Chapter 1

Introduction

...

Chapter 2

The SpiCE Corpus

2.1 Introduction

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

Conversational speech allows for richer and more accurate empirical description of spoken language, as it represents more realistic and natural productions than scripted laboratory speech, whether compared to isolated word production or scripted connected speech. It enables the study of speech style, style 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).

The discrepancies between results for conversational and lab speech have been found for monolingual (English) speech, but are likely be found with bilingual speech as well. Resources to query bilingual conversational speech are limited, however, as the necessary resources permitting this type of inquiry are relatively rare. Table 2.1 provides a sample of prominent bilingual speech corpora, summarzing 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 represents the focus 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). As will become apparent later in this chapter, the SpiCE corpus focuses on early bilingualism. In light of this, Table 2.1 only includes speech corpora that involve similar populations (as opposed to late bilinguals and language learners).

Corpus	**Language balance**	**Size**	**Style**	**Within-talker**	**Phonetic analysis**
Bangor Miami (Deuchar et al., 2014)	63% English 34% Spanish	CHECK	Conversational, code-switching	Yes, most talkers	Limited
Bangor Patagonia (Deuchar et al., 2014)	78% Welsh 17% Spanish <0.5% English	CHECK	Conversational, code-switching	CHECK	Limited
Bangor Siarad (Deuchar et al., 2014)	84% Welsh 4% English	CHECK	Conversational, code-switching	CHECK	Limited

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]

The corpus design is based on key aspects of widely used existing corpora, such as the Buckeye corpus of conversational speech (Pitt et al., 2005). 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 with respect to interview style and recording quality.

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—limiting 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 balanced for the languages involved. 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, 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, with samples in both languages from the same set of individuals. Again, there are similarities in what SpiCE aims to accomplish, but it nonetheless occupies a different niche in the speech sciences.

The primary motivation for collecting this corpus was to have comparable high-quality recordings of conversational speech from early bilinguals in two languages, which in turn enables large scale phonetic analysis on a within-speaker basis. 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,

being slightly smaller in size than the Buckeye corpus (Pitt et al., 2005). Both of these are purpose-built corpora that are recorded in person. Truly large corpora tend to be collected from existing recordings (radio, YouTube, audiobooks, etc.; e.g., Librispeech: Panayotov et al., 2015), crowdsourced online (e.g. Mozilla Common Voice: Ardila et al., 2020), via phone (e.g., SWITCHBOARD: 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 our knowledge, this type of resource does not yet exist for any pair of languages, much less for a typologically distinct pair like Cantonese (Sino-Tibetan) and English (Indo-European). Furthermore, Cantonese is a relatively understudied language, despite there being approximately 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 design and collection procedures, a description of the speakers, and the transcription and annotation pipeline. It concludes with descriptive statistics.

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.

2.2.1 Recruitment

Participants were recruited for the SpiCE corpus through a variety of methods at the University of British Columbia. This inlucded word of mouth, the Linguistics Human Subject Pool, the Psychology Paid Studies list, advertisements in depart-

ment email lists, advertisements in linguistics courses, printed flyers, and posts on various club forums.

The recuitment 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 langauges 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.

Prior to 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 cancelled and they were contacted with an expalantion.

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, 17 of which are female, and 17 of which are male. 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 prior to the recording session, as well as a copy of the survey itself. Basic summary information is included in Table 2.2, and in visualizations throughout this chapter.

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 the purposes of this corpus, an early bilingual is someone who began learning both Cantonese and English

No.	ID	Order	Age	Gender	Age Began Learning English	Age Began Learning 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, including age, gender, age of acquisition (AoA), and the order the interviews occurred.

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 (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.

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 born and raised. Figure 2.2 does exactly this. 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 (as in Figure 2.2), the numbers are 27 and 18, respectively.

Additionally, calling an individual a bilingual does not preclude knowledge of additional languages. In fact, all but on 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 language 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.

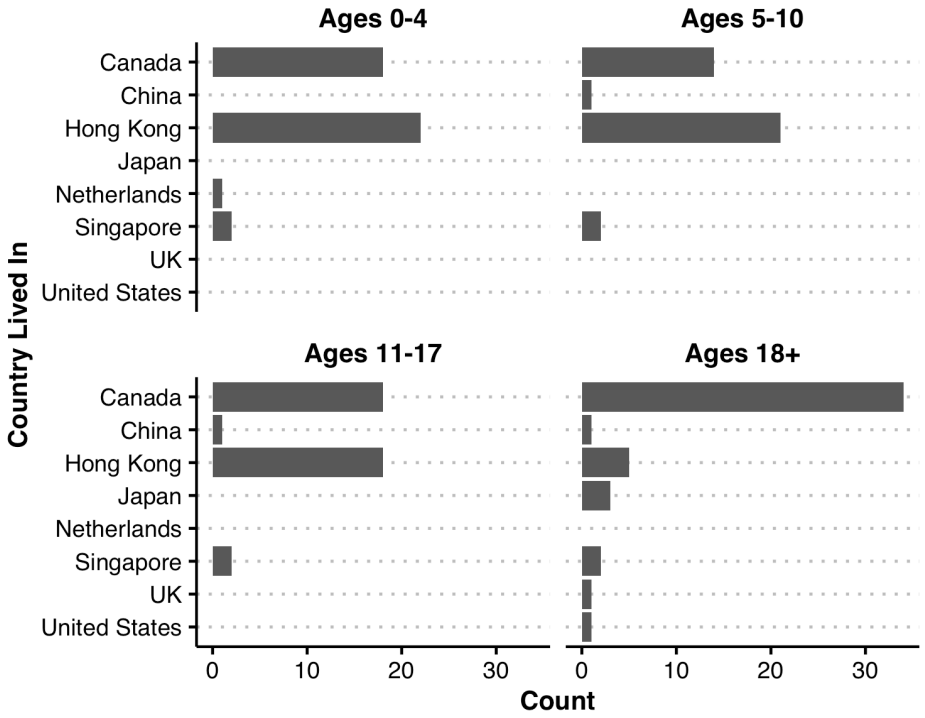

Figure 2.1: This four panel bar chart summarizes where the SpiCE participants lived during different portions of their lives.

