Crosslinguistic Similarity and Structured Variation in Cantonese-English Bilingual Speech Production

by

Khia Anne Johnson

B.A. Linguistics, University of Washington, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Linguistics)

The University of British Columbia
(Vancouver)

December 2021

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

Crosslinguistic Similarity and Structured Variation in Cantonese-English Bilingual Speech Production

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

Examining Committee:

Molly Babel, Linguistics, UBC
Supervisor

Kathleen Currie Hall, Linguistics, UBC
Supervisory Committee Member

Márton Sóskuthy, Linguistics, UBC
Supervisory Committee Member

TBD, Linguistics, UBC
University Examiner

TBD, Department, UBC
University Examiner

TBD, Department
External Examiner

Abstract

Bilingual speech production is highly variable. This variability arises for numerous sources, ranging from the heterogeneity of linguistic experiences to crosslinguistic influence and more. This area has historically been challenging to study, given the relative lack of high-quality bilingual speech corpora and scientific inquiry that such resources enable. This dissertation introduces the SpiCE corpus of bilingual speech in Cantonese and English and describes two corpus studies assessing crosslinguistic similarity. Chapter 2 describes how the SpiCE corpus was designed, collected, transcribed, and annotated. Broadly, it comprises recordings of 34 early Cantonese-English bilinguals conversing in both languages, hand-corrected orthographic transcripts, and force-aligned phone level annotations. Chapters 3 and 4 are motivated by a desire to understand how crosslinguistic similarity in the speech signal facilitates multilingual talker identification and discrimination.

Chapter 3 addresses this question at the level of voice quality. Using 24 filter and source-based acoustic measurements over all voiced speech in the interviews, principal components and canonical redundancy analyses demonstrate that while talkers vary in the degree to which they have the same "voice" across languages, all talkers show strong similarity with themselves. To a lesser extent, talkers exhibit similarities with one another, providing further support for prototype models of voice.

Chapter 4 pivots to the level of sound categories. Prior work in this area emphasizes detecting crosslinguistic influence for phonetically distinct yet phonolog-

ically similar sounds. This chapter leverages the uniformity framework to assess underlying phonetic similarity for the long-lag stop series in Cantonese and English. Results indicate moderate patterns of uniformity within each language and weak patterns across languages. These weak patterns were further problematized by clear crosslinguistic differences for two of the sounds, which were apparent despite their proximity in the long-lag space. Yet, at the same time, more of the overall variation seems to derive from individual-specific differences.

Together, Chapters 3 and 4 provide evidence for talker identification and discrimination based on voice quality and category similarity. Altogether, this dissertation provides a novel resource and highlights the necessity of doing corpus phonetics research, both for understanding productive processes and in speculating about the bases of different mechanisms in perception.

Lay Summary

Bilingual speech is highly variable—one major source of variability arises from how bilinguals' languages influence one another. This dissertation sheds light on how languages influence each other by analyzing conversations with Cantonese-English bilinguals. In addition to contributing a new open-access data set, this dissertation examines similarity across languages. The first question deals with voice: Do bilinguals have the same voice in each language? Are voices like auditory faces? In short—yes. The second question addresses whether this same group shares P, T, and K sounds across languages—that is, do bilinguals say K the same way in English and Cantonese. The answer to this question is less clear, with variability arising from the language and the person. Together, these studies clarify which aspects of speech can be used to recognize individuals speaking more than one language and give insight into how languages do and do not interact in the mind.

Preface

This dissertation is original work, and I am the primary author of each chapter. Additionally, I am the sole author of chapters 1, 4, and 5. All work in this dissertation was covered by the Behavioural Research and Ethics Board at the University of British Columbia under certificate H18-02017.

Chapter 2 was a collaborative effort, and I conceptualized, designed, and led all parts of the corpus development process. The corpus itself was collected by Nancy Yiu, Ivan Fong, and myself. Various members of the Speech-in-Context Lab supported transcription and annotation. The writing in Chapter 2 is based on a paper published in the proceedings of the *12th Language Resources and Evaluation Conference* (Johnson et al., 2020a), for which I did the vast majority of the writing.

Chapter 3 is based on a paper published in the *Proceedings of Interspeech 2020* (Johnson et al., 2020b). Molly Babel contributed to the conceptualization, design, writing, and revisions. Robert A. Fuhrman advised on the methods and suggested the addition of the canonical correlation analyses.

Chapter 4 is based on a solo-authored paper published in the *Proceedings of Interspeech 2021* (Johnson, 2021a). Molly Babel provided early input regarding the study's design and feedback on a prior version of the paper.

Table of Contents

List of Tables

List of Figures

Acknowledgments

This is a pre-defense draft of the acknowledgements, and as a result, will likely change so that I can thank the rest of the examining committee.

I am very fortunate to have had a deep support network throughout my degree at UBC and while writing this dissertation. First and foremost, my gratitude goes to Molly Babel—my advisor, mentor, collaborator, and friend. Molly supported me intellectually, creatively, and financially at all stages of my graduate career and dissertation writing adventure. I somehow stumbled into the perfect advising relationship, and I am so glad I did. I must also extend my gratitude to my committee members. Kathleen Currie Hall and Márton Sóskuthy both inspire me in different ways—encouraging me to think deeper, pursue open science practices, develop my technical expertise, and so much more. I know that my dissertation has less phonology in it than Kathleen may have hoped for, but I deeply appreciate her support for my research program, regardless of how closely it lined up with hers. This is the kind of committee I didn't know to hope for.

Other faculty in linguistics have inspired and pushed me to excel in many different ways. Carla Hudson Kam helped me through the qualifying papers process by pushing for a high standard of work while letting my not-quite-publishable study serve its purpose: learning. Henry Davis tried to get me to do a fieldwork dissertation, and while unsuccessful, it made me feel confident in what I did choose.

In fall 2017, two seminars helped shape my research program in lasting ways. I am grateful to John Alderete and Henny Yeung at Simon Fraser University for their excellent seminar on the psycholinguistics of Chinese languages. In the same

term, I also took an advanced phonetics seminar with Molly. Together, these seminars pushed me to think deeply about sound and psycholinguistics, and I certainly wouldn't have written this thesis in this way without them.

This dissertation included creating a speech corpus from scratch. I would not have been able to do this without a fantastic team of research assistants. Ivan Fong and Nancy Yiu were active during all stages of the project—designing, recruiting, recording, and transcribing. They did it all. I am also deeply grateful to everyone who contributed to the corpus, including Katherine Lee, Kristy Chan, Natália Oliveira Ferreira, Michelle To, Rachel Ching Fung Wong, Christina Sen, Ariana Zattera, and Rachel Soo. Creating the corpus would not have been possible without support from the UBC Public Scholars Initiative. I was able to submit and ultimately win an award, largely thanks to Serbulent Turan's tireless support of public scholars and Molly and Kathleen's encouragement to dream big. (And their support of my application, too!)

I would also like to thank everyone in the Speech-in-Context Lab for giving me feedback on earlier versions of my chapters, engaging with my ideas, and being the first group of people to use the corpus! I know that the lab will be the perfect home for the corpus over time. My dissertation also benefitted from feedback from the audiences of LREC, Interspeech, ASA meetings, and the Cantonese psycholinguistics workshop organized by John Alderete.

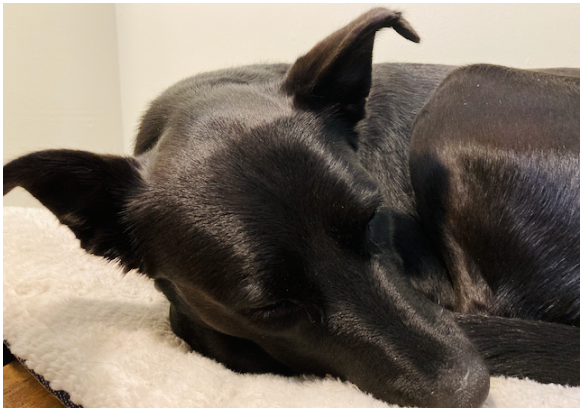
Writing is hard and having a community helps. Thank you to Gloria Mellesmoen for our many writing sessions, even those where we didn't write. It was the support I clearly needed, and it helped me keep a level head throughout the grad school roller coaster. While Gloria was the only person I managed regular writing time with, having an online writing community via 100 Days and Twitter was one of the more pleasant surprises of the pandemic times. Knowing that we were all in it together made it all better.

Thank you to the grad students ahead of me in the linguistics program for showing me the ropes, answering my questions, going out for beers, and being all-around awesome people: Megan Keough, Oksana Tkachman, Emily Sadlier-

Brown, Adriana Osa Gomez del Campo, Alexis Black, Natalie Weber, Michael Schwan, and many many others. And to my cohort—Gloria (again!), Roger Yu-Hsiang Lo, Wendy Amoako, Bruno Andreotti, and Daniel Reisinger.

And lastly, to my friends and family. My parents have always supported me in following my linguistics career, and supporting me in this Ph.D. was no exception. Little did they know back when I started how this would launch my career. Thank you to my dad for all the erudite arguments over dinner and relentless word games. Thank you to my mom for always wanting to read my papers and for leading by example—we're both Dr. Johnson now!

Thank you, in particular, to my amazing partner and husband, Brennan, for supporting me even when that meant tearing me away from a midnight writing deadline for fancy takeout. That extra mile means the world to me. And to Pepper, dog and writing companion extraordinaire—she was impeccable in reminding me to take writing breaks, ideally outside. 15/10 work-from-home pal.



Chapter 1

Introduction

Bilingual and monolingual linguistic experience differs drastically. This sentiment is captured by the often-repeated observation that a "bilingual is NOT the sum of two complete or incomplete monolinguals; rather, [they have] a unique and specific linguistic configuration...a different but complete linguistic entity" (Grosjean, 1989, p. 6). One of the defining characteristics of a bilingual's linguistic configuration is a shared phonetic space where both languages are produced and perceived (Flege & Bohn, 2021). Broadly, this dissertation is concerned with the implications of such a space and what aspects of sound are shared across languages. And, given that speech usually occurs in a communicative context, what is available in the bilingual speech signal to facilitate processes like talker identification and processing?

The studies presented in this dissertation approach this larger question at different levels in the phonetic space—voice quality (Chapter 3) and sound categories (Chapter 4). Yet, the studies share motivation in speech perception, as various levels of phonetic variation have been proposed to account for how listeners can identify and track a talker across languages. This introduction briefly sets up that shared motivation, tying two otherwise quite different studies together. Given the unique angle of each study, most of the literature is reviewed in the relevant chapters.

The introduction proceeds as follows. Section 1.1 gives context to the study of bilingualism in phonetics in broad strokes—that is, who is considered to be bilingual and what are some of the key characteristics that define them. The goal is to set up later chapters rather than provide a comprehensive discussion. Section 1.2 reviews some of the literature on how multilingual phonetic variation is perceived, emphasizing talker identification and how multilingual listeners process code-switching. Section 1.3 motivates the need to attend to speaking style and argues that spontaneous speech corpora greatly facilitate the study of multilingual phonetic variation. Lastly, Section 1.4 provides the specific goal or research question for each of the main content chapters—2, 3, and 4.

1.1 Bilingualism

The population of interest in this dissertation comprises early bilinguals who are comfortable speaking and comprehending their two main languages—Chapter 2 provides a detailed description. In the most general sense, a bilingual is someone with knowledge of two or more languages (Grosjean, 1989). This incredibly broad definition includes a diverse range of types of bilinguals. Different kinds of bilinguals are perhaps best described on a continuum from first language (L1) to second language (L2) dominance, the bookends of which are monolinguals and replacive bilinguals, with learners, attriters, and balanced bilinguals in the middle (Gertken et al., 2014). Using a continuum in this way reflects the heterogeneous nature of bilingualism, even if it only captures a particular facet of bilingual competence. Dominance—and other aspects like patterns of use—are affected by factors such as age of acquisition, immersion environment, frequency, social and communicative context (Gertken et al., 2014).

While a spectrum may better reflect the reality of bilingualism, much of the literature focuses on more clearly defined groups at discrete points of the spectrum, such as language learners or early (balanced) bilinguals. Typically, early bilinguals have learned both languages from their first years of life. A common cutoff is age five, or the age at which children begin regularly attending primary school

(Amengual, 2017), as this marks a qualitative change in the kind of linguistic input bilinguals receive. Regardless of when bilinguals acquire a language, they do not necessarily use their languages to the same extent across different domains. For example, a Cantonese-English bilingual in Vancouver, BC, Canada, might use English at school and Cantonese at home. Bilingual language experience varies is still other ways, including code-switching (Fricke et al., 2016a), immersion environment (Sancier & Fowler, 1997), and formal instruction (Fricke et al., 2019). Each of these factors has a demonstrated effect on speech production. Such variety leads to markedly different linguistic experiences across groups of bilinguals, and as a result, markedly different patterns in speech production.

Across different kinds of phonetics research in bilingualism, there is a common trend of comparing bilinguals to "closely matched" monolingual populations. Given the sheer heterogeneity within and across bilingual populations, there may not always be an appropriate monolingual comparison group. Further, Grosjean (1989) and many others have argued that such comparisons are often inappropriate. As a result, drawing comparisons between monolinguals and bilinguals may not always be fruitful or necessary, depending on the circumstances. This is reflected by a shift in the literature towards examining bilinguals on a within-population (e.g., Chan et al., 2020) or within-talker basis (e.g., Simonet & Amengual, 2019), or by comparing separate bilingual populations with different characteristics (e.g., Brown & Harper, 2009). This range of study designs will be apparent in the literature reviews of the following chapters. While there remains a broad range of comparisons in the literature, in all cases, there is a strong push to consider bilinguals as the complete speakers they are.

1.2 Processing bilingual talkers

Communicating in more than one language doesn't just involve the language produced by bilingual talkers; it is also impacted how listeners perceive those talkers. As noted early in this introduction, one of the major consequences of bilingualism is a shared phonetic space (Flege & Bohn, 2021), in which bilinguals presumably

(i.e., are hypothesized to) use similar voice quality to produce similar sound categories.[1] This shared phonetic space thus also impacts the perception of bilingual talkers, whether the listener is a fellow bilingual or not.

While bilingual speech perception is a large and multifaceted field (Ingvalson et al., 2014), the clearest motivation for the studies in this dissertation comes from the advantage that multilingualism offers in talker identification. Orena et al. (2019) report on a talker identification study with French-English bilingual talkers, in which bilingual listeners—particularly those with language mixing experience—were better able to generalize talker-indexical information learned in English to French and vice versa when compared to monolingual English listeners. However, all groups in the study were above chance, suggesting both linguistic and non-linguistic components to talker identification. Orena et al. offer several potential explanations for this advantage: "that there are systematic changes in indexical information...[or] systematic consistencies in linguistic information across bilingual speech" (2019, p. EL308). Bilingual listeners are highly sensitive to subtle differences in acoustic input (Ju & Luce, 2004). As a result, the presence of systematicity in both talker-indexical and linguistic information—however subtle—would be accessible to bilingual listeners, particularly those with language mixing experience. Orena et al. also suggest that the results could be explained because the bilinguals "were familiar with both languages...while the monolinguals were only familiar with one of the languages" (2019, p. EL309), though this account would be difficult to separate from the previous two.

Regardless of the particular explanation, the bilingual advantage in bilingual talker identification likely arises from their deep familiarity in listening to how talkers vary within and across languages. While these accounts emphasize the linguistic aspect of bilingual competence, there is more to the picture than that. Bullock & Toribio (2009) highlight the integral role that sociolinguistic competence plays in accounting for variability in production and that bilinguals have an

[1]The speech production literature discussing similarity in voice quality and sound categories will be reviewed in greater detail within Chapters 3 and 4, respectively.

expanded repertoire of forms. That is, a bilingual can produce forms canonical to either language, and they can also show divergence, convergence, interference, or hypercorrection depending on the social and cognitive circumstances. This competence is echoed by Kleinschmidt et al. (2018) in their integrated account of how social, acoustic, and linguistic variation are perceived. Learning the structure and systematicity of variation, then, is both a linguistic, talker-indexical, and social venture. This dissertation investigates the role of the former two but acknowledges the importance of attending to and considering the social component as well.

While Orena et al. (2019) point to some prior work supporting the talker-indexical and linguistic accounts of bilingual talker identification, convincing evidence remains scarce. This dissertation directly addresses these accounts from the perspective of documenting the speech signal. Chapter 3 examines voice variation—generally considered to reflect talker-indexicality. Chapter 4 focuses on the structure of phonetic category variation—a reasonably clear example of linguistic information. While using distinct methods and addressing different aspects of phonetics, each chapter represents an aspect of the signal that may facilitate crosslinguistic talker identification.

1.3 Variability in conversational speech

One of the primary goals of this dissertation is to document and investigate the structure of phonetic variation. While variability is inherent to the speech signal (cf. the lack of invariance problem Liberman et al., 1967), spontaneous speech encompasses a greater degree of variability than other speaking styles (e.g., reduction phenomena: Johnson, 2004). Spontaneous speech—in the form of conversations—is thus the focus of study in this dissertation.

Additionally, as the motivation for the studies in Chapters 3 and 4 stems from listeners' ability to identify talkers in more than one language, using conversational speech also supports the external validity of this dissertation. Conversational speech better reflects the range of forms that people use and perceive in their daily lives (cite something—ideas?). Additionally, given the potential range of variabil-

ity, it is also necessary to study a large enough sample such that it comprises the range of variation for the particular communicative situation. For similar reasons, Tanner et al. (2020) argue that large-scale corpus studies are uniquely valuable for understanding phonetic variation. In this vein, this dissertation leverages the study of conversational speech data from a sufficiently large speech corpus. The corpus, along with further motivation for its use, is described in Chapter 2.

1.4 Thesis goals and research questions

While each of the main content chapters in this dissertation is united by common motivation, each has a unique focus or research question. These are as follows:

Chapter 2 expands on the motivation behind studying spontaneous speech and introduces the SpiCE corpus of spontaneous bilingual **Sp**eech **i**n **C**antonese and **E**nglish (Johnson, 2021b). The development and dissemination of this corpus comprise a substantial portion of this dissertation.

Chapter 3 focuses on the structure of voice variation. In broad terms, it asks: do bilinguals have the same voice in each of their languages? More specifically, do bilinguals exhibit similar spectral properties and lower-dimensional structure in their voice across each language they speak? Chapter 3 also addresses a methodological question regarding the sample size necessary for the methods used.

Chapter 4 focuses on the structure of sound categories. In broad terms, it asks: Do bilinguals produce long-lag stops in the same way in each of their languages? More specifically, it describes the structure and sources of variation in how bilinguals produce voice-onset time in conversational speech.

Chapter 2

The SpiCE Corpus

2.1 Introduction

Much of our formal knowledge about the phonetics of spoken language and speech processing comes from monolingual individuals producing scripted speech in laboratory settings. While far from the only source of knowledge, especially as areas like sociophonetics and corpus phonetics continue to grow, laboratory speech perhaps retains an outsized role. Monolingual lab speech allows researchers to exercise tight control over the linguistic backgrounds of the speakers and the linguistic material (e.g., reading or repeating sounds, words, or sentences). While highly informative, these controlled monolingual speech samples represent a minority of the contexts in which spoken languages are used around the world. Bilingualism is the norm, not the exception, and individuals regularly make creative linguistic choices in their spontaneous speech.[1]

Conversational speech allows for a richer empirical description of spoken language compared to—or at the very least, in addition to—laboratory elicited speech. It provides a more realistic representation of how individuals produce language in

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

everyday contexts that isolated word production and sentence reading do not faithfully capture. It enables and facilitates the study of non-formal speech styles, style-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 & Babel, 2021).

Discrepancies between results for conversational and lab speech have been found for monolingual (English) speech but are likely to be found with bilingual speech as well. Research on bilingual conversational speech is limited, however, as the resources needed for this type of inquiry are relatively rare. Furthermore, the corpora that do exist have typically focused on bilinguals of two European languages.

As a step towards filling this gap, this chapter introduces the **SpiCE** corpus of conversational bilingual **Sp**eech **i**n **C**antonese and **E**nglish (Johnson, 2021b). In contributing to filling this gap, SpiCE will allow researchers to address a set of research questions that were previously not possible, using both conversational, bilingual speech, and sophisticated phonetic measurements at scale. To preview the end product before diving into the details—SpiCE is a corpus of bilingual speech in Cantonese and English, comprising high-quality recordings of 34 early Cantonese-English bilinguals. The participants were young adult members of the heterogeneous bilingual speech community in the Vancouver, BC, Canada, area. Each participant completed a few different tasks—reading sentences, narrating a cartoon storyboard, and conversing freely in a semi-structured interview with a bilingual peer as the interviewer. All of the recordings were manually transcribed at the word level and force-aligned at the phone level. This chapter describes each of these components in detail and offers justification for the decisions where warranted.

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

Given the bilingual design, SpiCE crucially includes speech from the same individual in more than one language. Inspiration in this regard is drawn from the Bangor corpora of Spanish-English, Welsh-English, and Welsh-Spanish bilingual speech (Deuchar et al., 2014). The Bangor corpora include speech from the same individual in more than one language but largely comprise field recordings, some of which are noisy. For example, many of the recordings in the Spanish-English Bangor corpus were made with a lapel microphone worn on the participant's belt and others with a radio microphone placed on a table. This variable—and often noisy—recording quality limits the scope of phonetics research using the corpora. Additionally, the Bangor corpora were designed for understanding code-switching in everyday situations. While this facilitates understanding broad patterns of language use, it also means that the corpora are not necessarily balanced for the languages involved—people do not necessarily use their languages in equal proportions. So while these corpora are incredibly valuable for linguistics research, there are nonetheless limitations. Compared to these corpora, SpiCE uses a more controlled and balanced recording setup, which allows for more nuanced acoustic-phonetic measurements. This is, however, at the expense of other criteria (e.g., naturalness), in which the Bangor corpora excel.

The SpiCE corpus was initially designed in mid-2018, with recording beginning in the fall of 2018. At that time, there were relatively few corpora comprising multilingual speech. There has been a notable uptick in the development of such corpora since that time. For example, there was a session at the 2018 Langauge Resources and Evaluation Conference on "Bilingual Speech Corpora and Code-

switching."[2] In a similar vein, Microsoft hosted the "First Workshop on Speech Technologies for Code-switching in Multilingual Communities" in conjunction with Interspeech 2020.[3] These workshops are examples of the growing interest in working with multilingual speech data at larger scales. It parallels the similarly large growth of corpus phonetics as a field (Liberman, 2019; Grieve, 2021).

SpiCE is also unique in the population it represents. Many of the resources available to researchers on sites like BilingBank, ELRA, and elsewhere feature late bilinguals and second language learners and vary widely in task and recording quality. One example of a Cantonese-English resource that fits this description is the ShefCE corpus (Ng et al., 2017). ShefCE is a parallel corpus featuring L1 Hong Kong Cantonese and L2 English read speech, where participants read lectures in each language one sentence at a time. While there are similarities with what SpiCE aims to accomplish (e.g., promoting research with Cantonese-English bilingual speech), ShefCE occupies a different niche in the speech sciences—it was designed for L2 pronunciation assessment and training speech recognition models. Another resource focused on bilinguals L1 Cantonese and L2 English is the ALLSSTAR corpus—Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings—of which Cantonese L1 talkers represent 14 of 140 people for whom English is their L2 (Bradlow et al., 2011).

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

[2]http://www.lrec-conf.org/proceedings/lrec2018/sessions.html

[3]https://www.microsoft.com/en-us/research/event/workshop-on-speech-technologies-for-code-switching-2020/

audiobooks, etc.; e.g., Librispeech, 1000 hours: Panayotov et al., 2015), crowd-sourced online (e.g. Mozilla Common Voice, 2500 hours: Ardila et al., 2020), via phone (e.g., SWITCHBOARD, 260 hours: Godfrey et al., 1992), and other similar more scalable methods. The reason? High-quality, purpose-built corpora are expensive and time-consuming to create.

To my knowledge, this type of resource does not yet exist for any pair of languages, much less for a typologically distinct pair like Cantonese (Sino-Tibetan) and English (Indo-European). Furthermore, Cantonese is a relatively understudied language, despite there being approximately 85 million speakers around the world (Ethnologue, 2021), though this is changing with new Cantonese language corpora (Luke & Wong, 2015; Leung & Law, 2001; Winterstein et al., 2020; Alderete et al., 2019), natural language processing tools (Lee, 2018; Yau, 2019), and support in speech technology applications (Google, 2019).

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

2.2 Corpus design and creation

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

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

2.2.1 Recruitment

Participants were recruited for the SpiCE corpus through a variety of methods at the University of British Columbia. This included word of mouth, the Linguistics Human Subject Pool, the Psychology Paid Studies list, advertisements in department email lists, advertisements in linguistics courses, printed flyers, and posts on various club forums.

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

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

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

2.2.2 Participants

The recordings in SpiCE comprise the speech of 34 early Cantonese-English bilinguals. Throughout this chapter and the corpus, participants are identified by participant IDs. The IDs are designed to provide basic information about the participant. For example, VF19A indicates that the participant was recorded in **V**ancouver, identified as **F**emale, and was 19 years old at the time of recording. The letter at the end distinguishes participants of the same age and gender. There were 17 participants who self-identified as female and 17 as male. Participants ranged in age from 19 to 34 years old at the time of recording. Apart from one talker who reported mild high-frequency hearing loss (VM25A), all participants reported normal speech and hearing. Additionally, all participants resided in the Metro Van-

couver, BC, Canada, area at the time of recording. The SpiCE corpus also includes a detailed summary extracted from an extensive language background survey administered before the recording session (without the researchers present), as well as a copy of the survey itself. Basic summary information is included in Table 2.1, and in visualizations throughout this chapter. All participant information is based on self-reported participant data from the survey.

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

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

This corpus reflects the diverse nature of Cantonese-English bilingualism in Vancouver, as it includes Canadian-born heritage speakers, recent immigrants from Hong Kong, Cantonese speakers from other parts of the Cantonese diaspora, and individuals who do not neatly fit into these particular categories. As a result, while all speakers are early bilinguals, various dialects are represented. Figure 2.1 depicts where SpiCE participants reported living during different age intervals. These intervals were selected after reviewing freeform participant responses comprising when they lived in different places. Specifically, Figure 2.1 reports the

No.	ID	Order	Age	Gender	Age of Acquisition	
					English	Cantonese
1	VF19A	E → C	19	F	0	0
2	VF19B	E → C	19	F	0	0
3	VF19C	E → C	19	F	3	0
4	VF19D	C → E	19	F	2	0
5	VF20A	C → E	20	F	4	0
6	VF20B	C → E	20	F	5	0
7	VF21A	E → C	21	F	0	0
8	VF21B	C → E	21	F	3	0
9	VF21C	C → E	21	F	4	0
10	VF21D	E → C	21	F	0	0
11	VF22A	C → E	22	F	0	0
12	VF23B	E → C	23	F	2	0
13	VF23C	C → E	23	F	0	0
14	VF26A	C → E	26	F	0	0
15	VF27A	E → C	27	F	0	0
16	VF32A	C → E	32	F	3	0
17	VF33B	C → E	33	F	0	0
18	VM19A	E → C	19	M	0	0
19	VM19B	C → E	19	M	2	0
20	VM19C	E → C	19	M	0	0
21	VM19D	C → E	18	M	1	1
22	VM20B	E → C	20	M	0	0
23	VM21A	E → C	21	M	0	0
24	VM21B	E → C	21	M	0	0
25	VM21C	C → E	21	M	0	0
26	VM21D	C → E	21	M	0	0
27	VM21E	C → E	21	M	5	0
28	VM22A	C → E	22	M	4	0
29	VM22B	E → C	22	M	0	0
30	VM23A	E → C	23	M	0	0
31	VM24A	E → C	24	M	3	0
32	VM25A	E → C	25	M	4	0
33	VM25B	E → C	25	M	0	0
34	VM34A	C → E	34	M	0	0

Table 2.1: Basic participant information from the language background survey, including age, gender (M for male and F for female), age of acquisition (phrased as "age began learning"), and the order the interviews occurred (E for English and C for Cantonese). See Section 2.2.4 for information about interview order.

number of participants who indicated that they lived in a given country during the age ranged for the panel. For example, if a participant moved from Hong Kong to Canada at age 7, they would be counted in both bars in that panel.

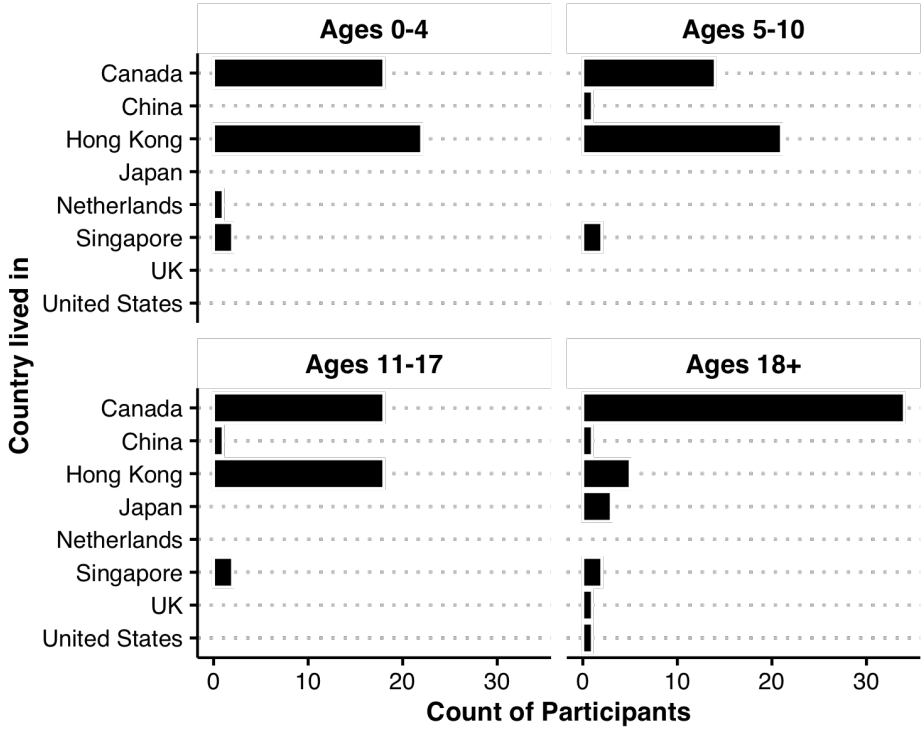


Figure 2.1: This four panel bar chart summarizes where the SpiCE participants lived during different portions of their lives.

Directly soliciting Cantonese dialect information would have been challenging, as many of the participants in the corpus would not have straightforward dialect classifications. This is especially true for individuals who were born and/or raised in the Cantonese diaspora, but to Hong Kongers as well, given the extent of multilingualism and globalization in Hong Kong (Bolton et al., 2020). In light of this, it is useful to summarize where the SpiCE participants' caretakers were primarily raised. Figure 2.2 does exactly this. The most well-represented group is Hong Kong, as 29 of 34 participants reported having at least one caretaker who was

primarily raised in Hong Kong. Of these, 20 report only having caretakers raised in Hong Kong. If caretaker birth location is considered instead, the numbers are 27 and 18, respectively.

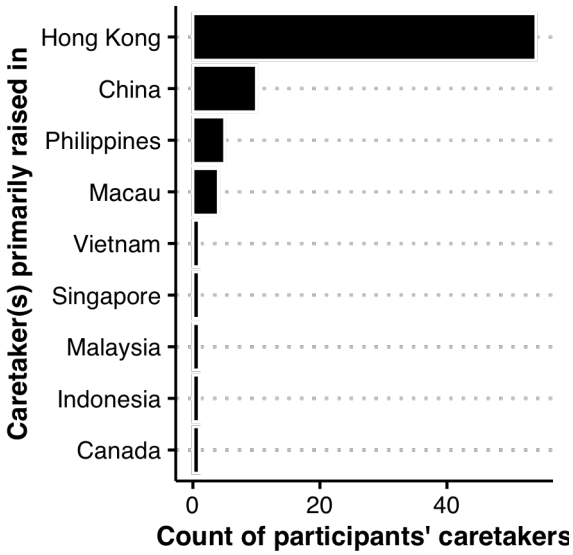


Figure 2.2: This bar chart summarizes the number of caretakers who were raised in various locations. Note that the number of caretakers reported by individual participants varies.

