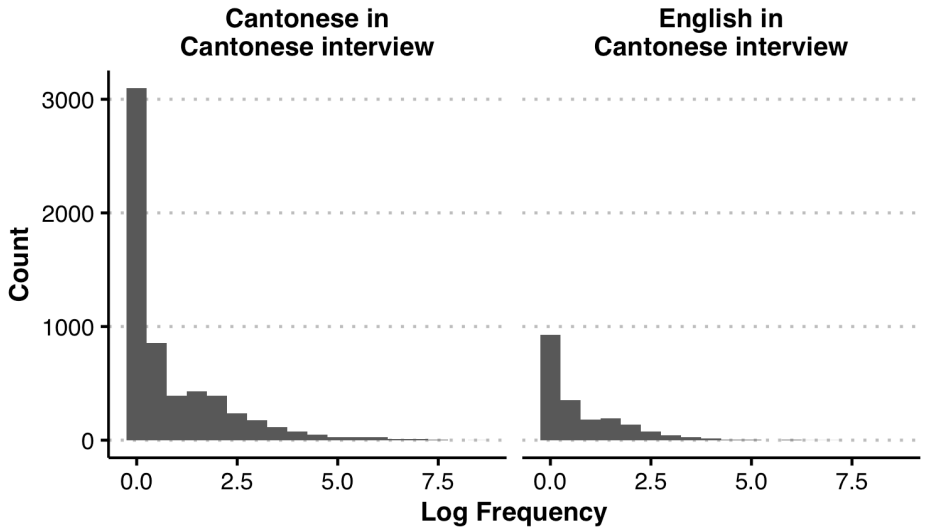

Figure 2.8: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

mum=904) and token count of 3,077 (SD=701, minimum=1,907, maximum=4,240). Figure 2.10 shows the split of these categories on a by-participant basis within the English interview sessions. Unlike the Cantonese interviews, there were relatively few switches to Cantonese, with 12 of the 34 participants producing fewer than 10 Cantonese words during the English sessions.

The distribution of log word frequency for both Cantonese and English words in the English interviews is portrayed in Figure 2.9. Word frequency follows a similar pattern to Cantonese word frequency, with most words occurring infrequently, and a smaller proportion occurring very frequently.

2.5 SpiCE Corpus Release

The SpiCE corpus was publicly released in May 2021 through the Scholars Portal Dataverse platform under a Creative Commons Attribution 4.0 International License.[10] In addition to the corpus itself, documentation is available online—the

[10]https://creativecommons.org/licenses/by/4.0/

t[h]

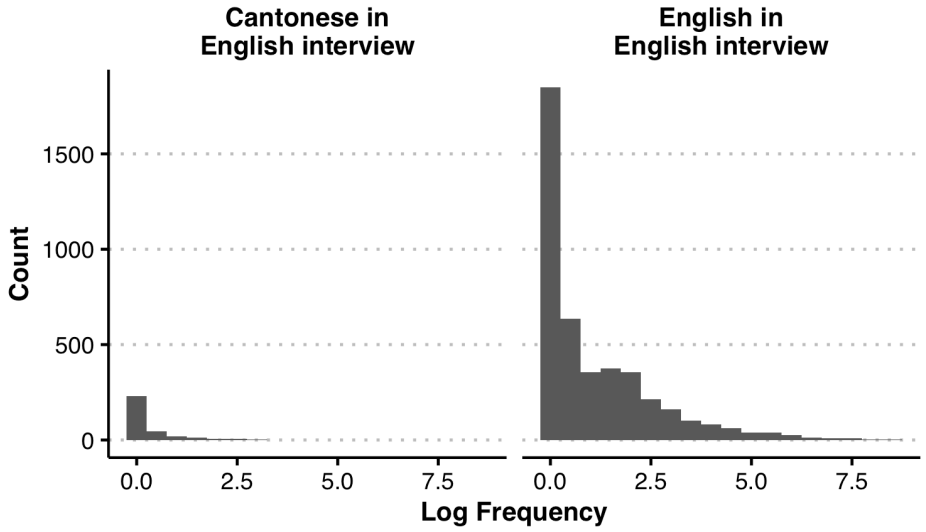

Figure 2.9: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

URLs are given in Section 2.2.

2.6 Discussion & Conclusion

While various bilingual corpora exist, they lack in different ways *for the purpose of doing corpus phonetics*. 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 in such resources limits the scope of phonetic research. With the release of SpiCE and its high-quality recordings, scholars can ask and answer empirically and theoretically motivated research questions within the speech and language sciences using more sophisticated phonetic measurement techniques (e.g., spectral measures, in addition to temporal measures). This presents substantial potential for increasing our understanding of bilingual spoken language from both phonetic and psycholinguistic perspectives. While the recording quality of this cor-

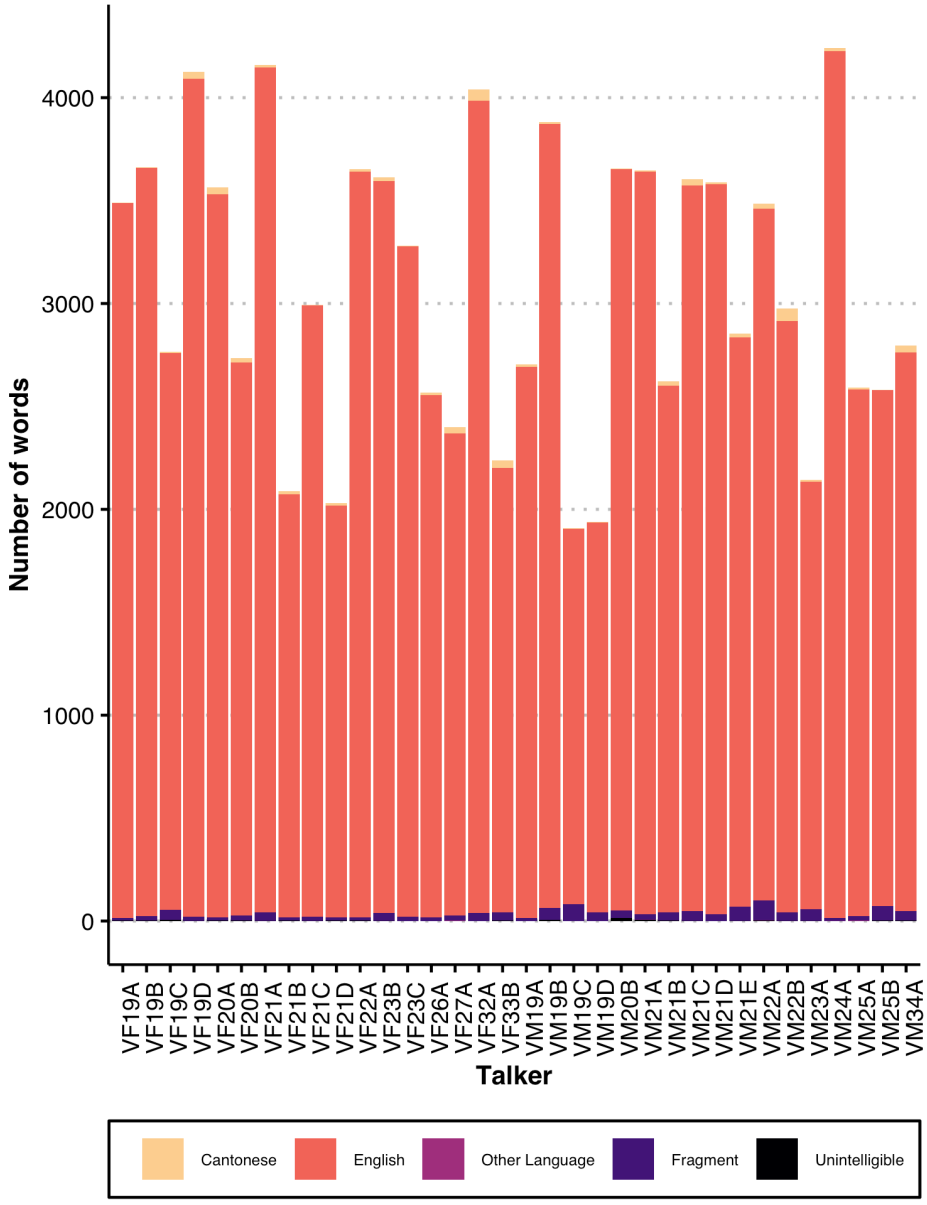

Figure 2.10: The total word count for each participant's English interview task is represented by bar height. Color indicates the kind of item counted.

pus 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 in bilingual speech

3.1 Introduction

Voices provide a lot of information about the person talking, ranging from their current physical and emotional state to talker indexical features that help listeners identify who they are. In this context, voices can be described as auditory faces—they are uniquely individual yet share basic characteristics with the broader population (Belin et al., 2004). Voices convey this rich array of information along with the message being communicated. Understanding the structure of a voice is no small feat, as it means understanding how listeners leverage different vocal dimensions to process talker-indexical, affective, social, and linguistic information. The processing challenge arises from the sheer variability across voices.

Though voices share some attributes, they also vary in unique ways (Lee et al., 2019). From the perspective of voice perception, the balance between shared and idiosyncratic characteristics makes sense. The shared dimensions allow listeners to recognize the sound they hear as a voice. They also help listeners perceive, classify, and understand new voices. Idiosyncrasies, on the other hand, enable identification and discrimination between different voices. While this makes sense conceptually, understanding the structure of voice variation in speech production and its complement in listeners' ability to process that information remains an active area of

research. The focus of this chapter is acoustic voice variability, and the emphasis on describing and processing variation echoes one of the big puzzles in phonetics: the "lack of invariance" problem (Liberman et al., 1967). That is, given the ubiquity of variation, how do perceivers efficiently extract relevant and important information from the communicative signal? This chapter foregrounds the speech signal itself, asking what is available in the speech production for listeners to use.

While variation is indeed wide-ranging, it remains far from random. Some of the most prevalent accounts of how individuals understand and process variation emphasize its structure. While this chapter looks at the structure of voices and the following chapter examines sound category structure, both attempt to elucidate structure in the speech signal that may be beneficial for listeners in understanding and processing new talkers.

In the domain of voice, Kreiman and colleagues have synthesized work from various areas and put forth a psychoacoustic model of voice quality (Kreiman et al., 2014). This model features a minimal set of acoustic dimensions necessary to encode and thus reproduce voice quality. While there are numerous dimensions in the model, extensive experimental work has validated the inclusion of each one (Kreiman et al., 2021, and references therein). As a result, Kreiman and colleagues argue that this set of dimensions is both sufficient and necessary to capture a wide range of normal and disordered voices. The psychoacoustic model of voice quality includes acoustic dimensions that capture vocal tract anatomy, pitch, loudness, harmonic voice source, and inharmonic voice source characteristics. While each dimension in the model could be considered independently by researchers—some of which are studied in isolation—Kreiman and colleagues argue that these dimensions are more than the sum of their parts. The measures covary and conspire together to form a multidimensional percept of voice. While this model establishes a set of acoustic dimensions, it does not arbitrate between them in a way that establishes what matters for perceiving and processing a particular voice in a particular language.

There is a large body of literature focused on understanding differences in variability across populations for a small set of acoustic measurements. Such studies typically compare summary statistics for fundamental frequency (F0) and a handful of spectral measures. This body of work is summarized below in the context

of crosslinguistic comparisons. Before summarizing this work, it is important to highlight that very little of it dives into the structure of voice variability, which is a relatively new area spearhead by Lee and colleagues (Lee et al., 2019; Lee and Kreiman, 2019, 2020). In this set of studies examining acoustic voice variation in different languages and speech styles, Lee and colleagues leverage the psychoacoustic model of voice quality (Kreiman et al., 2014) and adapt methods from the domain of face variability and perception (Burton et al., 2016). Their driving question is one of understanding the structure of acoustic information in the speech signal. In many ways, this is the first step towards understanding which aspects of voice are available to listeners and thus useable in perceptual processes, particularly when coarse summary statistics do not indicate cross-talker differences.