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)

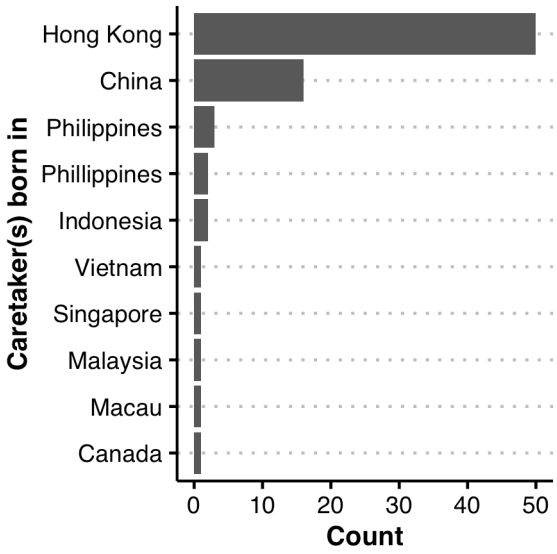

Figure 2.2: This bar chart summarizes the number of caretakers who were born in various locations. Note that the number of caretakers reported by individual participants varies.

on a PC laptop, and saved with a 44.1 kHz sampling rate, and 16-bit resolution.[1]

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 of the window of time in which they would be able to withdraw their data. Subsequently, participants were asked to provide written consent. Upon consent, participants completed a session in English, and a session in Cantonese. The order of languages was counterbalanced across participants (see Table 2.2). Each session consisted of three tasks—sentence reading, storyboard narration, and a conversational interview—described in the following sections. Each of these three tasks were recorded in the same audio file, though there are separate recordings

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

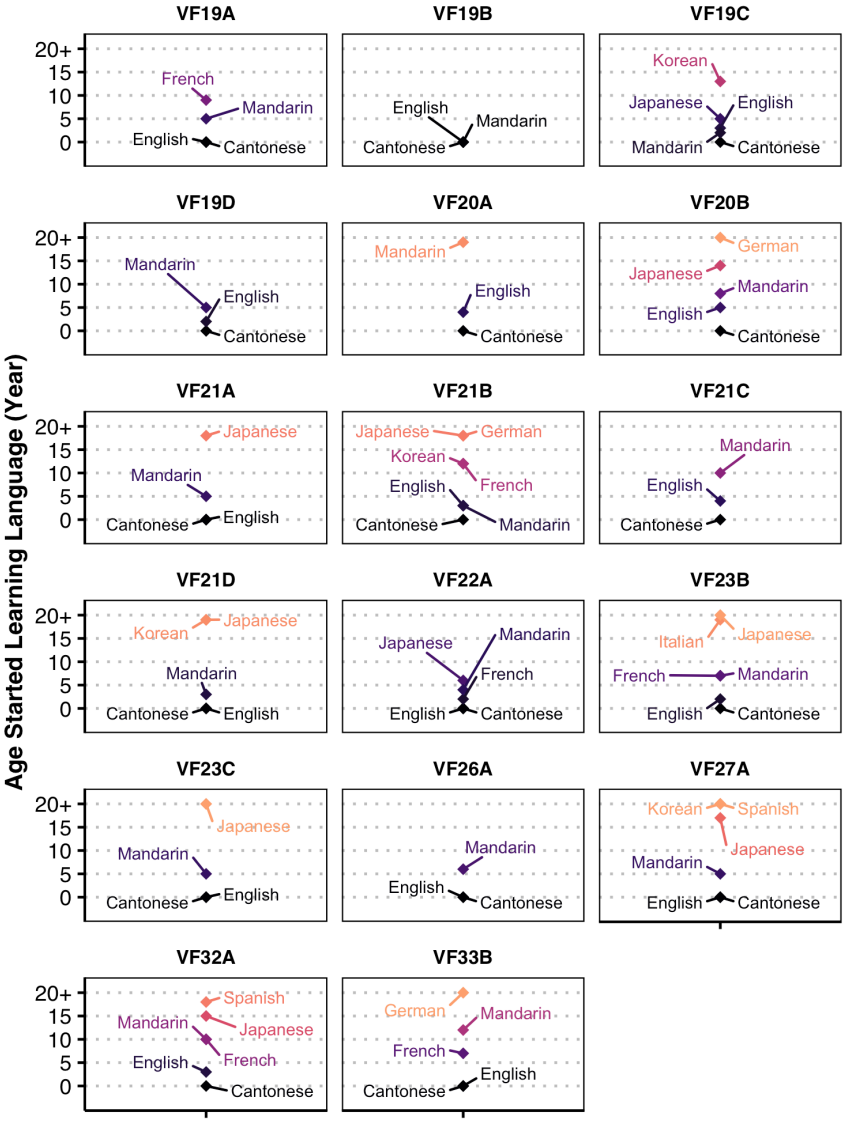

Figure 2.3: Multilingualism for the female participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

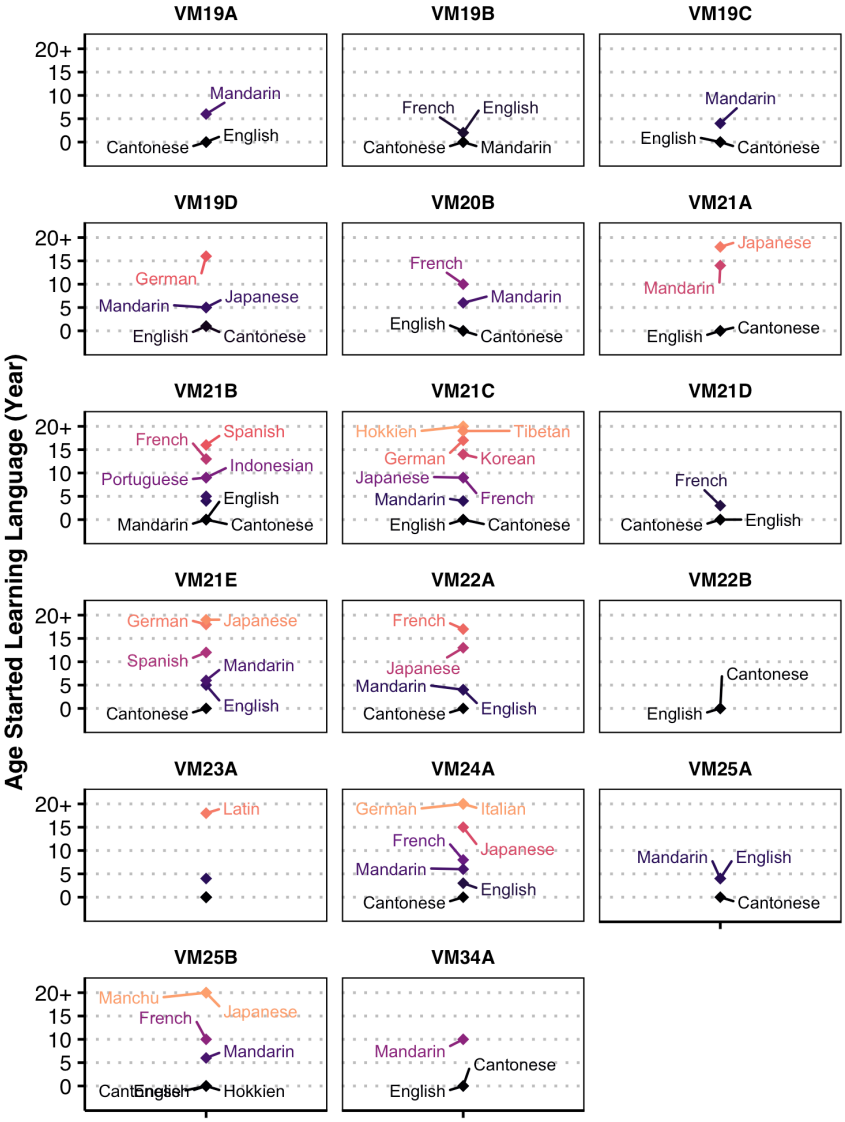

Figure 2.4: Multilingualism for the male participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

for each of the sessions. That is, each participant has a Cantonese session and an English session. Together, these three tasks took 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

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 both Cantonese characters and the Jyutping romanization.[2] If necessary, participants could make use of the phrase's English translation. The Cantonese sentences are well-known declarative phrases, typically associated with Chinese New Year. While a more explicitly balanced set of sentences could have been used, participants' familiarity was deemed more important, as many Cantonese-English bilinguals in Canada are not literate in Cantonese. The English sentences included the Harvard Sentences list number 60 (IEEE, 1969), as well as series of holiday-themed declarative sentences to better match the content of the Cantonese sentences. This task was relatively formal, and typically lasted less than one minute.