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

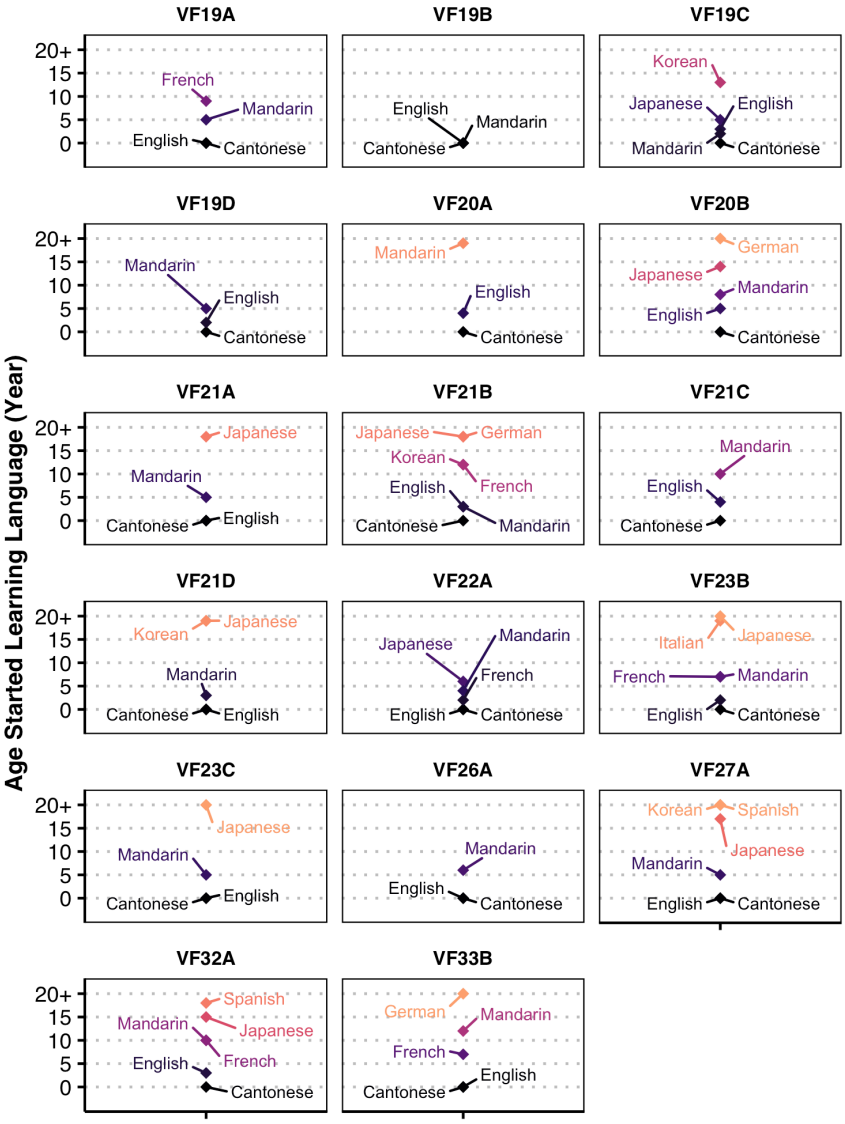


Figure 2.3: Multilingualism for the female participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

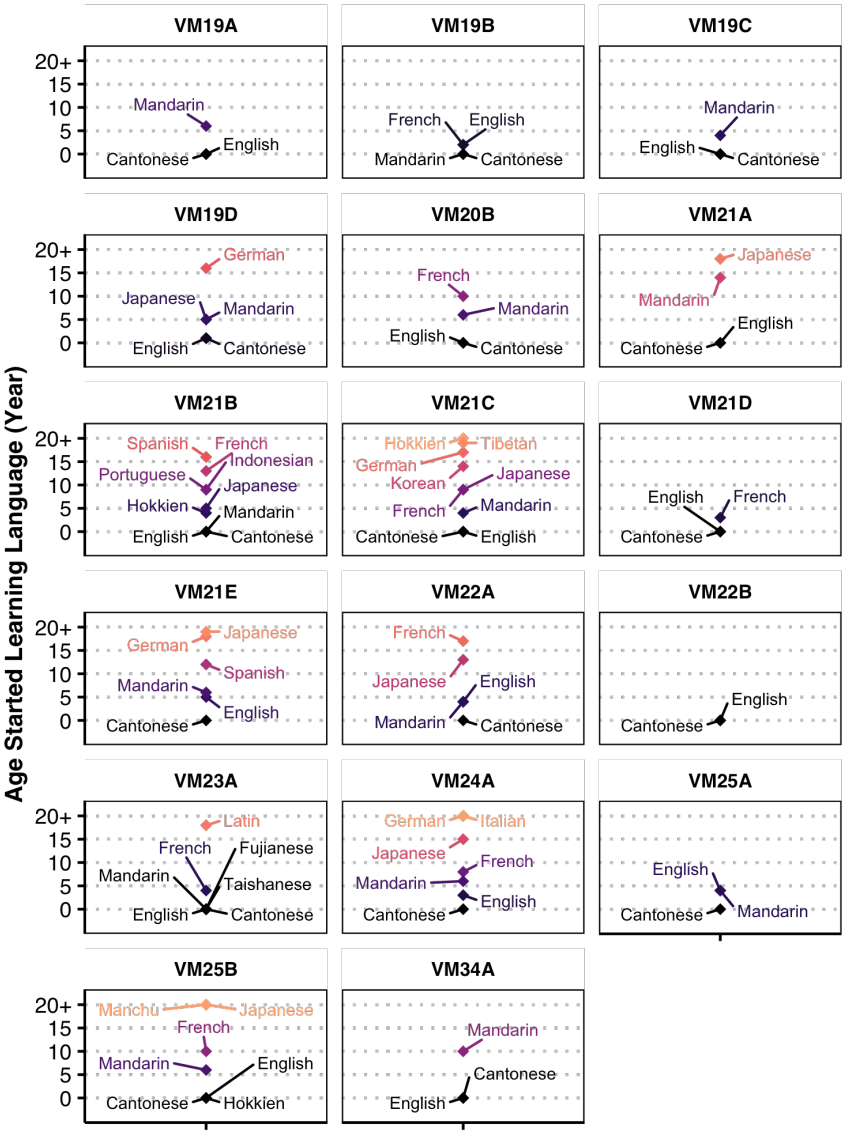


Figure 2.4: Multilingualism for the male participants in the SpiCE corpus. Points represent the age that a participant began learning the language indicated in the label. Color is redundant with age, such that earlier ages are darker in color.

2.2.3 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 who grew up in Hong Kong and moved to Vancouver for university. The interviewer and participant were outfitted with AKG C520 head-mounted microphones positioned approximately 3 cm from the corner of the mouth. The microphones were connected to separate channels on a Sound Devices USBPre2 Portable Audio Interface. Stereo recordings were made with Audacity 2.2.2 (Audacity Team, 2018) on a PC laptop and saved with a 44.1 kHz sampling rate and 16-bit resolution.[4]

2.2.4 Recording procedure

Upon arrival, participants were provided with an overview of the recording session procedures and informed of the corpus publication process. This included informing participants that they would be able to withdraw their data up until the SpiCE corpus' public release and that they would receive notice at least 30 days before publication.[5] Subsequently, participants were asked to provide written consent. Upon consent, participants completed a set of tasks in English and the same set of tasks in Cantonese—all within the same session. The order of languages was counterbalanced across participants (see Table 2.1). This counterbalancing did not extend to other participant characteristics, and as a result, a higher proportion of the female participants completed the Cantonese part of the session before the

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

[5]No participants withdrew their data. When participants were notified of the upcoming corpus release, they were also encouraged to let the research team know if there were any portions of the interview they would like silenced from the published version.

English part and vice versa for the male participants.

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

Sentence reading

Sentence reading was included in the session to ensure that different participants produced a set of identical items, considering the core of the session was an unscripted, conversational interview (described in Section 2.2.4). While these sentences do not exhaustively reflect the sound systems of Cantonese and English, they provide samples of identical items for all individuals, which is advantageous for future analyses or projects that require matched utterances.

Participants first read the sentences listed in Table 2.3 and Table 2.2 aloud, pausing between sentences. Participants completed a single repetition and were not instructed to speak in a particular style. As participants had varying levels of Cantonese reading ability, they were simultaneously presented with the Cantonese characters, Jyutping romanization, and English translation.[6] The Cantonese sentences were well-known declarative phrases, typically associated with Chinese New Year.[7] While a more explicitly balanced set of sentences could have been

[6]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).

[7]It is possible that familiarity and high frequency of some of these phrases led to them being

used, participants' familiarity was deemed more important, as many Cantonese-English bilinguals in Canada are not literate in Cantonese. The English sentences included the Harvard Sentences list number 60 (IEEE, 1969), as well as a series of holiday-themed declarative sentences to better match the content of the Cantonese sentences. This task was relatively formal and typically lasted less than one minute in each language.

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

Storyboard narration

For the second task, participants narrated a short story from a cartoon storyboard originally developed for linguistic fieldwork (Littell, 2010). The storyboard followed a simple plot about receiving gifts and writing thank-you notes to family members and friends—a topic that Cantonese-English bilinguals in the corpus were expected to be familiar with in both languages. A reproduction of the storyboard is available with the corpus download. This task was less formal than the sentence reading task and ensured that different participants produced some of the same words in a more spontaneous context. Participants varied in how they approached this task, with some treating it as a series of picture description tasks and others taking a more narrative approach. Despite this difference, this task may be useful for future analyses or projects that require utterances in a matched semantic space, as participants narrated the same cartoon in each language. This ensured that some of the same content was conveyed in each language (e.g., productions of *mother* in both languages). The storyboard narration lasted 4–5 minutes in each

produced with reduction patterns not present in a typical reading style—this is a limitation of the sentences.

No.	English
1	Stop whistling and watch the boys march
2	Jerk the cord, and out tumbles the gold
3	Slide the tray across the glass top
4	The cloud moved in a stately way and was gone
5	Light maple makes for a swell room
6	Set the piece here and say nothing
7	Dull stories make her laugh
8	A stiff cord will do to fasten your shoe
9	Get the trust fund to the bank early
10	Choose between the high road and the low
11	Wish on every candle for your birthday
12	Deck the halls with boughs of holly
13	Ring in the new year with a kiss
14	Have a spooky Halloween
15	Enjoy the vacation with your loved ones
16	Be filled with joy and peace during this time
17	Relax on your holiday break

Table 2.2: Sentences 1–10 comprise the Harvard Sentences List 60. Sentences 11–17 are holiday-themed imperatives created for this corpus to match the Cantonese sentences thematically.

session and allowed participants time to continue getting used to the recording setup. As with the sentences, the storyboard narration also facilitated participants getting into the language mode of the session before the conversational interview. This is important because language mode is known to affect the degree of crosslinguistic influence in speech production (Simonet & 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

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

the personal details typically elicited in a sociolinguistic interview (Nagy, 2011). A major goal was for participants—who knew they were being recorded for linguistic inquiry—to feel at ease and freely discuss the questions. Questions were loosely laid out under general topic headings, with optional follow-up questions. While the English and Cantonese interviews had the same structure and general topic areas, the particular questions differed. While within each language, the possible sequence of questions was the same, each interview took its own course, guided by what the participant wanted to talk about. This means that that the total number of general topics covered ranged from three to six. The interview materials are included with the corpus download. As a result, the speech samples from each language are comparable, but the specific questions differ between interviews and across participants.

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

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

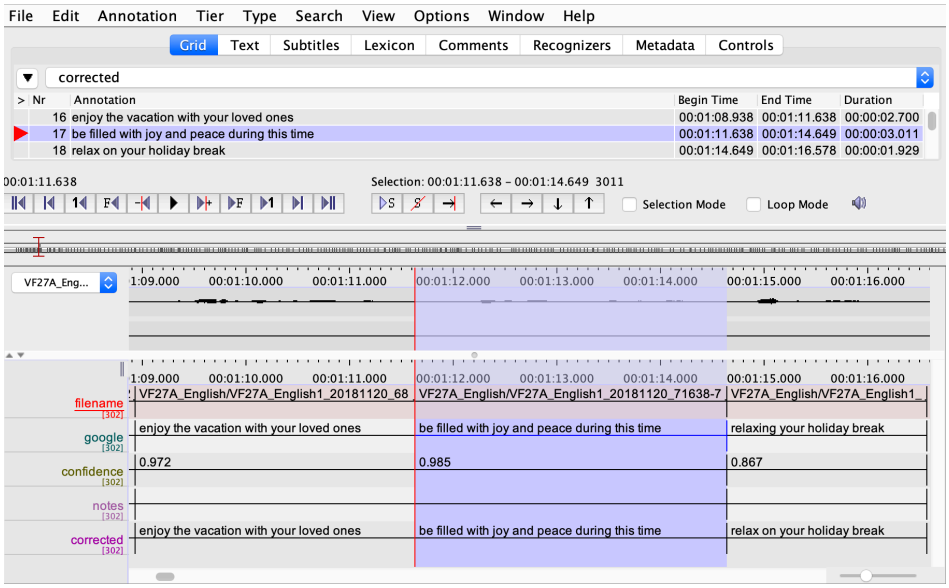


Figure 2.5: This screenshot from ELAN shows a sample of hand-corrected English from the sentence reading task for participant VF27A. The audio waveform is displayed in two channels, with one for the participant (top) and the other for the interviewer (bottom). The annotation tiers include (1) the short audio chunk's filename, (2) the raw speech-to-text transcript, (3) the speech-to-text confidence rating, (4) space for transcriber notes, if any, and (5) the corrected transcript. Note that "relaxing" was corrected to "relax on" in the rightmost section displayed.

2.3 Annotation

All recordings were processed according to the pipeline outlined in this section. As much as possible, automatic tools were leveraged to expedite manual correction.

2.3.1 Cloud speech-to-text

Google Cloud Speech-to-Text was used to produce an initial transcript of the interviews (Google, 2019). This was done using the Short Audio option, with the language variety set to Canadian English (en-CA) or Hong Kong Cantonese (yue-Hant-HK). To use this speech recognition product, the participant's speech was extracted from the participant's channel and segmented into short chunks, typically under 15 seconds in duration.[8] No attention was paid to constituents at this point; rather, breaks were placed at breaths and other pauses. Short chunks were necessary to use the speech recognition product with locally stored files, which was important for data privacy reasons. The short chunks would also prove useful for transcribers in the subsequent hand correction phase. With the audio files prepared in this way, speech recognition was completed using the Python client library for Google Cloud Speech-to-Text. The output included both a transcript and a confidence rating for each audio chunk. While the transcripts generated in this fashion were far from perfect, they served the function of expediting the hand-correction process.

2.3.2 Orthographic transcription hand-correction

The automatically generated transcripts were converted into multi-tiered ELAN transcription files (Sloetjes & 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. Fig-

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

ure 2.5 depicts an example of corrected English transcriptions in ELAN (Sloetjes & Wittenburg, 2008). Direct identifiers (e.g., names) were marked during this phase and silenced from the recordings prior to release. Transcriber guidelines were adapted from the multilingual Heritage Language Variation and Change 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 than English or Cantonese are transcribed in the language with, for example, "@m" appended to the end of each form for Mandarin (simplified characters), "@j" for Japanese, "@ml" for Malaysian, and "@i" for Indonesian.

Cantonese-specific conventions include:

- Where possible, transcription is in characters.
- Words without a standard character are transcribed in the Jyutping romanization system (e.g., *jyut6ping3*).
- Fully lexicalized syllable fusion is transcribed with the smallest number of characters representing what was produced by the talker.[9] For example, when fully fused, 乜嘢 (*mat1 ye5*, "what") is transcribed as 咩 (*me1*). In

[9]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.

some instances, an intermediate form is produced. For this lexical item, the intermediate form would be transcribed as 咩嘢 (*me1 ye5*). Cases of fully lexicalized syllable fusion tend to be relatively clear to identify.

- Non-lexicalized (or ambiguous) cases of syllable fusion are transcribed with the full number of characters present in the un-fused form but with brackets identifying which syllables are fused. For example, 朝頭早 ("morning") is pronounced *ziu1 tau4 zou2* in its full form but can be fused to *ziau14 zou2*—this fused form would be transcribed as 【朝頭】早.

- Filled pauses are transcribed with the character 啲 (*e6*), or using Jyutping if different (e.g., *m6*).

- Transcribers followed a shared set of guidelines for transcribing sentence-final particles. This includes the following common particles:

  - 呀 is the sentence-final particle used at the end of lists and for exclamations and questions.
  - 呢 was used for both *ne1* and *le1* in marking questions.
  - 囉 was used as a sentence-final particle for marking emphasis.
  - 嚟 was the final particle used to expresses something being done or completed.
  - 咗 was the particle used after verbs to mark past tense.
  - While not a *final* particle, 吖 was consistently used as a filler in the words 吖嗎 "obviously" and 吖嘛 "isn't it."

English-specific conventions include:

- Standard spelling is used.[10]

[10]While the goal was to use standard Canadian English spelling, given the range of transcriber backgrounds, the corpus includes some cases of American English spelling. This is relevant for a very small set of lexical items in the corpus (e.g., theater/theatre, favor/favour, etc.)

- Proper nouns are capitalized (e.g., “British Columbia”).
- Filled pauses are transcribed with “um”, “er”, “uh”, and other similar, non-elongated forms.
- Numbers are written out in word form (e.g., “one hundred”).

2.3.3 Forced alignment

Force-aligned transcripts were produced with the Montreal Forced Aligner (McAuliffe et al., 2017), using the hand-corrected orthographic transcripts. The output of the forced alignment process was phone-level annotations for each audio file. Phones in this kind of transcript reflect a relatively broad level of transcription and are often referred to as phonemes in the context of acoustic models. While the definition of phoneme from each camp certainly has overlap between acoustic modeling and phonology, the two should not be equated.

In Cantonese, forced alignment was completed with the Train-and-Align option, as there was no pre-trained model available for Cantonese. As Cantonese orthography does not separate words with spaces, words segmentation was done using the *jieba* Python library (Sun, 2020), along with a Cantonese word segmentation dictionary designed for use with *jieba*.[11] While using an automated tool such as this is likely an imperfect solution, it has the benefit of reproducibility and consistency. This is important, as it can be difficult to define wordhood in Cantonese (e.g., see Wong, 2006).

The Cantonese pronunciation dictionary was generated using the *PyCantonese* Python library (Lee, 2018). Pronunciations were identified by getting the Jyutping romanization from each character or when transcription was done in Jyutping, using that existing Jyutping transcription. Next, the Jyutping was separated into segments, and the tone number was appended to the syllable nucleus (i.e., vowel

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

or syllabic nasal). Research assistants supplemented the dictionary with alternative pronunciations for words that participated in syllable fusion. This approach bears some similarity to that of Tse (2019) but differs in that it also includes tonal information—which has been shown to improve forced alignment as long as there are not too many tone-nucleus combinations (Ćavar et al., 2016; Yuan et al., 2014).

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

The word and phone output of the forced alignment process were included in a Praat TextGrid for each audio recording, along with annotation tiers for the task (sentences, storyboard, and interview) and utterance (the short chunks). In both sessions, any material not in the main language of the session was not force-aligned and appears as "<unk>" in the word tier and "spn" in the phone tier, representing unknown words and spoken noise respectively.[12] The force-aligned transcripts were not manually corrected or checked. This means that any short chunk with code-switching or unintelligible speech will likely have poorer alignment because the model does not have a representation for that span of speech, either in the phone set or the pronunciation dictionary. As a result, it is advisable to use stringent exclusionary criteria or perform checks before analyzing data from the corpus.

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

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

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

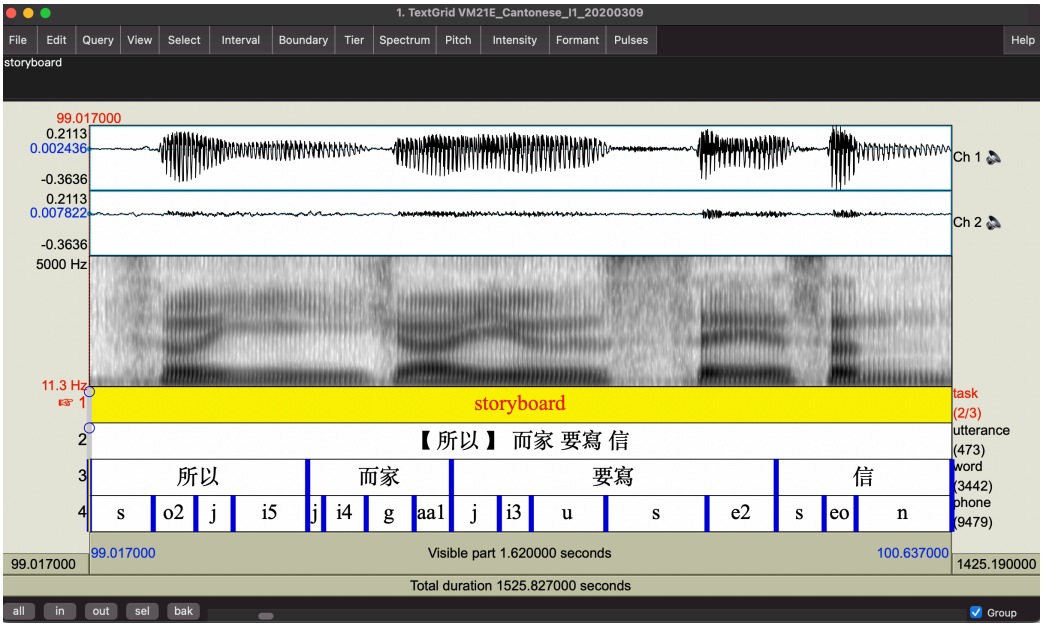


Figure 2.6: This screenshot from Praat shows what the final transcript looks like for a small portion of a Cantonese interview.

data as described in the previous section, specifically the hand-corrected utterance tier and the force-aligned phone tier. Additionally, this section only reports on participant speech, though the interviewer's speech is included in its own channel in the stereo audio files.

2.4.1 Cantonese interviews

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

In the Cantonese interview sessions, there were a total of 8,112 word types and 90,512 word tokens. The number of words varies substantially across participants,

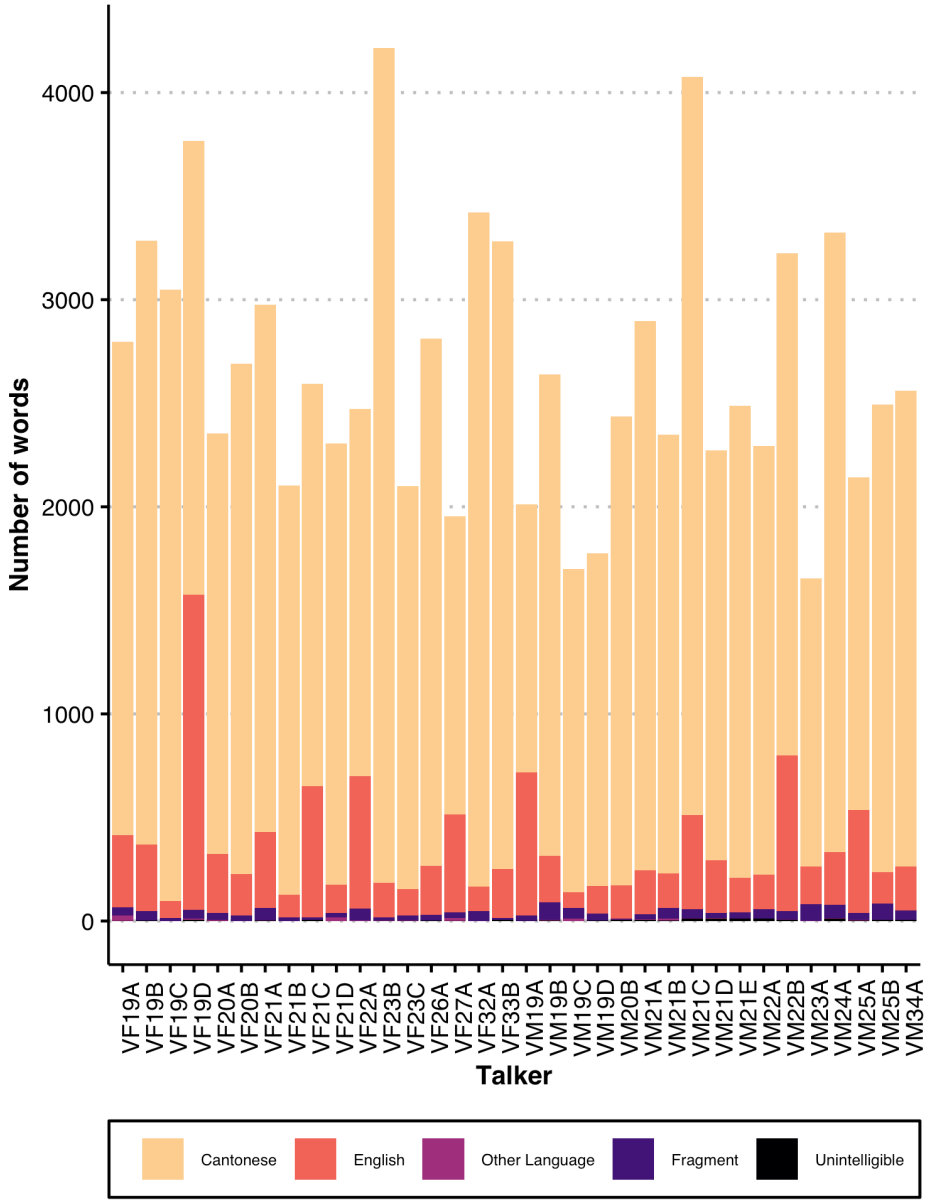


Figure 2.7: The total word count for each participant's Cantonese interview task is represented by bar height. Color indicates the kind of item counted.

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

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

2.4.2 English interviews

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

The English interviews include a total of 4,972 word types and 104,618 word tokens. As in the Cantonese interviews, the number of words varies substantially by participant, with a mean word type count of 609 (SD=119, minimum=434, maximum=904) and token count of 3,077 (SD=701, minimum=1,907, maximum=4,240). Figure 2.10 shows the split of these categories on a by-participant basis within the English interview sessions. Unlike the Cantonese interviews, there were relatively few switches to Cantonese, with 12 of the 34 participants producing fewer than 10 Cantonese words during the English sessions.

The distribution of log word frequency for both Cantonese and English words

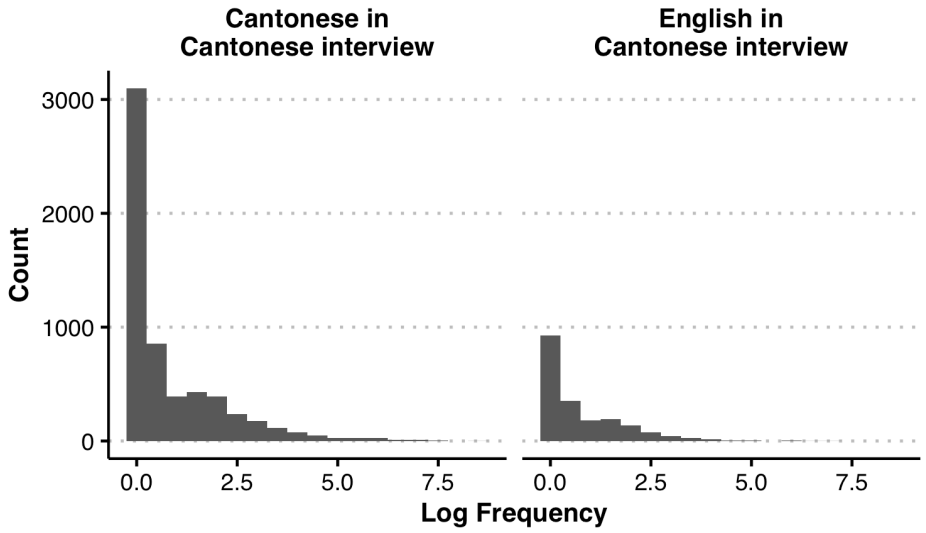


Figure 2.8: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

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

2.5 SpiCE corpus release

The SpiCE corpus was publicly released in May 2021 through the Scholars Portal Dataverse platform under a Creative Commons Attribution 4.0 International License.[13] In addition to the corpus itself, documentation is available online—the URLs are given in Section 2.2.

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

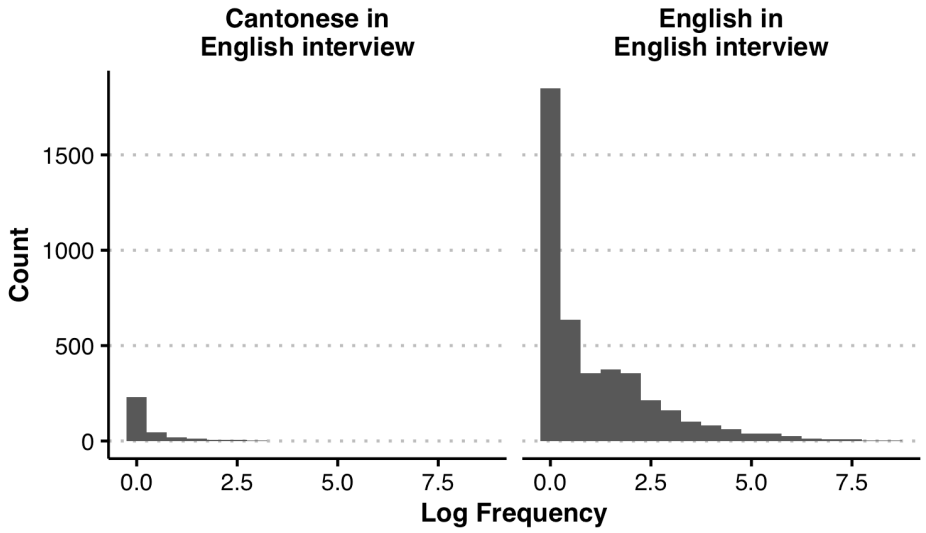


Figure 2.9: The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.

2.6 Discussion and conclusion

While various bilingual corpora exist, they lack in different ways *for the purpose of doing corpus phonetics*. The SpiCE corpus described here enables within-speaker phonetic comparisons across languages. While this would be possible with some of the bilingual speakers in resources like the Bangor corpora (Deuchar et al., 2014), the recording quality in such resources limits the scope of phonetic research. With the release of SpiCE and its high-quality recordings, scholars can ask and answer empirically and theoretically motivated research questions within the speech and language sciences using more sophisticated phonetic measurement techniques (e.g., spectral measures, in addition to temporal measures). This presents substantial potential for increasing our understanding of bilingual spoken language from both phonetic and psycholinguistic perspectives. While the recording quality of this corpus offers these particular advantages, SpiCE is also suitable for any other standard corpus-based inquiry with conversational speech, whether linguis-

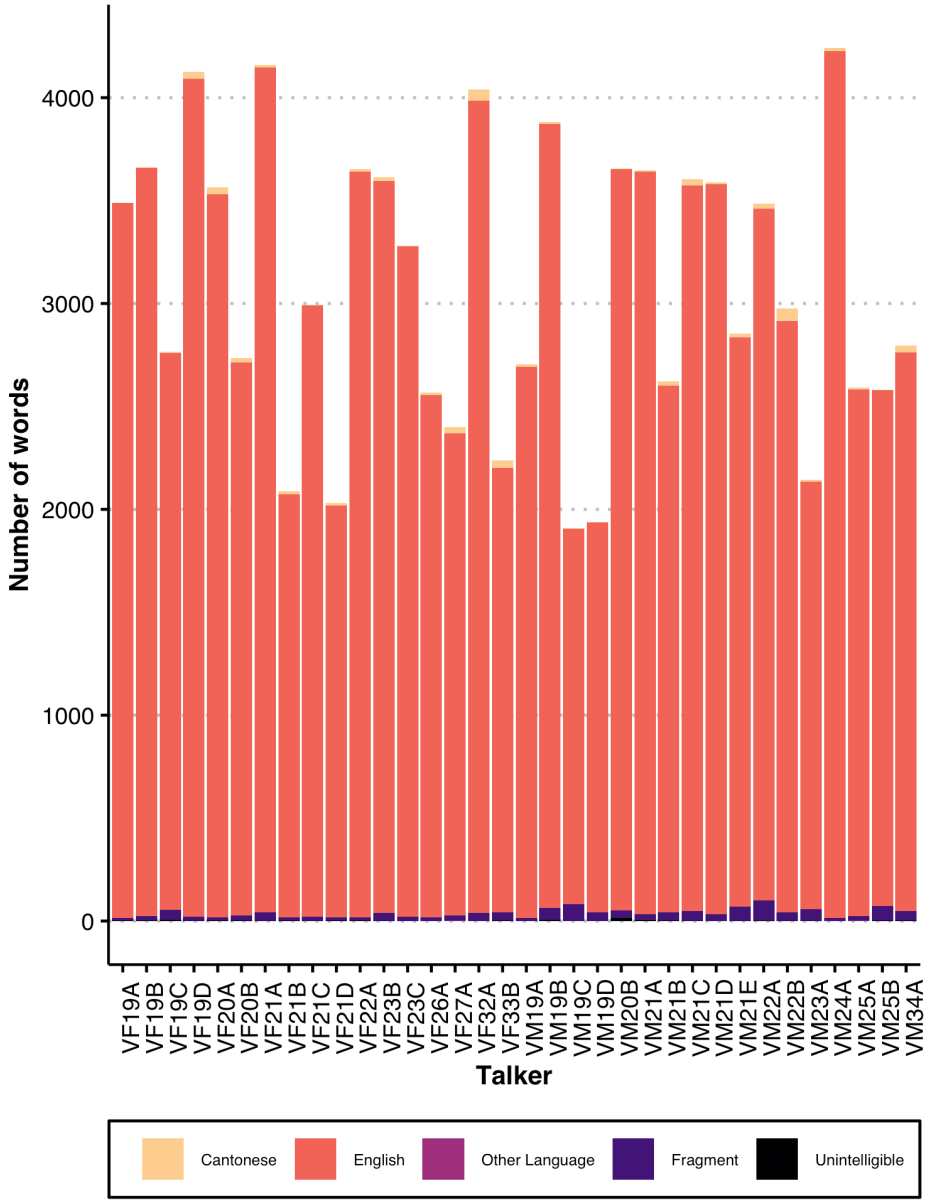


Figure 2.10: The total word count for each participant's English interview task is represented by bar height. Color indicates the kind of item counted.

tic or paralinguistic in nature. The opportunities made available with SpiCE are especially important given the typological difference between the languages under consideration and the fact that Cantonese is an understudied language.