To drill down into the structure of voice variability, Lee et al. (2019) use a series of principal components analyses to investigate how acoustic measurements pattern with one another. The techniques used in this study will be described in greater detail in the Methods section of this chapter. In their original paper, Lee examines the structure of variability on a within-talker basis as well as across the larger speech community represented within the University of California, Los Angeles Speaker Variability Database (Keating et al., 2019). Crucially for the comparison with their later work, this study focused on relatively small samples of sentence reading.

The takeaway from this work is that different voices share a handful of dimensions with one another and the group as a whole. Despite this shared structure, however, much of the way a voice varies is idiosyncratic. Commonly shared dimensions were spectral shape and noise parameters in the higher frequencies, the fourth formant, and formant dispersion. These spectral measures are associated with vocal breathiness or brightness. The formant-based measures are typically associated with speaker identity and vocal tract size. Lee and Kreiman (2019) replicates this work with short samples of spontaneous speech from the same database. The results were similar, with the exception that F0 emerges as a shared relevant dimension. This result arguably reflects the difference between reading and spontaneous spoken English, with reading tending to be more monotonous and spontaneous speech exhibiting more affective qualities. Lee and Kreiman (2020) replicates this work again with sentence reading in Seoul Korean, again finding minimal differences that are explained readily by typological differences from English. Unlike English,

F0 and variability in the lower formants emerged as relevant dimensions in read Korean speech. The authors argue that this reflects phrasal intonation patterns that occur in Korean reading.

Conceptualizing what these dimensions mean and how to think about acoustic voice variability in this way is challenging, as many of the acoustic dimensions considered do not map neatly onto a single percept. F0 is a straightforward example, given its clear relationship to pitch. Many of the spectral measures, both harmonic and noise-based, are much more challenging to interpret without considering multiple measures simultaneously. The domain of faces thus provides a useful analogy for thinking about what shared structure looks like compared to idiosyncratic aspects of the structure. Burton et al. (2016) found that all faces share dimensions of variability related to things like lighting and viewing angle (i.e., looking up, down, or to the side). Idiosyncratic variation in face structure arose from things like facial hairstyle, makeup, and expressions. Burton et al. (2016) discuss their results as supporting a prototype model of faces. As with the face literature, Lee and colleagues argue that the structure of voice spaces supports a prototype model of voice perception (Lavner et al., 2001; Latinus and Belin, 2011), in which novel individual voices are perceived in the context of a speech community average, or prototype. Their results add to the argument that faces and voices share processing mechanisms (Yovel and Belin, 2013).

In any case, Lee et al. (2019) argue that familiarity with a voice arises from learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions. And indeed, familiarity with a voice pays off—listeners are good at identifying familiar voices but perform poorly on the same tasks with unfamiliar voices (Nygaard and Pisoni, 1998). The prototype model merely proposes a mechanism by which listeners learn a novel talker's voice.

The literature on voice perception has approached the question of what listeners use in voice identification, discrimination, and learning through the lens of familiarity (Levi, 2019; Perrachione, 2018). This body of experimental work pairs different combinations of listeners, talkers, languages, and stimuli manipulations to probe how listeners identify and discriminate among talkers. While identification and discrimination are often talked about in conjunction with one another, the processes are likely supported by different perceptual mechanisms (Perrachione et al.,

2019). One of the biggest takeaway points from this literature is the Language Familiarity Effect (LFE), which encompasses a broad range of findings where listeners are better at identifying talkers in a familiar language (for a recent review, see Perrachione, 2018). Bilinguals are especially good at this kind of task and show evidence of generalizing across languages (Orena et al., 2019).

Very little of this work identifies what listeners use in the signal, and as such, claims about the relative importance of linguistic or talker-indexical information should be tempered. However, there are exceptions to this. For example, Perrachione et al. (2019) collected perceptual voice (dis)similarity ratings for Mandarin and English voices by Mandarin and English native listeners and report on the relationship between several acoustic measurements and rating data. Perrachione et al. (2019) found that when the talker was the same, regardless of the manipulations used in the study (language and time-reversal), all listeners rated stimuli pairs as highly similar. This result highlights that listeners are sensitive to low-level acoustic information present in voices, regardless of whether they know the language or understand the stimuli. Additionally, Perrachione et al. (2019) found that some acoustic measurements predict similarity ratings, while others do not. F0 was the most prominent measure, which is unsurprising given its salience, and how much the voice variability literature has focused on it (e.g., Keating and Kuo, 2012). Other measures predicting similarity were the harmonics-to-noise ratio and formant dispersion, which are associated with voice quality and vocal tract size, respectively. That listeners appear to use these measures is of direct relevance to the study presented in this chapter, and represents a point that will be returned to in this chapter's discussion.

In light of this perceptual work on the language familiarity effect and the complicated interactions that abound between different listener and talker populations, it makes sense that Lee et al. (2019) restricted variability while introducing a novel set of methods. Their extension to spontaneous English and Seoul Korean demonstrates that this method replicates well and that it also presumably allows for observing typological differences across languages that can affect voice quality. This chapter builds on Lee and colleagues' body of work by extending their methods to the case of spontaneous bilingual speech.

Describing and analyzing acoustic voice variation in bilingual speech has moti-

vation in both perception and production. As apparent from the language familiarity effect literature, listeners are capable of learning and identifying voices in one language and then generalizing across languages. Listeners are better at identification and discrimination when they have more familiarity with the language, but performance on such tasks tends to be well above chance (e.g., Orena et al., 2019). In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic information in the voice—or so the argument goes (Perrachione et al., 2019). Understanding the structure of that variability brings us one step closer to understanding what listeners are using from the signal to process speech, as it limits the hypothesis space. On the production side of things, bilingual speech presents an ideal test case for the argument that voices function like auditory faces. If the structure of variability from each of a bilingual's languages is matched, then voices can be straightforwardly thought of as auditory faces.

Additionally, examining the structure of the same talker's voice in each language lends additional validation to the arguments made by Lee and Kreiman (2020) for the differences between English and Seoul Korean sentence reading. In comparing these studies, Lee and colleagues argue that both language and biological factors contribute to the structure of voice variation. Bilingual speech, again, presents an ideal test ground for disentangling biological and linguistic factors from one another. While common in the literature, this dichotomy is somewhat misleading. Voices ultimately have biological constraints, such as vocal tract lengths or pathologies. Yet at the same time, individuals nonetheless exert remarkable and wide-ranging control over their voice space and are highly capable of manipulating factors that are not linguistically important but which signal social and contextual information. This applies across all aspects of an individual's linguistic repertoire (Bullock and Toribio, 2009; Wei, 2018). Thus in the case of bilinguals, the only aspect we can be truly confident in being held constant across languages is the biological part. The same "hardware" can be used for drastically different ends.

In this chapter, I examine how voice varies across a bilingual's two languages. Some differences are expected, despite the characterization of voices as auditory faces. While all languages have consonants and vowels, they differ in distribution, articulation, and acoustics (e.g., Munson et al., 2010). Suprasegmental and prosodic properties also vary. Languages differ in terms of whether a suprasegmental di-

mension is made use of at all. For example, does a language encode lexical tone contrastively? Another way languages vary in this respect is in how they carve up the suprasegmental space. For example, how many lexical tones are there? What shapes of tone are present? This particular question is relevant in the present case, where the languages considered are Cantonese (a language with lexical tone) and English (a language without lexical tone). Segmental and suprasegmental differences both have cascading effects on voice quality.

The following paragraphs detail comparisons that have been made between English and Cantonese in the literature thus far. As there is an additional body of work comparing English and Mandarin Chinese—typologically similar to Cantonese—comparisons between English and Mandarin are also summarized. While the most relevant comparisons for the present work are those made within bilinguals, some of the relevant literature compares separate populations. What this work has in common, is that it paints with relatively broad strokes—crosslinguistic comparisons are often made with summary statistics for a small set of spectral measurements. Results have been decidedly mixed.

In a small study of Cantonese-English bilingual (n=9), Russain-English bilingual (n=9), and English monolingual (n=10) young women, Altenberg and Ferrand (2006) examined F0 patterns in conversational speech across the different languages and populations. As some languages reportedly have different mean F0 (e.g., Keating and Kuo, 2012), Altenberg and Ferrand (2006) focused on whether F0 shifts when an individual switches languages and whether different languages have different baselines. Ultimately, Russian-English bilinguals exhibited differences in mean F0, and Cantonese-English bilinguals did not. Though, they did produce a wider F0 range in Cantonese compared to their English. While the results in Altenberg and Ferrand (2006) ultimately paint a coarse picture of bilingual F0 production with a small sample size, they highlight an important point of departure—bilinguals can differ in F0 across languages.

In a larger study of Cantonese-English bilinguals reading passages (n=40), Ng et al. (2012) examined a variety of different voice measures with both male and female talkers. Results were based on Long-Term Average Spectral (LTAS) measures. Female talkers exhibited higher F0 in English than Cantonese, but males did not. 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. The LTAS measure of the first spectral peak did not differ across languages, suggesting that vocal stiffness remained consistent in the bilinguals' two languages.

Ng et al. (2010) examined F0 in spontaneous speech from 86 Cantonese-English bilingual children and found it to be lower in Cantonese compared to English. This corroborates Ng et al. (2012), and goes against the nonsignificant difference in (Altenberg and Ferrand, 2006). This mixed bag of results could ultimately be attributed to differences in sample sizes, the quantity of speech analyzed, or in language backgrounds of the bilinguals studied. While the picture regarding voice quality measures appears clearer and more consistent, those conclusions arise from a single study. In any case, these three studies offer reason to expect that Cantonese and English might differ in measures associated with pitch and phonation type.

The authors of these studies speculate that Cantonese's status as a tone language may account for some of these differences compared to English. Though it is important to emphasize that this explanation is pure speculation. In this light, it is also relevant to consider the larger body of research comparing voice quality for Mandarin and English. Lee and Sidtis (2017) compare F0, speech rate, and intensity in a small group of Mandarin-English bilinguals (n=11) across three different tasks. They report a higher mean F0 for Mandarin reading compared to English, but no differences in the other tasks (picture description and monologue). Additionally, there were no differences in F0 variability across languages or tasks. Lastly, while there were no differences in intensity, the bilinguals spoke faster in Mandarin. Lee and Sidtis (2017) speculate that Mandarin's status as a tone language may account for the higher mean F0 in reading, as it echoes some prior work with separate populations of English and Mandarin speakers, in which Mandarin tends to have higher and more variable F0 (Keating and Kuo, 2012). This finding may be strongly associated with the type of bilinguals studied. Xue et al. (2002) found that Mandarin-English bilinguals aged 22-35 produced higher F0 in English than Mandarin. This group differed from the participants in Lee and Sidtis (2017), in that they are described as non-native English speakers. Producing higher F0 in a

non-native language arguably reflects factors like stress or confidence (Järvinen et al., 2013; Lee and Sidtis, 2017).

The speculation that higher F0 is a feature of tone languages does not align with the observation in Ng et al. (2012), who argued the opposite for Cantonese: that lower F0 could be accounted for by lexical tone. While the tone inventories for Cantonese and Mandarin have substantial differences, it seems clear that appealing to the presence or absence of lexical tone is too simplistic of an account. Alternatively—or perhaps, concurrently—talkers may be expressing different social and cultural identities in each of their languages (Loveday, 1981; Voigt et al., 2016). Regardless of whether language, experiential, or social factors drive differences across languages, this body of work highlights the importance of comparing within the same task.