Sentence reading was included in the session to insure that different participants produced a set of identical items, considering the core of the session was unscripted conversational interview (described in Section 2.2.4). 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 for future analyses or projects that require matched utterances.

In practice, the utility of these sentences may be somewhat limited, as sentences with speech errors were not neccessarily 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 language mode prior to the upcoming interview. As such, they can be considered a warmup task.

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

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 original imperatives, designed to thematically match the Cantonese sentences.

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.

Storyboard Narration

For the second task, participants narrated a short story from a cartoon storyboard originally developed for linguistic field work (Littell, 2010). The storyboard followed a simple plot about receiving gifts and writing thank you notes to family members and friends—a topic that Cantonese-English bilinguals in the corpus were expected to be familiar with in both languages. This task was less formal than the sentence reading task, and ensured that different participants produced some of the same words in a more spontaneous context. Participants varied in how they approached this task, with some treating it like a serious 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 matched utterances, 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 facilited participants getting into the language mode of the session prior the 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. Furthermore, each interview took its own shape, and was guided by what the participant wanted to talk about, anywhere from three to six topic areas covered—the planned sequence of questions is included in the Appendix. As a result, the speech samples from each language are comparable, but the specific questions differ between interviews and across participants.

Participants were encouraged to code-switch between languages by the interviewer, who included code-switches in some of her questions, and asked about topics that encouraged switches (e.g., Chinese foods in English; university course work in Cantonese). While code-switching was encouraged, it was not a primary focus for the session. As will become apparent later in this chapter, there was substantially more code-switching in the Cantonese part of the session.

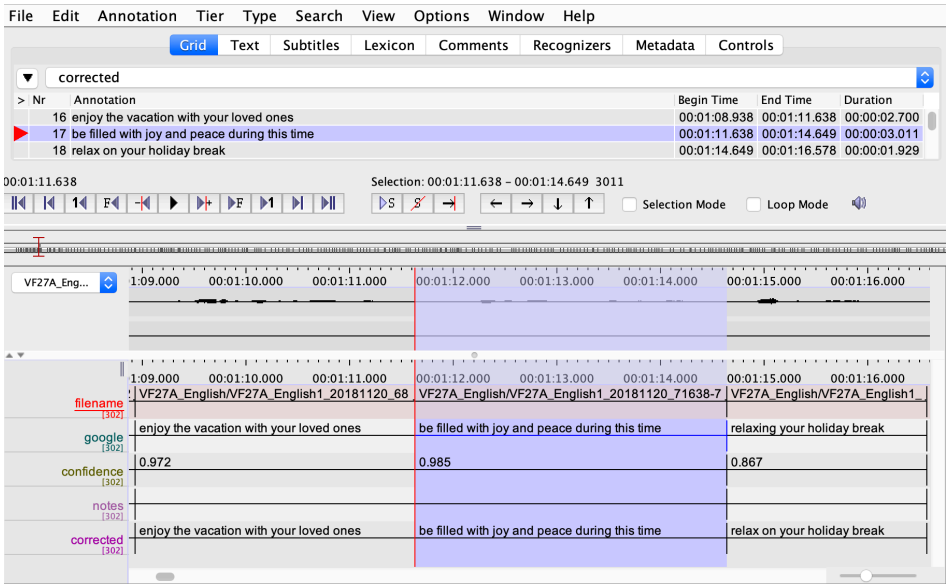

Figure 2.5: This screenshot from ELAN shows a sample of hand-corrected English from the sentence reading task for participant VF27A. The audio waveform is displayed in two channels, with one for the participant (top) and the other for the interviewer (bottom). The annotation tiers include (1) the short 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 hand correction.

2.3.1 Cloud Speech-to-Text

Google Cloud Speech-to-Text was used to produce an initial transcript of the interviews (Google, 2019). This was done using the Short Audio option, with the language variety set to Canadian English (en-CA) or Hong Kong Cantonese (yue-Hant-HK). In order to use this speech recognition product, the participant's speech was extracted from the participant's channel and segmented into short chunks, typically under 15 seconds in duration.[3] No attention was paid to constituents at this point; rather, breaks were placed at breaths and other pauses. Short chunks were necessary in order 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

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

were adapted from the multilingual Heritage Language Variation and Change corpus, 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, and similar.

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 is transcribed with the typical smaller number of characters (e.g., 咩 *me1* is a fully fused version of 乜嘢 *mat1 ye5*, and intermediate version 咩嘢 *me1 ye5*—all translate to "what").[4]
- Non-lexicalized (or ambiguous) cases of syllable fusion are transcribed with the full number of characters fused (e.g., 朝頭早, "morning," is fully pronounced as *ziu1 tau4 zou2*, but can be fused to 【朝頭】早 *ziau14 zou2*. Brackets indicate the fused syllables).
- Filled pauses are transcribed with the character 啲 (*e6*), or using Jyutping if different (e.g., *m6*).

[4]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.

- Transcribers followed a shared set of guidelines for transcribing sentence final particles.

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.

In Cantonese, forced alignment was completed with the Train-and-Align option, as there was no pretrained model available for Cantonese. As Cantonese orthpgraphy does not separate words with spaces, words segmentation was done using the *jieba* Python library (Sun, 2020), along with a Cantonese dictionary.[5] 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 using the Jyutping transcribed), separating it into segments, and appending the tone number 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).

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

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

The word and phone output of the forced alignment process was included in a TextGrid for each audio recording, along with annotation tiers for task (sentences, storyboard, or interview), 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>" for unknown in the word tier and "spn" in the phone tier. 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. As a result, it is advisable to use stringent exclusionary criteria or perform checks prior to analyzing data from the corpus.

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 stero 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. 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 by participant, with a mean of 2,662 word tokens per interview (SD=637, minimum=1,654, maximum=4,212), and 749 word types (SD=157, minimum=483, maximum=1081).