Chapter 3

The Structure of Acoustic Voice Variation in Bilingual Speech

3.1 Introduction

How does voice vary across a bilingual's two languages? This question sets the stage for this chapter. But first, it is important to consider what voice is and does. Voices provide considerable information about the person talking, ranging from their current physical and emotional state to talker indexical features that help listeners identify who they are (Podesva & Callier, 2015). In this context, voices can be described as auditory faces—they are uniquely individual yet share basic characteristics with the broader population (Belin et al., 2004). Where faces share an overall shape and composition of features (e.g., eyes, nose, etc.), voices share the acoustic consequences of similar vocal anatomy. Yet, at the same time, when you see a familiar face or hear the voice of a person you know, you can often immediately recognize who it is, as well as ascertain some information about their present state. In this way, both voices and faces signal identity along with aspects of the individual's physical and emotional state.

Along with all of this information, voices simultaneously convey a communicative message. Podesva & Callier discuss voice as a "bridge between body

and language" (2015, p. 175), though "body" is perhaps slightly misleading, as it is used to evoke identity in addition to the physical body. Disentangling such information and understanding the structure of voices is no small feat—it means understanding how listeners leverage variable vocal dimensions to process talker-indexical, affective, social, and linguistic information. This feat also presents a processing challenge—one that arises from the sheer variability within and across voices.

Though voices share some attributes—such as in how spectral shape, noise, and formants pattern—they also vary in unique ways (Lee et al., 2019). From the perspective of voice perception, the balance between shared and idiosyncratic characteristics makes some amount of sense. The shared dimensions allow listeners to recognize the sound they hear as a voice, and they also help listeners perceive, classify, and understand new voices. Idiosyncrasies, on the other hand, enable identification and discrimination between different voices. While this makes sense conceptually, understanding the structure of voice variation in speech production and its complement in listeners' ability to process that information remains an active area of research. This topic was touched upon in Chapter 1 and will be revisited in Chapter 5.

The focus of this chapter is acoustic voice variability. The emphasis on describing variation echoes one of the enduring puzzles in phonetics—the "lack of invariance" problem (Liberman et al., 1967). Given the ubiquity of variation, the lack of invariance problem asks how perceivers can efficiently extract relevant and important information from the communicative signal—whether that information relates to the talker, message, or some other dimension (Kleinschmidt et al., 2018). This chapter focuses on variation in speech production, particularly as it relates to talker identity and the analogy of voices as auditory faces. By foregrounding the speech signal itself, this chapter effectively asks what is available in the speech signal for listeners to use in voice identification. While this dissertation does not include perception research—and as a result, will not be able to comment on what listeners use—it generates hypotheses about perception that are grounded in the in-

formation comprising the speech signal. These hypotheses are outlined in Section 5.4.

While variation is indeed wide-ranging, it remains far from random. Some prevalent accounts of how individuals understand and process variation emphasize its structure (e.g., Chodroff & Wilson, 2017; Lee et al., 2019). While this chapter looks at the structure of voices and the following chapter examines sound category structure, both attempt to elucidate structure in the speech signal that may be beneficial for listeners in understanding and processing new talkers.

The introduction to this chapter proceeds as follows. Literature about voice and voice quality is briefly summarized in Section 3.1.1 and is followed by a summary of work on the *structure* of voice quality variation in Section 3.1.2. Then, Section 3.1.3 covers some relevant background in the domain of voice perception, echoing what was discussed in Section 1.2 back in the introduction to this dissertation. Section 3.1.4 provides a lengthy review of the literature comparing specific aspects of voice and voice quality for Cantonese and English, as well as relevant findings from other pairs of languages. Lastly, Section 3.1.5 outlines the specific research questions at hand and provides a roadmap for the rest of the chapter.

3.1.1 Voice and voice quality

While the introduction to the chapter thus far sets the stage for voice and variation broadly, this section approaches voice from the bottom up. In the narrowest sense, voice has been defined by the behavior of the vocal folds—the source—though Garellek (2019) acknowledges that the acoustic and perceptual consequences of the source cannot be entirely separated from supralaryngeal factors (i.e., the filter).

Voices are multidimensional and can vary on a vast array of different dimensions. While early work focused on how different articulatory settings correspond to voice quality (Laver, 1980; Pittam, 1987), more recent accounts advocate for a psychoacoustically informed model (Kreiman et al., 2014). The rationale for this shift arises from the observation that there is not a one-to-one mapping from articulation to perception via acoustics, as discussed in detail by Garellek (2019), in his

recent review chapter on the phonetics of voice. In any case, both approaches reflect the wide range of dimensions. The behavior of fundamental frequency (F0)—including its absence—captures Garellek's (2019) articulatory categories of vocal fold approximation, voicing, and rate of vibration. Voice quality is captured acoustically by spectral shape and noise parameters (to be reviewed in detail in Section 3.2.2) and articulatorily by constriction degree, irregularity, and tension.

While Garellek's (2019) chapter focuses on voice and voice quality in the domain of linguistic contrast and variation, the psychoacoustic model of voice quality he references (Kreiman et al., 2014) accounts for voice more broadly—it also captures filter behavior via formants.

As touched on in the first paragraph of this chapter, there is also a large body of work highlighting the many different things that voice indexes—affect, stance, psychological states, behavior, physical characteristics, and identity (Podesva & Callier, 2015). Identity here includes the idea of "linguistic identity," which stems from early work summarizing how "phonetic settings" vary across languages and dialects (see Podesva & Callier, 2015; Pittam, 1987; Mennen et al., 2010). So while the acoustic and articulatory dimensions noted above vary for linguistic reasons, the same set can also vary for non-linguistic reasons. Acoustic dimensions thus index a multitude of different things simultaneously. This observation is especially relevant in light of Kreiman and colleagues' argument that their perceptually validated set of dimensions are more than the sum of their parts (Kreiman et al., 2014, 2021). Voice quality and what it indexes thus form a many-to-many relationship, where measures covary and conspire together to form a multidimensional percept of voice.

3.1.2 Structure in voice quality variation

There is a large body of literature focused on understanding differences in variability across populations for a small set of these acoustic measurements. Such studies typically compare summary statistics for F0 and a handful of spectral measures. This body of work is summarized in Section 3.1.4 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 spearheaded by Lee and colleagues (Lee et al., 2019; Lee & Kreiman, 2019, 2020). In this set of studies examining acoustic voice variation in different languages and speech styles, Lee and colleagues leverage the psychoacoustic model of voice quality (Kreiman et al., 2014) and adapt methods from the domain of face variability and perception (Burton et al., 2016). Their driving question is one of understanding the structure of acoustic information in the speech signal. As noted in Section 3.1.1, acoustic dimensions do not behave in isolation; rather, they pattern together in complex ways. Part of what is novel about Lee et al.'s approach is that it gets at how these variables behave in relation to one another—that is, how covariation is 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, particularly when coarse summary statistics (i.e., means, ranges, and standard deviations) do not indicate cross-talker differences.

To drill down into the structure of voice variability, Lee et al. (2019) use a series of principal components analyses (PCAs) to investigate how acoustic measurements pattern with one another. PCA is a dimensionality reduction technique—that is, a large set of variables are distilled into components that reflect covarying bundles of variables. The methods used in this study will be described in greater detail in Section 3.2.5. In their original paper, Lee et al. (2019) examined 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). This database includes English recordings and force-aligned transcripts of 201 talkers completing 12 different tasks ranging from scripted to unscripted. Talkers were all UCLA students, varying in their language background (i.e., whether or not English is their L1) and sex (here, male or female). Crucially for the comparison with their later work, Lee et al. (2019) focused on relatively small samples of sentence reading from within this corpus.

The takeaway from this work is that different voices share structure with each

other and the group as a whole. Lee et al. (2019) reach this conclusion by analyzing the configurations and variance accounted for by components across talkers. Shared structure is characterized by the same set of variables covarying and together accounting for comparable amounts of the overall variation. The most commonly shared component in Lee et al. (2019) consisted of higher spectral slope and noise variables and accounted for approximately 20% of the overall variance. These variables are associated with vocal breathiness or brightness. The next most commonly shared component comprised higher formant variables and accounted for approximately 10% of the overall variance. These variables are typically associated with vocal tract size and speaker identity. Despite this shared structure, however, Lee et al. (2019) argue that the rest of voice structure variation is largely idiosyncratic.

Lee & Kreiman (2019) replicate this work with short samples of spontaneous speech from the same database. The results were similar, with the exception that F0 emerged as a shared relevant dimension. This result arguably reflects the difference between reading and spontaneous spoken English, with reading tending to be more monotonous and spontaneous speech exhibiting more affective qualities. In spontaneous speech, F0 varies along with the higher source spectral shape and noise parameters. In read English speech, F0 likely varies quite a bit less. Lee & Kreiman (2020) replicate this work again with sentence reading in Seoul Korean, again finding some differences that are small and readily explained by typological differences from English. Unlike English, F0 and variability in the lower formants emerged as relevant dimensions in read Korean speech. The authors argue that this reflects phrasal intonation patterns that occur in Korean reading.

Conceptualizing what these dimensions mean and how to think about acoustic voice variability in this way is challenging, as many of the acoustic dimensions considered do not map neatly onto a single percept. F0 is a straightforward example, given its clear relationship to pitch. Many other spectral measures, both harmonic and noise-based, are much more challenging to interpret without considering multiple measures simultaneously. Garellek (2019) gives an example of this

in how spectral shape needs to be interpreted in the context of spectral noise. For example, lower spectral shape indicates a more creaky voice quality, while higher spectral shape indicates a more breathy voice. This correspondence, however, falls on a spectrum. Without knowing the value of a variable like the harmonics-to-noise ratio (HNR), it is not possible to objectively say where on the spectrum a particular item is located with spectral shape alone. HNR thus provides the necessary context for interpreting spectral shape.

The domain of faces thus provides a useful analogy for thinking about what shared structure looks like compared to idiosyncratic structure. Burton et al. (2016) found that all faces share dimensions of variability related to things like lighting and viewing angle (i.e., looking up, down, or to the side). Understanding how a face changes according to light or angle is useful structural knowledge that can be transferred any new face. The dimension is shared *because* it applies to all faces. Idiosyncratic variation in face structure arose from things like facial hairstyle, makeup, and expressions. While these variables may be shared by a subset of faces, understanding how something like the application of makeup varies is not applicable to all faces.

Returning to voice, Lee et al. (2019) argue that the structure of voice spaces supports a prototype model of voice perception (Lavner et al., 2001; Latinus & Belin, 2011) in which novel individual voices are perceived in the context of one or more prototypes housed in listeners' memory. Lavner et al. define a prototype as a pattern comprising "an ensemble of acoustic features, related to the language, the accent, the phonemes and allophones, and to the voice production system...[reflecting] the average of speakers' features or a very common voice" (2001, p. 64). The authors then argue that new voices are perceived in the context of this prototype, such that "only those features that significantly deviate from the prototype are stored (memorized) for the long term, and identification of familiar voices is based on searching and locating the voice, using only those features deviating from the prototype" (Lavner et al., 2001, p. 64).

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. This familiarity could easily be characterized in terms of the extent and manner that a voice deviates from a prototype.

3.1.3 Voice perception

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

Very little of this work identifies what parts of the signal listeners use, and as such, claims about the relative importance of linguistic or talker-indexical information should be tempered. However, there are exceptions to this. For example, Perrachione et al. (2019) collected perceptual voice (dis)similarity ratings for Mandarin and English voices by Mandarin and English native listeners and reported on the relationship between several acoustic measurements and rating data. Perrachione et al. (2019) found that when the talker was the same, regardless of the manipulations used in the study (language and time-reversal), all listeners rated stimuli pairs as highly similar. This result highlights that listeners are sensitive to low-level acoustic information present in voices, regardless of whether they know the language or understand the stimuli. Additionally, Perrachione et al. (2019) found that some acoustic measurements predict similarity ratings, while others do

not. F0 was the most prominent measure, which is unsurprising given its salience, how much the voice variability literature has focused on it, and the extent to which researchers treat it as an important variable (e.g., Keating & Kuo, 2012). Other measures predicting similarity were HNR and formant dispersion, which are associated with non-modal phonation and vocal tract size, respectively. That listeners appear to use these measures is of direct relevance to the study presented in this chapter, as it signals their importance in perception and processing—this represents a point that will be returned to in this chapter's discussion (Section 3.3).

3.1.4 Bilingual voices

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

Describing and analyzing acoustic voice variation in bilingual speech has motivation in both perception and production. As apparent from the LFE 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 above chance even for listeners who lack familiarity with the language (e.g., Orena et al., 2019). Knowledge of the language used in the experiment lends the greatest advantage; and, knowledge of a related language also provides a benefit (Zarate et al., 2015). Presumably, listeners in the latter situation can extract some degree of linguistic information given overlap in the sound structure of the two languages. In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic acoustic/auditory information in

the voice (Perrachione et al., 2019). Understanding the structure of that variability brings us one step closer to understanding what listeners are using from the signal to process speech, as it limits the hypothesis space.

On the production side of things, bilingual speech presents an ideal test case for the argument that voices function like auditory faces. If the structure of variability from each of a bilingual's languages is well matched—comparatively speaking—then voices can be straightforwardly thought of as auditory faces. While "well-matched" is a vague term, its use reiterates that the meaningful threshold for comparison is not some absolute value but rather how structure is shared within and across languages for between-talker comparisons. While this characterization may seem unsatisfying, it is worth noting that face variability is not identical across languages. A small body of work illustrates that language identification is possible using only lip movements by both humans (Soto-Faraco et al., 2007) and machines (Afouras et al., 2020), indicating that there are indeed language-specific patterns in facial postures for face perception. An example of this might be the case of languages—like Cantonese and English—with different distributions of lip rounding in their segmental inventories (cf. Tables 3.1 and 3.2).

Additionally, examining the structure of the same talker's voice in each language lends additional validation to the arguments made by Lee & Kreiman (2020) for the differences between English and Seoul Korean sentence reading. In comparing these studies, Lee and colleagues argue that both language and biological factors contribute to the structure of voice variation. Bilingual speech, again, presents an ideal test ground for disentangling biological and linguistic factors from one another. While common in the literature, the language versus biology dichotomy is somewhat misleading. Voices ultimately have biological constraints due to physical and physiological limitations (e.g., vocal tract length, vocal fold mass) or pathologies. Yet, at the same time, individuals exert remarkable and wide-ranging control over their voice space and are highly capable of manipulating factors that are not linguistically important but which signal social and contextual information. This applies across all aspects of an individual's linguistic repertoire

(Bullock & Toribio, 2009; Wei, 2018). Thus in the case of bilinguals, the only aspect we can be truly confident in being held constant across languages is the biological part. The same "hardware" can be used for drastically different ends.

English and Cantonese

This chapter examines how voice varies across Cantonese-English bilinguals' two languages. Some differences are expected, despite the characterization of voices as auditory faces. While all languages have consonants and vowels, they differ in distribution, articulation, and acoustics (e.g., Munson et al., 2010). An overview of the inventories of Cantonese and English is provided in Tables 3.1 and 3.2. Additionally, Suprasegmental and prosodic properties also vary. Languages differ in terms of whether a suprasegmental dimension is made use of in distinguishing linguistic contrasts. For example, does a language encode lexical tone contrastively? Another way languages vary in this respect is in how they carve up the suprasegmental linguistic space. For example, how many lexical tones are there? What shapes of tone are present? The question of tone and how it impacts voice variability is relevant in the present case, where the languages considered are Cantonese (a language with lexical tone) and English (a language without lexical tone). Cantonese has six lexical tones, which are often referred to by numbers one through six: (1) high level, (2) high rising, (3) mid level, (4) low falling, (5) low rising, and (6) low level. It is important to highlight these differences, as both segmental and suprasegmental differences have cascading effects on voice quality.

The following paragraphs detail voice quality comparisons that have been made between English and Cantonese in the literature thus far. As there is an additional body of work comparing English and Mandarin Chinese—typologically similar to Cantonese—comparisons between English and Mandarin are also summarized in the next section. While the most relevant comparisons for this chapter are those made within bilinguals, some of the relevant literature compares separate populations. What this body of literature has in common—whether within- or between-talker—is that it paints with relatively broad strokes—crosslinguistic comparisons

Table 3.1: The Cantonese segmental inventory as described by Matthews et al. (2013). Note that Cantonese vowels combine into many different diphthongs.

Consonants	**Nasal**	**Stop/Affricate**	**Fricative**	**Approximant**
Bilabial	m	p / p^h		
Labiodental			f	
Dental	n	t / t^h	s	l
Alveolar		ts / ts^h		
Velar	ŋ	k / k^h		
Labiovelar		k^w / k^{wh}		
Glottal			h	

Vowels	**Front**	**Central**	**Back**
High	i / y		u
Mid	ɛ / œ		ɔ
Low		ɐ / aː	

Table 3.2: The English segmental inventory as described by Wilson & Mihalicek (2011), with [ʔ ɾ ˷w] excluded. Note that some English vowels combine into diphthongs.

Consonants	**Nasal**	**Stop/Affricate**	**Fricative**	**Approximant**
Labial	m	p / b	f / v	
Dental			θ / ð	
Alveolar	n	t / d	s / z	l
Palatal		tʃ / dʒ	ʃ / ʒ	ɹ
Velar	ŋ	k / g		j
Glottal			h	w

Vowels	**Front**	**Central**	**Back**
High	i / ɪ		u / ʊ
Mid	e / ɛ	ə / ʌ	ɔ
Low	æ		ɑ

are often made with summary statistics for a small set of spectral measurements. With such methods, results have been decidedly mixed.

In a small study of Cantonese-English bilingual (n=9), Russian-English bilingual (n=9), and English monolingual (n=10) young women, Altenberg & 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 & Kuo, 2012), Altenberg & Ferrand (2006) focused on whether F0 shifts when an individual switches languages and whether different languages have different baselines. Ultimately, Russian-English bilinguals exhibited differences in mean F0 across their two languages, 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 & Ferrand (2006) ultimately paint a coarse picture of bilingual F0 production with a small sample size, they highlight an important point of departure—bilinguals can differ in F0 across languages.

In a larger study of Cantonese-English bilinguals reading passages (n=40), Ng et al. (2012) examined a variety of different voice measures with both male and female talkers. Results were based on Long-Term Average Spectral (LTAS) measures. Female talkers exhibited lower F0 in Cantonese than English, but males did not. In the same study, all participants had greater mean spectral energy values (mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese (Ng et al., 2012). Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English. The LTAS measure of the first spectral peak did not differ across languages, suggesting that vocal stiffness remained consistent in the bilinguals' two languages.

Ng et al. (2010) examined F0 in spontaneous speech from 86 Cantonese-English bilingual children and found it to be lower in Cantonese compared to English. This corroborates Ng et al. (2012), and goes against the nonsignificant difference in Altenberg & Ferrand (2006). This mixed bag of results could ultimately be attributed to differences in sample sizes, the quantity of speech analyzed, or the language

backgrounds of the bilinguals studied. While the picture regarding voice quality measures appears clearer and more consistent, those conclusions arise from a single study. In any case, these three studies offer reason to expect that Cantonese and English might differ in measures associated with pitch and phonation type.

English and other languages

The authors of these studies speculate that Cantonese's status as a tone language may account for some of these differences compared to English. It is important to emphasize that this explanation is pure speculation. In this light, it is also relevant to consider the larger body of research comparing voice quality for Mandarin and English. Lee & Sidtis (2017) compare F0, speech rate, and intensity in a small group of Mandarin-English bilinguals (n=11) across three different tasks. They report a higher mean F0 for Mandarin reading compared to English, but no differences in the other tasks (picture description and monologue). Additionally, there were no differences in F0 variability across languages or tasks. Lastly, while there were no differences in intensity, the bilinguals spoke faster in Mandarin. Lee & 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 & Kuo, 2012). This finding may be strongly associated with the type of bilinguals studied. Xue et al. (2002) found that Mandarin-English bilinguals aged 22-35 produced lower F0 in Mandarin than English. This group differed from the participants in Lee & Sidtis (2017), in that they are described as non-native English speakers. Producing higher F0 in a non-native language arguably reflects factors like stress or confidence (Järvinen et al., 2013; Lee & 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 a simple

appeal to the presence or absence of lexical tone does not present a substantive argument. While an account that invokes the distribution of lexical tones would be somewhat more compelling, it would also need to account for all of the other ways that F0 varies in language production (e.g., prodosy). Alternatively, talkers may be expressing different social and cultural identities in each of their languages (Loveday, 1981; Voigt et al., 2016). Regardless of whether language, experiential, or social factors drive differences across languages, this body of work highlights the importance of comparing within the same task (i.e., isolated word production, reading, spontaneous speech, etc.).

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

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

Russian was faster. This result goes against the findings for the bilinguals studied in Altenberg & 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 & Mennen (2017)—they demonstrate differences in F0 range and level across languages for female Welsh-English bilinguals in a reading task, for whom Welsh had a higher and wider F0 range. This result did not hold for males from the same population, who varied more in their F0 level and range. The authors argue that the crosslinguistic difference is likely to be sociocultural in this case, as different patterns were observed for male and female speakers on a within-speaker basis. Ordin & Mennen (2017) argue that if a difference in F0 stemmed purely from language differences, that males and females would both show the pattern. Because this is not the case, they argue that the result is unlikely to be due to anatomical or purely linguistic reasons. While this argument does not necessarily disentangle social from linguistic, it emphasizes the need to consider social dimensions.

Considering these studies together, a few key observations are especially relevant to the present chapter. While studying bilingual talkers provides a clear path to disambiguating the role of anatomical differences in voices, it does not necessarily facilitate disentangling linguistic and sociocultural factors from one another. Most likely, both contribute simultaneously to the differences in voice patterns across languages—and may or may not be disentangle-able. 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 & Sidtis (2017). On the other hand, Ordin & Mennen (2017) show social rather than linguistic stratification via gendered patterns in Welsh-English bilinguals. While these studies examine different populations, they nonetheless highlight different sources of variation.

This body of work mostly focuses on linguistic and social differences. While some of it dives into individual differences, between-talker variability should per-

haps be given more of a spotlight. In work with speech rate, Bradlow et al. (2017) found that some talkers are fast and others are slow and that some languages are fast while others are slower. Crucially, these relationships held across talkers in various languages. That is, if someone was a fast talker in their dominant language, they were also a fast talker in their non-dominant language, and likewise for slow talkers. In this sense, both talker-indexical and linguistic (or sociocultural) factors contribute to speech rate behavior. It is not a particularly big leap to suggest that other speech signal variables might pattern in the same way. Adding to this picture of variability across individuals, it is important to remember that bilinguals are sophisticated social actors and are fully capable of tailoring their speech behavior to a wide variety of contexts (Bullock & Toribio, 2009).

3.1.5 The present study

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 across-talker voice similarity ratings. Yet, at the same time, there is so much more to voice than pitch, particularly if the characterization of voices as auditory faces holds up.

This chapter brings together work describing crosslinguistic voice differences and the structure of acoustic voice variation to provide a more comprehensive picture of how voices vary across languages. Using the corpus introduced in Chapter 2, this chapter describes the behavior of various spectral properties (e.g. Ng et al., 2012), and also examines how acoustic variation is structured, following the work of 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 size both within and across talkers and languages. It also extends their methods by introducing a mechanism to assess structural similarity within and between individuals and languages.

3.2 Methods and results

The methods and results of each part of this analysis are reported in tandem, as the various parts build on one another. This section proceeds as follows. Section 3.2.1 describes how the SpiCE corpus was used in this chapter. Section 3.2.2 justifies and defines the acoustic measurements upon which the rest of the chapter rests, and Section 3.2.3 provides an account of how measurement error was dealt with. The first of the analyses, described in Section 3.2.4, compares overall distributions for each of the acoustic measurements across languages, largely following the tradition of the research described in Section 3.1.4. The second analysis reports on the structure of PCAs within and across languages in Section 3.2.5. The third part of the analysis—Section 3.2.6—builds on the PCAs, and introduces canonical redundancy as a method to compare PCAs. Lastly, the analysis of sample size in Section 3.2.7 serves to validate decisions made earlier in the chapter and offer guidance for future studies in this domain.

3.2.1 Data

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

While prior work by Lee and colleagues (e.g., Lee et al., 2019) uses relatively short chunks of speech, the present analysis is focused on longer stretches of spontaneous speech from conversational interviews. While it would have been possible to include the sentence reading and storyboard task recordings from each participant, there are practical reasons for excluding them from the analysis. The sentence sets were overall quite short and thus unlikely to be sufficiently representative on their own. Additionally, as many of the SpiCE talkers were not confident in their

Cantonese reading, there was a wide range of familiarity with the materials represented. Some talkers knew all of the sentences, and others struggled with some of them. This variability renders the sentences less comparable to their English counterparts in the SpiCE corpus. There are also imbalances in the storyboard task. As talkers narrated the same story in both languages, they were often more confident the second time around. Excluding both of these tasks is motivated by prior work that highlights how confidence (Järvinen et al., 2013) and speaking style (Lee & Sidtis, 2017) impact voice quality.

As discussed in the previous chapter, the recordings are high-quality, with a 44.1 kHz sampling rate, 16-bit resolution, and minimal background noise. Recall that both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. For the analysis in this chapter, the participant channel was extracted from the stereo recordings, including any code-switches they made during the interview. While it would be possible to exclude items not produced in the primary language of the interview, this was not done. The driving reason for keeping code-switches in the analysis is that such code-switches are representative of the particular talker's language behavior. Further, just because someone switches languages does not mean that they fully and immediately switch language modes (e.g., Fricke et al., 2016b). For example, individual words may be borrowed and pronounced with the phonology of the interview's primary language (cf. 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 & Weenink, 2021), implemented with the Parselmouth Python package (Jadoul et al., 2018). The pitch range settings used with *Point Process (periodic, cc)* were 100–500 Hz for female talkers and 75–300 for male talkers. These settings reflect a balance between known differences between male and female pitch (Simpson, 2009) and the wide range of F0 variability in spontaneous speech while guarding against the pitch estimation issues of doubling and halving. While speech from the interviewer can occasion-

ally be heard in the participant channel, it is quiet enough to have been ignored by the Praat algorithms and likely did not influence the results.[1] This method of identifying voiced portions of the speech signal captures vowels, approximants, and some voiced obstruents. As a result, this process differs slightly from the methods described in Lee et al. (2019), the paper on which the methods of this chapter were modeled. Lee et al. (2019) examined only vowels and approximants.

3.2.2 Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by Lee et al. (2019), except formant dispersion, which was excluded given its very strong correlation with the measured value of F4 in this chapter (following the exclusionary criteria in Section 3.2.3: Pearson's $r = 0.94$, df $= 3071734$, $p < 0.001$). The choice of measurements in Lee et al. (2019) is based on Kreiman et al.'s (2014) psychoacoustic voice quality model, as well as the availability of algorithms in the software used to extract measurements. Measurements were made every 5 ms during voiced segments in VoiceSauce (Shue et al., 2011).[2] 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). It is one of the more widely studied variables on this list, as evidenced by the literature cited in Section 3.1.4.

F1, F2, and F3 The first three formant frequencies—also measured in Hertz—

[1] Note, however, that the volume of the interviewer's speech was not examined in either channel of the stereo recordings.

[2] The version of VoiceSauce was $\geq$ v1.28, which is relevant for F0 tracking with the STRAIGHT algorithm. I do not currently have access to the computer in the Speech-in-Context lab to verify the exact version used due to COVID-19 access restrictions.

are typically discussed for linguistic contrasts, particularly with vowels and sonorant consonants. Recall the differences in Cantonese and English inventories described in Tables 3.1 and 3.2. All formants were estimated using the Snack Sound Toolkit method Sjölander (2004), with the default settings of 0.96 pre-emphasis, 25 ms window length, and 1 ms frameshift.

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

H1*–H2* The corrected amplitude difference between the first two harmonics is one of four primary measures used to characterize source spectral shape—also called spectral tilt—in the psychoacoustic model of voice quality (Kreiman et al., 2014). It is typically associated with phonation type, such that lower values fall on the creakier end of the spectrum and higher values on the breathier end of the spectrum (Garellek, 2019). Interpretation is less than straightforward, however, as determining where on the spectrum from creaky to breathy a particular observation falls depends on a measure of spectral noise (e.g., CPP below; Garellek, 2019). Additionally, H1*–H2* can be confounded by nasality (Munson & Babel, 2019). The asterisks here—and in the following spectral shape measures—indicate that the value has been corrected (Iseli et al., 2007), to account for the amplifying impact of nearby formants on the amplitudes of harmonics. This allows for different vowels and other voiced segments to be compared with one another. This amplitude difference is measured in dB. Note that this measure—along with the following three spectral shape measures—depends on an accurate F0 measurement.

H2*–H4* The corrected amplitude difference between the second and fourth harmonics is the second of four measures capturing spectral shape. Like H1*–

H2*, 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 2,000 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 in a similar manner to H1*–H2*. Like the other spectral shape measures, it is in dB.

H2kHz*–H5kHz The amplitude difference between the harmonics closest to 2,000 Hz (corrected) and 5,000 Hz (uncorrected) is a measure of harmonic spectral shape that does not depend on F0. The amplitude of the harmonic nearest 5,000 Hz is not corrected by VoiceSauce, given inaccuracies in the correction algorithm at higher frequencies. It captures the highest frequency band of the four shape measures, reflects phonation type as H1*–H2* does, and is measured in dB.

CPP Cepstral Peak Prominence measures the degree of harmonic regularity in voicing, and as such, it is associated with non-modal phonation types. VoiceSauce computes CPP according to the algorithm in Hillenbrand et al. (1994). Specifically, CPP measures the difference between the amplitude of the peak in a cepstrum and the value at the same quefrency on the regression line for that cepstrum.[3] 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 a perceptual correlate of volume or loudness. It 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

[3]For details and definitions of terms like *cepstrum* and *quefrency*, please refer to Hillenbrand et al. (1994).

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.[4]

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 H5kHz. Observations were also excluded if Energy was more than three standard deviations above the grand 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. As nonzero values for CPP reflect a range of modal and nonmodal phonation, a zero for CPP likely reflects either a lack of voicing or an erroneous F0 measurement. Lastly, only the spectral measure for H5kHz was used in filtering (uncorrected, and not the difference used in the analysis), as erroneous values tended to co-occur on the same observation. The distribution of H5kHz did not span zero, except for a spike of erroneous values equal to zero. This operationalization minimizes the removal of correctly measured zero values, which occurred with all of the other spectral shape parameters, whether corrected

[4]Note that this repository is currently private.

or uncorrected. In aggregate, these filtering criteria led to the removal of 37% of the original set of observations.