Treating Mandarin and Cantonese as similar just because they are both tone languages may not be appropriate, though there is little in the way of conclusive research on the topic. In a study with 12 Cantonese-Mandarin bilinguals who are Cantonese-dominant, Yang et al. (2020) found no differences in their F0 profiles across languages. F0 profiles were characterized by F0 minimum, maximum, range, and mean. The authors also examined a Mandarin-dominant group and reported clear differences between the two populations' F0 profiles in Mandarin. The Mandarin-dominant individuals produced higher F0 with a narrower range. While the conclusions from this study are tenuous given the small sample size, it nonetheless highlights an important point: that typologically related tone languages may not necessarily behave comparably.

While the studies reviewed thus far provide a mixed picture of voice differences across language pairs, there is a strong focus on F0. Both the F0-centricity and variable outcomes are apparent in work on other language pairs as well. For example, Cheng (2020) finds that Korean has consistently higher F0 than English, regardless of whether they were early sequential or simultaneous bilinguals, but that differences in F0 range differed for cisgender males and females. This result builds on the findings for Korean-English bilinguals (Lee and Sidtis, 2017). While the results for Korean-English bilinguals seem to be straightforward, the same cannot be said for other language pairs. For example, Ryabov et al. (2016) look at rate, duration, and F0 for Russian-English bilinguals, finding no F0 differences, but that

Russian was faster. This result goes against the findings for the bilinguals studied in Altenberg and Ferrand (2006), where Russian exhibited consistently higher F0 than English. While higher F0 and slower speech rates can be characteristics of speech by non-native or non-dominant speakers (Järvinen et al., 2013), such an explanation cannot account for both outcomes.

Another example of less than clear-cut results comes from Ordin and Mennen (2017). They demonstrate differences in F0 range and level across languages for female Welsh-English bilinguals in a reading task, for whom Welsh had a higher and wider F0 range. This result did not hold for males from the same population, who varied more in their F0 level and range. The authors argue that the crosslinguistic difference is likely to be sociocultural in this case, as different patterns were observed for male and female speakers on a within-speaker basis. This gender difference means that the result is unlikely to be due to anatomical or purely linguistic reasons.

Considering these studies together, a few key observations are especially relevant to the present chapter. While studying bilingual talkers provides a clear path to disambiguating the role of anatomical differences in voices, it does not necessarily facilitate disentangling linguistic and sociocultural factors from one another. Most likely, both contribute simultaneously to the differences in voice patterns across languages. For example, there is clear evidence that Korean has a higher F0 than English, given results from two studies with different populations of bilinguals Cheng (2020); Lee and Sidtis (2017). Conversely, Ordin and Mennen (2017) show social rather than linguistic stratification.

This body of work mostly focuses on linguistic and social differences. While some of it dives into individual differences, between-talker variability should perhaps be given more of a spotlight. In work with speech rate, Bradlow et al. (2017) found that some talkers are fast and others are slow and that some languages are fast while others are slower. Crucially, these relationships held across talkers in various languages. That is, if someone was a fast talker in their dominant language, they were also a fast talker in their non-dominant language, and likewise for slow talkers. In this sense, both talker-indexical and linguistic (or sociocultural) factors contribute to speech rate behavior. It is not a particularly big leap to suggest that other speech signal variables might pattern in the same way. Adding to this picture

of variability across individuals, it is important to remember that bilinguals are sophisticated social actors and are fully capable of tailoring their speech behavior to a wide variety of contexts (Bullock and Toribio, 2009).

While this body of work highlights important points, it is limited by its laser focus on F0, with occasional forays into speech rate, intensity, and other spectral measures. The focus on F0 is not without reason—Perrachione et al. (2019) found it to be the most important perceptual dimension for voice similarity ratings. Yet at the same time, there is so much more to voice than pitch, particularly if the characterization of voices as auditory faces holds up.

This chapter brings together work describing crosslinguistic voice differences and work describing the structure of acoustic voice variation, to provide a more comprehensive picture of how voices vary across languages. Using the corpus introduced in 2, I describe various spectral properties Ng et al. (e.g. 2012), and also examine how acoustic variation is structured, following the work of Kreiman, Lee, and colleagues (Kreiman et al., 2014; Lee et al., 2019). This chapter builds on Lee et al. (2019) in a handful of ways: it extends the methods to the case of bilinguals, considers longer samples, and addresses the role of sample duration both within and across talkers and languages. I also extend their methods by introducing a mechanism to assess structural similarity within and between individuals and languages.

3.2 Methods & Results

3.2.1 Data

The data used in this analysis comes from the conversational interviews in the SpiCE corpus described in Chapter 2. The analysis uses both Cantonese and English interviews. As noted before, the 34 talkers studied here are all early Cantonese-English bilinguals from a heterogeneous speech community (Liang, 2015). For additional information about the participants, please refer to sections 2.2.2 and 2.4 in the previous chapter.

While prior work by Lee and colleagues (e.g., Lee et al., 2019) uses relatively short chunks of speech, the present analysis is focused on longer stretches of spontaneous speech. While it would have been possible to include the sentence reading

and storyboard task recordings from each participant, there are practical reasons for excluding them from the analysis. The sentence sets were overall quite short and thus unlikely to be sufficiently representative on their own. Additionally, as many of the SpiCE talkers were not confident in their Cantonese reading, there was a wide range of familiarity with the materials represented. Some talkers knew all of the sentences, and others struggled with some of them. This renders the sentences less comparable to their English counterparts in the SpiCE corpus. There are also imbalances in the storyboard task. As talkers narrated the same story in both languages, they were often more confident the second time around. Excluding both of these tasks is motivated by prior work that highlights how confidence (Järvinen et al., 2013) and speaking style (Lee and Sidtis, 2017) impact voice quality.

As discussed in the previous chapter, the recordings are high-quality, with a 44.1 kHz sampling rate, 16-bit resolution, and minimal background noise. Recall that both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. For the analysis in this chapter, the participant channel was extracted from the stereo recordings, including any code-switches they made during the interview. While it would be possible to exclude items not produced in the primary language of the interview, this was not done. The driving reason for keeping code-switches in the analysis is that such code-switches are representative of the particular talker's language behavior. Further, just because someone switches languages, does not mean that they fully and immediately switch language modes (e.g., Fricke et al., 2016). For example, individual words may be borrowed and pronounced with the phonology of the interview's primary language (c.f.., the matrix language in code-switching Myers-Scotton, 2011).

All voiced segments were identified with the *Point Process (periodic, cc)* and *To TextGrid (vuv)* Praat algorithms (Boersma and Weenink, 2021), implemented with the Parselmouth Python package (Jadoul et al., 2018). The pitch range settings used with *Point Process (periodic, cc)* were 100–500 Hz for female talkers and 75–300 for male talkers. 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, and likely exerted little to no influence on the results. This method of identifying voiced portions of the speech signal captures vowels, approximants, and

some voiced obstruents. This result of this process differs slightly from the methods described in Lee et al. (2019), the paper on which the methods of this chapter were modeled. Lee et al. (2019) examined only vowels and approximants.

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), except formant dispersion, which was excluded given its near-perfect correlation with the measured value of F4. The choice of measurements in Lee et al. (2019) is based on the psychoacoustic voice quality model described in the introduction to this chapter (Kreiman et al., 2014), as well as the availability of algorithms in the software used to extract measurements. Measurements were made every 5 ms during voiced segments in VoiceSauce Version 1.28? (Shue et al., 2011). The measurements are described below. Note that the shorthand name for each measurement is presented in boldface, and will be used throughout the rest of the chapter.

F0 Fundamental frequency is a correlate of pitch and is associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics. F0 was measured in Hertz using the STRAIGHT algorithm (Kawahara et al., 2016), which is regarded to be more accurate than other options in VoiceSauce. It is one of the more widely studied variables on this list, as evidenced by the literature cited in the introduction.

F1, F2, and F3 The first three formant frequencies—also measured in Hertz—are typically discussed for linguistic contrasts, particularly with vowels and sonorant consonants. A total of four formants were estimated using the Snack Sound Toolkit method Sjölander (2004), with the default settings of 0.96 pre-emphasis, 25 ms window length, and 1 ms frameshift.

F4 The fourth formant frequency is not typically discussed in linguistic contexts and is instead associated with talker characteristics, such as vocal tract length. In this light, it is not particularly surprising that it was highly correlated with formant dispersion. F4 is also measured in Hertz. It was calculated along with the first three formants, using the same settings.

H1*–H2* The corrected amplitude difference between the first two harmonics is one of four primary measures used to characterize source spectral shape—also called spectral tilt—in the psychoacoustic model of voice quality (Kreiman et al., 2014). It is typically associated with phonation type but can be confounded by nasality (Garellek, 2019; Munson and Babel, 2019). The asterisks here—and in the following spectral shape measures—indicate that the value has been corrected (Iseli et al., 2007), to account for the amplifying impact of nearby formants on the amplitudes of harmonics. This allows for different vowels and other voiced segments to be compared with one another. This amplitude difference is measured in dB. Note that this measure—along with the following three spectral shape measures—depends on an accurate F0 measurement.

H2*–H4* The corrected amplitude difference between the second and fourth harmonics is the second of four measures capturing spectral shape. It is associated with phonation type and is measured in dB.

H4*–H2kHz* The corrected amplitude difference between the fourth harmonic and the harmonic closest to 2000 Hz is the third spectral shape measure. Unlike the previous two, one of the harmonics depends on F0, while the other does not. It captures shape in a higher frequency range and is also associated with phonation type. Like the other spectral shape measures, it is in dB.

H2kHz*–H5kHz* The amplitude difference between the harmonics closest to 2000 Hz (corrected) and 5000 Hz (uncorrected) is a measure of harmonic spectral shape that does not depend on F0. The amplitude of the harmonic nearest 5000 Hz is not corrected by VoiceSauce, given inaccuracies in the correction algorithm at higher amplitudes. It captures the highest frequency band of the four shape measures, reflects phonation type and is measured in dB.

CPP Cepstral Peak Prominence measures the degree of harmonic regularity in voicing, and as such, it is associated with non-modal phonation types. VoiceSauce computes CPP according to the algorithm in Hillenbrand et al. (1994). Specifically, CPP measures the difference between the amplitude of the peak

in a cepstrum and the value at the same quefrency on the regression line for that cepstrum. It is measured in dB.

Energy Root Mean Square (RMS) Energy is a measure of spectral noise that reflects overall amplitude and is calculated over a window comprising five pitch periods. Energy is measured in dB.

SHR The subharmonics-harmonics amplitude ratio is a measure of spectral noise associated with period-doubling or irregularities in phonation. VoiceSauce's implementation is based on the algorithm described in Sun (2002). While based on amplitude, this ratio is unitless.

The raw VoiceSauce output used in this chapter is available in a repository on the Open Science Framework, in the data subfolder at https://osf.io/9ptk4/. The analysis code used for the following sections is available on GitHub, at https://github.com/khiajohnson/dissertation. Note that the diss repo is currently private!

3.2.3 Exclusionary criteria and post-processing

Given the nature of the corpus and the level of automation in the methods thus far, there is reason to expect a sizable number of erroneous measurements. To filter these out before analysis, measurements were subjected to exclusionary criteria focused on identifying impossible values. Observations were excluded in cases where any of the following measurements had a value of zero: F0, F1, F2, F3, F4, CPP, or uncorrected H5kHz. Observations were also excluded if Energy was more than three standard deviations above the mean. This may exclude some valid measurements but removes the long right tail of likely erroneous measures, as humans can only produce speech so loud.