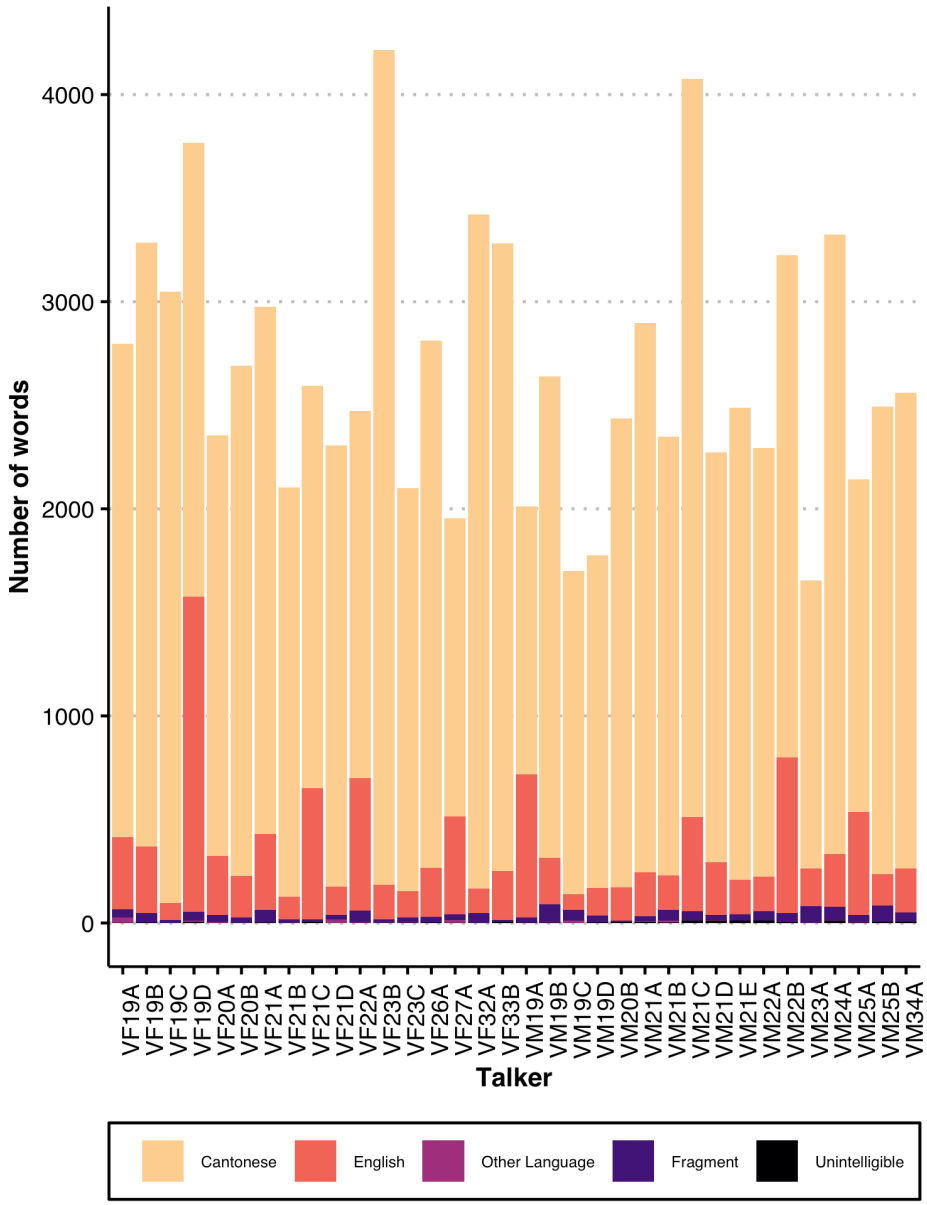

Figure 2.6: The total word count for each participant's Cantonese interview task is represented by bar height. Color indicates the kind of item counted.

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 2.6 shows split of these categories on a by-participant basis within the Cantonese interview sessions. Figure 2.6 indicates that all participants switch into 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 2.7. 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.

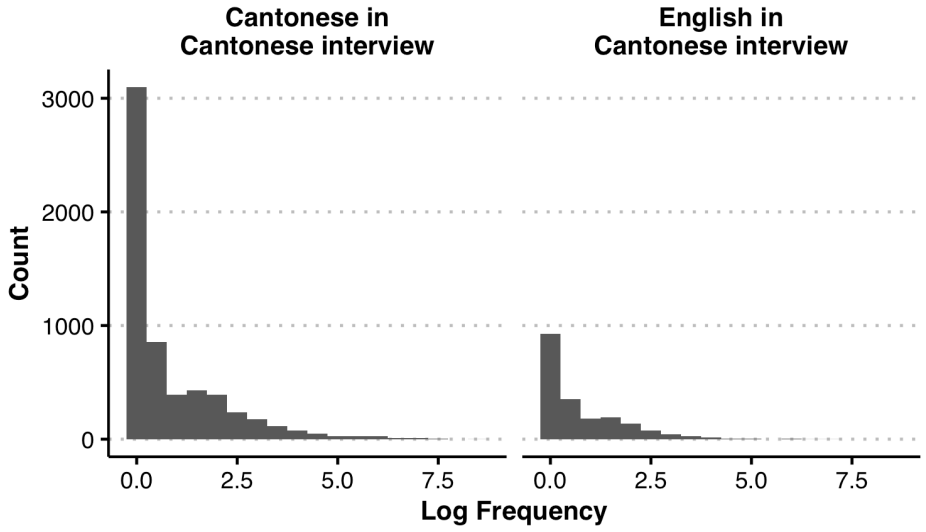

Figure 2.7: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

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 token count of 3,077 (SD=701, minimum=1,907, maximum=4,240), and word type count of 609 (SD=119, minimum=434, maximum=904). Figure 2.9 shows split of these categories on a by-participant basis within the English interview sessions. Unlike the Cantonese interviews, there were relatively few switches into 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 2.8. Word frequency follows a similar pattern to Cantonese word frequency, with most words occurring infrequently, and a smaller proportion occurring very frequently.