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). The rationale for this difference is that all variables were scaled before inclusion in the PCAs described in the next section, and as a result, there should not be any undue effect on the outcome as the transformation from standard deviation to coefficient of variation is a scaling transformation. 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 may have an impact on the PCA based analyses in Sections 3.2.5 and 3.2.6—this is expanded upon in the next paragraph.

To control for the impact of sample size in that part of the analysis, the number of samples for each talker was capped to include only the first 20,124 samples for each interview. This value was selected as it represents the interview with the fewest observations. Put simply, differences in sample size reflect the variability in how much different individuals in the corpus talked. Those who produced longer

passages of speech ultimately had more observations of voiced speech. Passage length was expected to impact the analysis, given how much affect and style can vary within a single conversation. Over time, individuals cover more of their range of variation, and as such, a regression to the mean is expected over time. That is, PCAs based on shorter stretches of speech would be subject to greater variability, while those based on longer stretches would converge on a structure. To level the playing field in this first analysis, the sample size was controlled. At the end of this chapter, in Section 3.2.7, a follow-up analysis validates this assumption. To preview those results, 20,000 samples appear sufficient for capturing the range of variability in acoustic voice variation.

Following this last winnowing step, there were 1,368,432 total observations (34 talkers $\times$ 2 interviews $\times$ 20,124 observations per interview). While the winnowing process removed a substantial amount of the data, the total number of samples per talker is still much larger than the approximately 5,000 used in Lee et al. (2019).

3.2.4 Crosslinguistic comparison of acoustic measurements

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

For each acoustic measurement and talker, Cohen's *d* was calculated using the *lsr* package (Navarro, 2015) in R (R Core Team, 2020); this provides a high-level assessment of whether variable means differed across the two languages. These

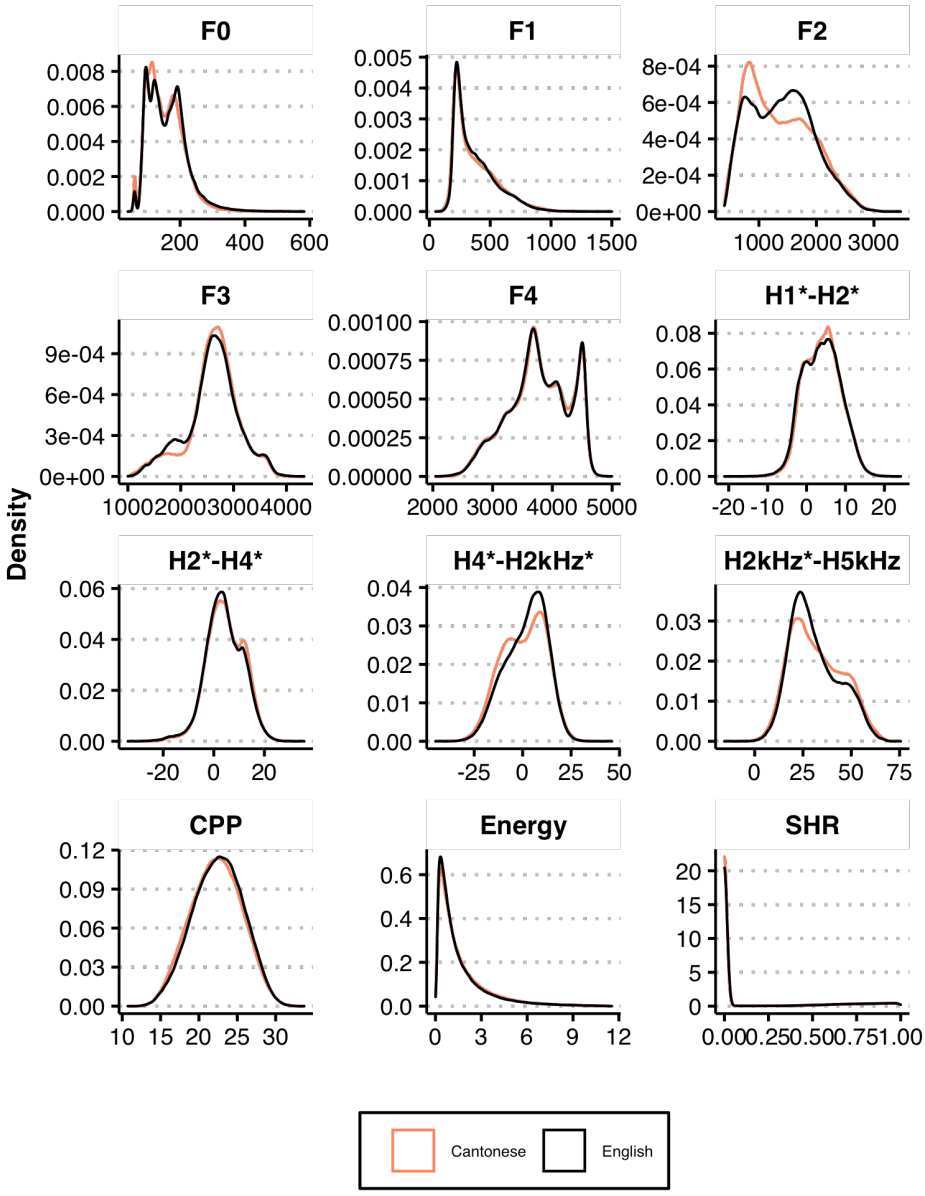


Figure 3.1: Each panel depicts a density plot that pools measurments from all talkers together to show the range of values for that measure. The x-axes each have their own scale. Language is separated out by color.

comparisons have no bearing on how a given variable *varies*. Table 3.3 reports counts of talkers by effect size. Notably, across all talkers and variables, only 21.1% yielded non-trivial Cohen's *d* values, though most talkers (32/34) had at least one non-trivial comparison. The distribution of these counts is depicted in Figure 3.2. Additionally, Figures 3.3 and 3.4 depict the relationship between the difference of means across languages and Cohen's *d* for all of the measures. While redundant, these figures facilitate visual identification of the trends in the data.

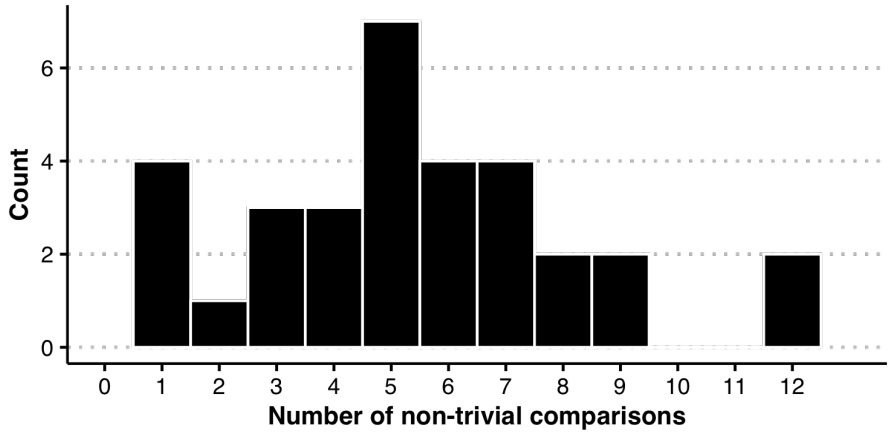


Figure 3.2: A histogram summary of the number of non-trivial comparisons from Table 3.3 across the 34 talkers.

For the non-trivial comparisons, there were consistent patterns across languages for a handful of the variables, including F0, H4*–H2kHz*, and to a lesser extent, H1*–H2*. If there was a non-trivial difference in F0 across languages, then Cantonese had a lower mean F0 than English (13/34; Female = 7), though most talkers did not exhibit a difference (21/34). This is consistent with prior findings that when a difference between English and Cantonese was found, Cantonese had a lower mean F0 for females (Ng et al., 2012; Altenberg & Ferrand, 2006). This difference is occurs at similar rates for female and male talkers in the SpiCE corpus.

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

Table 3.3: This table reports counts of Cohen's *d* for crosslinguistic comparisons of each of the acoustic measurements by talker. For most talkers and variables, the difference in means was trivial, which is reflected in that column's high counts.

	Cohen's *d*		
Variable	**Trivial** ***0.0–0.2***	**Small** ***0.2–0.5***	**Medium** ***0.5–0.8***
F0	21	10	3
F0 s.d.	34	-	-
F1	24	9	1
F1 s.d.	29	5	-
F2	26	8	-
F2 s.d.	32	2	-
F3	24	9	1
F3 s.d.	29	5	-
F4	30	3	1
F4 s.d.	28	6	-
H1*–H2*	18	15	1
H1*–H2* s.d.	32	2	-
H2*–H4*	25	9	-
H2*–H4* s.d.	31	3	-
H4*–H2kHz*	25	8	1
H4*–H2kHz* s.d.	34	-	-
H2kHz*–H5kHz	23	10	1
H2kHz*–H5kHz s.d.	31	3	-
CPP	21	10	3
CPP s.d.	32	2	-
Energy	17	14	3
Energy s.d.	18	16	-
SHR	31	3	-
SHR s.d.	29	5	-

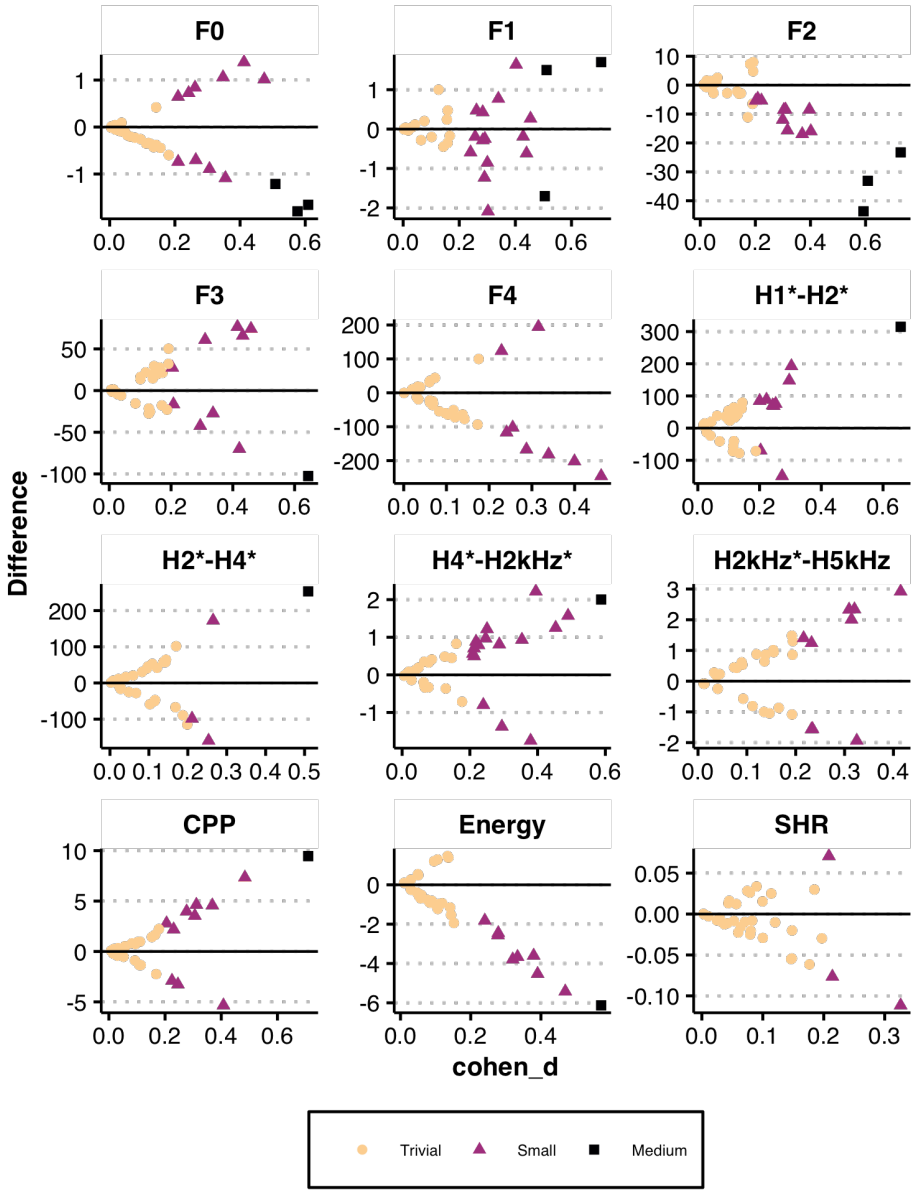


Figure 3.3: Each panel plots Cohen's *d* on the x-axis (scales differ) and the difference between language means on the y-axis. Positive values indicate a higher mean in Cantonese than English. The color reflects the levels of interpreation for Cohen's *d*. Each point represents a talker.

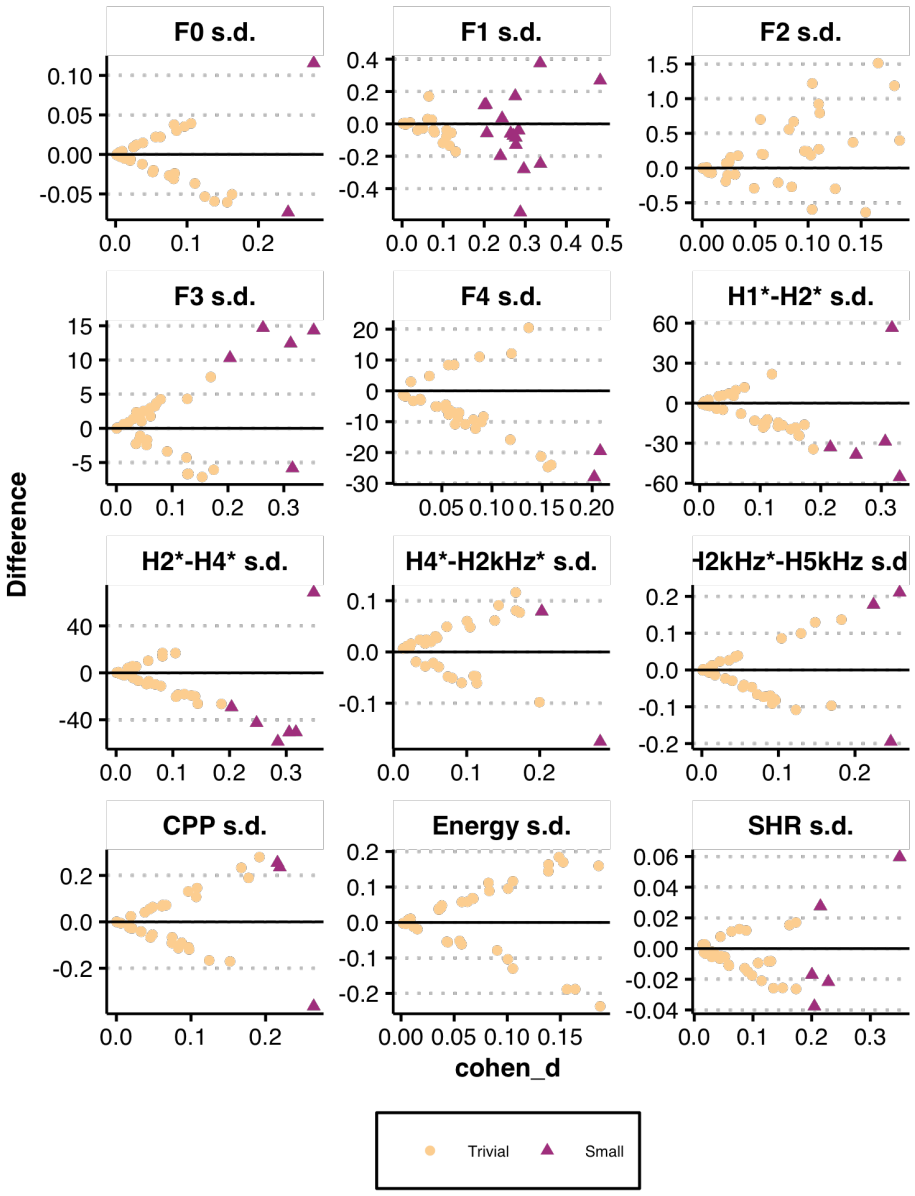


Figure 3.4: This figure uses the format of 3.3, but reports on the standard deviation measures.

not exhibit a difference on this measure. H1*–H2* was significantly higher in Cantonese for a relatively large subset of the talkers (13/34), lower for a small number (3/34), but trivial for most (18/34). While based on different measures than (Ng et al., 2012), the H1*–H2* results are consistent with the finding that Cantonese tends to be breathier (or English creakier)—the current analysis does not distinguish between these interpretations. The H4*–H2kHz* results are not consistent with Ng et al. (2012), yet for both spectral shape measures, it is important to reiterate that they are difficult to interpret on their own. The interpretation here should be taken with a grain of salt.

For the remaining variables, while some talkers exhibited a difference in mean values, the direction of the difference varied, or relatively few talkers exhibited the difference. For example, a variable like F4 would be unlikely to vary across languages within the same talker, given its association with vocal tract size. This interpretation is reflected in the relatively low count of talkers with a non-trivial difference across languages for F4.

Another example of variable behavior comes from F2, which also stands out because of the stark difference in Figure 3.1. While the distribution is perhaps not surprising, given the extent to which English high and mid back vowels have fronted across numerous varieties of English, including those in Western North America (Labov et al., 2008), the figure seems to be at odds with the Cohen's *d* results in Table 3.3—of eight non-trivial comparisons, two were positive and six were negative. The difference in the figure, however, seems to be driven by individual differences in F2 behavior, as the stark difference is not reflected on an individual level. Instead, talkers tend to have either a wide spread of F2 values or a strongly skewed distribution with a long right tail in both languages—this indicates that vowel fronting varies by individual.

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

CPP also exhibits a split between positive (n=6) and negative (7). Higher CPP values are associated with both breathy or creaky non-modal phonation types. In this sense, a positive difference would indicate that Cantonese was more non-modal, while a negative difference would indicate that English was more non-modal. Interpreting CPP is not so straightforward, however, as it is not immediately clear which type of non-modal phonation the measure entails. Given the H1*–H2* results, it seems clear that knowing where on the creaky-modal-breathy spectrum a given speaker falls is pertinent to interpreting this measure. CPP would likely corroborate that outcome on a by-observation basis. In any case, listener assessments would help pinpoint how spectral shape and noise parameters map onto voice quality.

Overall, while talkers show some clear across-language differences, these are far outnumbered by instances with no meaningful difference. The variability observed here fits in with the variable outcomes of previous work. Yet, at the same time, the results in this section do not neatly compare to clearcut differences between male and female talkers found in prior work.

3.2.5 Principal components analysis

Methods

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

be used to generate perception prediction and interpret the results of perception research, as outlined in Chapter 1.

Often, the goal of PCA is to take a large number of dimensions and extract a much smaller set to use for some additional purpose (e.g., linear regression). The focus in this chapter is on the internal structure of the components. That is, it delves into what makes up components for different talkers and whether an individual's voice structure varies (or not) across languages.

This chapter adapts methods from work on voices (Lee et al., 2019; Lee & Kreiman, 2020) and faces (Burton et al., 2016; Turk & Pentland, 1991). The goal of this analysis is to capture similarities or differences in the structure of each talker's voice across languages. As such, there are 68 PCAs—one for each talker and language combination—and the results of each talker's English and Cantonese PCAs are compared. All 24 measures were standardized on a by-PCA basis before the analysis. PCAs were implemented with the *parameters* package (Lüdecke et al., 2020) in R (R Core Team, 2020), using an oblique *promax* rotation to simplify the factor structure, as the measurements reported in the previous section were expected to be somewhat correlated given prior findings (Lee et al., 2019) and a broader understanding of how different acoustic measures align with one another (Kreiman et al., 2014, 2021).

A crucial step in a PCA is determining the number of components. As PCA is a dimensionality reduction technique, this number is crucially smaller than the total number of components. There are many different methods for setting the number, and in this analysis, each PCA included the number of components for which all resulting eigenvalues were greater than 0.7 times the mean eigenvalue, following Jolliffe's (2002) recommended adjustment to the Kaiser-Guttman rule. This rule was used in place of a more sophisticated test (e.g., broken sticks), as it is not detrimental to this exploratory analysis to err on the side of including marginal components (i.e., those that account for relatively minimal amounts of the overall variance).

Additionally, across each of the components, only loadings with an absolute

value of 0.45 or higher were interpreted. While Lee et al. (2019) use a threshold of 0.32, Tabachnick & Fidell (2013) note that higher loadings indicate that a particular variable is a better measure of the component, with 0.32 corresponding to poor (but still interpretable) overlap between the variable and the component. The guidelines in Tabachnick & Fidell (2013) indicate that loadings of 0.45 correspond to fair, 0.55 to good, 0.63 to very good, and 0.71 and above to excellent. Given the large number of components and loadings in this analysis, only loadings greater than the fair threshold are interpreted. This methodological decision facilitates interpreting meaningful loadings on components.

Results

The PCAs across both languages for all 34 talkers resulted in 10–14 components and accounted for 74.9–82.7% of the total variation. Half of the talkers had the same number of components for each language (17 of 34). Of the remainder, 16 of the 34 talkers had a difference of one in the number components, and only one had a difference of two. Talkers had 4–11 identical component configurations across their languages (M=7.82)—that is, the same variables loaded on the components above the fair threshold (though loading values varied). These shared components represent 33.3%–91.7% of the total components for talkers (M=66.7%). The numbers comprising these summary statistics are provided in Table 3.4. While this already indicates a substantial amount of shared lower-dimensional structure across languages, it likely underestimates the actual shared structure. The reason is that similarity of component structure is not taken into account—for example, a component with loadings above the fair threshold for F2, F3, and F4 versus a component with just F2 and F3. This similarity will be taken into account in the next part of the analysis in Section 3.2.6.

To assess whether talkers exhibit the same structure in voice variability across their languages, patterns present across the different PCAs are considered. This provides context for understanding what unique structural characteristics in talkers' voices look like. To this end, this section briefly summarizes common patterns

Table 3.4: The number of components, variance accounted for, and number of identical components across languages for each PCA.

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

across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 3.5 shows all of the components of participant VF32A's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages. Figure 3.5 can be interpreted as follows. The left column visualizes the VF32A's Cantonese PCA, and the right column English. Each panel depicts a single component, and the components are numbered along the right in order by the amount of variance accounted for in the PCA.

VF32A provides a clear illustration of how components compare across languages in different ways. The most straightforward comparison is one where the same variables make up a component in the same position—as is the case for the first component of each language in the figure. While the loadings and the variance accounted for differ, VF32A's first component is formed of F2, H2kHz*–H5kHz, and H4*–H2kHz* in both languages. This type of similarity would have been identified under a stricter replication of prior methods (Lee et al., 2019). Another kind of straightforward comparison is where the same component structure occurs in both languages but in a different ordinal position. Consider, for example, VF32A's component 3 in Cantonese and component 6 in English. Both components comprise F3 and F4 exclusively and account for 7.6% and 7.5% of the overall variance in the respective PCAs. These components are extremely similar to one another in every way but the ordering of components.

The remaining types of comparisons are somewhat less straightforward but still relevant. For example, VF32A's English component 5 (F0 s.d. and CPP) consists of a subset of the variables in her Cantonese component 2 (H2*–H4*, F0, F0 s.d., and CPP). And lastly, sometimes variables just pattern differently—in English, F1 and F1 s.d. pattern with F0 s.d., while in Cantonese, they pattern with H1*–H2* and H2*–H4*. While an in-depth analysis of each component of each PCA is beyond the scope of this chapter, examining VF32A's components in this way highlights the importance of not attributing too much value to the ordering of components. Instead, it is more appropriate to attend to component composition

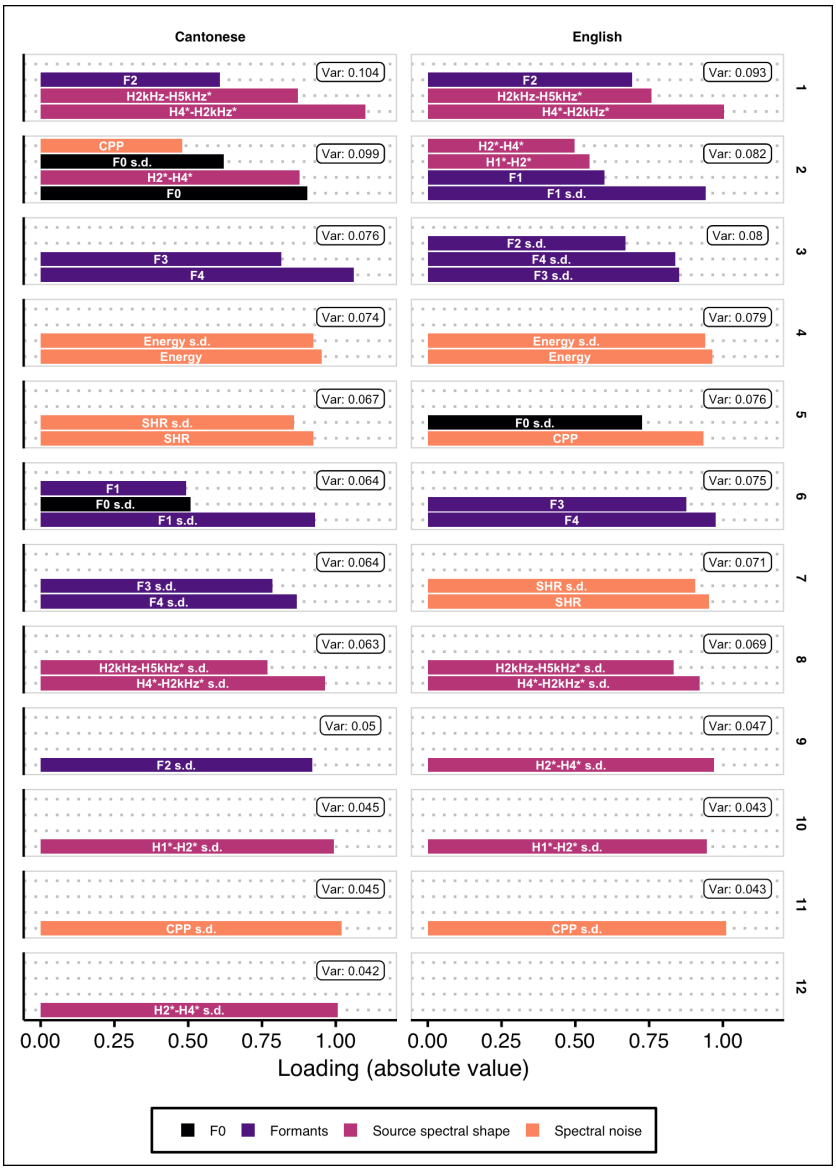


Figure 3.5: In this depiction of the components of the Cantonese and English PCAs for VF32A—a single talker from the corpus taken as an example. Loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings. The component's variance accounted for is superimposed.

and the variance accounted for by different components.

Broadly, there were many similarities in component composition across talkers and languages. The following paragraphs summarize the components that were present in every PCA, regardless of talker or language.

The shared component accounting for the most variation across talkers had a core structure consisting of F2 and H4*–H2kHz*. These usually went along with H2kHz*–H5kHz (Cantonese = 34, English = 31), and occasionally with F3 and F4 (Cantonese = 3, English = 3). While a concise summarization of what this component means is tricky, it includes both higher spectral shape parameters and up to three formants. Respectively, these variables are typically associated with phonation type and vowel quality (or other aspects of the filter). This component thus reflects how some variables that are often studied in isolation, in fact, covary (for a cautionary tale of interpreting F3 and voice quality in the context of sound change, see Sóskuthy & Stuart-Smith, 2020).

In a similar vein, all talkers had a component consisting of H4*–H2kHz* s.d. and H2kHz*–H5kHz s.d., though it accounted for a smaller proportion of the total variation. While not shared by as many talkers, there was a similar component with a different spectral shape variable. H2*–H4* s.d. commonly occurred alone (Cantonese = 18, English = 18) or in combination with H1*–H2* s.d. (Cantonese = 13, English = 14). These components reflect variability in non-modal voice quality and the timbre of the voice—described as brightness in Lee et al. (2019).

Formant s.d. parameters often co-occurred. In both languages, this component typically consisted of F3 s.d. and F4 s.d. (Cantonese = 32, English = 26), though a subset of these cases also included F2 s.d. (Cantonese = 6, English = 10). That formant variability dimensions pattern together likely reflects how formants move in concert across coarticulatory processes. Constantly moving articulators simultaneously impact all of the formants, leading to the covariation observed here.

While the formant and spectral shape moving standard deviations often exhibited these common patterns, variables in these categories were just as likely to pattern in more idiosyncratic ways, loading alongside each other, F0, formants,

and spectral measures. This kind of variability is not readily summarizable.

The spectral noise parameters had a relatively consistent component structure across talkers and languages. Energy and Energy s.d. consistently loaded on the same component as each other—order aside—and were sometimes accompanied by F0 (Cantonese = 6, English = 2) and F0 s.d. (Cantonese = 1). This configuration indicates that volume and volume variability covaried and that in many cases, they covaried along with pitch.

CPP s.d. occurred consistently on its own component for all English PCAs, and 31 of the Cantonese PCAs. In the remaining three Cantonese PCAs, CPP s.d. was accompanied by CPP (n=1) or H1*–H2* s.d. (n=2). CPP patterned less consistently but was most often accompanied by F0 s.d. (Cantonese = 19, English = 14). These two components reflect the relative independence of CPP and how it varies, which measures regularity in the harmonic structure (i.e., degree of modal phonation). That CPP often loads with F0 s.d. makes sense, as an increase in local F0 variation could simply be another way to say there is less regularity in the pitch periods. These components thus likely reflect non-modal phonation.

SHR and SHR s.d. exclusively loaded together for 31 talkers in each language and SHR by itself for a single talker per language. The pair was sometimes accompanied by H1*–H2* (Cantonese = 2, English = 2), H2*–H4* (English = 1), or F0 (English = 3). SHR is associated with period-doubling and irregularities in phonation. That these two parameters occur most often alone suggests that this is a meaningful dimension in voice quality.

While this covers many of the variables that went into the PCAs, F0 is notably sparse in the above paragraphs. While F0 s.d. was fairly consistent in emerging either with CPP (Cantonese = 21, English = 17) or alone (Cantonese = 9, English = 10), the same cannot be said for F0. No particular component structure with F0 occurred more than six times, and across the wide range of configurations, F0 was accompanied by all kinds of variables: F0 s.d., H1*–H2*, H1*–H2* s.d., H2*–H4*, F1 s.d., F4 s.d., CPP, Energy, Energy s.d., and SHR, SHR s.d. The lack of consistency in F0 across talkers is notable for a few reasons. First, F0 plays

a major role in prior work on voice production and perception, given its salience as an acoustic dimension (Perrachione et al., 2019). A second reason for it being notable comes from Lee and colleagues' work, where F0 emerged as an important feature of acoustic voice variation structure in English spontaneous speech (Lee & Kreiman, 2019) and Korean sentence reading (Lee & Kreiman, 2020). In both studies, it consistently covaried with spectral shape and noise variables on the first and second components. This consistent pattern was not present in English sentence reading (Lee et al., 2019).

While several variables are often loaded on the same component, the same variable rarely had a *complex loading pattern*—that is, it was rare for a variable to load on multiple components at the same time. There were some exceptions to this. Three talkers had complex loading structures for H2*–H4* in both languages. Across talkers, only three had complex loading structures for H2*–H4* in each language. F0 and F0 s.d. participated in complex loadings for a single English PCA and twice in the Cantonese PCAs. The remaining variables that participated in complex loading structures only occurred in one or two PCAs across all talkers and languages. This means that for a given PCA, the interpretation of components is reasonably straightforward, even if drawing generalizations over the full group is not.

There were additional components (not reported here) that were shared by less than half of the talkers. A full list of component configurations, along with the number of occurrences and range of variation accounted for is provided in the GitHub repository for this dissertation at https://github.com/khiajohnson/dissertation.

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

the PCAs ranges from 17.3%–25.1%, the unaccounted for variability may also be idiosyncratic in nature.

3.2.6 Canonical redundancy analysis

Methods

The goal of the analysis in this section is to provide a numerical comparison of PCAs in a pairwise fashion to assess the extent of similarity in lower-dimensional structure within and across languages and talkers. The analysis accomplishes this by comparing PCAs using a technique called a *canonical correlation analysis*, which provides a metric of redundancy (i.e., overlap) between the two PCAs compared. A benefit of this method is that the resulting metric is easy to interpret.