Filtering based on F0 and the four formant frequencies reflects the observation that zero measurements are not possible for voiced portions of the speech signal. The interpretation for zero in CPP would indicate there is no cepstral peak, that is, no regularity in the voicing. In this sense, a zero for CPP likely also reflects either a lack of voicing or an erroneous F0 measurement. Lastly, only the uncorrected spectral measure for H5kHz was used in filtering, as erroneous values tended to co-occur on the same observation. The distribution of H5kHz did not span zero, except

for a spike of erroneous values equal to zero. This operationalization minimizes the removal of correctly measured zero values, which occurred with all of the other spectral shape parameters, whether corrected or uncorrected.

Moving standard deviations were calculated for each of the 12 measures using a centered 50 ms window, such that each window includes approximately ten observations. The moving standard deviations capture 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). This analysis uses moving standard deviations, as opposed to the coefficients of variation used by Lee et al. (2019). This should not have any undue effect on the outcome, as all variables were scaled before inclusion in the principal components analysis described in the next section. The last round of exclusionary criteria uses these moving standard deviations. If an observation was missing a moving standard deviation value, it was removed. Given the centered window, this means that observations falling less than 25 ms away from a voicing boundary were not included.

There were 24 total measures, with a measured value and a moving standard deviation for each of the acoustic measurements listed above. These 24 measures were used in the analyses described in the following sections. Across the 34 talkers, there were 3,071,736 observations after winnowing the data from an initial count of 6,560,403 observations. These observations were not evenly distributed across talkers and languages. While this full set of observations is perfectly valid for the crosslinguistic comparison in Section 3.2.4 and is used there, sample size is likely to have an impact on the principal components based analysis in Sections 3.2.5 and 3.2.6.

To control for the impact of sample size in that part of the analysis, the number of samples for each talker was capped to include only the first 20,124 samples for each interview. This value was selected as it represents the interview with the fewest observations. Put simply, differences in sample size reflect the variability in how much different individuals in the corpus talked. Those who produced longer passages of speech ultimately had more observations of voiced speech. Passage length was expected to impact the analysis, given how much affect and style can vary within a single conversation. Over time, individuals cover more of their range of variation, and as such, a regression to the mean is expected over time. To level

the playing field in this first analysis, the sample size was controlled. At the end of this chapter, in Section 3.2.7, a follow-up analysis validates this assumption. To preview those results, 20,000 samples appear sufficient for capturing the range of variability in acoustic voice variation.

Following this last winnowing step, there were 1,368,432 total observations. While the winnowing process removed a substantial amount of the data, the total number of samples per talker is still much larger than the approximate 5,000 used in Lee et al. (2019).

3.2.4 Crosslinguistic comparison of acoustic measurements

Following prior work, the first step in this analysis is a crosslinguistic comparison for each talker and measure. As discussed in the introduction to this chapter, there are some commonly found—though inconsistent—differences between Cantonese and English. Prior work has found that speakers sometimes produce lower and more variable F0 in Cantonese (Altenberg and Ferrand, 2006; Ng et al., 2012, 2010). Additionally, Ng et al. (2012) also report on spectral measurements that indicate Cantonese has a generally more breathy (or less creaky) phonation quality compared to English. Other measures were either inconclusive, non-significant, or not considered by the researchers. Figure 3.1 depicts the distribution of values for each of the acoustic measurements across languages, with all talkers pooled together.

For each acoustic measurement and talker, I conducted a Student's *t*-test and calculated Cohen's *d* using the *lsr* package (Navarro, 2015) in R (R Core Team, 2020); this provides 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, though most talkers (32/34) had at least one non-trivial comparison. The distribution of these counts is depicted in Figure 3.2.

For the non-trivial comparisons, there were consistent patterns across languages for a handful of the variables, including F0, H4*–H2kHz, and to a lesser extent, H1*–H2**. If there was a non-trivial difference in F0 across languages, then Can-

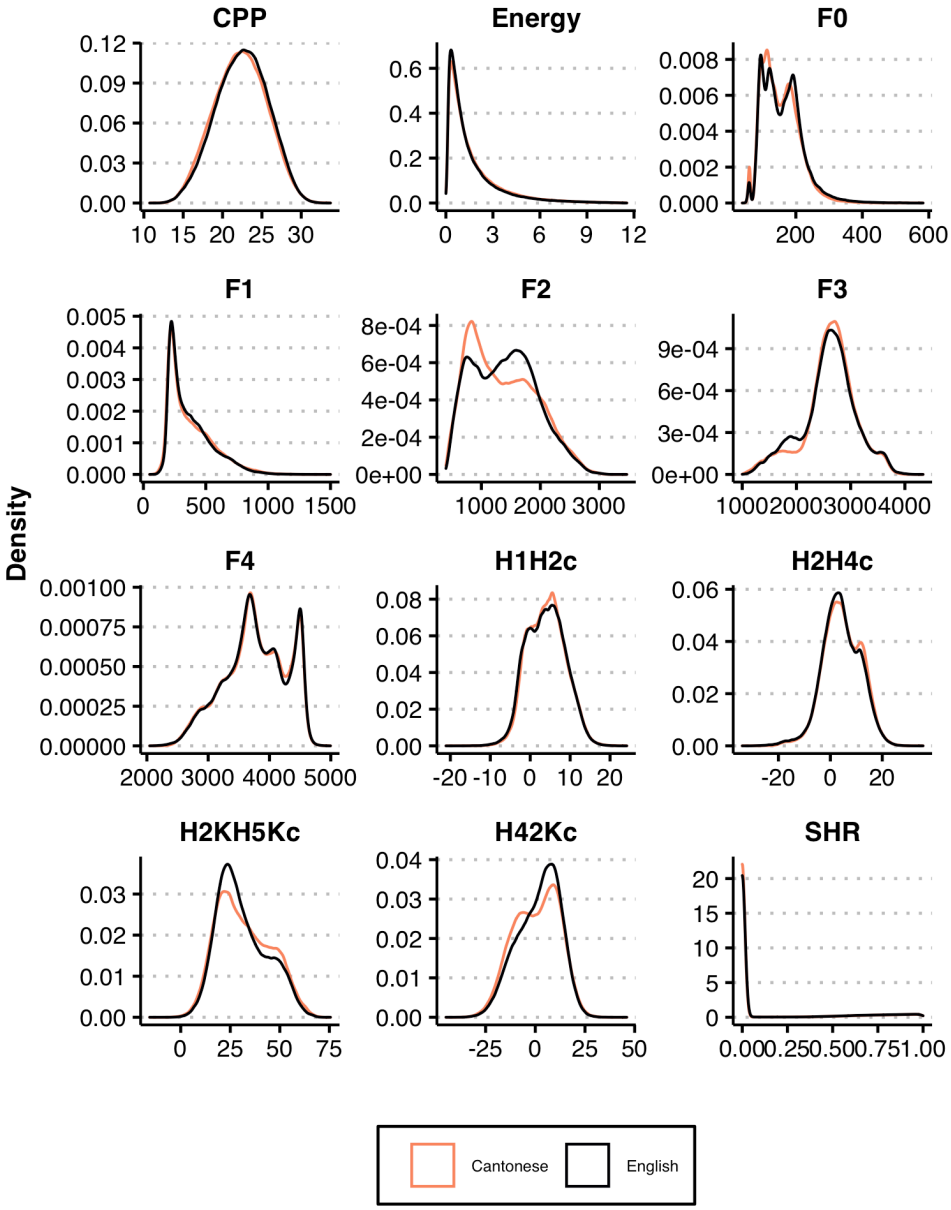

Figure 3.1: Each panel depicts a density plot that pool measurments from all talkers together to show the range of values for that measure. The x-axes each have their own scale. Language is separated out by color.

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

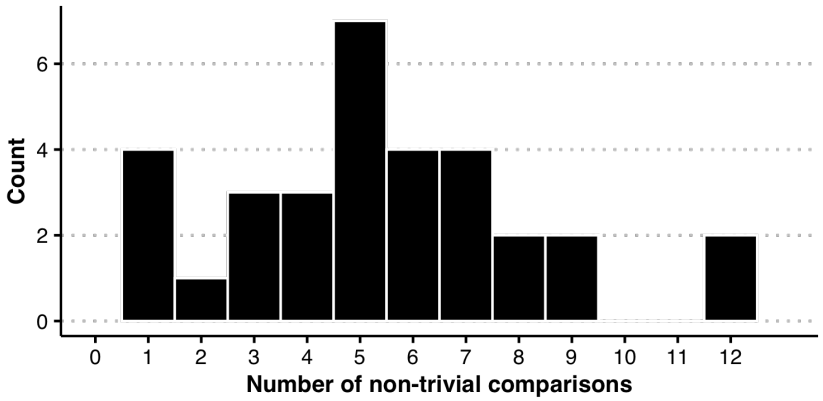

Figure 3.2: A histogram summary of the number of non-trivial comparisons from Table 3.1 across the 34 talkers.

tonese 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). I also observe this difference for a small number of males.

As for the two spectral shape measures, H4*–H2kHz was consistently lower in Cantonese when the comparison was not trivial (n=9), though most talkers did not exhibit a difference on this measure. 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.

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. For example, a variable like F4 would be unlikely to vary across languages, given its association with vocal tract size. This is reflected in the relatively low count of talkers with a non-trivial difference across languages for F4.

Other measures, such as Energy, have a high number of nontrivial comparisons

but show a relatively even split for direction (Positive = 7, Negative = 10). The large spread for Energy may reflect things like speaking confidence in the two languages, which likely varies by individual (Järvinen et al., 2013).

CPP also exhibits a split between positive (n=6) and negative (7). Higher CPP values are associated with both breathy or creaky non-modal phonation types. In this sense, a positive difference would indicate that Cantonese was more non-modal, while a negative difference would indicate that English was more non-modal. Interpreting CPP is not so straightforward, however, as it is not immediately clear which type of non-modal phonation the measure entails. Given the results of H1*–H2**, it seems clear that knowing where on the creaky-modal-breathy spectrum a given speaker falls is pertinent to interpreting this measure. CPP would likely corroborate that outcome. How much more interpretation should I add in here?

Overall, while talkers show some clear across-language differences, these are far outnumbered by instances with no difference. The variability observed here fits in with the variable outcomes of previous work but does not necessarily fall neatly along the lines prior work would suggest that male and female talkers fall along.

3.2.5 Principal components analysis

Methods

Principal components analysis (PCA) is a dimensionality reduction technique appropriate for data with many potentially correlated variables. In the case of voices, distilling numerous acoustic dimensions into a smaller number of components facilitates identifying and describing the structure of voice variability. PCA provides insight into how variables pattern together in a data set. This feature of PCA is especially relevant here, as voice perception research has made it clear that individual acoustic measurements may be necessary to capture and encode a voice but may not be perceptually meaningful to listeners. What matters is how the different pieces conspire together to form a percept.