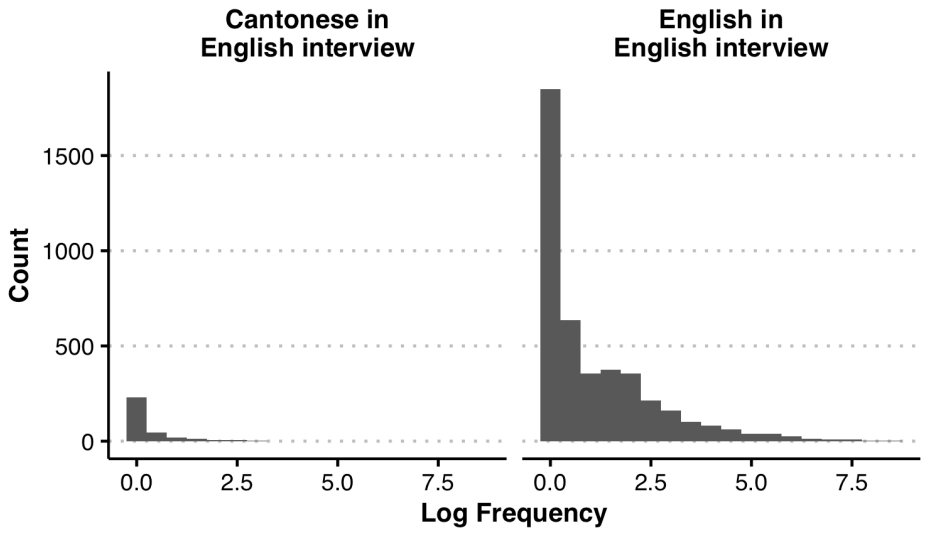

Figure 2.8: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

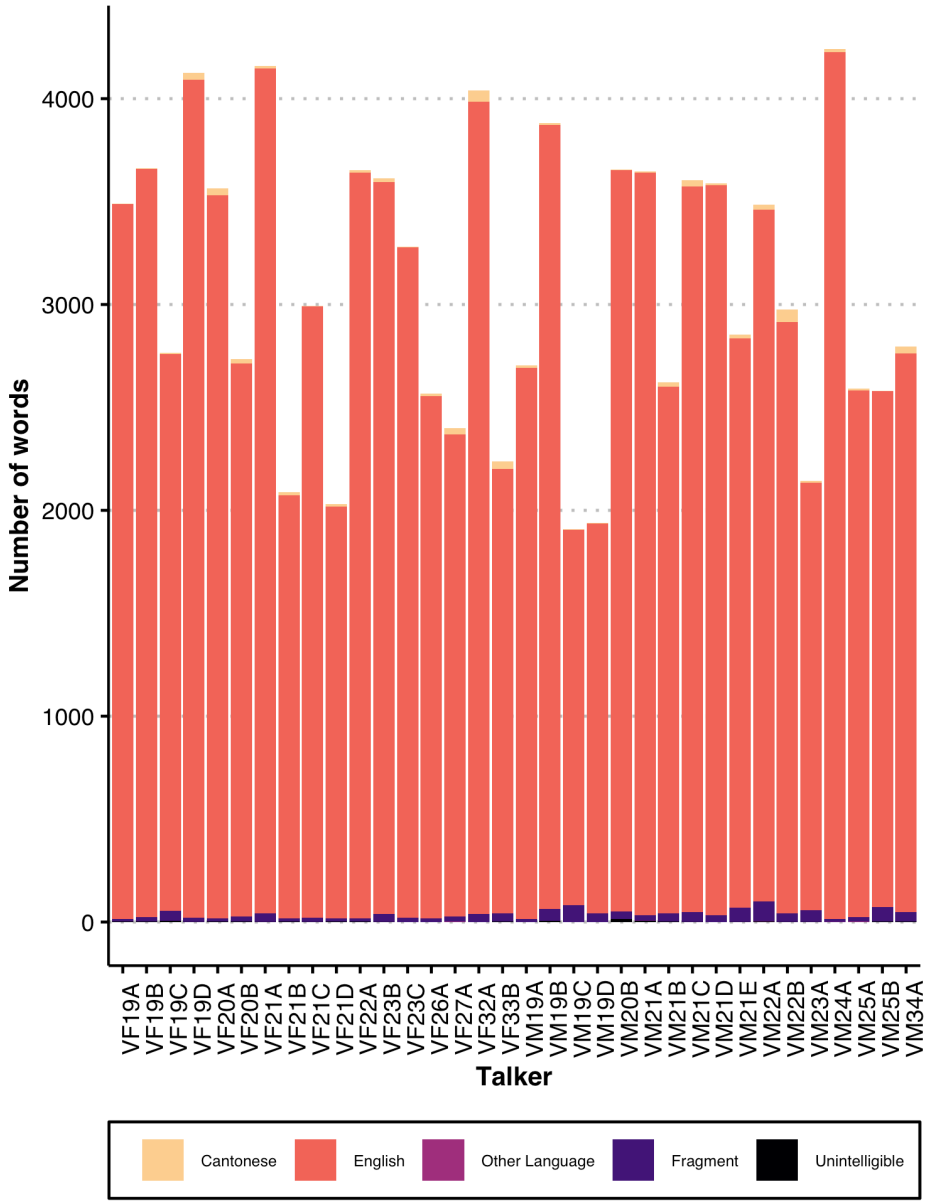

Figure 2.9: The total word count for each participant's English interview task is represented by bar height. Color indicates the kind of item counted.

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.[6] In addition to the corpus itself, documentation is available online.[7]

2.6 Discussion & Conclusion

While various bilingual corpora exist, they lack in different ways. The SpiCE corpus described here enables within-speaker phonetic comparisons across languages. While this would be possible with some of the bilingual speakers in resources like the Bangor corpora (Deuchar et al., 2014), the recording quality in such resources limits the scope of phonetic research. With the release of SpiCE and its high-quality recordings, scholars have the ability to 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 offers substantial potential for increasing our understanding of bilingual spoken language from both phonetic and psycholinguistic perspectives. While the recording quality of this corpus offers these particular advantages, SpiCE is also suitable for any other standard corpus-based inquiry with conversational speech, whether linguistic or paralinguistic in nature. The opportunities made available with SpiCE are especially important given the typological difference between the languages under consideration, and the fact that Cantonese is an understudied language.

[6] https://creativecommons.org/licenses/by/4.0/
[7] https://spice-corpus.readthedocs.io/

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, in that 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 is understanding how listeners use different dimensions within the voice in processing talker indexical, affective, social, and linguistic information. The difficulty here arises from the sheer variability across voices. While voices share attributes and relevant acoustic dimensions, much of the variation across voices appears idiosyncratic (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 perceive, classify, and understand new voices, while the idiosyncrasies enable identification and discrimination between 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. While the focus of this chapter is acoustic

voice variability, 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 signal itself, asking what is available in the signal 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 that variation in speech production is highly structured. This chapter looks at the structure of voices, and the following chapter examines structure for sound categories—both attempt to elucidate what exists in the signal for listeners to use. In the domain of voice quality, Jody 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 dimension (Kreiman et al., 2021, and references therein). As a result, Kreiman and colleagues argue that this set is both sufficient and necessary to capture a wide range of normal and disordered voices. This model includes acoustic dimensions that capture harmonic and inharmonic voice source, pitch, loudness, and vocal tract characteristics. While each dimension in the model could be considered independently by researchers, Kreiman and colleagues argue that these dimensions are more than the sum of their parts. The measures covary and conspire together to form a percept. While this model establishes a set of acoustic dimensions, it does not arbitrate between them in a way that establishes what matters for a given voice in a given 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 will be 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, the authors 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). The driving question for Lee and colleagues is one of understanding what information exists in the signal and how it's structured. In many ways, this is the first step towards understanding which aspects of voice are available to listeners and thus useable in perceptual processes.

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. Typically shared dimensions were spectral shape and noise parameters in the higher frequencies, the fourth formant, and formant dispersion. The spectral measures are associated with vocal breathiness or brightness, and the formant-based measures with speaker identity and vocal tract size. Lee and Kreiman (2019) replicates this work with short samples of spontaneous speech from the same database, with the exception that F0 emerges as a shared relevant dimension. This arguably reflects the difference between read and spontaneous speech in English, with reading tending to be more monotone and spontaneous speech more affective. 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 reading.

Conceptualizing what these dimensions mean and how to think about acous-

tic voice variability in this way is somewhat of a challenge, given the abstractness of these measurements. 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 angle (i.e., looking up, down, or to the side) as well as lighting. Idiosyncratic variation in structure arose from things like facial hairstyle, makeup, and expressions. 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), in which novel individual voices are perceived in relation to a speech community average.

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. 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 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 must be tempered. However, there are exceptions to this. For example, Perrachione et al. (2019) collected perceptual voice (dis)similarity ratings for Mandarin and En-

glish 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. F0 was the most prominent measure, which is unsurprising given 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 the 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 affect voice quality. This chapter builds on this body of work, by extending the methods introduced by Lee and colleagues to the case of bilingual spontaneous speech.