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

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

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

Results

Redundancy indices for within-talker comparisons ranged from 0.80 to 0.97, (Mdn = 0.92, M = 0.91, SD = 0.04) and are displayed in Figure 3.6, with the two redundancy indices for a given pairwise comparison plotted against one another. Comparisons across talkers within-language ranged from 0.64 to 0.96 (Mdn = 0.83, M = 0.83, SD = 0.5). Comparisons across both talkers and languages ranged from 0.64 to 0.97 (Mdn = 0.83, M = 0.83, SD = 0.5). Within-talker values were confirmed to be higher than across-talker comparisons, per a Welch's t-test ($t(70.93) = -17.35$, $p < 0.001$, d = 1.77)—this result indicates that regardless of language, talkers are more similar to themselves than talkers are to each other. A second Welch's t-test testing the same versus different language for the across-talker comparisons did not find a difference between those groups ($t(4485.9) = -1.53$, $p = 0.13$, d = 0.05). This result demonstrates that language is not a delineating factor, or at the very least, the role of language is eclipsed by the role of talker. This interpretation makes sense, given the high degree of within-talker similarity demonstrated in the first Welch's t-test.

While the across-talker comparisons were generally lower than the within-talker ones, the redundancy indices are overall still relatively high. The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded components almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.2.5 , there were a substantial number of commonly occurring patterns across talkers and languages. Together, this supports the con-

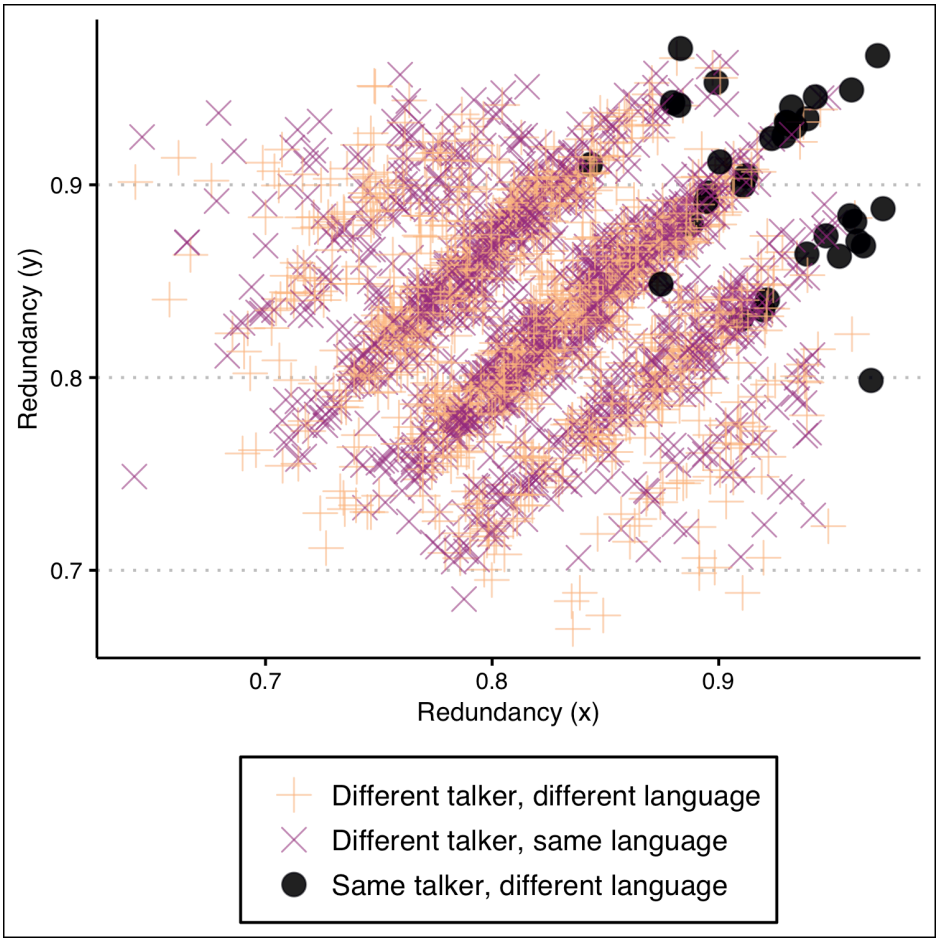


Figure 3.6: This plot depicts the relationship between the two redundancy indices for three different types of comparisons. Across-talker comparisons represented by orange "+" (different language) and pink "x" (same language) overlap in their entirety. Within-talker comparisons are represented by the black circles and are clearly clustered at the top right.

ceptualization of a voice space comprising a shared structure—as in the case of the prototype account—where voices can only deviate from one another so much.

3.2.7 Passage length analysis

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

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

Figure 3.7 plots lines reflecting the redundancy indices for each interview, with superimposed mean GAM smooths. The x-axis represents the sample size of the shorter passage length in the comparison. The y-axis represents an average of the

two redundancy indices. The vertical line at 5,000 represents the average sample size from Lee et al. (2019). The vertical line at 20,124 represents the sample size used in Sections 3.2.5 and 3.2.5 . While there are some gains in sample sizes above the second vertical line, they are comparatively small. The leveling-off point falls somewhere between 10,000 and 15,000 samples.

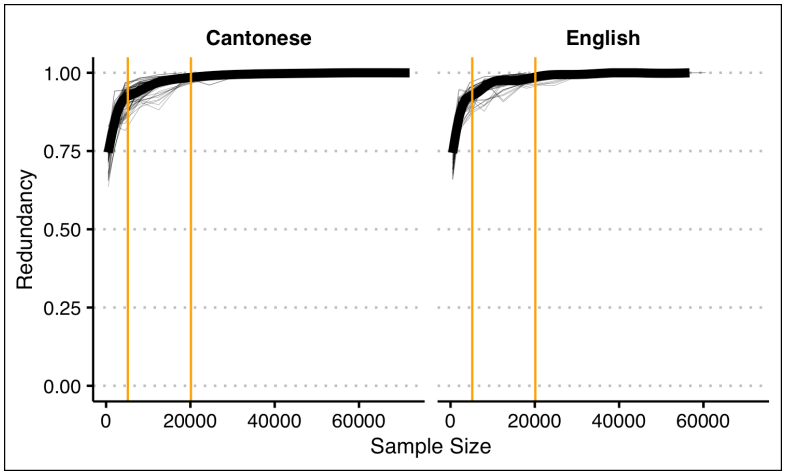


Figure 3.7: Passage length redundancy indices are plotted against the sample size of the smaller PCA. Smoothed curves show a rapid increase in redundancy followed by a levelling off between the vertical orange lines, which represent the sample sizes used in prior work (x = 5,000) and the present study (x = 20,124).

It is readily apparent from this plot that the sample size used for PCAs in this chapter was sufficient to capture most of the range of talkers' within-interview variability. Additionally, given how sample size seems to impact redundancy, this analysis confirms that fixing the sample size in the previous sections was an appropriate decision. As the leveling-off point likely varies across speech styles, it is not immediately apparent whether the sample size in Lee et al. (2019) sufficiently captured the range of talker variability and thus may not adequately capture the structure of their variability.

3.3 Discussion and conclusion

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

A clear result in this chapter is that the bilinguals studied here exhibit similar spectral properties and similar lower-dimensional structure in their acoustic voice variation. This similarity is most apparent on a within-talker basis but still present across talkers and languages, despite substantial segmental and suprasegmental differences across English and Cantonese (Matthews et al., 2013; Wilson & Mihalicek, 2011). In this sense, the SpiCE corpus talkers appear to have the same "voice" in each of the two languages. This outcome supports the characterization of voices as auditory faces. The face-voice comparison is especially apt if you take into account findings that talkers' facial postures vary across languages, as evidenced by work demonstrating that lip movement patterns alone are sufficient for humans and machines to identify and discriminate between spoken languages (Afouras et al., 2020; Soto-Faraco et al., 2007). Voices and faces are highly similar across languages but are not necessarily identical—this leaves room for individuals who are familiar with both the individuals and languages in question to excel

at perceptual tasks in both domains.

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

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

The overlap in component structure between this chapter and prior work (Lee et al., 2019; Lee & Kreiman, 2019, 2020) adds credibility to the idea of a prototype model in voice (Lavner et al., 2001; Latinus & Belin, 2011). In this body of work, a prototype is typically thought of as a speech community average. That there are similarities across disparate populations and languages (e.g., this chapter and Lee & Kreiman, 2020), suggests that such a prototype may extend beyond tightly defined speech communties.

The PCA analysis in this chapter also adds additional commonly occurring components to the mix, suggesting that is yet more lower dimensional structure shared by voices. Examples of this include separate components that put each of

the spectral noise dimensions at center stage—SHR, Energy, and CPP (with or without FO s.d.). That these components emerge in the form that they do validates the use of these measures for describing how voices vary—each is capturing unique variability in the structure of the voice. Conversely, the spectral shape variables tend to covary in more complicated ways—this reflects a more general understanding of how the four spectral shape parameters tell us about the shape of a spectrum in aggregate, and how they are more challenging to interpret on their own (Garellek, 2019). The addition set of shared components serves to flesh out the structure of what a protoypical voice might look like.

This high degree of similarity does not preclude crosslinguistic differences on a within-talker basis but rather suggests that such differences occur on a more global level. This is apparent in Figure 3.8, which depicts the relationship between within-talker, across-language redundancy (averaged) from Section 3.2.6 and the difference between the mean values for each of the acoustic measurements in Section 3.2.4. If there were clear relationships between large crosslinguistic differences and redundancy, the regression lines should be strongly negative—this does not seem to be the case. Instead, this figure demonstrates that there is not much of a relationship Cohen's *d* and redundancy. This suggests that the mean differences are not exerting much influence on the redundancy analysis. Coarse summary statistics and the structure of variability thus give very different, and likely independent, views into the how voices vary.

Such high similarity in the PCAs was not entirely expected, given the results of Lee et al. (2019), where a handful of shared components were evident but were complemented by numerous idiosyncratic components. At face value, the results in this chapter suggest that a heterogeneous bilingual population has more across-talker similarity than a tightly controlled group of monolingual English speakers. Several analysis decisions may have contributed to this apparent difference. Similar components were compared independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that comparisons are made among like items. Any downside to this methodolog-

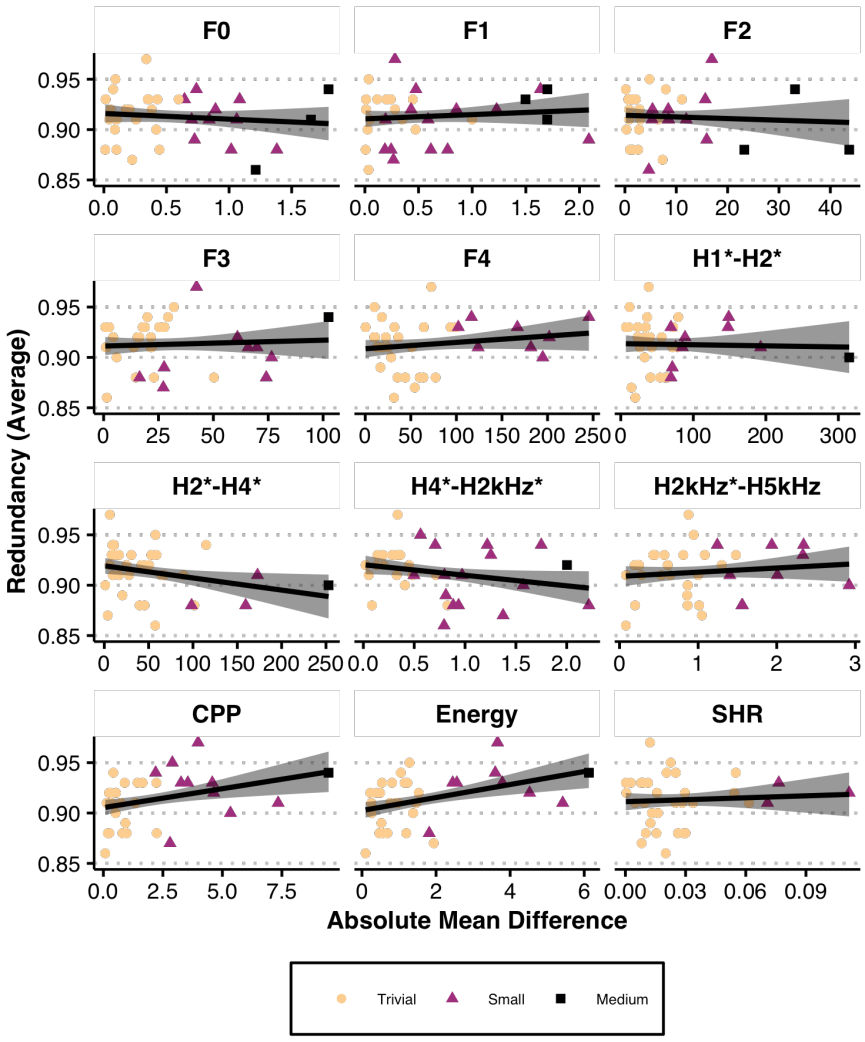


Figure 3.8: The average redundancy value for each talker is plotted against the absolute value of the difference of means across languages for that talker. Color and shape indicate the size of Cohens' *d*. The superimposed regression line summarizes the relationship between these values.

ical decision is mitigated by the fact that most components made relatively small contributions in how much of the overall variance they accounted for (see Table 3.4). As such, I predict that increased across-talker similarity would be found in a reanalysis of the UCLA Speaker Variability Database (Keating et al., 2019) using the adapted methods of this chapter.

While methodological choices may account for some part of these results, the data differences between the current chapter and previous studies are also pertinent. This chapter uses substantially longer passages than the short samples in Lee et al. (2019). Larger speech samples allow for a stable underlying structure to emerge. Smaller samples, conversely, may reflect more ephemeral variation in a talker's voice, and thus not be representative of the talker's full range. The passage length analysis in this chapter shows that the number of samples needed for stabilization is substantially larger than the 5,000 samples used in Lee et al. (2019). This does not necessarily discount their work, however, as the current chapter uses spontaneous speech, which is arguably more variable than read speech.[5] It's plausible that an analysis of sentence reading would not need as much data to cover talkers' range of variability in reading aloud. The body of literature in the introduction establishes differences in voice quality across speaking styles (e.g., Lee & Sidtis, 2017). As such, the threshold suggested here may only be appropriate for the speaking style of peer-to-peer conversational interviews. In any case, the methods presented here offer a tool for researchers to use in assessing whether their sample size is representative of a larger whole. Understanding how this interacts with speaking style is left for future directions.

Ultimately, the goal of this line of research is to understand how the acoustic variability and structure of talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination. Turning to listener and behavioral data will help in deciphering what is meaningful variation within

[5]While it is true that Lee & Kreiman (2019) examined spontaneous speech, the poster only states that two minutes of speech were used for each participant. By this estimation, the sample size was likely on the lower side, compared to the 20-25 minute interviews in the SpiCE corpus. However, it is not possible to make a direct comparison without knowing the number of samples.

a voice from low-level noise that cannot be attributed to a particular vocal signature. Verification from listener performance will help adjudicate which statistical choices present an acoustic voice space that matches listener organization. The results of this chapter set up predictions for that work, some of which is currently underway (Lloy et al., 2020, 2021). These predictions will be revisited in general discussion, in Section 5.4.

Chapter 4

The Structure of Voice Onset Time Variation in Bilingual Long-lag Stops

4.1 Introduction

One of the primary goals in this chapter is to investigate what languages can share in the mental representation of similar speech sound categories.[1] The idea of representation is intended here in the manner typically meant by psycholinguists (e.g., Llompart & Reinisch, 2018), exemplar theory proponents (e.g., Amengual, 2018), and Flege & Bohn (2021) in their revised Speech Learning Model (SLM-r). These groups use similar language to describe representation, emphasizing distributions of sensory experiences rather than theoretical linguistic descriptions. For example, Flege & Bohn describe the units of a multilingual segment inventory as categories comprising input distributions of exemplars: "the sensory stimulation associated with...speech sounds that are heard and seen during production by others...in meaningful conversations" (2021, p. 32). The SLM-r also posits that the speech sounds

[1] Note that "similar" is an often ill-defined concept that will be grappled with in Section 4.1.1.

of a bilingual's languages exist in a shared phonetic space, regardless of what they share in their representations. In starting with the SLM-r—where distributions of exemplars for different categories cohabit the same phonetic space—this chapter addresses the extent to which languages share representation(s). Much like Chapter 3, the approach here is one of leveraging the structure of variation to understand the system. Additionally, as with the preceding chapters, this chapter focuses on the speech of early bilinguals, as discussed in Section 1.1. This is an important reminder, as the conceptual model referenced throughout this chapter—Flege & Bohn's (2021) SLM-r—focuses more on the speech of second language (L2) learners and late bilinguals, as opposed to the early bilinguals in this dissertation.

There are many pieces to this puzzle, and the literature has already addressed some of them. The introduction to this chapter proceeds as follows—Section 4.1.1 addresses which sound categories are candidates for shared representation for bilinguals in the first place and addresses key terminology, like "similarity." Section 4.1.2 builds on this by briefly summarizing the relevant crosslinguistic influence literature, addressing assimilation, dissimilation, and how they reflect on the idea of shared representation for sound categories. Section 4.1.3 identifies a limitation of the existing paradigms in crosslinguistic influence and proposes adapting the uniformity framework as a way to fill the gap. This framework offers a way to interpret the structure of variation for a given acoustic dimension. Section 4.1.4 introduces the focus of this particular study—long-lag stops in Cantonese and English—and outlines the specific research questions and hypotheses. After Section 4.1.4, the chapter moves onto methods and results in Sections 4.2 and 4.3.

4.1.1 Identifying "links" across bilinguals' languages

At first glance, the best candidates for shared representation are sound categories that are "linked" together. The definition of links, however, can be frustratingly vague in the multilingualism literature. In a handbook chapter on bilingual phonetics and phonology, Simonet describes "links or connections of one sort or another

between the phonetic categories" (2016, p. 10). Despite being vaguely defined, links nonetheless represent a crucial concept. In the most basic sense, links are defined by the behavior they account for—they exist between sound categories that exert influence on one another under *some* set of circumstances. This definition is somewhat problematic, as it would fail to identify (real) links in cases where influence is less likely to occur (cf. Grosjean, 2011). Links behave dynamically, as such, Simonet also notes that "these connections...are transiently strengthened in contexts that induce the activation of both languages and inhibited in contexts that favor the use of only one of the languages" (2016, p. 10). Arguably, such links must exist because crosslinguistic influence can be observed (this body of literature will be reviewed in Section 4.1.2). While there may be alternative explanations (i.e., global influence), the concept of links is widely assumed in accounting for bilinguals' behavior.

Flege & Bohn (2021) expand on the idea of links in the SLM-r by providing a framework for predicting which sound categories will be linked together. The proposal is simple—namely, sound categories will be linked to the closest category in the other language. Determining which categories pair up, however, remains an empirical challenge from the perspective of speech production, and as a result, Flege & Bohn (2021) rely on perceptual metrics. This is not to say that either kind of metric is inherently better or worse, but rather, that perception- and production-based metrics offer different types of insight into the linguistic system. The reason for the challenge of pairing up sound categories is that perception and production do not always line up neatly. Flege & Bohn assert that similarity "must be assessed perceptually rather than acoustically because acoustic measures sometimes diverge from what listeners perceive" (2021, p. 33). The disconnect between the two processes arises from both linguistic and physiological bases.

In an overview chapter on similarity in crosslinguistic influence, Chang (2015) suggests an alternative—in accounting for behavior, similarity is best captured abstractly. Chang states that crosslinguistic influence at the segmental level tends to occur between sounds that share "(1) similar positions in the respective phone-

mic inventories (when considering the contrastive feature oppositions—or, more broadly, the 'relative phonetics'—of the sounds in relation to other sounds in the inventory), and (2) similar distributional facts" (2015, p. 201). While "distributional facts" seems to be intended in a broad sense, the example Chang gives is co-occurrence restrictions. This approach to similarity emphasizes a general role for abstraction but does not necessarily invite a formal phonological analysis. Developing such an analysis would likely constitute a dissertation in itself—Mielke (2012) highlights the challenges of applying phonological features across languages, given the sheer variety of phonetics-phonology mappings in the world's languages.

While Flege & Bohn (2021) and Chang (2015) take different approaches—perceptual ratings and relative phonetics—they ultimately accomplish a similar goal, by accounting for abstraction and nonlinearity in listeners' mental representations. In sum, abstract similarity in the mind seems to be a prerequisite for the emergence of a link between two sound categories, given how it does a better job of accounting for when and where crosslinguistic influence occurs. The presence of abstract similarity, however, does not address what happens next. It does not entail any particular outcome, and it does not directly address how representation is structured for the sound categories in question.

4.1.2 Crosslinguistic influence and representation

The next step in the puzzle is understanding what happens to linked sound categories. The SLM-r outlines two primary outcomes for sound categories in a shared system—assimilation and dissimilation (Flege & Bohn, 2021). Assimilation is a merging of phonetic properties and arguably occurs when bilinguals and learners do not "discern a phonetic difference (or differences) between the realizations" of sound categories (Flege & Bohn, 2021, p. 40). It is not entirely clear at what level discernment between sound categories occurs, though it is most likely intended to be a linguistic level rather than a purely auditory one. Dissimilation, then, is the reverse—a diverging of phonetic properties that occurs when a difference is

detected but is too small to maintain. These two processes are defined in rather absolute terms, which are tempered to some extent in the following paragraphs' continued discussion of the SLM-r.

Notably, these processes need not impact all phonetic properties in the same way. For example, in a study of coronal stops produced by simultaneous French-English bilinguals, Sundara et al. (2006) found that bilinguals produced differences across languages for voice onset time (VOT) and the standard deviation of burst frequency. These bilinguals did not differentiate based on other spectral moments of the burst that monolingual comparison populations did, and as a result, Sundara et al. (2006) illustrate divergence on some properties and convergence on others.

The motivation for the outcomes of assimilation and dissimilation arises from two simple constraints from the production and perception systems—effectively, do not get too close to each other in perception, and do not get too complicated in production (Guion, 2003; Lindblom & Maddieson, 1988; Flege & Bohn, 2021). These constraints lead SLM-r to posit that proximity leads to instability, even if what counts as close for early bilinguals remains unclear. Bilinguals can, after all, perceive subtle acoustic differences between similar sound categories (Ju & Luce, 2004). In SLM-r, the potential outcomes of instability are assimilation and dissimilation. Considering that bilinguals are fully capable of distinguishing highly similar sound categories across languages (e.g., Sundara et al., 2006; Lein et al., 2016; Casillas, 2021), this is not a trivial point to make. Yet, given SLM-r's primary focus on L2 learning, it is perhaps not a surprising approach. It may thus be more appropriate in the case of early bilinguals to also consider contrast maintenance alongside dissimilation. That is, early bilinguals may not need to assimilate or dissimilate a pair of similar sounds but rather simply maintain the subtle contrast as is.

Following what SLM-r posits, a relatively simple account is that dissimilation for similar sound categories would lead to distinct representations of those categories and that assimilation leads to a shared representation. The picture is complicated, however, by the idea of imperfect assimilation and what Flege &

Bohn term *composite categories*. Suppose sound categories from two languages are phonetically close to each other but do not fully assimilate. In that case, the SLM-r proposes that they will remain linked in a composite category "defined by the statistical regularities present in the combined distributions of the perceptually linked...sounds" (Flege & Bohn, 2021, p. 41). This scenario might be characterized as imperfect or partially shared representation, where certain dimensions are kept apart, and others overlap. For example, the place of articulation may be shared across languages even if VOT differs—this scenario is observed by Sundara et al. (2006), described above. Alternatively, the lack of clear-cut examples of assimilation in the literature may instead indicate that assimilation and dissimilation might be better cast as ends of a spectrum for a gradient, context-sensitive phenomenon. This re-conceptualization makes room for things like contrast maintenance and composite categories.

There are a few potential reasons for the lack of clear-cut assimilation outside of late bilingual and L2 speech. First, true assimilation might just be rare in bilingual speech. This reason is supported by a recent meta-analysis of crosslinguistic influence for Spanish and English initial stop consonants (Casillas, 2021). In this environment, English long-lag stops and Spanish short-lag stops are linked to one another (Fricke et al., 2016b; Goldrick et al., 2014; Bullock & Toribio, 2009; Olson, 2016). Casillas (2021) found that early bilinguals did not produce "compromise" stop categories. That is, early Spanish-English bilinguals did not, on the whole, produce VOT that was somehow intermediate to the canonical productions by monolinguals of either language. Instead, crosslinguistic influence can be fully attributed to what Casillas describes as "performance category mismatches that result from dynamic phonetic interactions associated with language activation" (2021, p. 16). In this sense, the production of each category was influenced by task demands and factors such as social context. So while some assimilation occurs, it is far from the only process at play. This finding echoes arguments made by Bullock & Toribio (2009) on the sophistication and control that bilinguals exert over their range of possible forms. So while there is clear evidence of a link

between the two sounds—and perhaps even evidence for a composite category—"compromise" seems inappropriate for capturing behavior. Instead, bilinguals produce a wide range of forms appropriate to and influenced by different contexts. Without considering task and context factors, it is perhaps not surprising for two sound categories to masquerade as a single composite category.

A second reason for the rarity of complete assimilation arises from the experimental and corpus-based approaches typically used to study crosslinguistic influence. Experimental approaches to crosslinguistic influence use paradigms such as sentence reading, isolated word production, or picture naming—each paired with various common manipulations. At a more global scale, some studies set the language mode of the full session and compare individuals across sessions (Grosjean, 2011; Simonet & Amengual, 2019; Sancier & Fowler, 1997), or compare groups of individuals in different language mode conditions (Antoniou et al., 2010). At a more local level, some experimental designs leverage language switching across blocks (Sundara et al., 2006), across trials (Goldrick et al., 2014), or within trials (i.e., prompted code-switching; Bullock & Toribio, 2009; Antoniou et al., 2011; Olson, 2016). Both types of experimental studies often include both cognate and non-cognate items as a focus or manipulation (e.g., Goldrick et al., 2014). Corpus-based approaches similarly tend to focus on proximity to code-switching (Fricke et al., 2016b; Balukas & Koops, 2015) and cognate production (Brown & Amengual, 2015). Across all study types, there are common findings. Typically, cognates, words occurring before a language switch, and words produced in more bilingual modes show increased convergence. Conversely, unilingual modes, non-cognates, and words occurring far from a code-switch tend to show a greater degree of contrast maintenance (or divergence). While there is a tendency for convergence, Bullock & Toribio (2009) demonstrate that proximity to a code-switch in a formal experimental setting leads some individuals to exaggerate the difference between English and Spanish VOT. This kind of linguistic behavior, arguably, reflects deep metalinguistic knowledge on behalf of bilinguals like those studied in Bullock & Toribio (2009).

In these approaches, the ability to examine crosslinguistic influence for any given pair of sounds hinges on the presence of an observable acoustic difference under some set of conditions. Arguably, for this reason, most prior work in crosslinguistic influence has focused on sounds that are phonologically similar (i.e., abstract, relative phonetics) yet phonetically distinct. A common example of this arises from languages that differ in their initial stop voicing contrasts. For example, as discussed at the beginning of this section, North American English contrasts long- and short-lag stops in initial position. Conversely, Spanish contrasts short-lag and prevoiced initial stops. Despite the clear difference in how languages encode a laryngeal timing contrast, there is nonetheless strong evidence for a crosslinguistic link between English long-lag and Spanish short-lag stops (Casillas, 2021; Fricke et al., 2016b; Goldrick et al., 2014; Bullock & Toribio, 2009; Olson, 2016).

These studies demonstrate phonetic convergence—or variable assimilation—in two ways. First, VOT is shorter for English initial stops produced by bilinguals when compared to monolingual control groups. This result is attributed to the influence on English long-lag stops from the short-lag category in the other language (Olson, 2016; Johnson & Babel, 2021). Similarly, French-English bilinguals are more likely to produce lead voicing in initial English voiced stops compared to English monolinguals (Sundara et al., 2006). Second, evidence of crosslinguistic influence can also come from comparing bilinguals to themselves across different circumstances. For example, Fricke et al. (2016b) use a spontaneous speech corpus to demonstrate that Spanish-English bilinguals produce shorter, more Spanish-like VOT in the lead up to an English-to-Spanish code switch (Fricke et al., 2016b). An experimental example comes from Simonet & Amengual (2019), where individuals participate in multiple sessions in which language mode is carefully controlled. While this body of work makes the presence of a link clear, it also highlights that there are distinct aspects of how these sound categories are represented in the bilingual mind (Casillas, 2021). In the SLM-r, these examples might be considered composite categories. Alternatively, they might be examples of contrasts being maintained in the face of proximity.

In any case, this focus presents a conundrum. By using methods where observing degrees of similarity hinges on the ability to detect a difference, researchers often preemptively exclude some of the best candidates for shared representation—those that share both abstract *and* acoustic similarity. While focusing on both is rare, it is not entirely absent from the literature. One example comparing highly similar sound categories in the early bilingualism literature comes from a lab-based study of Mandarin-English bilingual children (Yang, 2019). The authors found that highly proficient bilingual 5 to 6-year-olds produced equivalent VOT for Mandarin and English long-lag stops, even though the monolingual comparison groups were consistently different. Yang's result suggests that the difference is either too small to maintain or that 5 to 6-year-old children have not yet mastered it. These claims should be tempered, however, as Yang (2019) did not control for language mode, and adult bilingual behavior was not considered.

Despite some inroads, there is nonetheless a distinct paucity of work examining highly phonetically similar speech sounds across languages, even when such a connection would make sense. A recent study of crosslinguistic influence by Tsui et al. (2019) compares English long-lag and Cantonese short-lag stop production in bilinguals. The study used a picture naming task in the context of a language switching design, where participants named pictures in both languages. The crucial comparison in the study is between trials that occurred immediately after a trial of the same language or the other language (i.e., whether there was a switch). While the design closely mirrored Goldrick et al. (2014), the results were murky—balanced bilinguals showed no evidence of crosslinguistic influence. The decision to use English long-lag and Cantonese short-lag stops as the stimuli could explain this outcome. This comparison reflects the need for stimuli to be acoustically distinct beforehand—as noted above—yet, it glosses over the fact that both languages contrast short-lag and long-lag VOT in initial position. The best candidates for links—and accompanying crosslinguistic influence—would be the long-lag stops in each language. The long-lag stops occupy the same relative position in their respective inventories and bear resemblance physically (e.g., see references

in Section 4.1.4). Tsui et al.'s (2019) null result with balanced bilinguals is thus unsurprising.

This criticism suggests that (Tsui et al., 2019) would have gotten more insightful results by comparing Cantonese *long-lag* stops to English long-lag stops. More importantly, however, it highlights the design constraints of the paradigms used in crosslinguistic influence research. Such methods are better suited for detecting links between acoustically distinct sound categories and documenting the circumstances that undergird parallel activation of languages. In Grosjean's (2011) terms, these methods are best suited to detecting *interference*, as opposed to *transfer*. Interference is the kind of crosslinguistic influence observed between simultaneously activated mental representations—it is ephemeral, occurring "online." Transfer, on the other hand, occurs on a longer time scale and affects the representations themselves. While Grosjean (2011) argues that disentangling the two types of influence is difficult, the methods described above seem tailored more towards interference, given the way that they promote activation of both languages.

A good example of interference comes from Catalan-Spanish bilinguals' vowel production. Simonet & Amengual (2019) compare vowels on a within-talker basis from two separate sessions—unilingual Catalan and bilingual Spanish-Catalan—and found that Catalan /a/ was produced more like its Spanish counterpart in the bilingual session. While this result is straightforward, it is unique in that the authors show a dynamic within-talker process facilitated by language mode. In a monolingual setting, talkers maintain a contrast. However, the same talkers show partial assimilation in the bilingual setting. When both languages are activated, Catalan interferes with Spanish, leading to the observed outcome of phonetic convergence. Simonet & Amengual (2019) argue that these sounds are linked and thus simultaneously activated but ultimately have separate representations in long-term memory (i.e., do not reflect transfer). In its discussion of category formation, SLM-r seems more concerned with assimilation and dissimilation at the level of long-term representations (i.e., transfer; Flege & Bohn, 2021), even though more of the literature reviewed in support of the model uses designs that center inter-

ference. Notably, however, transfer and interference are difficult to disentangle (Grosjean, 2011).