Often, the goal of PCA is to take a large number of dimensions and extract a much smaller set to use for some additional purpose (e.g., linear regression). The

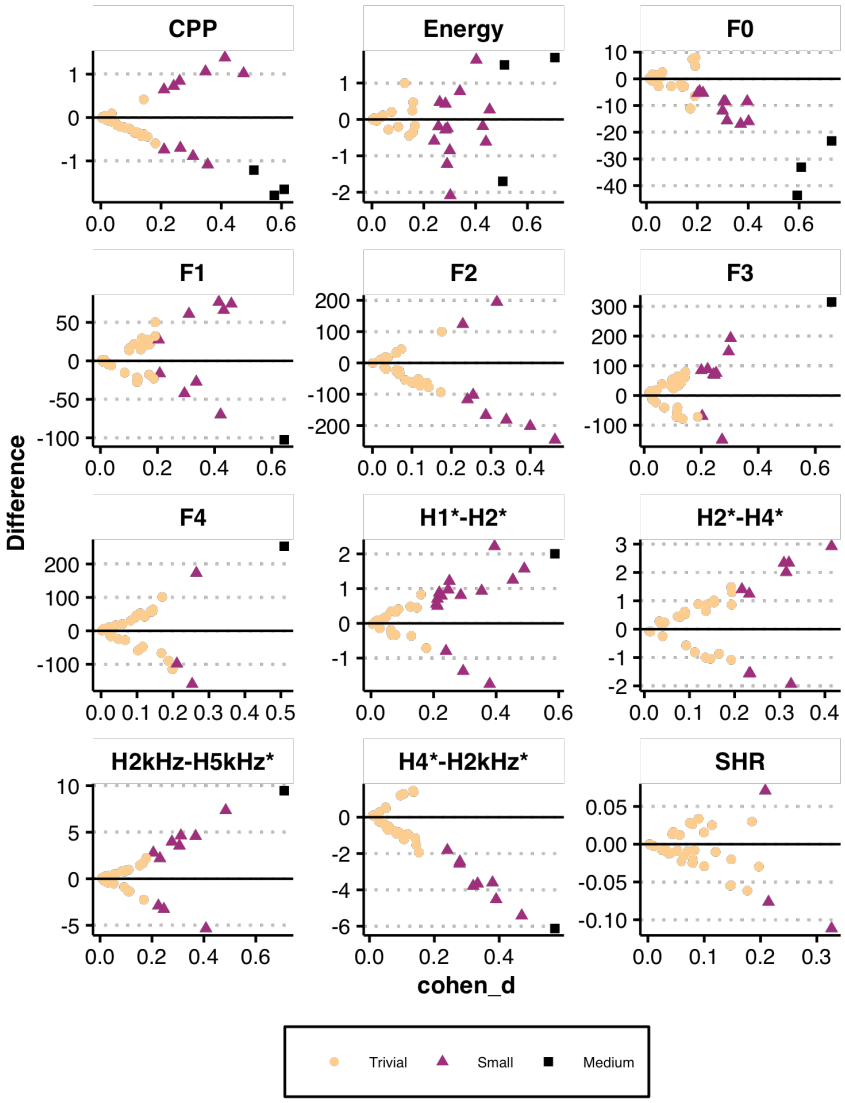

Figure 3.3: Each panel plots Cohen's *d* on the x-axis, and the difference between means from the t-tests on the y-axis. Positive values indicate a higher mean in Cantonese than English. The color reflect the levels of interpreations for Cohen's *d*.

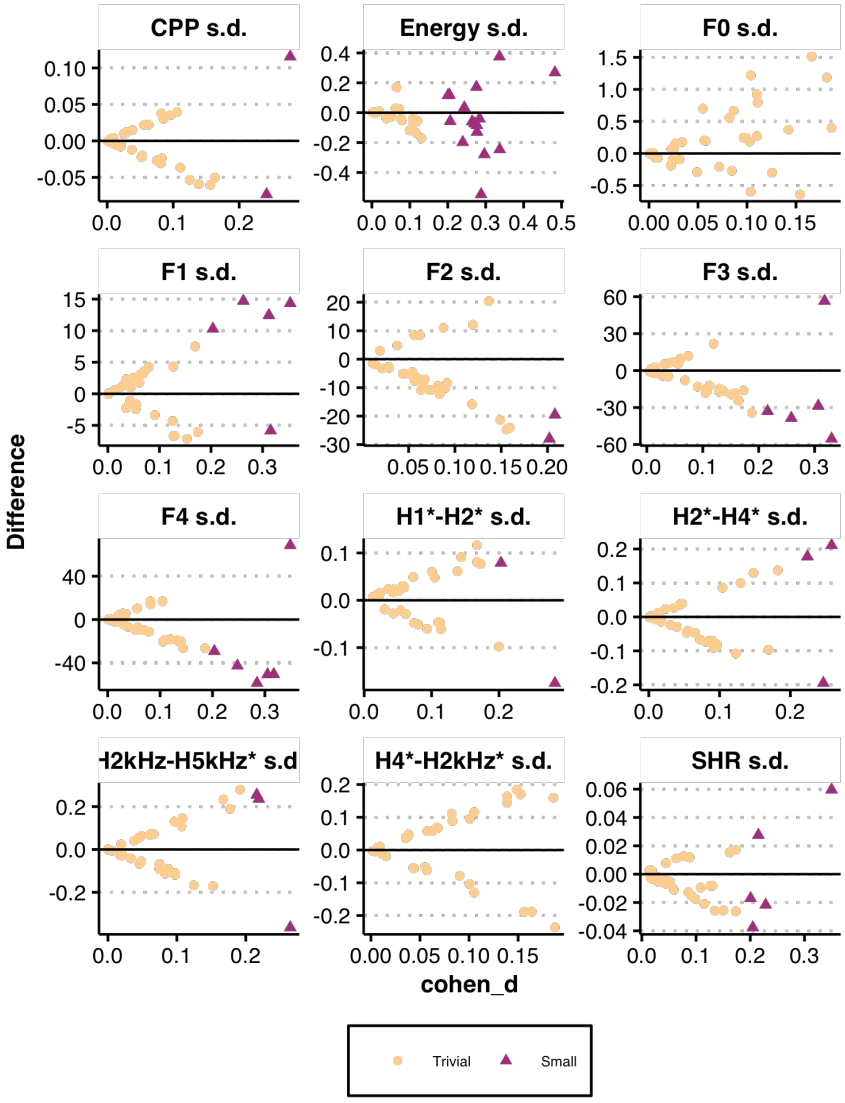

Figure 3.4: This figure is a continuation of 3.3.

focus in this chapter is on the internal structure of the components. That is, I examine what makes up components for different talkers and whether an individual's voice structure varies (or not) across languages.

I adapt methods from work on voices (Lee et al., 2019; Lee and Kreiman, 2020) and faces (Burton et al., 2016; Turk and Pentland, 1991). The goal of this analysis is to capture similarities or differences in the structure of each talker's voice across languages. As such, I conducted PCAs separately for each talker-language pair and compared the results of each talker's English and Cantonese PCAs. All 24 measures were standardized on a by-PCA basis before 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 given prior findings (Lee et al., 2019) and a broader understanding of how different acoustic measures align with one another (Kreiman et al., 2014, 2021).

Each PCA included the number of components for which all resulting eigenvalues were greater than 0.7 times the mean eigenvalue, following Jolliffe's (2002) recommended adjustment to the Kaiser-Guttman rule. This rule was used in place of a more sophisticated test (e.g., broken sticks), as it is not detrimental to this exploratory analysis to err on the side of including marginal components. Additionally, across each of the components, only loadings with an absolute value of 0.45 or higher were interpreted (Lee et al., 2019; Tabachnick and Fidell, 2013). While Lee et al. (2019) use a threshold of 0.32, Tabachnick and Fidell (2013) note that higher loadings indicate that a particular variable is a better measure of the component, with 0.32 corresponding to poor (but still interpretable) overlap between the variable and the component. The guidelines in Tabachnick and Fidell (2013) indicate that loadings of 0.45 correspond to fair, 0.55 to good, 0.63 to very good, and 0.71 and above to excellent. Given the large number of components and loadings in this analysis, only loadings over the fair threshold are interpreted.

Results

The PCAs across both languages for all 34 talkers resulted in 10–14 components and accounted for 74.9–82.7% of the total variation. Half of the talkers had the

same number of components for each language (17 of 34). Of the remainder, 16 of the 34 talkers had a difference of one in the number components, and only one had a difference of two. Talkers had 4–11 identical component configurations across their languages (M=7.82). These shared components represent 33.3%–91.7% of the total components for talkers (M=66.7%). The numbers comprising these summary statistics are provided in Table 3.2. While this already indicates a substantial amount of shared lower-dimensional structure across languages, it likely underestimates the actual shared structure. The reason is that similarity of component structure is not taken into account (i.e., a component of F2, F3, and F4 versus a component with just F2 and F3), as is the case in Section 3.2.6.

To assess whether talkers exhibit the same structure in voice variability across their languages, I first consider the patterns present across the different PCAs. This provides context for understating what unique structural characteristics in talkers' voices looks like. To this end, I briefly summarize common patterns across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 3.5 shows all of the components of participant VF32A's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages. It also highlights the importance of not attributing too much value to the ordering of components, but rather to their composition and variance accounted for.

Broadly, there were many similarities in component composition across talkers and languages. The following paragraphs summarize the components that were present in every PCA, regardless of talker or language. The shared component accounting for the most variation across talkers had a core structure consisting of F2 and H4*-H2kHz*. These usually went along with H2kHz*-H5kHz* (Cantonese = 34, English = 31), and occasionally with F3 and F4 (Cantonese = 3, English = 3). In a similar vein, all talkers had a component consisting of H4*-H2kHz* s.d. and H2kHz*–H5kHz* s.d., though it accounted for a smaller proportion of the total variation. Should I describe what these mean here? Or in the discussion?

In the case of the moving standard deviation parameters, there were a few common configurations. Formant s.d. parameters often co-occurred. In both languages, the component typically consisted of F3 s.d. and F4 s.d. (Cantonese = 32, English = 26), though a subset of these cases also included F2 s.d. (Cantonese = 6, En-

Table 3.2: The number of components and the variance accounted for is listed for each PCA. The last column indicates the number of identical components across langauges.

	Cantonese		English		
Talker	**N**	**Variance**	**N**	**Variance**	**Identical N**
VF19A	11	0.77	12	0.80	8
VF19B	12	0.78	12	0.78	8
VF19C	12	0.78	12	0.79	9
VF19D	13	0.81	13	0.78	9
VF20A	11	0.78	12	0.79	6
VF20B	13	0.81	12	0.82	8
VF21A	12	0.78	12	0.80	6
VF21B	12	0.78	12	0.80	8
VF21C	14	0.83	13	0.83	10
VF21D	12	0.79	12	0.81	9
VF22A	11	0.78	12	0.80	7
VF23B	12	0.78	12	0.78	8
VF23C	12	0.79	12	0.80	7
VF26A	12	0.78	13	0.80	7
VF27A	11	0.79	11	0.77	8
VF32A	12	0.78	11	0.76	8
VF33B	12	0.77	12	0.79	9
VM19A	12	0.78	11	0.76	5
VM19B	11	0.80	12	0.80	6
VM19C	11	0.76	11	0.78	6
VM19D	13	0.80	14	0.82	10
VM20B	12	0.80	11	0.76	9
VM21A	10	0.78	11	0.79	5
VM21B	11	0.79	11	0.76	9
VM21C	12	0.80	12	0.77	9
VM21D	11	0.75	12	0.77	7
VM21E	10	0.74	12	0.80	7
VM22A	12	0.77	13	0.83	11
VM22B	12	0.79	12	0.79	7
VM23A	12	0.81	12	0.79	4
VM24A	11	0.77	11	0.76	8
VM25A	12	0.81	12	0.77	11
VM25B	11	0.74	12	0.76	7
VM34A	11	0.77	12	0.81	10