Describing and analyzing acoustic voice variation in bilingual speech has motivation 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. In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic information in the voice. Understanding the structure of that variability brings us one step closer to understanding what listeners are using from the signal to process speech. On the production side of things, bilingual speech presents an ideal test case for the designation of voices as 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, understanding 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, a cross-study comparison of different populations. Across each of their 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. It is important to note that this dichotomy is somewhat misleading. While there ultimately are biological constraints on a voice (e.g., vocal tract length, pathologies, etc.), 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 both within languages (), as well as across languages in the case of bilinguals (Bullock and Toribio, 2009). 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. 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 across languages. Languages can differ in terms of whether a suprasegmental dimension is exploited 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 bilingual speech is considered in 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 (which is typologically similar to Can-

tonese), comparisons between English and Mandarin are also summarized. While the most relevant comparisons for the present work are those made on 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 focused on 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) are primarily concerned with whether F0 shifts when an individual switches languages and with 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 theme—bilinguals can differ in F0 across languages.

In a 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. Based on Long-Term Average Spectral measures, females 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) examine F0 in a 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 back-

grounds of the bilinguals studied. While the picture regarding voice quality measures appears clearer and more consistent, the conclusions arise from a single study.

The authors of these studies speculate that Cantonese's status as a tone language may account for some of these differences compared to English. In this light, it is also relevant to consider the larger body of research comparing voice quality for Mandarin and English. Additional language pairs also offer insight into voice comparisons for typologically distinct languages.

Lee and Sidtis (2017) compare F0, speech rate, and intensity in a small group of late 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. 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 seems to reflect more balanced and proficient bilinguals. Xue et al. (2002) found that Mandarin-English bilinguals aged 22-35 years produced higher F0 in English. Lee and Sidtis (2017) argue that the difference in results can be attributed to the language backgrounds of the respective groups studied. Xue et al. (2002) looked at non-native English speakers, who arguably produce higher pitch speech for reasons related to 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 answer. Alternatively —or perhaps, concurrently—talkers may be expressing different cultural identities in each of their languages (see Loveday, 1981). 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 research to say either way. 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 minimum, maximum, range, and mean. The authors also examined a Mandarin-dominant group and report 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 in comparable ways.

While the studies reviewed thus far provide a mixed picture of voice differences across different language pairs, there is a strong focus on F0. Both the F0-centricity and variable outcomes are apparent in work on other language pairs. For example, Cheng (2020) finds that Korean has consistently higher F0 than English, regardless of whether they were early sequential or simultaneous bilingual, 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.

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 has a higher and wider F0 range. This result did not hold for males from the same population, who were more variable. The authors argue that the cross-linguistic 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 means that the result cannot

be due to anatomical or purely linguistic reasons.

Considering these studies together, a few key observations are especially relevant to the study described in this 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 stratification, rather than linguistic.

This body of work mostly focuses on linguistic and social differences, and while some of it dives into individual differences, individual differences 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. 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.

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 is to hold 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 voice varies across languages. Using the corpus described in 2, I describe 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 come from the conversational interviews in the SpiCE corpus described in the previous chapter. Both Cantonese and English interviews are considered. As noted before, the 34 talkers studied here are all early Cantonese-English bilinguals from a heterogeneous population (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 () uses relatively short chunks of speech, the present analysis is focused on longer stretches of spontaneous speech. While it would certainly have been possible to include the sentence reading and storyboard task recordings from each participant, there are practical reasons for excluding them in this 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 is a wide range of familiarity with the materials represented. Some talkers knew all of the sentences, and others struggled. This renders them less comparable in relation 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 main interview language from the final sample using the time-aligned transcripts, 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 in and pronounced with the phonology of the main language (i.e., 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 set to 100–500 Hz for female talkers, and to 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 largely ignored by the Praat algorithms, and likely exterted 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 differs slightly from the methods described in Lee et al. (2019), the paper on which the methods of this chapter were modeled.

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) comes from the psychoacoustic voice quality model described in the introduction to this chapter (Kreiman et al., 2014). Measurements were made every 5 ms during voiced segments, as in Lee et al. (2019), using 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 that the alternative choices in VoiceSauce. It is one of the more widely studied variables on this list, as evidenced by the literature cited in the introduction (e.g., Cheng, 2020; Ng et al., 2012).

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

F4 The fourth formant frequency is not typically discussed in linguistic contexts, and is instead associated with talker characteristics. In this light, it is not particularly surpriseing that it was highly correlated with formant dispersion. Both measures reflect talker characteristics such as vocal tract length. F4 is 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 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 tilt 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. The amplitude difference is measured in dB. Note that this measure—along with the following three spectral tilt 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 tilt 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 tilt 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 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 would have occurred with any of the other spectral shape parameters (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 impact the principal components based analysis in Sections 3.2.5 and 3.2.7. 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 were considered for each interview. This value was selected as it represents the interview with the smallest observation count across all talkers and languages.

Passage length was expected to have an impact, given the variability in affect and style-shifting 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 3.3, a follow-up analysis validates this assumption. To preview those results, 20,000 samples appear sufficient for the stabilization of the results.

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

3.2.4 Crosslinguistic comparison of acoustic measurements

Following from 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 often (but not always) found differences for 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 measurments across languages, with all talkers pooled together.

For each of the 12 acoustic measurements (but not the moving standard deviations), a separate Bayesian mixed-effects model was used to compare the distribution of values by talkers across languages. The models were implemented with *brms* () in R (). The brms package offers a simple R-based interface for fitting Bayesian models with the Stan probabilistic programming language ().

Bayesian models provide many advantages... For more information, see Vasishth et al. (2018)

A robust linear mixed-effects model was used for each standardized acoustic measure, with a Student's t distributed dependent variable. This decision was motivated by the strong possibility that outliers—real or erroneous—remained in the data. It also follows from Kruschke's (2013) paper arguing that robust Bayesian estimation provides better group comparison estimates. The exception to this was for SHR. As SHR is bounded by zero and one and contains many real and meaningful zero measurements, SHR was modeled as zero-inflated beta distribution. Apart from this, the models share structure and specifications. In all cases, the formula in 3.1 was used, where Measure refers to the particular standardized acoustic measure for the model.

$$\text{Measure} \sim 0 + \text{Language} + (0 + \text{Language} \mid \text{Talker}) \tag{3.1}$$

Weakly informative priors were used in all cases, following guidelines from Gelman for standardized data. For the robust linear regression models, they were as follows:

Each model consisted of four chains with XXX iterations, half of which were warm-up iterations, for a total of XXX samples after warm-up.