To summarize, most work in crosslinguistic influence has focused on phonologically similar yet phonetically distinct pairs of segments and how they interfere with one another during the process of producing speech. These pairs are not strong candidates for transfer and shared mental representation (as defined at the beginning of this chapter). This widespread focus likely arises for several different reasons. The established paradigms—which greatly facilitate research—tend to require a detectable difference. It is also possible that assimilation in long-term mental representations is rare for early bilinguals, which would limit the options for studying such a phenomenon. Lastly, comparisons of categories that already exhibit both abstract and phonetic similarity may be taken for granted and not considered an interesting problem to focus on, despite the nature of the mental representation of sound categories being a key focus in psycholinguistics (Samuel, 2020).

While many psycholinguists are indeed concerned with representation, processing seems to have taken center stage in the psycholinguistics of bilingualism. In a prominent example of this, Fricke et al. argue that "bilingualism has the potential to reveal the fundamental breadth and underlying nature of variation in language processing" (2019, p. 204). This chapter foregrounds the argument that bilingualism also offers a window into understanding the nature of mental representation. In the interest of understanding it, the best category candidates would be the hardest to distinguish using only surface forms.

4.1.3 Adapting the uniformity framework

The study described in this chapter focuses on assessing whether phonetically similar sounds share a mental representation or not. Recall that mental representations are defined in psycholinguistic terms, as outlined at the beginning of this chapter. Unlike prior work focusing on variable convergence and divergence, this chapter addresses whether a single category is used in both languages or whether each lan-

guage carries a separate representation of similar categories. Testing directly for shared structure in this way means that the set of methods that rely on detecting and modulating differences is not appropriate. To this end, this chapter extends the articulatory uniformity framework to the study of multilingual segment inventories.

Articulatory uniformity is conceptualized as a constraint on within-talker phonetic variation, in which articulatory gestures or phonological primitives are implemented systematically in speech production (Chodroff & Wilson, 2017; Faytak, 2018; Ménard et al., 2008). The core idea of the articulatory uniformity framework is that phonetic variation is highly structured. While Chodroff & Wilson (2017) draw tight connections between uniformity and phonological features, Faytak (2018) instead emphasizes how talkers learn and reuse articulatory gestures. This articulatory account builds on earlier work by Ménard et al. (2008), who argue that the stability of the first formant in French vowel production is best accounted for by stability in the tongue height gesture (i.e., reuse of the gesture). While the specific theoretical accounts vary somewhat by author, there is nothing to suggest that such accounts are incompatible with one another. Both articulatory and phonological explanations may be valid and even related to one another. Given the focus of this chapter on phonetic and psycholinguistic accounts of category formation and representation (rather than phonological), the articulatory account—with its accompanying acoustic consequences—is likely more appropriate, even if this dissertation does not directly engage with articulatory phonetics.

In this light, if a set of segments share an attribute (i.e., share a description such as "long-lag" or belong to the same natural class), then talkers should implement the segments with the same phonetic target or articulatory gesture. This systematicity has been observed for vowel height (Ménard et al., 2008), tongue shape (Faytak, 2018), fricative peak frequency (Chodroff & Wilson, in press), and stop consonant VOT (Chodroff & Wilson, 2017). In the case of VOT in particular, the relationship between a laryngeal gesture and its acoustic consequence is clear. This allows for the extension of Ménard et al.'s (2008) argument regarding

F1 and tongue height to VOT and its corresponding laryngeal gesture. Reusing the gesture across sounds that share the relevant attribute "may simplify the somatosensory feedback needed to control the speech task" (Ménard et al., 2008, p. 26). In simple terms, reusing gestures is easier than the alternative—using different gestures—in the case of high vowels. The same argument could easily be extended to long-lag stops.

Findings for within-language stop consonant uniformity appear to be quite robust. Chodroff & Wilson (2017) report consistent results across a lab study based on reading a list of CVC words and a corpus study comprising connected read speech. Chodroff & Baese-Berk (2019) replicate the uniformity findings for stop consonants with connected read speech samples from 140 non-native English speakers with a wide range of native languages in the ALLSSTAR corpus (Bradlow et al., 2011). While Chodroff & Baese-Berk (2019) found a greater degree of between-talker variability with non-native speakers compared to the prior monolingual work (Chodroff & Wilson, 2017), the within-talker structure was robust. However, the uniformity framework has not yet been extended to early bilingual speech to compare how bilinguals produce phonetically similar sounds in each language. Extending the framework across languages follows the framing of uniformity as arising from articulatory reuse (Faytak, 2018), effectively asking whether or not reuse extends across languages.

There is also motivation for uniformity in perception. As outlined in Section 1.2, Orena et al. (2019) speculated that a bilingual advantage at generalizing across languages in talker identification might derive from sensitivity to crosslinguistic structural similarity. While this remains speculation on behalf of crosslinguistic generalization, there is evidence that within-language uniformity facilitates talker identification, above and beyond typical talker-indexical components of a voice (Ganugapati & Theodore, 2019). It follows that this boon would also extend to bilingual talker identification, provided that phonetic variation across languages also exhibits uniform structure.

4.1.4 Long-lag stops in Cantonese and English

English and Cantonese initial long-lag stops are strong candidates for shared mental representation because they exhibit both relative and physical phonetic similarity, akin to the difference for Mandarin and English in Yang (2019).[2] Consider the initial stop [k^h]—in citation speech—with a mean VOT of 80 ms in American English (Lisker & Abramson, 1964) and 91 ms in Hong Kong Cantonese (Clumeck et al., 1981). While these values are objectively different—though based on small sample sizes—it seems that using the same laryngeal timing gesture would be advantageous given the small difference across monolingual populations (that may or may not be perceptible). There is ample work documenting long-lag VOT across different varieties of English and speaking styles, with values as low as the 30–50 ms in spontaneous speech (Stuart-Smith et al., 2015). There is far less work documenting Cantonese long-lag VOT; nonetheless, descriptive work casts it as having generic long-lag aspiration similar to English (Matthews et al., 2013; Bauer & Benedict, 1997; Chan & Li, 2000; Mielke & Nielsen, 2018). For example, Matthews et al. (2013) describe initial stops in both English and Cantonese as voiceless and aspirated, even though they differ in their phonological features—English is typically analyzed with a ±voicing distinction and Cantonese with an ±aspiration distinction (Matthews et al., 2013).

While the presence of articulatory reuse within Cantonese and across languages remains an empirical question, it aligns with the finding that bilingual Mandarin-English children did not distinguish between languages in VOT (Yang, 2019). Additionally, the predictions of the SLM-r (Flege & Bohn, 2021) suggest that long-lag items of minimally distinct VOT would assimilate or dissimilate but not be stable in such proximity. Thus, the present study asks: do Cantonese-English bilinguals uniformly produce long-lag stops within and across each of their languages? Leveraging the methodology from Chodroff and colleagues (Chodroff & Wilson, 2017, 2018; Chodroff & Baese-Berk, 2019) allows for a new perspec-

[2]Please refer to Tables 3.1 and 3.2 in the previous chapter for a summary of the segmental inventories in both languages.

tive on the structure of variation and nature of mental representation in bilinguals' segment inventories. It also facilitates the study of phonetically similar speech sounds in ways that other paradigms do not. The hypothesis in this chapter was that bilinguals would indeed exhibit crosslinguistic uniformity and leverage articulatory reuse across Cantonese and English.

4.2 Methods

4.2.1 Corpus

This study uses the conversational interview recordings from the SpiCE corpus described in Chapter 2. As a reminder, the corpus comprises recordings of 34 early Cantonese-English bilinguals in both languages. The analysis in this chapter builds on the force-aligned phone transcripts. Please refer to Chapter 2 for additional information about the talkers.

4.2.2 Segmentation and measurement

All instances of prevocalic word-initial /p t k/ were identified from the conversational interview portion of the SpiCE corpus' force-aligned Praat TextGrid transcripts. The identification of tokens was based on the phone tier of the transcripts, and as a result, derives from the forced aligner's identification of stops given a lexical item and its entry in the pronunciation dictionaries. For English, only words with initial stress were included in the initial sample (Lisker & Abramson, 1967).[3] Code-switches out of the interview's primary language were not aligned, and as a result, they do not appear in the phone tier of the TextGrids. This limitation of forced alignment means that Cantonese /p t k/ were only considered if they occurred in the predominantly Cantonese interviews, and likewise for English. The initial total count of /p t k/ across talkers and languages included 10,428 tokens.

[3]Chodroff & Wilson (2017) specifically excludes the extremely high-frequency English word "to." The initial stress requirement implicitly accomplishes this here—while "to" only has one syllable, the most commonly used pronunciation variant in the dictionary is unstressed.

While forced alignment performed reasonably well, anecdotally speaking, it was not perfect. Additionally, forced alignment includes both the closure and release of the stop in the marking of stop consonants. For these reasons, VOT estimates were refined using AutoVOT (Keshet et al., 2014)—a command-line software tool that facilitates automated measurement of positive VOT. AutoVOT identifies the onset and offset of positive VOT within a specified window and with a minimum duration. Here, the minimum allowed VOT was set to 15 ms. This value was selected as the stops under consideration are all long-lag stops, and aspiration values under 15 ms are typical of short-lag stops (Lieberman & Blumstein, 1988). The window used with AutoVOT was defined as the force-aligned segment boundaries plus or minus 31 ms (as recommended by Chodroff & Wilson, 2017). If stops were too close for a 31 ms buffer, the onset of the second stop's window was set as the offset of the preceding window, as TextGrids do not permit overlapping intervals and AutoVOT uses the full TextGrid. This would occur, for example, in cases where a short vowel separates two stops, as may be the case in a phrase like "too tall" in running speech.

After running AutoVOT, instances of /p t k/ were subjected to exclusionary criteria to catch errors and exclude tokens immediately after a code switch. Tokens were excluded if there was substantial enough misalignment such that the AutoVOT offset did not fall within the original force-aligned boundaries of the word (n = 567). Tokens were also excluded if the previous word was unknown (i.e., unintelligible or in a different language; n = 263), if VOT was equal to the minimum value of 15 ms (n = 446), or if tokens had a VOT more than 2.5 standard deviations above the grand mean ($>$ 129.5 ms; n = 191), as in Chodroff & Wilson (2017).

Of the initial sample, 14.1% was excluded, resulting in 8,961 stop tokens, summarized in Table 4.1. Talkers had a median of 97 Cantonese stops (range: 54–194) and 150.5 English stops (range: 73–540). Cantonese stops were culled at a slightly higher rate—they represent 43% of the initial sample, but only 38% of the final, post-exclusions sample. As there were comparable amounts of recorded speech

in each language, the higher number of English stops in both the initial and final sample is likely due primarily to lexical distributional reasons. In addition to reporting on token frequency, Table 4.1 also summarizes the number of word types for each of the segments in each language.

Additionally, English has a greater number of highly frequent /k/-initial word types, while Cantonese /p/ occurs in fewer, less frequent word types in the final sample (n = 60, max token frequency of 97) than English (n = 158, max token frequency of 215).

Table 4.1: The number of stop tokens (overall and range across talkers) and word types for each language and sound category.

Language	**Frequency**	**/p/**	**/t/**	**/k/**
Cantonese	Token (overall)	374	1373	1688
	Token (range)	0–32	17-79	19–116
	Type (overall)	60	157	68
English	Token (overall)	1035	1336	3155
	Range (tokens)	4–96	15–150	52–294
	Type (overall)	158	143	208

4.3 Analysis and results

The articulatory uniformity framework offers solid theoretical grounds for interpreting the structure of VOT variation within and across talkers. This analysis provides a qualitative description and quantifies that structure from a few different angles. Section 4.3.1 describes the ordinal relationship between each of the segments across talkers and languages (i.e., how they are ordered by VOT). Section 4.3.2 reports on a series of pairwise correlations of talker means for each of the three segments in each language. Lastly, Section 4.3.3 comprises the results of a Bayesian mixed-effects model aimed at elucidating the role of language while accounting for variables known to impact VOT.

4.3.1 Ordinal relationships

Prior work with lab and read speech strongly suggests an expected ordinal relationship for VOT across places of articulation, in which /p/ is consistently shorter than /k/ and where /t/ falls in the middle. The argument for this widely attested pattern is based on vocal tract aerodynamics and articulatory constraints (Cho & Ladefoged, 1999). One of the major contributions of Chodroff & Wilson (2017) is that these ordinal relationships are much more constrained than would be expected from a purely ordinal perspective. Ordinal relationships are a starting place, and they represent just one piece of the puzzle.

The results presented in this section suggest that *puzzle* is an appropriate characterization, as talkers largely did not adhere to the expected order. While there is some reason to expect coronals not to pattern accordingly in English, as Chodroff & Wilson (2017) review literature indicating coronal behavior to be more variable across dialects of English, the relationship between /p/ and /k/ is inconsistent across talkers in the SpiCE corpus. Table 4.2 reports the proportion of talkers whose mean VOT values followed the expected /p/ $<$ /t/ $<$ /k/ relationships. Note that one talker (VM25A) did not have any instances of Cantonese /p/ in the final sample. The unexpected results were as follows. Cantonese /t/ is typically longer than Cantonese /p/—the opposite holds for English. Cantonese /k/ tends to be longer than Cantonese /t/, but this is almost never the case for English. The ordering of /p/ and /k/ is a toss-up in both languages.

Prior work with English connected speech reports rates of adherence in the 80-90% range for all pairwise combinations, with the exception of /t/ $<$ /k/ being drastically lower for native English speakers in Chodroff & Baese-Berk (2019). While the English /t/ $<$ /k/ comparison is remarkably low here at 6%, only the English /p/ $<$ /t/ ordering falls in the range that prior work suggests, at 82%. This lack of adherence is apparent in the relative ordering or markers in Figures 4.1 and 4.2, which depict the mean and standard error of VOT for each segment, language, and talker. The goal of Figures 4.1 and 4.2 is to showcase the variety of patterns across individuals and to highlight that a single summary plot of means only would

be inappropriate. In many cases, the standard errors for the different segments in a given talker's panel overlap, as is the case for VM21B in Figure 4.2. Such overlap in the standard errors indicates that strict ordering may not be appropriate here, as there is not a great deal of confidence in the means' ordering. Additionally, talkers do not appear to be consistent across languages. For example, talker VF19B in Figure 4.1 exhibits a clear /p/ $<$ /t/ $<$ /k/ relationship in Cantonese, but a clear /p/ $<$ /k/ $<$ /t/ relationship in English. In fact, only three talkers in the corpus exhibit the same pattern of means across languages (VM21B, VM23A, and VM25A).

Table 4.2: Proportion of talker means that adhered to expected ordinal relationship for VOT: /p/ $<$ /t/ $<$ /k/ mean VOT durations. Note that talker VM25A has no instances of Cantonese /p/ in the final sample.

Language	**/p/ $<$ /t/**	**/t/ $<$ /k/**	**/p/ $<$ /k/**	**/p/ $<$ /t/ $<$ /k/**	**n**
Cantonese	0.24	0.61	0.39	0.15	33
English	0.82	0.06	0.47	0.00	34

4.3.2 Pairwise correlations

To examine the relationship between stops within and across languages, 15 pairwise Pearson's *r* correlations were calculated using talker means. Each correlation compares talkers means for two different segments. The full set of pairwise correlations includes three within English, three within Cantonese, and nine comparing English to Cantonese. These correlations are reported along with Holm-adjusted *p*-values to account for multiple comparisons. This analysis uses the *psych* (Revelle, 2021) package in R (R Core Team, 2020). As in Chodroff & Wilson (2017), this correlation analysis aims to elucidate within-talker invariance and between-talker variability. Tight correlations for between-talker means signals within-talker invariance, while a wide spread between points signals between-talker variability. While using means ignores information about within-category variability—a major shortcoming of this approach—prior work sets up strong, clear expectations about the pattern of mean values for long-lag VOT (Chodroff & Wilson, 2017; Cho

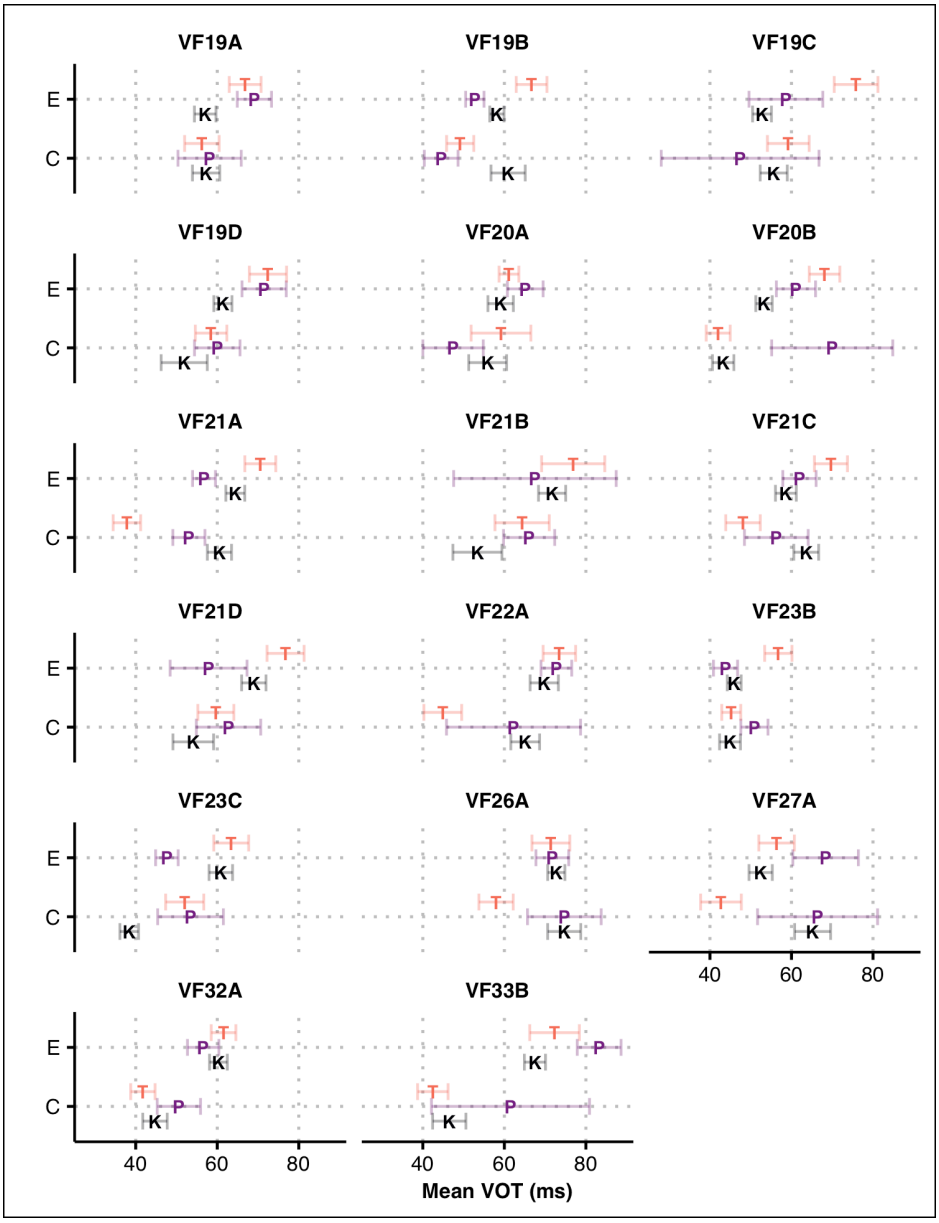


Figure 4.1: This figure depicts the ordinal relationships for the female talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows.

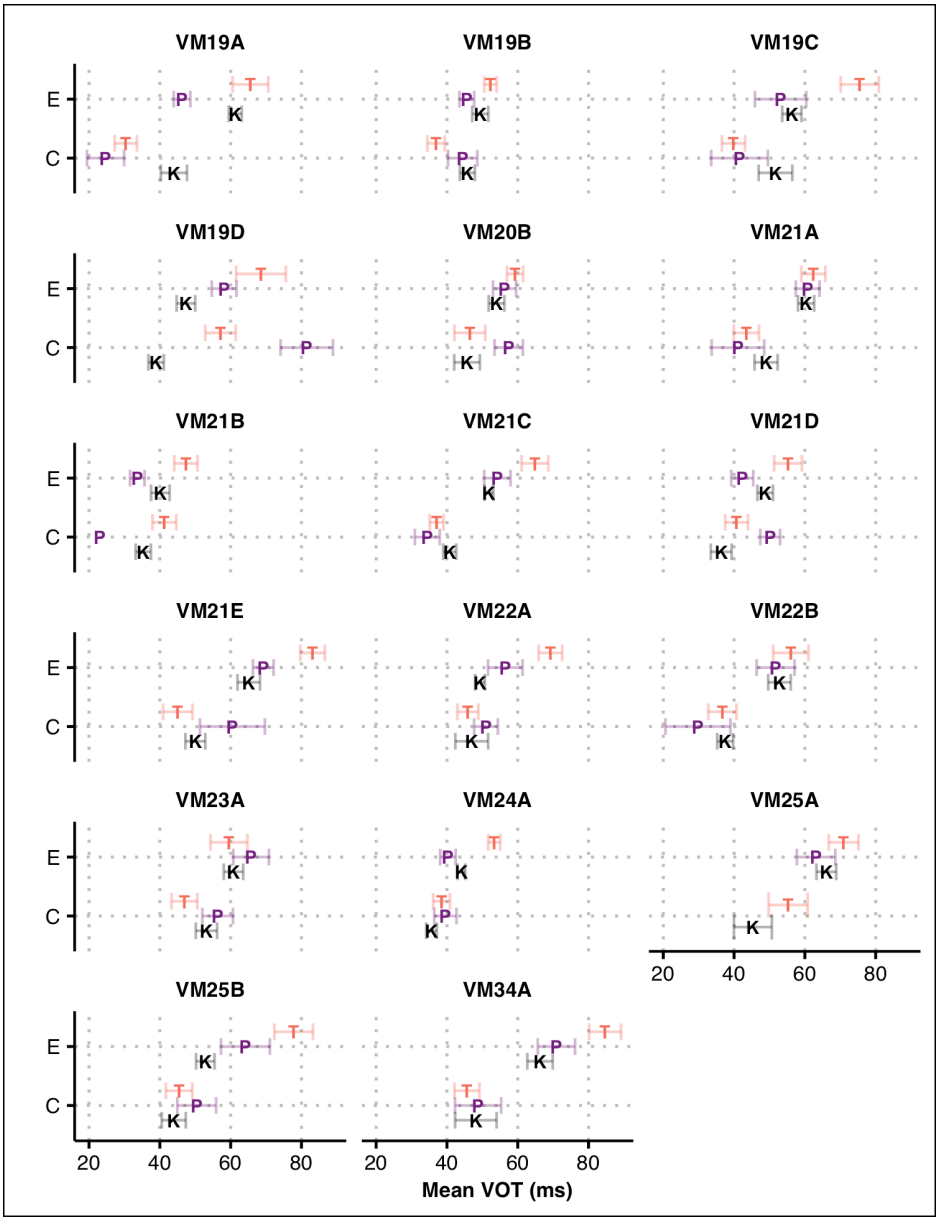


Figure 4.2: This figure depicts the ordinal relationships for the male talkers. Each panel depicts the mean VOT and standard error for VOT for each segment, with E(nglish) and C(antonese) in separate rows. VM25A had no /p/ tokens.

& Ladefoged, 1999). The mixed-effects analysis in the following section takes this variation into account.

Table 4.3 summarizes the output of all 15 correlations in text form. Figure 4.3 depicts the six within-language correlations and Figure 4.4 depicts the across-language correlations. While there is some evidence for both within- and across-language structured variation, the correlations reported here are considerably lower than prior work on English read speech, where within-language comparisons had $r > 0.7$ (Chodroff & Wilson, 2017; Chodroff & Baese-Berk, 2019). With the exception of the English /p/ $\sim$ /k/ ($r = 0.70$, $p < 0.001$), all of the correlations here were either moderate ($0.5 < r < 0.7$; $p < 0.01$; n = 6) or weak and non-significant (n = 8). Within-English correlations were the most consistent—all three had r at or above 0.65 ($p < 0.001$). Of the within-Cantonese correlations only /p/ $\sim$ /t/ was significant ($r = 0.59$; $p = 0.003$). This disparity across English and Cantonese for the same set of talkers highlights the need to study a variety of typologically distinct languages to understand how the structure of variation *varies*.

Two of three across-language correlations at the same place of articulation were significant, with moderate r values (/p/ $\sim$ /p/: $r = 0.62$, $p = 0.001$; /k/ $\sim$ /k/: $r = 0.57$, $p = 0.004$). Notably, the correlation for /t/ $\sim$ /t/ was not significant ($r = 0.40$, $p = 0.11$). Of the across-language comparisons that do not share a place of articulation, only one was significant—Cantonese /k/ $\sim$ English /p/ ($r = 0.58$, $p = 0.003$). Again, /t/ is absent here.

Chodroff & Wilson (2017) also repeat the correlation analysis in a way the coarsely accounts for speaking rate. This consideration is important, as the local speaking rate is known to influence long-lag VOT in spontaneous speech (Stuart-Smith et al., 2015) and because prior work demonstrates both talker and language effects on speech rate (Bradlow et al., 2017). In comparing the two versions of the correlation analysis, Chodroff & Wilson found that "the magnitudes of the correlations among voiceless stops did not deviate from the original magnitudes, demonstrating that differences among talkers in the realization of these sounds cannot be reduced to talker-specific speaking rates" (2017, p. 34).

Table 4.3: All 15 correlations are based on raw mean VOT—and separately, residual VOT after accounting for speaking rate—for each talker, language, and segment. Each row indicates the comparison, Pearson's r, and the Holm-adjusted p-value given 15 comparisons.

		Raw		**Residualized**	
Type	**Comparison**	r	p	r	p
Within-Cantonese	Cantonese /p/ ∼ Cantonese /t/	0.59	0.003	0.59	0.003
Within-Cantonese	Cantonese /p/ ∼ Cantonese /k/	0.44	0.08	0.55	0.01
Within-Cantonese	Cantonese /t/ ∼ Cantonese /k/	0.38	0.11	0.34	0.21
Within-English	English /p/ ∼ English /t/	0.65	<0.001	0.63	0.001
Within-English	English /p/ ∼ English /k/	0.70	<0.001	0.70	<0.001
Within-English	English /t/ ∼ English /k/	0.66	<0.001	0.60	0.002
Across-language	Cantonese /p/ ∼ English /p/	0.62	0.001	0.57	0.01
Across-language	Cantonese /t/ ∼ English /t/	0.40	0.11	0.35	0.21
Across-language	Cantonese /k/ ∼ English /k/	0.57	0.004	0.54	0.01
Across-language	Cantonese /p/ ∼ English /t/	0.41	0.11	0.29	0.31
Across-language	Cantonese /p/ ∼ English /k/	0.40	0.11	0.29	0.31
Across-language	Cantonese /t/ ∼ English /p/	0.43	0.08	0.37	0.20
Across-language	Cantonese /t/ ∼ English /k/	0.37	0.11	0.27	0.31
Across-language	Cantonese /k/ ∼ English /p/	0.58	0.003	0.59	0.003
Across-language	Cantonese /k/ ∼ English /t/	0.38	0.11	0.37	0.20

A similar analysis was done here, using means calculated over *residual* VOT values from a simple linear regression in which VOT was predicted by average phone duration within the word. Average phone duration is a proxy for speech rate. It was calculated as the difference between the word's AutoVOT-estimated onset and force-aligned offset, divided by the number of segments in the canonical form of the word.[4] The results—Pearson's r and Holm-adjusted p values—are reported in the rightmost columns of Table 4.3. Qualitatively, the results mostly mirror the correlations based on raw VOT, though there are some minor differences in significance and magnitude. Both versions of the analysis support a conclusion in which the patterns are weak overall.

[4]The canonical form was pulled from the pronunciation dictionaries used during forced alignment. In cases where there was more than one form in the dictionary, the entry with the higher number of phones was used.

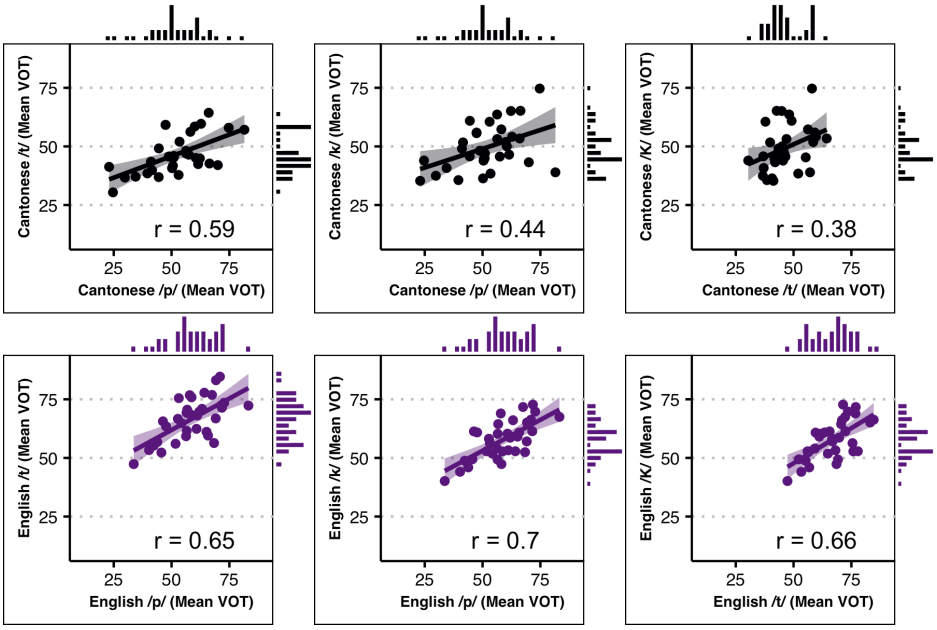


Figure 4.3: Correlations for within-language pairwise comparisons of raw mean VOT are depicted with points representing talker means for the segments on the x and y axes and superimposed regression lines. The margins display histograms for each of the axes. Within-Cantonese comparisons are depicted in black, and within English comparisons in purple. *Note that while some of the distributions in the margins appear different, they are not. This is an artifact of plotting the same distribution on different axes in different plots—they only appear mirrored.*

While these relationships indicate some degree of articulatory reuse, the overall picture is far from compelling, particularly when considered alongside the analysis of the ordinal relationships in Section 4.3.1. Compared to prior work, these correlations are less consistent and generally weaker, indicating that the uniformity constraint discussed in Section 4.1.3 may not be as robust as previously argued. This point will be returned to in this chapter's discussion (Section 4.4).

The next steps in Chodroff & Wilson's (2017) methods focus on validating the strength of the correlations. Their approach includes estimating confidence

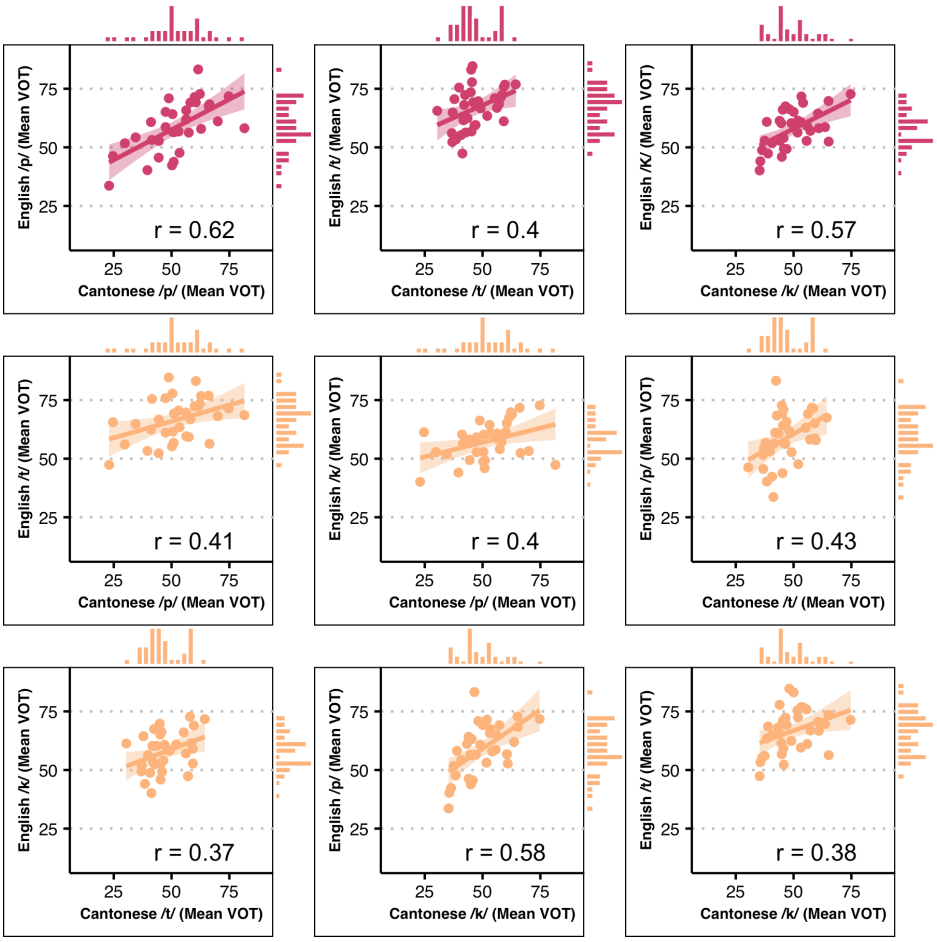


Figure 4.4: Correlations for the across-language comparisons of raw mean VOT are depcited in the same manner as Figure 4.3. Comparisons at the same place of articulation are depicted in pink, and comparisons at different places of articulation are in orange.

intervals for the correlations using a bootstrap procedure. In a later paper, Chodroff et al. (2019) simulate what would emerge from a purely ordinal relationship (i.e., a system where the only requirement was the relative VOT ordering of segments) between stops and demonstrate that the observed correlations are much stronger—ultimately arguing for a uniformity constraint on phonetic variation. Given that the correlations found in this chapter are drastically lower and largely do not adhere to the expected ordinal relationships, the remainder of this analysis takes a different approach.