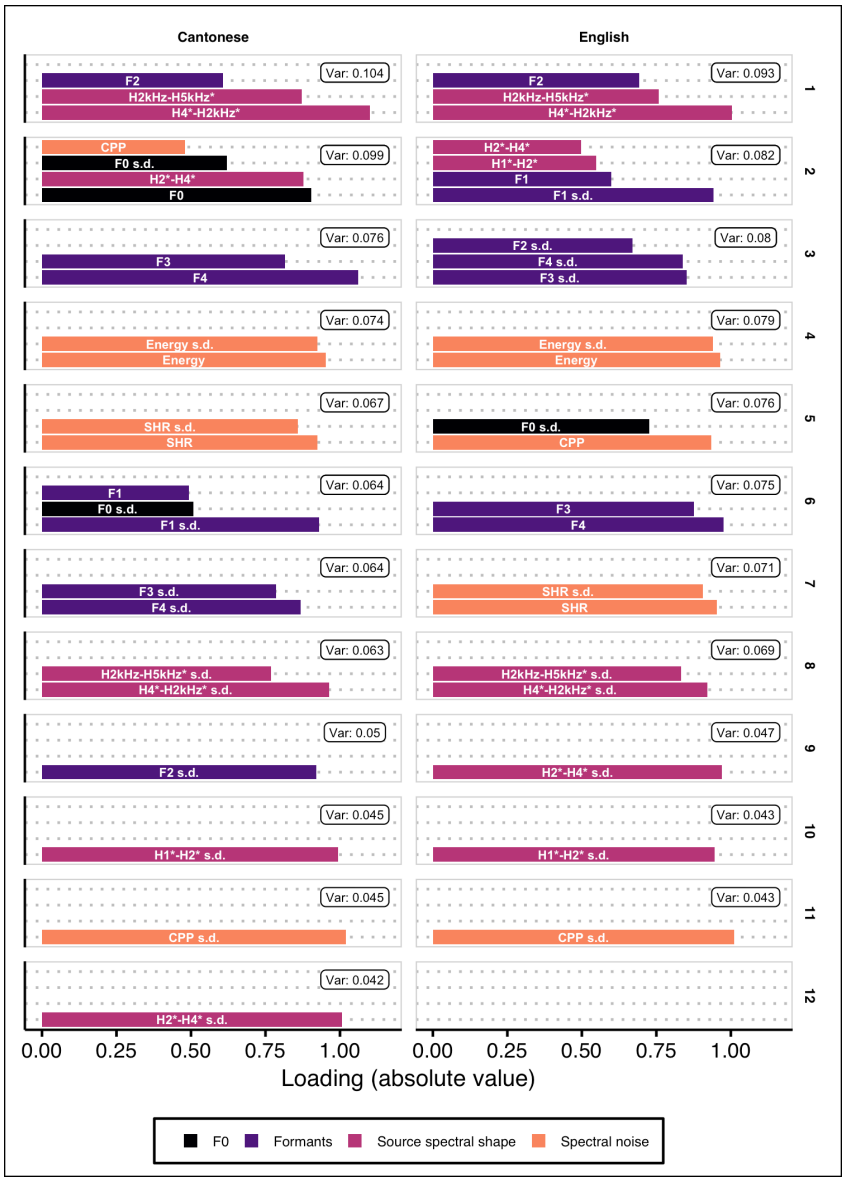

Figure 3.5: In this depiction of the components of VF32A's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

glish = 10). In the case of spectral shape, the variable for H2*–H4* s.d. commonly occurred alone (Cantonese = 18, English = 18) or in combination with H1*–H2* s.d. (Cantonese = 13, English = 14). While the formant and spectral shape moving standard deviations often exhibited these common patterns, variables in these categories were just as likely to pattern in more idiosyncratic ways, loading alongside each other, F0, formants, and spectral measures. This kind of variability is not readily summarizable.

The spectral noise parameters had a relatively consistent component structure across talkers and languages. Energy and Energy s.d. consistently loaded on the same component and were sometimes accompanied by F0 (Cantonese = 6, English = 2) and F0 s.d. (Cantonese = 1). CPP s.d. occurred consistently on its own component for all English PCAs, and 31 of the Cantonese PCAs. In the remaining three Cantonese PCAs, CPP s.d. was accompanied by CPP (n=1) or H1*-H2* s.d. (n=2). CPP patterned less consistently but was most often accompanied by F0 s.d. (Cantonese = 19, English = 14). SHR and SHR s.d. exclusively loaded together for 31 talkers in each language and SHR by itself for a single talker per language. The pair was sometimes accompanied by H1*-H2* (Cantonese = 2, English = 2), H2*-H4* (English = 1), or F0 (English = 3).

While this covers many of the variables that went into the PCAs, F0 is notably sparse in the above paragraphs. While F0 s.d. was fairly consistent in emerging either with CPP (Cantonese = 21, English = 17) or alone (Cantonese = 9, English = 10), the same cannot be said for F0. No particular component structure with F0 occurred more than six times, and across the wide range of configurations, F0 was accompanied by all kinds of variables: F0 s.d., H1*–H2*, H1*–H2* s.d., H2*–H4*, F1 s.d., F4 s.d., CPP, Energy, Energy s.d., and SHR, SHR s.d. The lack of consistency in F0 across talkers is notable for a few reasons. First, F0 plays a major role in prior work on voice production and perception, given its salience as an acoustic dimension (Perrachione et al., 2019). A second reason for it being notable comes from Lee and colleagues' work, where F0 emerged as an important feature of acoustic voice variation structure in English spontaneous speech (Lee and Kreiman, 2019) and Korean sentence reading (Lee and Kreiman, 2020), but not for English sentence reading (Lee et al., 2019).

On the whole, variables emerged on a single component. That is, very few

variables had complex loading structures. Across talkers, only three had complex loading structures for H2*–H4* in each language. F0 and F0 s.d. participated in complex loadings for a single English talker, and twice in the Cantonese PCAs. The remaining variables that participated in complex loading structures only occurred in one or two PCAs across all talkers and languages. This means that for a given PCA, the interpretation of components is reasonably straightforward, even if drawing generalizations over the full group is not.

There were additional components (not reported here) that were shared by less than half of the talkers. A full list of component configurations, along with the number of occurrences and range of variation accounted for is provided in the supplementary materials. Well... it will be! Also, would a table in this section help with interpretation? Or is prose plus supplementary materials good enough?

In summary, this PCA analysis found a greater amount of component structure overlap than was reported in Lee et al. (2019). At the same time, idiosyncratic variation was still readily apparent in the PCAs, both in how variables co-occur, as well as in how much variance is accounted for by the different components. Additionally, it is important to remember that these PCAs represent the lower dimensional structure of the voices they measure. Considering that the total variance unaccounted for by the PCAs ranges from 17.3%–25.1%, this unaccounted for variability may also be idiosyncratic in nature.

3.2.6 Canonical redundancy analysis

Methods

To assess whether variation in a talker's voice is structurally similar across both languages, I compare PCA output from both languages by calculating redundancy indices in a canonical correlation analysis (CCA Stewart and Love, 1968; Jolliffe, 2002). CCA is a statistical method used to explore how groups of variables relate 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. For example, the two redundancy 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). This is particularly relevant in cases where the PCAs comprise different numbers of components, as determined by the stopping rule described above. The PCA with more components will likely account for more of the variation in a PCA with fewer components than the reverse.

Redundancy indices were computed for all pairwise combinations, including cases where similar values were expected (same talker, different language) and cases where dissimilarity was anticipated (different talker and language). Considering that the PCA analyses capture the lower-dimensional structure within each language, these redundancy indices effectively reflect the degree to which the lower-dimensional structure of acoustic voice variability is shared across a talker's two languages.

Results

Redundancy indices for within-talker comparisons ranged from 0.80 to 0.97, (Mdn = 0.92, M = 0.91, SD = 0.04) and are displayed in Figure 3.6, with the two redundancy indices for a given pairwise comparison plotted against one another. Comparisons across talkers within-language ranged from 0.64 to 0.96 (Mdn = 0.83, M = 0.83, SD = 0.5). Comparisons across both talkers and languages ranged from 0.64 to 0.97 (Mdn = 0.83, M = 0.83, SD = 0.5). Within-talker values were confirmed to be higher than across-talker comparisons, per a Welch's t-test ($t(70.93) = -17.35$, $p < 0.001$, d = 1.77). A second Welch's t-test testing the same versus different language for the across talker comparisons did not find a difference between those groups ($t(4485.9) = -1.53$, $p = 0.13$, d = 0.05).

While the across-talker comparisons were generally lower than the within-talker ones, the redundancy indices are overall still relatively high. The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded com-

ponents almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.2.5, there were a substantial number of commonly occurring patterns across talkers and languages. Together, this supports the conceptualization of a voice space comprising a shared structure—as in the case of the prototype account—where voices can only deviate from one another so much.

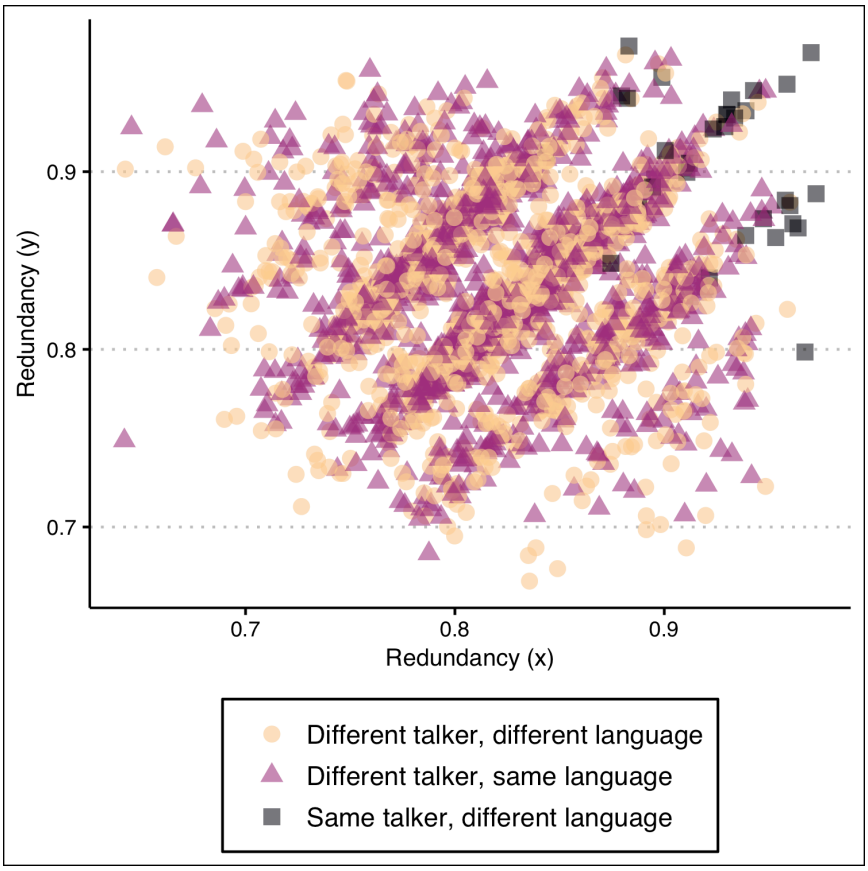

Figure 3.6: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are represented by the black squares and are clearly clustered at the top right.

3.2.7 Passage length analysis

As previewed in the introduction, passage length is an important consideration in the principal components and canonical redundancy analyses. It represents one possible reason why the results presented in this chapter differ from prior work. To examine the role of passage length, I conducted multiple PCAs for each talker and language combination, such that each PCA captured a progressively longer portion of the overall interview, using passage lengths comprising sample sizes of 500, 2000, 4500, 8000, 12500, 18000, 24500, 32000, 40500, 50000, 60500, and 72000. As the total number of samples per interview ranged from 20124 to 74638, there were six to 12 total PCAs per interview, depending on its maximum possible passage length. While the step sizes were somewhat arbitrarily selected, the goal was to give a more granular perspective on the lower end, while still covering the upper tail. Redundancy was expected to level off somewhere in the middle, as talkers should eventually cover their range of variability in a given style.