Divergent transitions, R-hat, ESS

These models give high-level insight into whether individuals differed for particular acoustic dimensions across language. The models also give insight into whether or not there is a common pattern across talkers. While this part of the analysis extends the work done primarily on F0 (e.g., by) to the present data set, it does

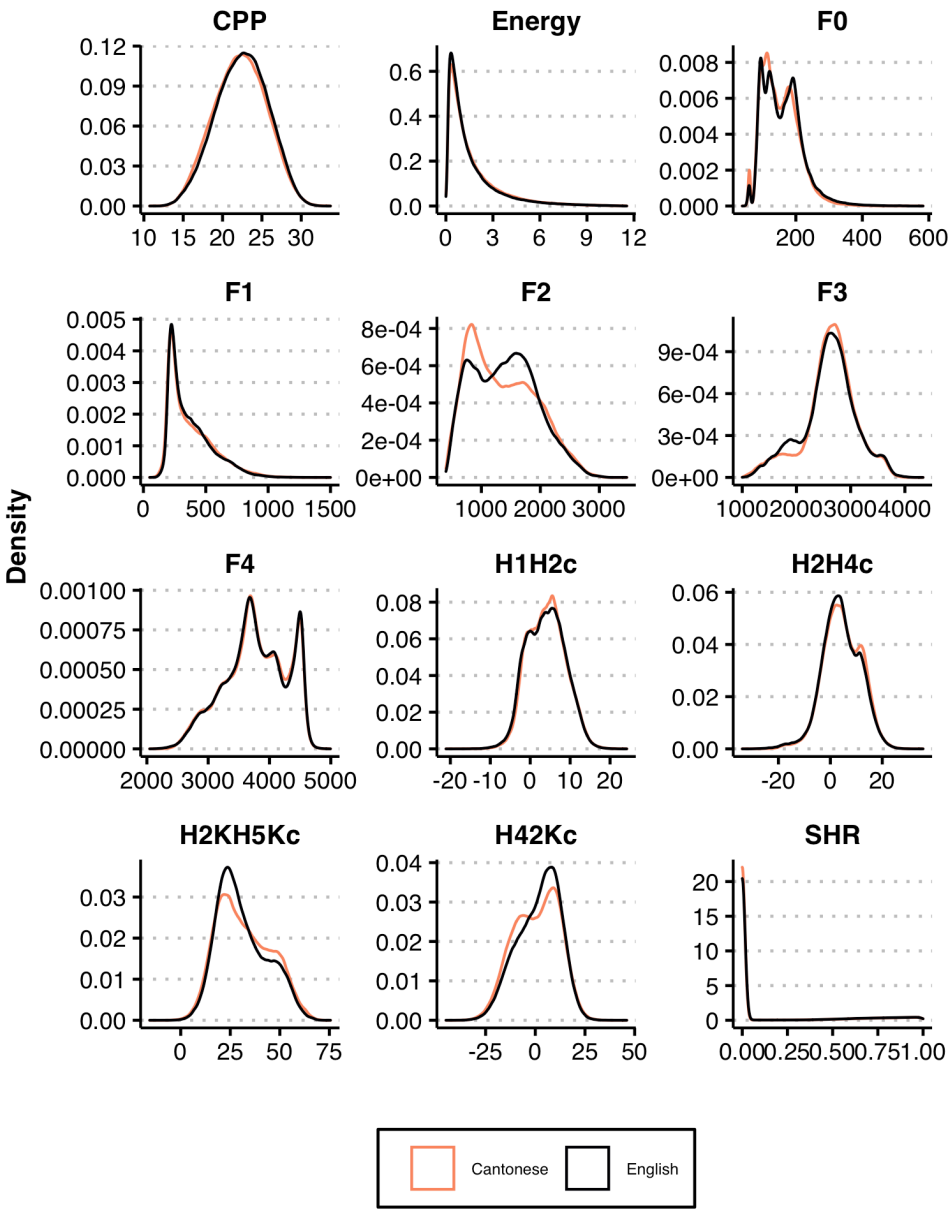

Figure 3.1: Distribution of all acoustic measures by language after filtering.

not elucidate how the different acoustic dimensions vary.

Inference in Bayesian statistics is based on the posterior distributions, which reflect the range and probability of credible values for parameters... For each of the models, a difference can be computed between the posterior distributions of Cantonese and English for each participant. This provides a magnitude of difference between the two languages, as well as how it varies across individuals.

In the case of SHR... how to interpret

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 not be perceptually meaningful to listeners. What matters is how the different aspects 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, such as regression modeling. The 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 is to the 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 normalized (z-scored) on by-PCA basis prior to 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 undesttanding 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 (Jolliffe, 2002) recommended adjustment to the Kaiser-Guttman rule. I used this rule, rather than a more sophisticated test (e.g., broken sticks), as it is not detrimental to our exploratory analysis to err on the side of including marginal components. Additionally, across each of the components, only loadings with an absolute value of 0.32 or higher were interpreted (Lee et al., 2019; Tabachnick and Fidell, 2013).

3.2.6 PCA results

The PCAs across both languages for all 34 talkers resulted in 10–15 components and accounted for 74.6–85.8% of the total variation. A slight majority of talkers had the same number of components for each of their languages (18/34). Of the remainder, most talkers had a difference of one in the number components (14/34), and far fewer differed by two (2/34). Table XX details the number of components and variance accounted for across all talkers and languages.

TABLE HERE

To assess whether talkers exhibit the same structure in voice variability across their languages, I first consider the patterns present across the different PCAs, as this provides context for understating what unique structural characteristics in talkers' voices looks like. To this end, I briefly summarize common patterns across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 3.2 shows the first four components of a single talker's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages. It also highlights the importance of not attributing to much value to the ordering of components, but rather to their composition and variance accounted for.

Broadly speaking, there were a lot of similarities in component composition across both talkers and languages, with the eight most commonly occurring com-

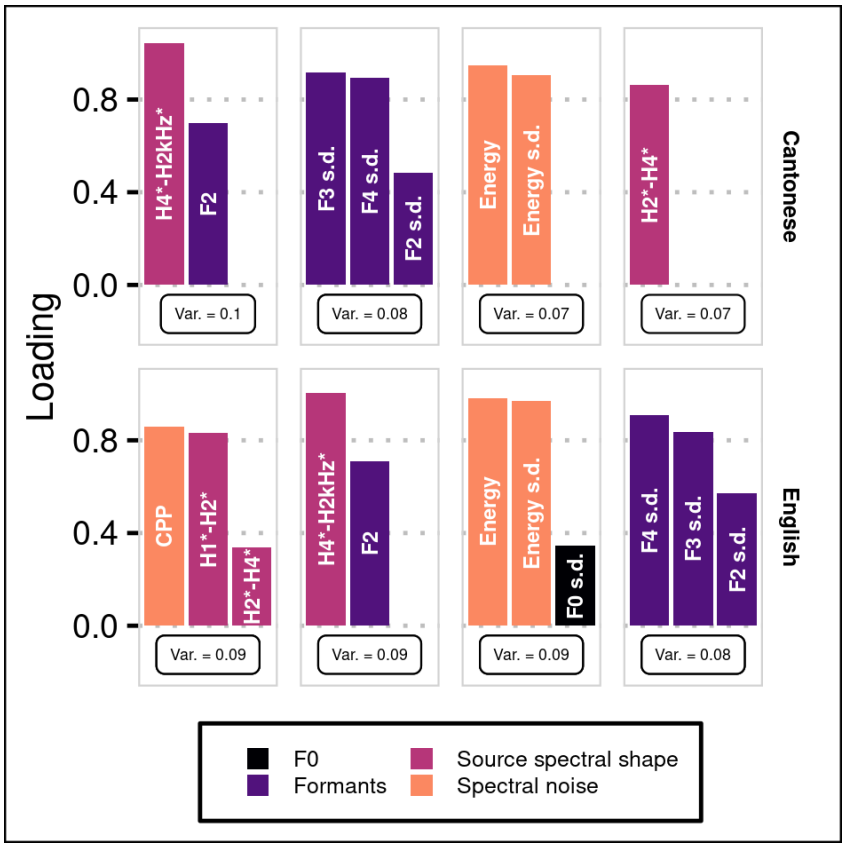

Figure 3.2: In the first four components of a talker's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

ponents summarized in Table 3.1. For context, recall that PCAs had anywhere from 10–15 components total. These eight components consisted of source spectral shape, spectral noise, as well as formant variables. On the other hand, F0 co-occurred with a wide variety of variables (often Energy), but in a manner that was less consistent across talkers. There were additional components (not reported here) that were shared by less than half of talkers. In summary, despite the greater amount of shared structure across PCAs than found in Lee et al. (2019), there is still ample room for idiosyncratic variation, both in terms of which variables co-occur, as well

as in how much variance different components account for.