4.3.3 Linear mixed-effects model

The analysis in this section leverages a Bayesian multilevel linear model to elucidate the sources of variation within and across talkers. As Bayesian modeling emphasizes the estimation of effect magnitudes, the model can be used to assess how talkers' sound categories compare to one another while simultaneously accounting for factors known to influence long-lag VOT, such as speaking rate and prosodic position. This section builds on the frequentist mixed-effects model analysis in Chodroff & Wilson (2017). Also, this modeling approach is more in line with the generative modeling approach advocated for by Haines et al. (2020)—in a way that the correlation and ordinal relationships analyzed in the preceding sections are not. Specifically, this approach retains the variation lost when working with means and also uses a response variable distribution that aligns with the constraints of the variable in question.

This section proceeds as follows. First, Bayesian modeling and Bayesian inference are described in broad terms. References are provided that point the reader to further reading on the topic. Second, the structure of the model used in this chapter is described and motivated. Lastly, the results of the model are reported. All code used in this analysis is available on GitHub, at https://github.com/khiajohnson/dissertation.[5]

[5]Note that this repository is currently private.

Bayesian inference

The corpus sample was analyzed with a Bayesian linear mixed-effects model using the brms package in R (Burkner, 2017; R Core Team, 2020). The brms package provides a simple, formula-based interface to Stan—a widely used probabilistic programming language for estimating Bayesian statistical models via Hamiltonian Monte Carlo and No-U-Turn Sampling (Stan Development Team, 2021). Bayesian models are desirable in the case of modeling multilingual VOT for both practical and theoretical reasons. Practically, they are not subject to the convergence problems that plague comparable frequentist models. Theoretically, they allow for graded statements regarding the strength of evidence for all parameters, both population-level (i.e., fixed effects) and group-level (i.e., random effects) parameters, as well as derived parameters (e.g., some combination of existing parameters). While there are many other benefits, readers are referred to Vasishth et al.'s (2018) recent in-depth tutorial paper on Bayesian modeling in the phonetic sciences for further argumentation.

Inference in Bayesian models is based on the posterior distributions of parameters in the model, which reflect the range and probability of credible values for parameters. The posterior combines information from prior knowledge and the likelihood of observing the data given the specified model. While some Bayesian models use detailed and specific prior knowledge, it is perhaps more common to use weakly informative, regularizing priors (Gelman et al., 2017), which constrain the parameter space to possible values and down weight extreme or unlikely values, while also not biasing the model toward any specific outcome. The model described in the next section uses regularizing priors.

While Bayesian modeling typically emphasizes parameter estimation in a probabilistic framework, there are decision criteria that facilitate hypothesis testing. One such technique is to use Kruschke's (2011) ROPE+HDI method. The ROPE is a "region of practical equivalence" surrounding the null value. HDI stands for highest density interval, and it is typically used to describe Bayesian posterior distributions. Kruschke's (2011) decision criterion is simple: if the HDI falls entirely

within the ROPE, then the null value can be accepted; if the HDI falls entirely outside the ROPE, then it can be rejected; if there is overlap, then a decision should be withheld. In the case of standardized data, Kruschke (2011) recommends the convention of setting a ROPE to be $[-0.1, 0.1]$—half the size of a small Cohen's *d* effect. This decision criterion provides a useful scaffolding for interpreting the magnitudes of standardized effects when presented alongside the posterior distributions.

Modeling multilingual VOT

VOT was modeled using a Bayesian linear mixed-effects model. The model used in this section is provided in Equation 4.1, below. While the model is not the maximal model, it instead follows guidelines for parsimonious model building, in which the parameters of direct interest are included as random slopes, and the controlling parameters are not (see: Barr et al., 2013; Bates et al., 2018). The controlling parameters are only included as population-level "fixed" effects.

One of the main benefits of multilevel modeling is partial pooling, where information for different levels of a variable is shared across those levels. McElreath argues that "any batch of parameters with exchangeable index values can and probably should be pooled [where exchangeable] just means the index values have no true ordering" (2020, p. 435). Pooling can be done for both intercepts and slopes, which in turn allows both to vary by group. While Bayesians tend to refer to the random effects structure in terms of partial pooling, it is important to note that partial pooling and random effects refer to the same thing, regardless of whether the analysis is frequentist of Bayesian.

The model was specified as follows. First, Equation 4.1 gives the formula used in *brms*. Immediately afterward is a description of the model's parameters and how they were specified.

$$\text{VOT} \sim 1 + \text{Place} \times \text{Language} + \text{Average Phone Duration} + \text{Pause} + (1 + \text{Place} \times \text{Language} \mid \text{Talker}) + (1 \mid \text{Word}) \tag{4.1}$$

VOT was the dependent variable—it was standardized (i.e., centered and scaled) in order to facilite the specification of priors and a ROPE.

Place encodes place of articulation for the stops and has three levels. Following Chodroff & Wilson (2017), Place was weighted effect coded in order to account for unequal sample sizes across the three levels and to faciliate the interpretation of the simple effects in light of the interaction term (Brehm & Alday, 2021). Specifically, weighted effect coding ensures that a simple effect is equivalent to what the main effect would be in a model without the interaction. Coding was implemented using the *wec* R package (Nieuwenhuis et al., 2017), and leads to reporting Place effects for T (weights: /p/ $= -1.92$, /t/ $= 1$, /k/ $= 0$) and K (weights: /p/ $= -3.44$, /t/ $= 0$, /k/ $= 1$).

Language is a binary variable that encodes whether the VOT measurement comes from an English or Cantonese word. As with Place, Language was also weighted effect coded (weights: Cantonese $= -1.61$, English $= 1$).

Average Phone Duration represents the average duration of phones within the word. It was calculated as the difference between the word's AutoVOT-estimated onset and the force-aligned word offset, divided by the number of segments in the canonical form of the word (see Footnote 4). As noted in Section 4.3.2, average phone duration serves as a proxy for local speaking rate. A word-internal measure is desirable here, as many tokens were preceded by a pause and thus lack the necessary preceding context to calculate the speaking rate. Average phone duration was standardized (i.e., centered and scaled).

(Preceding) Pause is a binary variable that indicates whether or not the token occurred after a pause or not. Pauses were identified using the force-aligned transcripts and include instances where the preceding phone was "sil" (silence) or "sp" (silent pause). Pause was also implemented with weighted effect coding (weights: False $= -0.33$, True $= 1$).

Word indicates the word that the VOT measurement comes from.

Talker indicates which of the 34 talkers produced the item.

The interaction for Language $\times$ Place was included in the model—it directly addresses the research question relating to whether or not bilinguals maintain a difference across languages for these sounds. Additionally, the model includes partial pooling (i.e., random intercepts) for Word and Talker, as well as for the Language, Place, and Language $\times$ Place terms (i.e., random slopes).

As noted above, the priors were set to be weakly informative and regularizing, motivated by the discussions in Gelman et al. (2017) and McElreath (2020). Specifically, the priors were set as follows.

Intercept Student's *t* distribution with $\nu = 3$, $\mu = 0$, and $\sigma = 2.5$

Population-level parameters Normal distribution with $\mu = 0$ and $\sigma = 1$

Group-level standard deviations Half Student's *t* distribution with $\nu = 3$, $\mu = 0$, and $\sigma = 2.5$

Group-level correlations LKJ distribution with $\eta = 2$

The model was fit using four chains with 5,000 iterations (2,500 warmup) for a total of 10,000 post-warmup samples. The chains were well-mixed, based on a visual inspection of trace plots, a lack of divergent transitions, and $\hat{R}$ values below 1.05. Additionally, the effective sample size was sufficiently large for all parameters (for discussion, see Vasishth et al., 2018).

Results

A summary of the model's population-level parameters is provided numerically in Table 4.4 and visually in Figure 4.5. The population parameters indicate that VOT is modulated by language, local speaking rate, and the presence of a preceding pause. Recall that the categorical population-level parameters were weighted effect coded—this facilitates the interpretation of simple effects in the presence of interaction terms.

The overall effect of Language indicates that English long-lag stops were produced with longer VOT than Cantonese ($\beta = 0.16$, 98.9% HDI outside ROPE). The effect of Place is not consistently modulated by Language, as both of the Place $\times$ Language interaction terms overlapped substantially with the ROPE. This interpretation is not, however, supported by the model predictions summarized in Figure 4.6, which shows the conditional effects for Place and Language—that is, the predicted means for each of the six combinations. The predictions look exactly as would be expected in the case of an interaction—the distance between means is absent for /p/, larger for /t/, and still larger for /k/. That this doesn't emerge in the parameter summary in Figure 4.5 and Table 4.4 may be due to how the categorical variables were coded. In cases such as these, McElreath (2020) argues that parameter summaries are often less informative (and useful) than model predictions.

To dig into this interaction and justify its inclusion in the model, a second model was fit without the interaction term. All other aspects were identical to the model described in Equation 4.1. The models with and without the interaction were then compared using the expected log pointwise predictive density (ELPD Vehtari et al., 2017), as implemented in the *loo* R package (Vehtari et al., 2020). The result of this comparison demonstrates the importance of the interaction term—it substantially improves the model's predictive accuracy (ELPD difference: -13.4, SE difference: 5.7). This result suggests that the interaction visible in the model's predictions in Figure 4.6 is valid, even if the parameterization of the population-level variables does not show such an outcome.

Returning to a summary of the original model, the control parameters behaved

as expected. VOT was longer when the local speaking rate was slower. This effect is captured by the relatively high posterior mean for Average Phone Duration ($\beta =$ 0.32, 100.0% HDI outside ROPE). VOT was also longer after a pause, though the effect size was considerably smaller than for speaking rate ($\beta = 0.12$, 94.0% HDI outside ROPE).

Table 4.4: Population parameter summary.

Parameter	**Est.**	**95% HDI**	**% Outside ROPE**
Intercept	0.18	[0.09, 0.28]	95.1
Place (T)	0.05	[-0.03, 0.13]	11.1
Place (K)	-0.02	[-0.07, 0.03]	0.1
Language (English)	0.16	[0.11, 0.21]	98.9
Average Phone Duration	0.32	[0.30, 0.34]	100.0
Preceding Pause (True)	0.12	[0.09, 0.16]	94.0
Place (T) $\times$ Language (English)	-0.01	[-0.07, 0.04]	0.2
Place (K) $\times$ Language (English)	0.04	[0.00, 0.08]	0.3

While the main takeaway from the population parameters is the difference in VOT across languages, the model also offers insight into the sources of variation in this population. A summary of the variability in the model's grouping parameters is provided numerically in Table 4.5 and visually in Figure 4.7. The largest source of variability in the model is in the Word intercepts ($\beta = 0.44$). The second-largest source of variability is in the Talker intercepts ($\beta = 0.25$). While there is variability across talkers in the random slopes, few are meaningfully different from the corresponding population-level parameters—this is evident in Figure 4.8, which depicts the by-Talker intercept and slope deviations from the model. The intercept can be interpreted as the model estimate when all parameters are at their zero value—in the case of continuous parameters, this is zero, while in the case of the categorical variables, it is the weighted grand mean. The intercepts in 4.8 thus reflect individuals' deviations from this overall intercept. A sizable plurality of talker intercept posterior distributions falls outside of the ROPE, while the vast majority of the by-talker slopes overlap substantially or fall entirely within the ROPE. In line with Chodroff & Wilson (2017), this result highlights between-talker variability

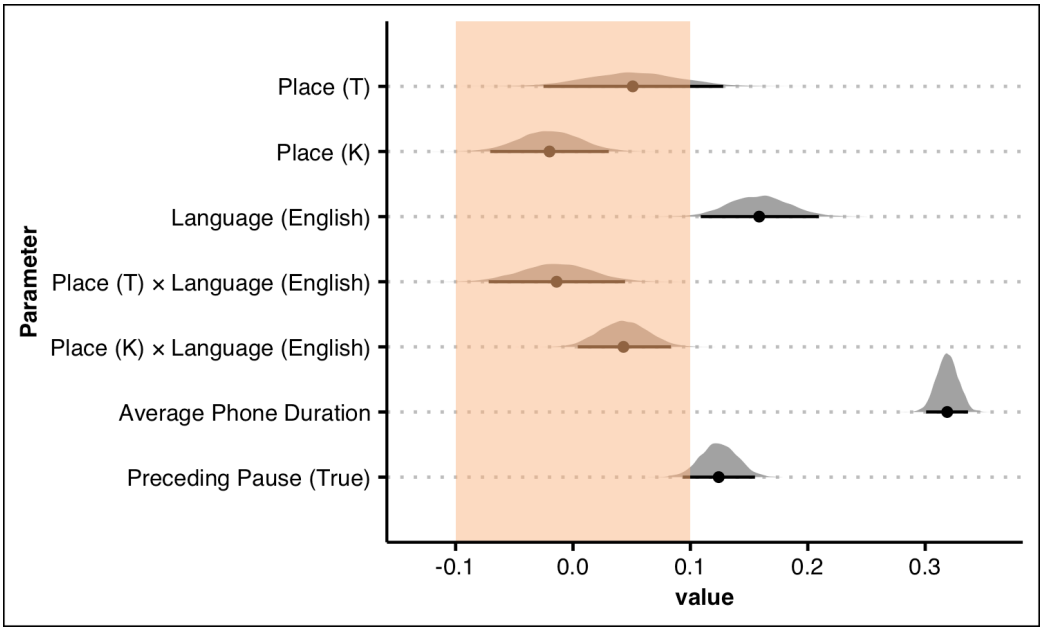


Figure 4.5: This figure depicts the 95% HDI posterior distributions for each of the population-level parameters, with the posterior mean indicated by the dot. The orange shaded section represents the ROPE. Recall how to interpret ROPEs—accept the null if posterior is fully within bounds and reject it if the posterior is fully outside ROPE; otherwise, withhold a decision.

and within-talker stability.

At first glance, the results of the mixed-effects model are puzzling in how they seem to contradict what is apparent in the raw data described and analyzed in Sections 4.3.1 and 4.3.2. While there seems to be a cross-language difference in VOT for /t/ and /k/, the model here suggests there is more uniformity than those analyses would support. Why? Given the uncontrolled and spontaneous nature of this speech data, and the large amount of variability captured in partial pooling for words, a simple answer to this question is that talkers simply use different words. To test this, a third model was fit without Word intercepts (but otherwise identical to Equation 4.1). Qualitatively, the exclusion of Word intercepts drastically

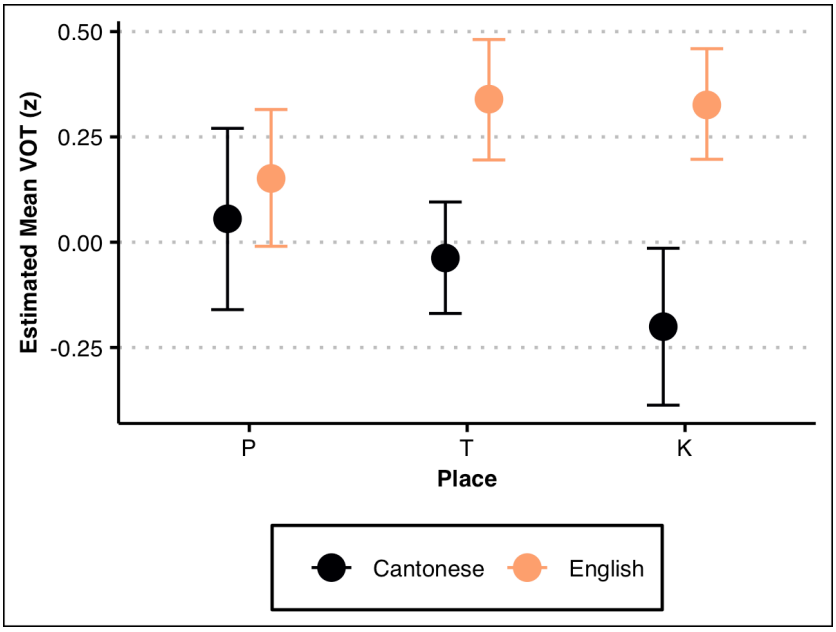


Figure 4.6: This figure depicts the model's predicted value and standard error of the predicted value for each of the places of articulation by language, using the fitted method in *brms*' conditional effects function. Notably, the error overlaps almost completely for /p/, but not at all for /t/ and /k/.

Table 4.5: Group parameter variability summary.

Group	**Parameter S.D.**	**Est.**	**95% HDI**
Word	Intercept	0.44	[0.40, 0.50]
Talker	Intercept	0.25	[0.19, 0.32]
Talker	Place (T)	0.09	[0.05, 0.14]
Talker	Place (K)	0.06	[0.04, 0.09]
Talker	Language (English)	0.08	[0.05, 0.11]
Talker	Place (T) × Language (English)	0.05	[0.02, 0.08]
Talker	Place (K) × Language (English)	0.04	[0.02, 0.06]

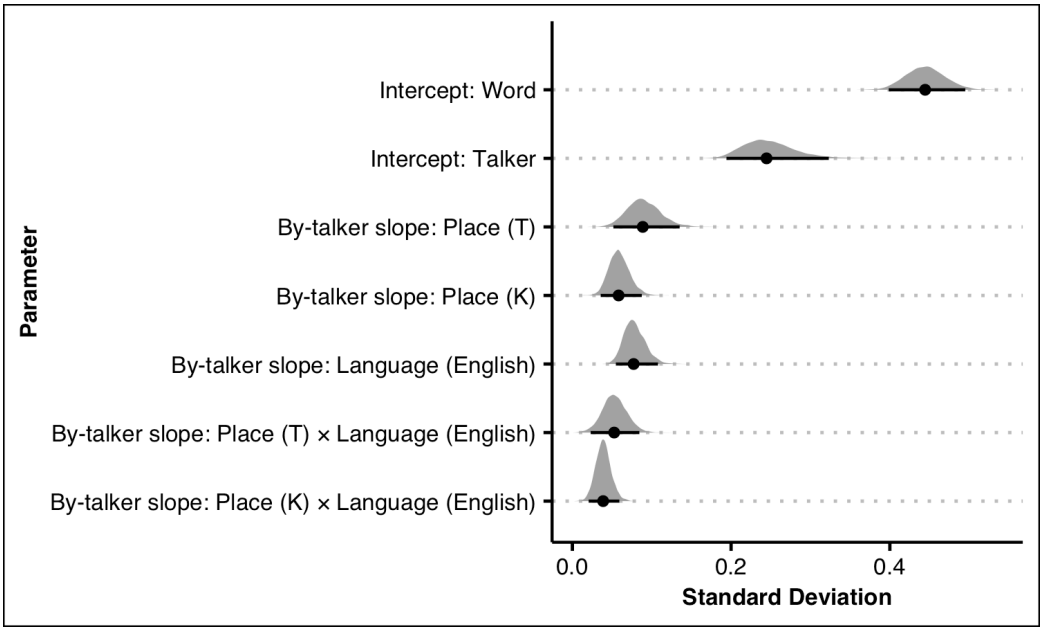


Figure 4.7: This figure depicts the posterior distributions for the standard deviation of each of the grouping parameters, both intercepts and slopes.

changes the model output. All of the remaining standard deviation parameters increased—the standard deviations for Talker intercepts and slopes for Place, Language, and Place $\times$ Language interaction. Additionally, the interaction between Place and Language predicted by the original model disappears. Without partial pooling for words, only the difference between English and Cantonese /t/ remains. These differences indicate that the apparent discrepancy between the mixed-effects model and the ordinal and correlational analyses can be explained via differences in word distributions across talkers. Essentially, talkers vary in the words they use, and this variation may render the results of Sections 4.3.1 and 4.3.2 somewhat misleading.

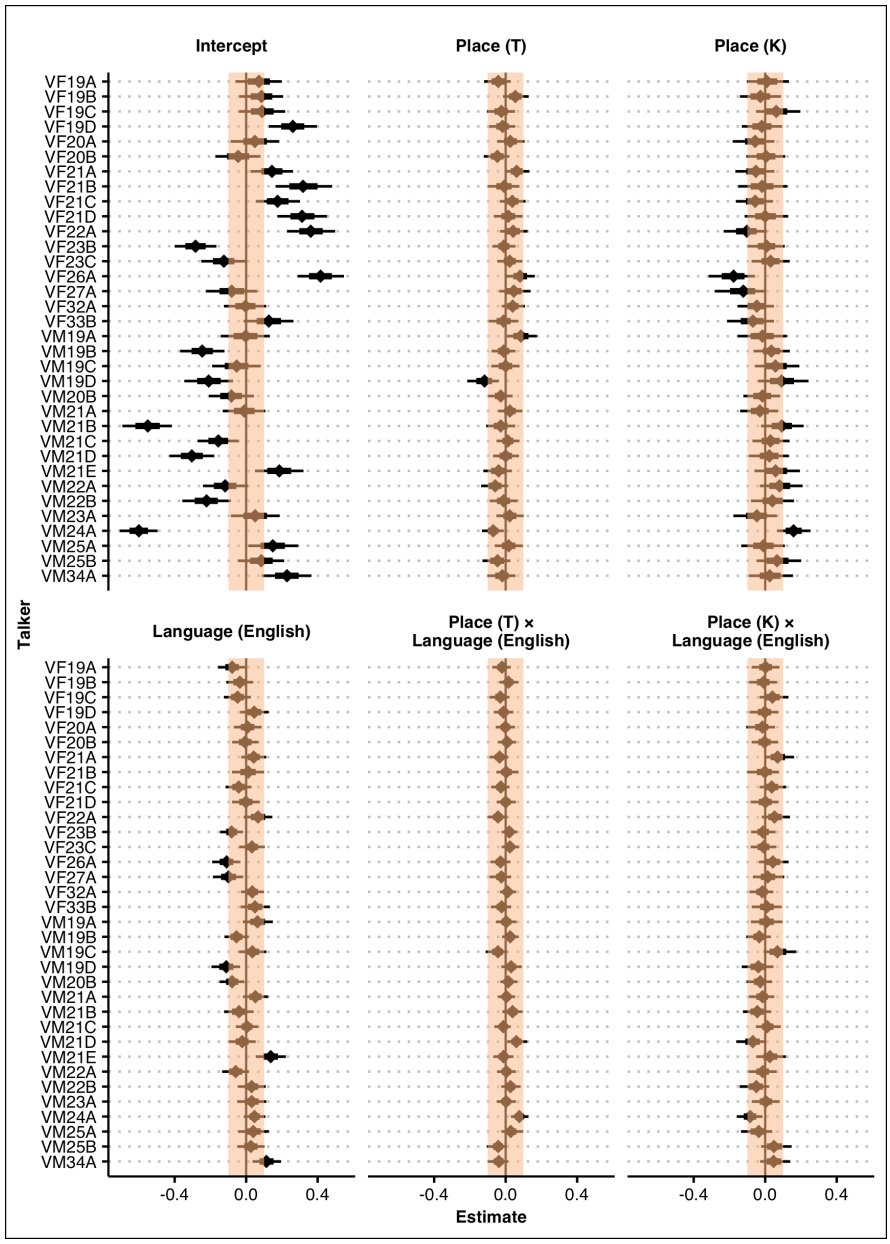


Figure 4.8: This figure depicts the 95% HDI for each talker across the talker intercepts and by-talker slope terms. The shaded orange interval represents the ROPE.

4.4 Discussion

This chapter reports on a study of long-lag stops in Cantonese-English bilingual speech from the SpiCE corpus described in Chapter 2. It leverages the uniformity framework to assess VOT similarity within and across languages from a few different angles—via ordinal relationships, pairwise correlations, and a Bayesian linear mixed-effects model. In broad strokes, the evidence for uniformity both within and across languages was somewhat mixed. Yet, uniformity was apparent in the mixed-effects model—along with a clear crosslinguistic difference—which is arguably the most robust and reliable of the methods used here (Haines et al., 2020).

An analysis of ordinal relationships between the duration of mean VOT for talkers in each language was inconclusive. Talkers largely did not adhere to the expected order, and further, talkers were not internally consistent across languages. This counters prior work establishing strong rates of adherence. However, the difference may be attributable to speaking style. Most of the work documenting ordinal relationships among mean stop VOT is based on isolated word production and read speech (e.g., Chodroff & Wilson, 2017; Cho & Ladefoged, 1999; Lisker & Abramson, 1964). In this body of work, Chodroff & Wilson (2017) found weaker correlations in reading compared to isolated word production, indicating that speech style plays a role. The yet weaker correlations in this chapter could be interpreted as extending that pattern to a more casual style. Another possible (complementary) account of the results in this chapter is one based on distributions of words produced by the talkers. Both isolated word production and reading offer tight control over the words produced, which means that the distribution of words produced is constant across talkers. The same cannot be said of spontaneous speech, in which talkers produce different words and utterances over the course of conversational interviews. The mixed-effects model in Section 4.3.3 highlights how this is a crucial difference across talkers in this chapter. In this sense, speaking style is an important factor, as it impacts both form and content.

The correlation analysis offers a slightly more nuanced take on structure, and in doing so, provides evidence for within-language and, to a lesser extent, across-

language uniformity. Across the board, the correlation magnitudes were weak or moderate, which differs from the strong and clear within-English patterns observed in prior studies (Chodroff & Wilson, 2017; Chodroff & Baese-Berk, 2019). The within-English comparisons were consistently moderate and significant, which replicates prior work. The within-Cantonese and across-language comparisons do not offer nearly as clear a picture. While there is some evidence of structure, particularly for /p/ and /k/ across languages, the evidence for tightly structured variation is far from compelling. This result was surprising, given the typological survey of VOT relationships in Chodroff et al. (2019), which found uniform structure at the level of languages. Recall that the prediction for this chapter included both within-language and across-language uniformity.

While the murky outcome of the ordinal relationships and correlational analyses was largely unexpected, observing a different outcome for a different speech style is not without precedent. For example, the correlation magnitudes that Chodroff & Wilson (2017) found for connected, read speech were not as strong as those for isolated word production. It follows that an even less formal connected speech style would lead to even weaker relationships—likely because of an increase in variability. Attending to speaking style is likely one of the main factors accounting for why lab and corpus results often differ (Gahl et al., 2012; Chodroff & Wilson, 2017); and, similarly, for corpus studies of monolingual and bilingual speech (Johnson, 2019).

The Bayesian mixed-effects model offers insight into sources of variation, including the specific role that language and place of articulation play in accounting for VOT variation. The model showed a clear difference in VOT by language, with English VOT being consistently longer for /t/ and /k/. While the introduction to this chapter reported on prior work indicating Cantonese might be longer (Clumeck et al., 1981; Lisker & Abramson, 1964), those numbers were not necessarily appropriate for the speaking style of the SpiCE corpus, particularly given the differences in English VOT across styles (Stuart-Smith et al., 2015). Interestingly, the model showed a consistent difference across languages for VOT, and very few talkers

deviated from this overall pattern in a meaningful way. This result suggests that the population-level account provides an appropriate generalization across talkers. While there is precedent for languages having different long-lag VOT settings (Chodroff et al., 2019), it is not immediately clear *why* English and Cantonese, in particular, would show different patterns on a within-speaker basis. It could be due to lexical distributional reasons, broad language timing differences, or underlying representational differences—such explanations would be mere speculation at this point. A much greater amount of variation in the model is captured by variable intercepts for word and talker. The model, then, supports the argument for uniformity—between-talker differences vary drastically, while talkers tend to be more internally consistent. Internal consistency, in this case, seems to be more on a "macro" level. That is, talkers with longer /p/ VOT tend to also have longer /t/ and /k/ VOT in both languages (as with speech rate in Bradlow et al., 2017). This kind of macro internal consistency says nothing about the specific ordinal relationships across languages or sound categories, just the general ballpark that VOT production falls into. This macro level, however, may ultimately be what listeners have access to for talker identification—this idea will be expanded upon in the next paragraph.

In light of the evidence for uniformity, the small but consistent difference across languages is worth some attention. While Section 4.3.3 models standardized VOT, back-transforming the value into the original units suggests a difference across languages of approximately 4 ms. This is a relatively small difference, yet, it is worth flagging that this kind of difference is often smallest in spontaneous speech compared to lab speech. Regardless, a difference of this magnitude is not likely to be perceptible in categorization (and similar) paradigms. Work by McMurray et al. (2002), however, demonstrates a gradient and fine-grained effect of processing VOT in increments as small as 5 ms. In this study, McMurray et al. monitor participants' eye moments in an experiment using the visual world paradigm to ascertain how small VOT differences impact lexical access. The crucial result is participants access the correct word, but that modulating within-category VOT

impacted processing difficulty. Further, as research on mergers in sound change has demonstrated, individuals do not always perceive differences that they produce (Yu & Zellou, 2019; Cheng et al., 2021). As such, perception may not be the best indicator of whether or not this size difference is meaningful in practice. If this difference is indeed meaningful and bears out in future work, it carries implications for how similar sound categories are represented and discussed in the literature. If talkers can maintain such small distinctions across languages, it would reiterate the rarity of assimilation for early bilinguals and necessitate a broader version of models like SLM-r, that account for a wider variety of multilingual backgrounds. This conclusion essentially questions whether full assimilation actually occurs in the speech of early bilinguals, and as a result, questions its utility for this kind of population. Partial and context-dependent assimilation (i.e., due to interference) seem to be more fruitful directions.

Another possibility is that the underlying laryngeal gesture is "the same" but subject to global language timing factors. That is, talker-internal and language-internal factors both influence how VOT manifests. The study in Bradlow et al. (2017) offers an example of this dual influence in the case of speech rate, using native and non-native speech from the ALLSSTAR corpus. In this study, Bradlow et al. demonstrate that talkers who speak faster in their first language (L1) tend to also speak similarly fast in their L2. As this study examined a wide variety of L1s (the L2 was always English), Bradlow et al. also demonstrate differences across languages, with some L1s tending to be slower and others faster. This interpretation could be applied fairly transparently to the study in this chapter: talkers with long VOT in one language would also have long VOT in the other language, even if they maintain a difference between languages.

Yet another possibility is that the VOT specification may be even more distinct underlyingly (i.e., $>$ 4 ms) but ultimately brought closer together by the bilingual language mode of the SpiCE interviews. While the interviews were set up such that sentence reading and storyboard narration tasks preceded the conversational interviews (see Section 2.2), and thus helped talkers get into the language mode of the

interview, the context is nonetheless bilingual—it promotes a bilingual language mode (see Grosjean, 2011). In this context, observing a meaningful difference across languages for VOT suggests that under different circumstances, such a difference might be more pronounced. While this account seems to contradict the one offered in the preceding paragraph, it does not specify where the difference arises from—it could be in the representation of VOT and/or in timing factors. The accounts are thus not necessarily contradictory.