In these PCAs, the number of components was fixed at 10, the lowest number found in Section 3.2.5. This was done to put the PCAs on more equal footing in the subsequent analysis, given the asymmetries in CCA when different numbers of components were present. For each interview, the canonical redundancy indices were calculated for each talker and language combination, comparing PCAs for each passage length to the PCA for the longest passage length. All of this was done on a within-language and within-talker basis. The final comparison thus has perfect redundancy, as the longest PCA for a given interview is compared to itself.

Figure 3.7 plots polynomial smooths for each interview, with superimposed mean smooths. The x-axis represents the sample size of the shorter passage length in the comparison. The y-axis represents an average of the two redundancy indices. The vertical line at 5000 represents the average sample size from Lee et al. (2019). The vertical line at 20124 represents the sample size used in Sections 3.2.5 and 3.2.5. While there are some gains in sample sizes above the second vertical line, they are comparatively small. It is readily apparent from this plot that the sample size used for PCAs in this chapter was sufficient to capture most of the range of talkers' within-interview variability. As the leveling-off point likely varies across speech styles, it is not immediately apparent whether the sample size in Lee et al.

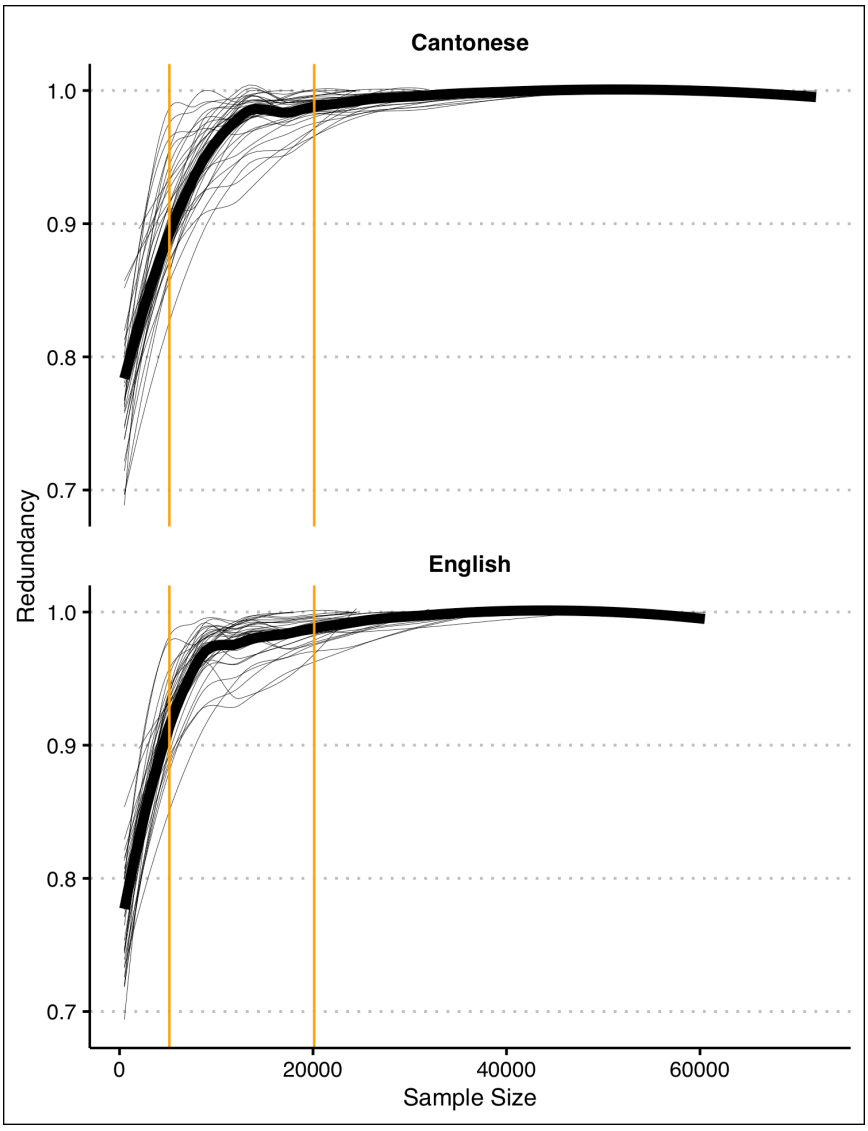

Figure 3.7: Passage length redundancy indices are plotted against the sample size of the smaller PCA. Smoothed curves show a rapid increase in redundancy followed by a levelling off between the vertical orange lines, which represent the sample sizes used in prior work (x = 5000) and the present study (x = 20124).

(2019) sufficiently captured the range of talker variability and thus may not adequately capture the structure of their variability.

3.3 Discussion and conclusion

This chapter examines spectral properties and structural similarities in an individual's voice across two languages. To this end, it uses conversational interviews from the SpiCE corpus of speech in Cantonese and English, described in Chapter 2. The analyses presented in this chapter cover three different exploratory approaches to the question of understanding crosslinguistic (dis)similarity in bilingual voices. Section 3.2.4 takes a coarse perspective, comparing overall distributions using *t*-tests and Cohen's *d* values. This approach follows from a body of literature focused on crosslinguistic comparisons of acoustic measurements—primarily F0—using means, ranges, and standard deviations to describe how voices differ (or not). Section 3.2.5 replicates Lee et al.'s (2019) methods for drilling down into the structure of acoustic voice variation using PCAs and extends it to the case of bilingual speech. Section 3.2.6 builds on the PCAs and introduces canonical redundancy as a metric for objectively assessing crosslinguistic similarity from the output of two PCAs. These methods are then extended in Section 3.2.7 to demonstrate that the analysis used a sufficiently large sample.

A clear result in this chapter is that the bilinguals studied here exhibit similar spectral properties and similar lower-dimensional structure in their acoustic voice variation. This similarity is most apparent on a within-talker basis but still present across talkers and languages, despite substantial segmental and suprasegmental differences across English and Cantonese (Matthews et al., 2013). In this sense, the SpiCE corpus talkers appear to have the same "voice" in each of the two languages. This outcome supports the characterization of voices as auditory faces. The face-voice comparison is especially apt if you take into account findings that talkers' faces vary across languages, as evidenced by work demonstrating that lip movement patterns alone are sufficient for humans and machines to identify and discriminate between spoken languages (Afouras et al., 2020; Soto-Faraco et al., 2007). Voices and faces are highly similar across languages but are not necessarily identical—this leaves room for individuals who are familiar with both the individ-

uals and languages in question to excel at perceptual tasks in both domains.

It is reassuring that the results from the first two approaches used here reflect prior findings. For example, when there was a difference for measures like F0 or H1*–H2*, it tended to mirror expectations from the literature that Cantonese tends to have lower pitch and breathier voice quality than English (Ng et al., 2012, 2010). At the same time, most talkers did not exhibit a meaningful difference, validating prior work that found no differences (Altenberg and Ferrand, 2006). The variability present in this particular sample of 34 talkers highlights the need to treat very small studies with some level of skepticism.

In the PCAs, similarity to prior work emerges in the structure of various components, including the ones that account for the most variability. Lee et al. (2019) report that three of the largest components captured lower-dimensional structure for (i) higher harmonic spectral shape variation, (ii) higher formants, and (iii) a combination of lower spectral shape with the lower formants. While the amount of overall variance accounted for differs here, these component structures also occurred for the SpiCE talkers. Respectively, they are associated with (i) perceived breathiness or brightness, (ii) vocal tract size or speaker identity, and (iii) a combination of phonation type and vocal tract configuration—perhaps reflecting shared linguistic variation. The cross-study overlap in component structure adds credibility to the idea of a prototype model in voice (Lavner et al., 2001; Latinus and Belin, 2011). Much like Lee et al. (2019), the key shared dimensions relate to the timbre, identity, and vocal tract configuration.

This high degree of similarity does not preclude crosslinguistic differences on a within-talker basis but rather suggests that such differences occur on a more global level. This is apparent in Figure 3.8, which depicts the relationship between talkers' average redundancy from Section 3.2.6 and the difference between the mean values for each of the acoustic measurements in Section 3.2.4. If there were clear relationships between large crosslinguistic differences and redundancy, the regression lines should be strongly negative—this does not seem to be the case.

Such high similarity in the PCAs was not entirely expected, given the results of Lee et al. (2019), where a handful of shared components were evident but were complemented by numerous idiosyncratic components. At face value, the results in this chapter suggest that a heterogeneous bilingual population has more across-

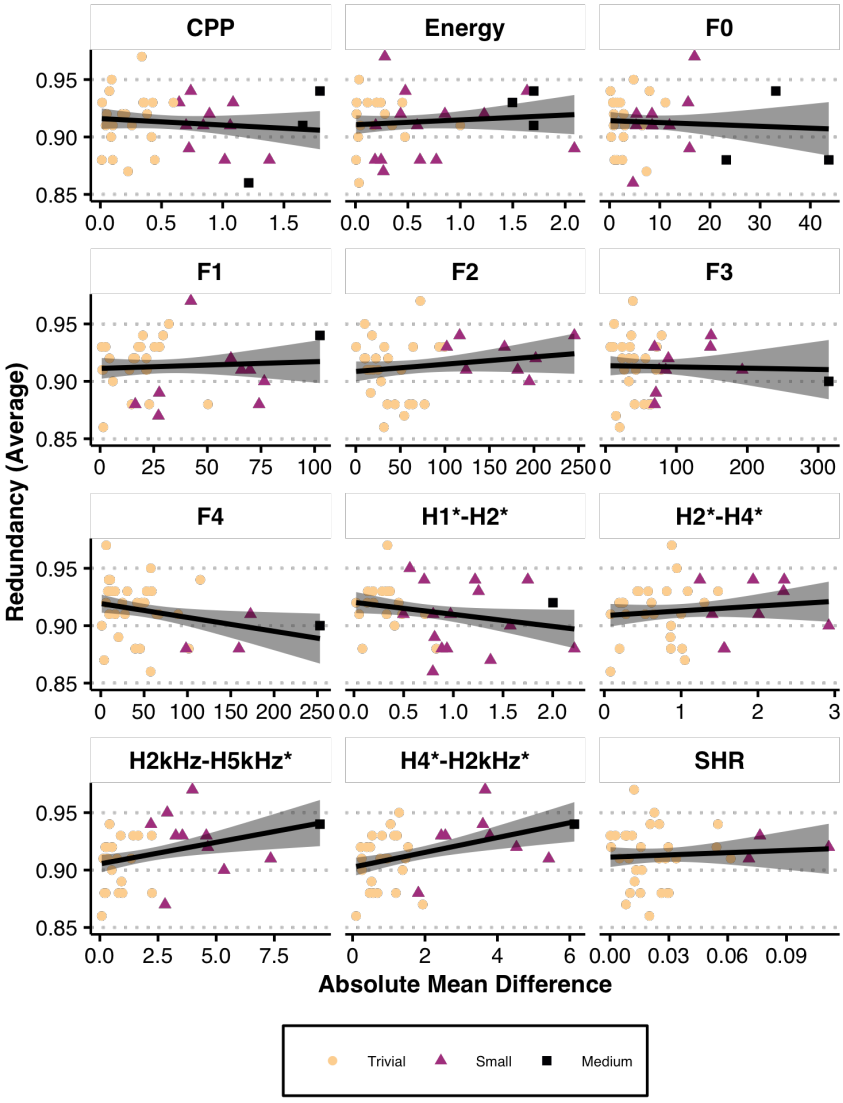

Figure 3.8: This plot depicts the relationship between the absolute value of the difference of means from the t-tests, plotted against the average redundancy value for the talker. Color and shape indicate whether the Cohens' *d* of the t-test was trivial, small, or medium. The superimposed regression line summarizes the relationship between these values.

talker similarity than a tightly controlled group of monolingual English speakers. Several analysis decisions may have contributed to this apparent difference. I compared similar components independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that comparisons are made among like items. Any downside to this methodological decision is mitigated by the fact that most components made relatively small contributions in how much of the overall variance they accounted for (see Table 3.2). As such, I predict that increased across-talker similarity would be found in a reanalysis of the UCLA Speaker Variability Database (Keating et al., 2019) using the adapted methods of this chapter.