Table 3.1: A summary of the most commonly occurring components across all PCAs. Variables are only included if |Loading| > 0.32. Italics indicate additional variables that were present on a component for a subset of talkers (i.e., an alternative but related configuration). *N* indicates the number of times a component occurred (out of 34), and *Var. %* gives the range of percent variance accounted for by the component.

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

3.2.7 Canonical redundancy analysis

Methods

In order 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 are related to one another. The two sets of variables are transformed such that the correlation between the rotated versions is maximized. This is useful here, as a

talker may have similar components in their English PCA and Cantonese PCA, but these components might not necessarily be in the same order, even if they account for comparable amounts of variance.

Redundancy is a relatively simple way to characterize the relationship between the loadings matrices of two PCAs—the two sets of variables under consideration here. For example, the two indices represent the amount of variation in a talker's Cantonese PCA output that can be accounted for via canonical variates by their English PCA output, and vice versa. Notably, the two redundancy indices are not symmetrical (Stewart and Love, 1968). This is particularly relevant in cases where the PCAs comprise different numbers of components, as determined by the stopping rule described above.

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

Results

Redundancy indices for within-talker comparisons ranged from 0.82 to 0.99, (*Mdn* = 0.93, *M* = 0.92, *SD* = 0.04), and are displayed in Figure 3.3, with the two redundancy indices for a given pair plotted against one another. Comparisons across talkers within-language (range: 0.63–0.98, *Mdn* = 0.84, *M* = 0.84, *SD* = 0.6) and across-language (range: 0.66–0.98, *Mdn* = 0.83, *M* = 0.84, *SD* = 0.6) are generally lower, but still relatively high. Within-talker values were confirmed to be higher than across-talker comparisons [*Welch's t*(71.36) = –17.83, $p < 0.001$, d = 1.76].

The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded components almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.2.6, there were a substantial number of commonly occurring patterns across talkers and languages.

3.3 Passage length analysis

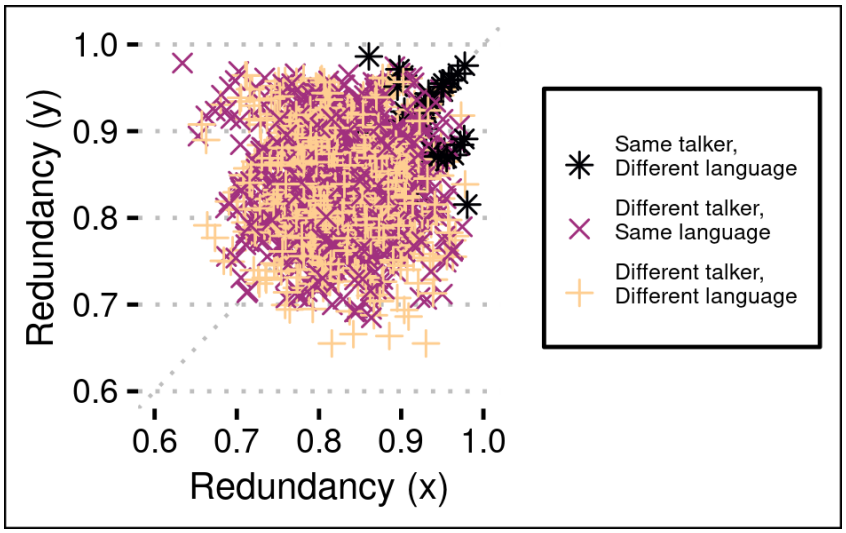

Figure 3.3: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are clustered at the top right.

3.4 Discussion and conclusion

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

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

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

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

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

Bibliography

Alderete, J., Chan, Q., and Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, 191:103952. → page 5

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 32, 34, 41

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 6

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 5

Audacity Team (2018). Audacity (R): Free audio editor and recorder. → page 9

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 26

Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Version 6.1.38. → page 37

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts l2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899. → page 35

Bullock, B. E. and Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, L., Winford, D., and deBot, K.,

editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam. → page 31

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 28, 29, 44

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 34, 35, 38

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., and Carter, D. (2014). Building bilingual corpora. In Thomas, E. M. and Mennen, I., editors, *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters. → pages 3, 4, 25

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 5

Fricke, M., Kroll, J. F., and Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, 89:110–137. → page 37

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806. → page 2

Garellek, M. (2019). The phonetics of voice. In Katz, W. F. and Assmann, P. F., editors, *The Routledge Handbook of Phonetics*. Routledge. → page 38

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 5

Google (2019). Cloud speech-to-text. → pages 5, 17

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 39

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246. → page 13

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 38

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 37

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 2

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → pages 45, 47

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 33, 34, 36

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 38

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. → page 28

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 30, 32, 33

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 27, 28, 35, 37, 38, 45

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 27, 45

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. → page 29

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 33, 34, 35, 36

Lee, J. L. (2018). PyCantonese [Version 2.2.0]. → pages 5, 19

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 26, 27, 28, 29, 30, 35, 37, 40, 41, 44, 45, 46, 49, 50

Lee, Y. and Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. → pages 27, 28

Lee, Y. and Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. 10.1121/1.5146847. → pages 28, 31, 44

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 5

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. → page 36

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 27

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 33

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 5

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 44

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → pages 13, 49

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner [Version 1.0.1]. → page 19

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 38

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 31

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 37

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, 43(1-2):65–84. → pages 13, 15, 18

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 32, 33, 35, 38, 41

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 32, 41

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 4

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376. → page 29

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. → pages 34, 35

Orena, A. J., Polka, L., and Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. *The Journal of the Acoustical Society of America*, 145(4):EL303–EL309. → page 29

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

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. → page 29

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 29, 30, 35

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, 15

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → page 44

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 34

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 37

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 15

Sjölander, K. (2004). The Snack Sound Toolkit. → page 38

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 17

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 47, 48

Sun, J. (2020). jieba [Version 0.42.1]. → page 19

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 39

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 45

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 13, 19

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 44

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 5

Wong, W. Y. P. (2006). *Syllable fusion in Hong Kong Cantonese connected speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. → pages 18, 19

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 33

Yang, Y., Chen, S., and Chen, X. (2020). F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Interspeech 2020*, pages 4163–4167. ISCA. → page 34

Yau, M. (2019). PyJyutping. → page 5

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 19

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 19