While methodological choices may account for some part of these results, the data differences between the current chapter and previous studies are also pertinent. This chapter uses substantially longer passages than the short samples in Lee et al. (2019). Larger speech samples clearly allow for a stable underlying structure to emerge. Smaller samples, conversely, may reflect more ephemeral variation in a talker's voice, and thus not be representative of the talker's full range. The passage length analysis in this chapter shows that the number of samples needed for stabilization is substantially larger than the 5000 samples used in Lee et al. (2019). This does not necessarily discount their work, however, as the current chapter uses spontaneous speech, which is arguably more variable than read speech.[1] It's plausible that an analysis of sentence reading would not need as much data to cover talkers' range of variability in reading aloud. The body of literature in the introduction establishes differences in voice quality across speaking styles (e.g., Lee and Sidtis, 2017). As such, the threshold suggested here may only be appropriate for the speaking style of peer-to-peer conversational interviews. In any case, the methods presented here offer a tool for researchers to use in assessing whether their sample size is representative of a larger whole. Understanding how this interacts with speaking style is left for future directions.

Ultimately, the goal of this line of research is to understand how the acoustic

[1] While it is true that examined spontaneous speech, the poster only states that two minutes of speech were used for each participant. By this estimation, the sample size was likely on the lower side, compared to the 20-25 minute interviews in the SpiCE corpus. However, it is not possible to make a direct comparison without knowing the number of samples.

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 behavioral 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. The results of this chapter set up predictions for that work. These predictions will be revisited in general discussion.

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 ($n = 2295$).

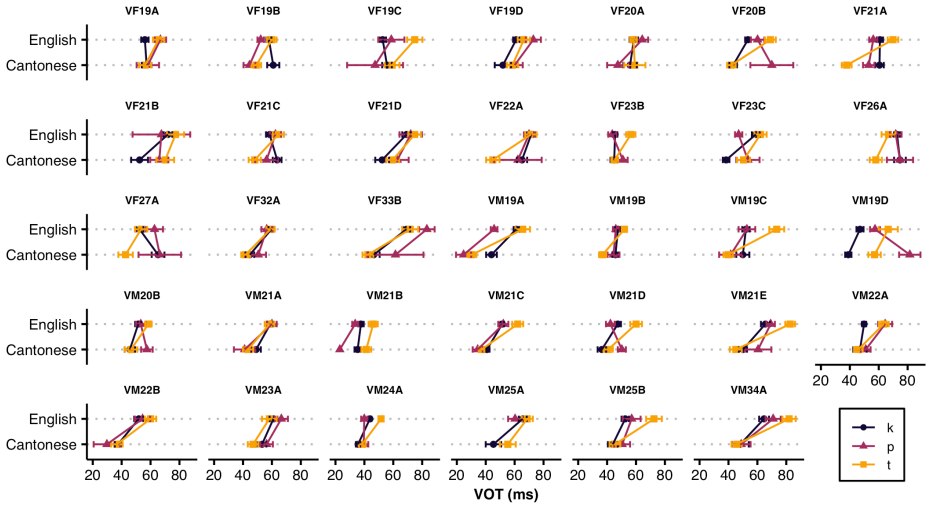

Figure 4.1: Mean and SE for VOT across place of articulation, language, and individuals in the SpiCE corpus.

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 (centered) was predicted by Average Phone Duration (standardized), Preceding Pause

(False$= -0.32$, True$= 1$), Language (Cantonese$= -1.75$, English$= 1$), Place of Articulation (Place T: /p/$= -1.91$, /t/$= 1$, /k/$= 0$; Place K: /p/$= -3.38$, /t/$= 10$, /k/$= 1$), and the Language $\times$ Place interaction. As likely apparent from the parenthetical values, all categorical fixed effects were weighted effect coded (following Chodroff and Wilson, 2017). Random intercepts for Talker and Word were 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

...

Bibliography

Afouras, T., Chung, J. S., and Zisserman, A. (2020). Now you' re speaking my language: Visual language identification. In *Proceedings of Interspeech 2020*, pages 2402–2406. ISCA. → page 63

Alderete, J., Chan, Q., and Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, 191:103952. → page 6

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 36, 37, 39, 46, 49, 64

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 8

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 69

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 68, 69

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association. → page 6

Audacity Team (2018). Audacity (R): Free audio editor and recorder. → page 11

Barlow, J. A. (2014). Age of acquisition and allophony in Spanish-English bilinguals. *Frontiers in Psychology*, 5. → page 69

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 76

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 30

Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Version 6.1.38. → page 41

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., and Mok, K. (2011). Language- and talker-dependent variation in global features of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 356–359, Hong Kong. → page 3

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. → pages 39, 75

Bullock, B. E. and Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, L., Winford, D., and deBot, K., editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam. → pages 35, 40, 69

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. → pages 32, 33, 53

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 70

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. → pages 38, 39

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 71, 72, 74, 75, 78

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 71, 72, 73, 74, 75, 77, 78

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 72

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 72, 78

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 4, 27

Ethnologue (2021). Chinese, Yue. In Eberhard, D. M., Simons, G. F., and Fennig, C. D., editors, *Ethnologue: Languages of the world*. SIL International, Dallas, TX, 24 edition. Online version. → page 6

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. → pages 71, 72

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 70

Flege, J. E. and Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In Wayland, R., editor, *Second Language Speech Learning: Theoretical and Empirical Progress*, pages 3–83. Cambridge University Press. → pages 68, 69, 70, 72, 78

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 41, 68, 69

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 3, 78

Garellek, M. (2019). The phonetics of voice. In Katz, W. F. and Assmann, P. F., editors, *The Routledge Handbook of Phonetics*. Routledge. → page 43

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE. → page 6

Goldrick, M., Runnqvist, E., and Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, 25(4):1031–1036. → page 68

Google (2019). Cloud speech-to-text. → pages 6, 19

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, 60(2):98–128. → pages 69, 70

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of speech and hearing research*, 37(4):769–778. → page 43

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246. → page 15

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 43

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 41

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 78

Johnson, K. A. (2021). SpiCE: Speech in Cantonese and English. V1. → page 3

Johnson, K. A. and Babel, M. (2021). Language contact within the speaker: Phonetic variation and crosslinguistic influence. Technical report, OSF Preprints. → page 3

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 72, 78

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → pages 53, 58

Järvinen, K., Laukkanen, A.-M., and Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, 38(2):47–51. → pages 38, 39, 41, 50

Kawahara, H., Agiomyrgiannakis, Y., and Zen, H. (2016). Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pages 221–228. → page 42

Keating, P., Kreiman, J., and Alwan, A. (2019). A new speech database for within- and between-speaker variability. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, pages 736–739, Melbourne, Australia. → pages 32, 66

Keating, P. and Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, 132(2):1050–1060. → pages 34, 36, 37

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 73

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 31, 32, 40, 42, 43, 53

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, 149(1):457–465. → pages 31, 53

Latinus, M. and Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2:175. → pages 33, 64

Lavner, Y., Rosenhouse, J., and Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, 4(1):63–74. → pages 33, 64

Lee, B. and Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, 20(3):174–185. → pages 37, 38, 39, 41, 66

Lee, J. L. (2018). PyCantonese [Version 2.2.0]. → pages 6, 22

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 30, 32, 33, 34, 40, 42, 45, 46, 53, 57, 58, 61, 63, 64, 66

Lee, Y. and Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. → pages 32, 57

Lee, Y. and Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. 10.1121/1.5146847. → pages 32, 35, 53, 57

Leung, M.-T. and Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6(2):305–325. → page 6

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science*, 10(2):e1483. → page 33

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. → page 40

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461. → page 31

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 70

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 72, 78

Littell, P. (2010). Thank-you notes [Version 1.0: Agent focus]. → page 15

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, 24(1):71–89. → page 38

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. → page 6

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 53

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → pages 15, 63

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner [Version 1.0.1]. → page 21

Munson, B. and Babel, M. (2019). The phonetics of sex and gender. In Katz, W. F. and Assmann, P. F., editors, *The Routledge Handbook of Phonetics*. Routledge. → page 43

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 35

Myers-Scotton, C. (2011). The matrix language frame model: Developments and responses. In *Codeswitching Worldwide*, volume 126 of *Trends in Linguistics. Studies and Monographs*. De Gruyter Mouton. → page 41

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 71

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, 43(1-2):65–84. → pages 15, 17, 20

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. University of Adelaide, Adelaide, Australia. R package version 0.5. → page 46

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 36, 37, 38, 40, 46, 49, 64

Ng, M. L., Hsueh, G., and Sam Leung, C.-S. (2010). Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, 12(3):230–236. → pages 37, 46, 64

Ng, R. W. M., Kwan, A. C., Lee, T., and Hain, T. (2017). ShefCE: A Cantonese-English bilingual speech corpus for pronunciation assessment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5825–5829. → page 5

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376. → page 33

Ordin, M. and Mennen, I. (2017). Cross-linguistic differences in bilinguals' fundamental frequency ranges. *Journal of Speech, Language, and Hearing Research*, 60(6):1493–1506. → page 39

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 34, 35, 79

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. → page 5

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 69

Perrachione, T. K. (2018). Recognizing speakers across languages. In Frühholz, S. and Belin, P., editors, *The Oxford Handbook of Voice Perception*, pages 514–538. Oxford University Press. → pages 33, 34

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, 146(5):3384–3399. → pages 33, 34, 35, 40, 57

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 4, 5, 17

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → pages 46, 53

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 79

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 38

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070. → page 71

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 42

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. → page 69

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 17

Sjölander, K. (2004). The Snack Sound Toolkit. → page 42

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 19

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., and Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, 69(2):218–231. → page 63

Statistics Canada (2017). Proportion of mother tongue responses for various regions in Canada, 2016 Census. → page 8

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163. → pages 58, 59

Sun, J. (2020). jieba [Version 0.42.1]. → page 22

Sun, X. (2002). Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–333–I–336. → page 44

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 68

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 53

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 15, 22

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 69, 70

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 53

Voigt, R., Jurafsky, D., and Sumner, M. (2016). Between- and within-speaker effects of bilingualism on f0 variation. In *Proceedings of Interspeech 2016*, pages 1122–1126, San Francisco, CA. → page 38

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, 39(1):9–30. → page 35

Winterstein, G., Tang, C., and Lai, R. (2020). CantoMap: A Hong Kong Cantonese MapTask corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 2906–2913, Marseille, France. European Language Resources Association. → page 6

Wong, W. Y. P. (2006). *Syllable fusion in Hong Kong Cantonese connected speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. → pages 20, 22

Xue, S. A., Hagstrom, F., and Hao, J. (2002). Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: a

functional system approach. *Asia Pacific Journal of Speech, Language and Hearing*, 7(1):55–62. → page 37

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820. → pages 69, 70, 72

Yang, Y., Chen, S., and Chen, X. (2020). F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Proceedings of Interspeech 2020*, pages 4163–4167. ISCA. → page 38

Yau, M. (2019). PyJyutping. → page 6

Yovel, G. and Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, 17(6):263–271. → page 33

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 8

Yuan, J., Ryant, N., and Liberman, M. (2014). Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2539–2543. → page 22

Ćavar, M., Ćavar, D., and Cruz, H. (2016). Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia. European Language Resources Association (ELRA). → page 22