The results presented in this chapter provide some support for a crosslinguistic uniformity constraint, in addition to providing an empirical description of bilingual long-lag stops. The weaker (or merely "macro") constraint on within-talker variability compared to prior work has implications for representation and perception. Tracking a uniformity-like pattern has been proposed as a mechanism for rapidly adapting to speech across languages (Reinisch et al., 2013), and in multilingual talker identification (Orena et al., 2019). Further, uniformity in structure is useful for talker identification within a single language (Ganugapati & Theodore, 2019). While these accounts are straightforward to interpret in the context of clear and strong relationships, the chapter raises questions about how useful fine-grained structure is in the case of spontaneous speech. It would be worth exploring in future work whether the "macro" structure discussed above is sufficient to confer a benefit in talker identification or if the tight structure of uniformity is necessary. Given the presence of macro structure in the results—as well as the importance of salient factors like pitch (over more subtle factors; Perrachione et al., 2019)—it seems that macro structure might be sufficient. If this interpretation stands in perception, it would lend insight and nuance into the utility of uniformity as a perceptual strategy in real communicative contexts.

Overall, this chapter highlights the need to study spontaneous speech and demonstrates the utility—and some limitations—of the uniformity framework for better understanding crosslinguistic similarity. This chapter also provides evidence for the speculations about what drives multilingual talker identification (Orena et al., 2019) outlined in Section 1.2.

Chapter 5

Discussion and Conclusion

What are the consequences of a shared phonetic space in the linguistic systems of bilinguals? What is shared? What is kept separate? And, how can methods couched in the study of crosslinguistic influence provide insight into these areas? These questions sit at the core of this dissertation and are approached from two different angles in Chapters 3 and 4, using the data set described in Chapter 2. While this dissertation focuses on describing and understanding the bilingual speech signal in production, the uniting motivation comes from how the signal is perceived. As such, this chapter proceeds as follows. Section 5.1 recapitulates the main points of the content chapters of this thesis, emphasizing the conclusions that are unique to each chapter. Section 5.2 dives into a more general discussion, highlighting how the studies conspire together to inform a broader understanding of how variation is structured in bilingual speech production. Additionally, implications for perception are considered. Section 5.3 makes note of limitations, and Section 5.4 highlights some of the hypotheses that this dissertation generates but does not answer. Lastly, Section 5.5 concludes by summarizing the key contributions of this dissertation to the fields of phonetics, psycholinguistics, and bilingualism.

5.1 Recap

Chapter 2 introduces a new speech corpus, developed as a part of this dissertation. The SpiCE corpus of Speech in Cantonese and English comprises high-quality recordings and transcripts of sentence reading, storyboard narration, and conversational interviews in each language. All talkers in the corpus were early Cantonese-English bilinguals and members of the heterogeneous bilingual speech community in Vancouver, BC, Canada. Chapter 2 documents the motivation, design, and procedures used in the creation of SpiCE. Additionally, a detailed description of the talkers is provided. SpiCE is an open-access corpus freely available to anyone interested in the data—researchers, developers, hobbyists, and the general public (Johnson, 2021b). On its own, SpiCE represents a major contribution to the study of bilingual speech production.

Chapter 3 describes a study on the structure of acoustic voice variation within and across languages for the talkers in the SpiCE corpus. Using a wide array of source and filter-based acoustic measurements on voiced speech in the conversational interviews, Chapter 3 investigates crosslinguistic similarity in three ways. First, the distributions of each measurement were compared across languages on a by-talker basis using Cohen's *d*. The vast majority of comparisons resulted in trivial differences, indicating that talkers were, for the most part, internally consistent. Where consistent differences emerged, they mostly aligned with prior work—Cantonese tended to have lower fundamental frequency and be associated with breathier (or less creaky) voice quality than English (Ng et al., 2012). Second, a series of principal components analyses (PCAs) were run for each talker and language pair. In broad terms, the PCAs bore remarkable similarities in component structure and variance accounted for, regardless of talker and language, given prior work in this domain (Lee et al., 2019; Lee & Kreiman, 2019, 2020). The PCAs were then subjected to canonical correlation analyses to elucidate how much of the lower dimensional structure in one PCA could be accounted for by the other PCA—that is, how much *redundancy* there is between two PCAs—and vice versa. The result of this analysis clearly demonstrates that talkers bear the most similarity

to themselves across languages, compared to across-talker comparisons within or across languages. While there is some variation in the degree of similarity, the takeaway from this chapter is that voices can largely be thought of as "auditory faces."

Chapter 4 presents a second corpus study, focused on describing and analyzing the structure of phonetic category variation within and across languages for long-lag stops in Cantonese and English. Leveraging the uniformity framework (Chodroff & Wilson, 2017), Chapter 4 demonstrates that there is some structure to the relationship between voice onset time (VOT) patterns, but that the account is far less compelling than prior work on English (Chodroff & Wilson, 2017; Chodroff & Baese-Berk, 2019). Talkers were wildly inconsistent with respect to the expected ordinal relationships between means for the stop categories within languages (Chodroff & Wilson, 2017; Cho & Ladefoged, 1999; Lisker & Abramson, 1964). That is, very few talkers produced /p/ with shorter VOT than /t/ or /k/, and likewise between /t/ and /k/. The second phase of the analysis considered pairwise correlations of category means for VOT within and across languages. Again, the results were far from compelling. While there were consistent moderate correlations for within-language comparisons—especially for English—the across-language correlations were weaker or non-significant compared to prior work (Chodroff & Wilson, 2017; Chodroff & Baese-Berk, 2019). Their presence indicates some degree of structure but does not make for tidy conclusions.

Chapter 4 ends with a Bayesian linear mixed-effects model with two primary goals. First, the model estimates the effect of language while accounting factors known to influence VOT, such as local speaking rate and position. Second, the model allows for evaluating the sources of variability within the model. There was a small but consistent effect of language, such that English stops are produced with longer VOT than their Cantonese counterparts. The model also indicates that differences between talkers and words account for far more variation than differences in language and place of articulation effects. Along with the ordinal relationships and correlation analysis, Chapter 4 depicts both talker and language influences on

the structure of VOT. The results corroborate the ideas in the uniformity framework, albeit in a somewhat weaker manner than anticipated.

5.2 General discussion

Each chapter in this dissertation includes a discussion that deals with topics central—and unique—to the study at hand. This general discussion focuses on the bigger picture and some of the implications that the two studies have for perception, as outlined in Chapter 1. While there is substantial similarity across languages in both voice variability and the structure of long-lag stops, each study leaves room for both language and talker-indexical influences.

5.2.1 Talker-indexical and linguistic influences

In Chapter 3, the dual influences of talker and language show up most clearly in the canonical correlation analysis. While talkers exhibit the greatest degree of redundancy (i.e., similarity metric for comparing two PCAs) when compared to themselves across languages, no talker exhibits perfect redundancy. That there is variability in this metric suggests influence from *non*-talker-indexical sources, which could be linguistic or reflect social factors. Yet, while Chapter 3 does not rule out any particular type of influence, the high decree of within-talker similarity across languages emphasizes a clear role for talker-indexical components. This observation is further reflected in the comparisons of each acoustic measurement via Cohen's *d*.

Just over a third of the talkers in the SpiCE corpus maintain a crosslinguistic difference—in the same direction—for measures associated with pitch and non-modal voice quality. Some prior work has attempted to account for differences in fundamental frequency via the presence or absence of lexical tone. These studies, however, are not consistent with one another—some suggest lexical tone leads to lower F0 (Ng et al., 2012), and others to higher F0 (Lee & Sidtis, 2017; Keating & Kuo, 2012). It may be the case that a particular tone system impacts a talker's F0

profile, but the evidence is not yet compelling for this argument. If the differences of the present study were due to linguistic reasons, then it would be surprising that only a third of talkers exhibited the difference. A more likely account is one that invokes social factors and how individuals express their identity in each language (Loveday, 1981; Voigt et al., 2016). The lack of a clear role for linguistic influences here is further compounded by the behavior of the acoustic measurements typically associated with linguistic attributes. For example, while the patterning of F2 differences across languages largely reflects expectations around the distributions of back vowels in English and Cantonese, the direction of the effect is not consistent across talkers. While the expected difference is present for some, variability on this front and the uncontrolled nature of spontaneous speech make it challenging to draw a conclusion that reflects the group as a whole.

The dual influences of talker and language are perhaps more apparent in Chapter 4. This outcome is not surprising, considering that Chapter 3 examined voice quality, while Chapter 4 homes in on speech sound categories. Voice can be used in both linguistic and non-linguistic manners, while the speech sound categories are a decidedly a linguistic level of structure. For the latter, influence from both talker and language shows up in the juxtaposition between moderate correlations and the difference in VOT across languages. The moderate correlations for homorganic cross-language pairs indicate some level of uniformity in production—talkers with longer VOT in one language tend to also have longer VOT in the other language. Yet, at the same time, the crosslinguistic difference whereby English is characterized by longer VOT reflects a language-based influence.

5.2.2 Shared structure and consequences for perception

Chapter 1 framed this dissertation as being concerned with the consequences of a shared phonetic space, particularly with regard to how the speech signal facilitates processing and identifying multilingual talkers. If talkers are to be consistently identified before and after a language switch, then the speech signals in each language must resemble one another to some extent. Chapters 3 and 4 grapple with

how that resemblance plays out in production. In terms of perception, recall the summary of Orena et al. (2019) in Chapter 1, in which bilingual listeners outperformed monolingual listeners in an identification study featuring bilingual talkers. Orena et al. proposed several accounts for this bilingual advantage, even if performance was above chance across the board. Two of the proposed accounts appeal to listener access to systematicity across languages—Orena et al. (2019) suggest that bilingual listeners are sensitive to systematic *changes* in talker-indexical information and systematic *consistencies* in the linguistic signal.

How, then, do these accounts fare in the context of this dissertation? Put differently, is the assumed (or proposed) systematicity present in the bilingual speech signal? Chapter 3 presents a strong case for the structure of acoustic voice variation and for a variety of specific acoustic measurements. Talkers show a high degree of internal redundancy across languages and overall similarity for the various acoustic measurements. Even in cases where a subset of talkers show a non-trivial difference across languages, few if any differences are large, especially when compared to across-talker differences. In essence, Chapter 3 supports the talker-indexical account, albeit one where there are more likely to be systematic similarities rather than changes. It does not, however, discount the linguistic account.

Chapter 4 offers support for both accounts, though again, it seems to be one of systematic consistencies in talker-indexical aspects and systematic differences in the linguistic system. Note, however, that while the small across-language difference is meaningful regarding what it tells us about the production system, it is not necessarily meaningful to listeners (needs a citation and maybe more elaboration). Such a small difference thus further supports the talker-indexical view of what listeners use to identify talkers across languages.

Returning to the broad question of what language share in phonetic space, this dissertation provides a peek behind the curtain. Languages appear to share a lot in voice quality, despite distributional differences in the segment inventories and different roles for suprasegmental linguistic components (Matthews et al., 2013). While such differences may be more readily apparent in shorter stretches of

speech—as suggested by the passage length analysis in Section 3.2.7—over time, talkers appear to cover their full range. This coverage indicates that different languages make similar use of an individual's full range of acoustic voice variation and exhibit similar patterns of variation in the long run. The task of matching a new utterance up with a familiar talker, then, is one of asking whether or not the new utterance is likely to have arisen from the known range of variation.

Languages also share some aspects of phonetic category structure, albeit to a lesser extent. Chapter 4 demonstrates that talkers with longer VOT in one language tend to have longer VOT in the other language, even though a small distinction between the two languages was maintained. This outcome echoes results for speech rate, in which late bilinguals who are fast in their first language also tend to be fast talkers in their second language (Bradlow et al., 2017). While this relationship between languages is not as simple as saying individuals use the same underlying category in each language, it does demonstrate a certain degree of shared structure; simultaneously, it highlights the complexity of factors conspiring together to produce the acoustic signal.

What is clear from Chapters 3 and 4 is that there is ample shared structure for listeners to use in identifying bilingual talkers. The bilingual advantage in this domain could stem from the variable degree of similarity for talkers across languages and bilinguals' familiarity with how voices might deviate due to social and linguistic reasons. Similarly, it could stem from bilinguals' familiarity with the variety of forms in which a particular category can be produced. While there are clear examples of this kind of sociolinguistically informed variation for initial stops in other language pairs (Bullock & Toribio, 2009), there is also evidence of metalinguistic knowledge for different sound categories in Cantonese-English bilinguals. A recent example juxtaposes the production of word-final stops by the talkers in the SpiCE corpus (Johnson & Babel, 2021) and a lab-based study of a similar population (Polinsky, 2018). The corpus study demonstrates variability that skews towards Cantonese-like unreleased stops. The lab-based study, conversely, gives evidence for hypercorrection towards longer releases. By adopting the perspective

of Bullock & Toribio (2009), this discrepancy is readily explainable via metalinguistic awareness and how bilinguals use their language in different ways when talking to their peers versus speaking in formal, monolingual, lab-based settings. The point of bringing this example up is to illustrate that bilingual listeners are not only sensitive to the fine-grained acoustics (Ju & Luce, 2004), they are also sensitive to how form varies by communicative context. In sum, there are both systematic similarities and differences available in the signals for listeners—and bilingual listeners in particular—to use in tasks like talker identification.

5.3 Limitations

As with any study, the results presented in this dissertation are necessarily tempered by some limitations and leave a substantial amount of variation unaccounted for. The simplest form of limitation arises from methodological decisions and is touched on in Chapters 3 and 4. Both studies use corpus methods with exclusionary criteria and minimal manual inspection. In Chapter 3 this takes the form of using an automated approach to identify voiced portions of speech and a set of exclusionary criteria to discard likely errors. Chapter 4 relies on forced alignment, refinement via automated methods, and exclusionary criteria. Such approaches allow for the studies to be done at a larger scale but also mean that some degree of error is inevitable. The samples may include items that do not reflect the target. For example, some erroneous VOT measurements may have evaded the exclusionary criteria in Chapter 4—without a rigorous manual check or manual transcription of the SpiCE corpus, the true extent of this problem will remain unknown. While it is outside the scope of this dissertation to perform such a check, prior corpus work with similar exclusionary criteria indicates that the error rate is relatively small (5%: Chodroff & Baese-Berk, 2019).

The population studied here also presents a limitation on the extent to which the results of this dissertation can be generalized to other groups. As summarized in Section 1.1, there is enormous variation between and among bilingual populations—the talkers in the SpiCE corpus are no exception. As described in

detail in Chapter 2, the population studied in this dissertation represents a heterogeneous group of early Cantonese-English bilinguals. While some factors were carefully controlled for in recruitment, others were not. On the one hand, talkers were 18-35 years old at recording, comfortable conversing in Cantonese and English, and began learning both languages before age five. On the other hand, most talkers had some knowledge of at least one additional language (e.g., Mandarin, French, etc.) and varied in their family's current or historical roots in a Cantonese-speaking homeland.

While variability is an inherent part of the Cantonese-speaking community in Vancouver, BC, Canada—and thus justifiably included in the corpus—it does make comparisons with other bilingual communities somewhat challenging. Such comparisons are also complicated by the rather unique position of the speech community. Cantonese is reported as the "mother tongue" of some eight percent of Metro Vancouver census respondents (Statistics Canada, 2017). Given the overall population, Cantonese is an incredibly visible minority language in the region. As argued in Chan et al. (2020), this population likely has more access to Cantonese than bilingual communities in other English-dominant societies (e.g., Bruggeman & Cutler, 2019). While there is much more to say on this topic, such detail is left for future work. Further, while this corpus could be used to explore sources of variation based on different aspects of the talkers' demographics, a lack of control on many of the potentially relevant parameters renders such approaches speculative.

5.4 Current and future directions

There is another kind of "limitation" baked into the framing of this dissertation. While the focus remains on describing and accounting for variation in speech production, the motivation arises from perception. This disjunction means that questions relating to speech perception are not answered here. Rather, the speech production results presented in this dissertation generate hypotheses for how bilingual talkers are perceived and identified. This section outlines what some of those hypotheses are. Each of the following paragraphs in this section begins with an ital-

icized hypothesis and is followed up by the result that motivates it. Any current research being done on the question will be noted. The first two hypotheses derive from Chapter 3 and the latter three from Chapter 4.

Increased within-talker canonical redundancy across languages will facilitate multilingual talker identification and discrimination. Concurrently, greater redundancy between talkers will lead to a higher chance of false alarms or confusability. This hypothesis emerges directly from the finding that some talkers are more similar to themselves than others in the structure of their acoustic voice variation. Variability in production is thus hypothesized to be mirrored in perception. Prior perceptual work in this area often only considers a handful of talkers and uses a handful of coarse checks on the voices (needs citation). As a result, spurious results may be treated as being unique to the talker. This hypothesis seeks to add a concrete account of why talkers differ in this way. Ongoing research on bilingual talker identification and discrimination in the Speech-in-Context Lab aims to address this hypothesis, along with an additional goal of better understanding the listeners' role (Lloy et al., 2020, 2021).

Global shifts in voice quality dimensions across languages will not disrupt talker identification and discrimination when they mirror consistent (when present) patterns in the speech signal. Conversely, shifts not present in the range of voice variation patterns will be disruptive. This hypothesis arises from the assumption that listeners are better at processing variation when they have experience with it. In the context of bilingualism, this is hinted at in a few ways. First, Orena et al. (2019) found that experience with code-switching led to increased performance at generalizing talker identification across languages. Second, anticipatory interference in the speech signal facilitates the processing of a switch from one language to another by listeners (Fricke et al., 2016b). In both cases, experience with how languages differ and experience hearing the two languages in close proximity led to improved processing of the multilingual speech signal. This hypothesis extends the assumption that experience with variation will lend yet another advantage in processing variation.

While all listeners will benefit from congruent VOT within-talker across languages, bilingual listeners will be more adept at learning systematic differences. This hypothesis emerges from Chapter 4, in which VOT differed slightly across languages but was also highly variable between talkers. While this hypothesis ultimately echoes the proposal suggested in Orena et al. (2019), it calls for more explicit manipulation of VOT (e.g., as in the experimental study in Fricke et al., 2016b).

Uniformity will decrease as speech style becomes progressively more formal. While this hypothesis is supported both by prior work (Chodroff & Wilson, 2017) and its comparison with the results in Chapter 4. However, a direct comparison of spontaneous speech to either of the styles represented in Chodroff & Wilson (2017) remains lacking, as the studies report on rather different populations and include many possible confounds.

Full category assimilation in early bilingual speech is exceedingly rare and possibly even non-existent. This hypothesis arises from the small difference maintained between long-lag stop categories across languages; additionally, it is supported by the arguments against compromise categories for bilinguals in Casillas (2021). This hypothesis merely extends that argument to long-lag stops that exhibit an even greater degree of crosslinguistic similarity from the outset.

5.5 Conclusion

Speech is variable, and learning a new talker can be characterized as learning how that talker varies. This dissertation focuses on comparing systematicity across languages to understand how such structure might facilitate processes like multilingual talker identification. The results presented here demonstrate the presence of systematicity at two levels—acoustic voice variation and how long-lag stop series manifest. This structure shows evidence of both talker-indexical and linguistic influences and generates a multitude of hypotheses for future work.

There is a balance between variation and structure at every level—talker, language, linguistic units, voice quality, and more. Working with spontaneous speech

corpora is one of the better ways to gain an appreciation for this observation. It would not be possible to do this kind of research without data, and one of the largest and lasting contributions of this dissertation is the SpiCE corpus. While this dissertation just scratches the surface of what can be done with SpiCE, making the data available will help push our understanding of bilingual speech production forward.

Bibliography

Afouras, T., Chung, J. S., & Zisserman, A. (2020). Now you're speaking my language: Visual language identification. In *Proceedings of Interspeech 2020*, (pp. 2402–2406). https://doi.org/10.21437/Interspeech.2020-2921 → pages 46, 82

Alderete, J., Chan, Q., & Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, *191*, 103952. https://doi.org/10.1016/j.cognition.2019.04.021 → page 11

Altenberg, E. P., & 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. https://doi.org/10.1016/j.jvoice.2005.01.005 → pages 49, 52, 61, 63, 83

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. https://doi.org/10.1177/1367006917741364 → pages 3, 13

Amengual, M. (2018). Asymmetrical interlingual influence in the production of Spanish and English laterals as a result of competing activation in bilingual language processing. *Journal of Phonetics*, *69*, 12–28. https://doi.org/10.1016/j.wocn.2018.04.002 → page 88

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, *38*(4), 640–653. https://doi.org/10.1016/j.wocn.2010.09.005 → page 94

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in

phonetic code-switching. *Journal of Phonetics*, *39*(4), 558–570. https://doi.org/10.1016/j.wocn.2011.03.001 → page 94

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 4218–4222). Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.520 → page 11

Audacity Team (2018). Audacity (R): Free audio editor and recorder. https://www.audacityteam.org/ → page 19

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443. https://doi.org/10.1177/1367006913516035 → page 94

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001 → page 115

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. *ArXiv Preprints*, (pp. 1–21). http://arxiv.org/abs/1506.04967 → page 115

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110823707 → page 101

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008 → page 37

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. Version 6.1.38. http://www.praat.org/ → page 55

Bolton, K., Bacon-Shone, J., & Lee, S.-L. (2020). Societal multilingualism in Hong Kong. In *Multilingual Global Cities*, (pp. 160–184). Routledge. https://doi.org/10.4324/9780429463860-12 → page 15

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., & Mok, K. (2011). Language- and talker-dependent variation in global features

of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, (pp. 356–359). Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Bradlow/Bradlow.pdf → pages 10, 100

Bradlow, A. R., Kim, M., & 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. https://doi.org/10.1121/1.4976044 → pages 53, 109, 126, 127, 135

Brehm, L., & Alday, P. M. (2021). A decade of mixed models: It's past time to set your contrasts. *OSF Preprints*. https://osf.io/3tgq6/ → page 116

Brown, E. L., & Amengual, M. (2015). Fine-grained and probabilistic cross-linguistic influence in the pronunciation of cognates: Evidence from corpus-based spontaneous conversation and experimentally elicited data. *Studies in Hispanic and Lusophone Linguistics*, *8*(1), 59–83. https://doi.org/10.1515/shll-2015-0003 → page 94

Brown, E. L., & Harper, D. (2009). Phonological evidence of interlingual exemplar connections. *Studies in Hispanic and Lusophone Linguistics*, *2*(2), 257–274. https://doi.org/10.1515/shll-2009-1052 → page 3

Bruggeman, L., & Cutler, A. (2019). No L1 privilege in talker adaptation. *Bilingualism: Language and Cognition*, (pp. 1–13). https://doi.org/10.1017/S1366728919000646 → page 137

Bullock, B. E., & Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In L. Isurin, D. Winford, & K. deBot (Eds.) *Studies in Bilingualism*, vol. 41, (pp. 189–206). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/sibil.41.12bul → pages 4, 47, 53, 93, 94, 95, 135, 136

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01 → page 114

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances.

Cognitive Science, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231 → pages 41, 43, 69

Casillas, J. V. (2021). Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages*, *6*(1), 9. https://doi.org/10.3390/languages6010009 → pages 92, 93, 95, 139

Ćavar, M., Ćavar, D., & Cruz, H. (2016). Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, (pp. 4004–4011). Portorož, Slovenia. https://aclanthology.org/L16-1632/ → page 29

Chan, A. Y. W., & Li, D. C. S. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems. *Language, Culture and Curriculum*, *13*(1), 67–85. https://doi.org/10.1080/07908310008666590 → page 101

Chan, L., Johnson, K. A., & Babel, M. (2020). Lexically-guided perceptual learning in early Cantonese-English bilinguals. *The Journal of the Acoustical Society of America*, *147*(3), EL277–EL282. https://doi.org/10.1121/10.0000942 → pages 3, 137

Chang, C. B. (2015). Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity. In E. Raimy, & C. E. Cairns (Eds.) *The Segment in Phonetics and Phonology*, (pp. 199–217). Chichester, UK: John Wiley & Sons, Inc., 1 ed. https://doi.org/10.1002/9781118555491.ch9 → pages 90, 91

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. https://doi.org/10.1121/10.0000498 → pages 51, 52

Cheng, L. S. P., Babel, M., & Yao, Y. (2021). Production and perception across three Hong Kong Cantonese consonant mergers: Community- and individual-level perspectives. Manuscript under review. → page 127

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. https://doi.org/10.1006/jpho.1999.0094 → pages 105, 106, 124, 131

Chodroff, E., & Baese-Berk, M. (2019). Constraints on variability in the voice onset time of L2 English stop consonants. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 661–665). Melbourne, Austrailia. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_710.pdf → pages 100, 101, 105, 109, 125, 131, 136

Chodroff, E., Golden, A., & Wilson, C. (2019). Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, *145*(1), EL109–EL115. https://doi.org/10.1121/1.5088035 → pages 113, 125, 126

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001 → pages 39, 99, 100, 101, 102, 103, 105, 106, 109, 111, 113, 116, 119, 124, 125, 131, 139

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2). https://doi.org/10.1515/lingvan_2017_0047 → page 101

Chodroff, E., & Wilson, C. (in press). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. *Language*. Expected publication in June 2022. https://eleanorchodroff.com/articles/ChodroffWilson_UniformitySibilants_Language_Accepted_2022.pdf → page 99

Clumeck, H., Barton, D., Macken, M. A., & Huntington, D. A. (1981). The aspiration contrast in Cantonese word-initial stops: Data from children and adults. *Journal of Chinese Linguistics*, *9*(2), 210–225. https://www.jstor.org/stable/23753507 → pages 101, 125

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., & Carter, D. (2014). Building bilingual corpora. In E. M. Thomas, & I. Mennen (Eds.) *Advances in the Study of Bilingualism*, (pp. 93–110). Multilingual Matters. https://doi.org/10.21832/9781783091713_008 → pages 9, 34

Ethnologue (2021). Chinese, Yue. In D. M. Eberhard, G. F. Simons, & C. D. Fennig (Eds.) *Ethnologue: Languages of the World*. Dallas, TX: SIL International, 24 ed. Online version. http://www.ethnologue.com → page 11

Faytak, M. D. (2018). *Articulatory Uniformity Through Articulatory Reuse: Insights from an Ultrasound Study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley. https://escholarship.org/uc/item/0jr0010h → pages 99, 100

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.) *Second Language Speech Learning: Theoretical and Empirical Progress*, (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002 → pages 1, 3, 88, 89, 90, 91, 92, 93, 97, 101

Fricke, M., Baese-Berk, M. M., & Goldrick, M. (2016a). Dimensions of similarity in the mental lexicon. *Language, Cognition and Neuroscience*, *31*(5), 639–645. https://doi.org/10.1080/23273798.2015.1130234 → page 3

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016b). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, *89*, 110–137. https://doi.org/10.1016/j.jml.2015.10.001 → pages 55, 93, 94, 95, 138, 139

Fricke, M., Zirnstein, M., Navarro-Torres, C., & Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, *22*(1), 200–207. https://doi.org/10.1017/S1366728918000482 → pages 3, 98

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006 → pages 8, 125

Ganugapati, D., & Theodore, R. M. (2019). Structured phonetic variation facilitates talker identification. *The Journal of the Acoustical Society of America*, *145*(6), EL469–EL475. https://doi.org/10.1121/1.5100166 → pages 100, 128

Garellek, M. (2019). The phonetics of voice. In W. F. Katz, & P. F. Assmann (Eds.) *The Routledge Handbook of Phonetics*. Routledge. https://doi.org/10.4324/9780429056253_5 → pages 39, 40, 42, 57, 84

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555 → pages 114, 117

Gertken, L. M., Amengual, M., & Birdsong, D. (2014). Assessing language dominance with the Bilingual Language Profile. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.) *Measuring L2 proficiency: Perspectives from SLA*, (pp. 208–225). Bristol, UK: Multilingual Matters. https://doi.org/10.21832/9781783092291-014 → page 2

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/icassp.1992.225858 → page 11

Goldrick, M., Runnqvist, E., & Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science*, *25*(4), 1031–1036. https://doi.org/10.1177/0956797613520014 → pages 93, 94, 95, 96

Google (2019). Cloud speech-to-text. V1. https://cloud.google.com/speech-to-text/ → pages 11, 25

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics*, *0*. https://doi.org/10.1515/ling-2021-0094 → page 10

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. https://doi.org/10.1016/0093_934X(89)90048_5 → pages 1, 2, 3

Grosjean, F. (2011). An attempt to isolate, and then differentiate, transfer and interference. *International Journal of Bilingualism*, *16*(1), 11–21. https://doi.org/10.1177/1367006911403210 → pages 90, 94, 97, 98, 128

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, *60*(2), 98–128. https://doi.org/10.1159/000071449 → page 92

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/xr7y3 → pages 113, 124

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of speech and hearing research*, *37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769 → page 58

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058 → page 21

Ingvalson, E. M., Ettlinger, M., & Wong, P. C. M. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism*, *18*(1), 35–47. https://doi.org/10.1177/1367006912456586 → page 4

Iseli, M., Shue, Y.-L., & 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. https://doi.org/10.1121/1.2697522 → page 57

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001 → page 55

Järvinen, K., Laukkanen, A.-M., & Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, *38*(2), 47–51. https://doi.org/10.3109/14015439.2012.687764 → pages 50, 52, 55, 67

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama, & K. Maekawa (Eds.) *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, (pp. 29–54). Toyko, Japan: The National International Institute for Japanese Language. https://linguistics.berkeley.edu/~kjohnson/papers/Massive.pdf → page 5

Johnson, K. A. (2019). Probabilistic reduction in Spanish-English bilingual speech. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 1263–1267). Melbourne, Australia. → page 125

Johnson, K. A. (2021a). Leveraging the uniformity framework to examine crosslinguistic similarity for long-lag stops in spontaneous Cantonese-English

bilingual speech. In *Proceedings of Interspeech 2021*, (pp. 2671–2675). https://doi.org/10.21437/Interspeech.2021-1780 → page vi

Johnson, K. A. (2021b). SpiCE: Speech in Cantonese and English. V1. https://doi.org/10.5683/SP2/MJOXP3 → pages 6, 8, 130

Johnson, K. A., & Babel, M. (2021). Language contact within the speaker: Phonetic variation and crosslinguistic influence. *OSF Preprints*. https://doi.org/10.31219/osf.io/jhsfc → pages 8, 95, 135

Johnson, K. A., Babel, M., Fong, I., & Yiu, N. (2020a). SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 4089–4095). Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.503 → page vi

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020b). Bilingual acoustic voice variation is similarly structured across languages. In *Proceedings of Interspeech 2020*, (pp. 2387–2391). https://doi.org/10.21437/Interspeech.2020-3095 → page vi

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer-Verlag, 2 ed. https://doi.org/10.1007/b98835 → pages 69, 77

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science*, *15*(5), 314–318. https://doi.org/10.1111/j.0956-7976.2004.00675.x → pages 4, 92, 136

Kawahara, H., Agiomyrgiannakis, Y., & 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*, (pp. 221–228). https://doi.org/10.21437/SSW.2016-36 → page 56

Keating, P., Kreiman, J., & Alwan, A. (2019). A new speech database for within- and between-speaker variability. In *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 736–739). Melbourne, Australia. https://www.assta.org/proceedings/ICPhS2019/papers/ICPhS_785.pdf → pages 41, 86

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*,

132(2), 1050–1060. https://doi.org/10.1121/1.4730893 → pages 45, 49, 50, 132

Keshet, J., Sonderegger, M., & Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction. Version 0.94. https://github.com/mlml/autovot/ → page 103

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. https://doi.org/https://doi.org/10.1111/tops.12331 → pages 5, 38

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1), e009. https://doi.org/10.3989/loquens.2014.009 → pages 39, 40, 41, 56, 57, 69

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, *149*(1), 457–465. https://doi.org/10.1121/10.0003331 → pages 40, 69

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925 → pages 114, 115

Labov, W., Ash, S., & Boberg, C. (2008). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton. https://doi.org/10.1515/9783110167467 → page 67

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 175. https://doi.org/10.3389/fpsyg.2011.00175 → pages 43, 83

Laver, J. (1980). *The phonetic description of voice quality*, vol. 31 of *Cambridge Studies in Linguistics*. New York: Cambridge University Press. https://www.cambridge.org/9780521108898 → page 39

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383 → pages 43, 83

Lee, B., & 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. https://doi.org/10.1080/2050571x.2016.1273572 → pages 50, 51, 52, 55, 86, 132

Lee, J. L. (2018). PyCantonese. Version 2.2.0. https://pycantonese.org/ → pages 11, 28

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, *146*(3), 1568–1579. https://doi.org/10.1121/1.5125134 → pages 38, 39, 41, 42, 43, 45, 53, 54, 56, 60, 61, 69, 70, 72, 74, 76, 81, 82, 83, 84, 86, 130

Lee, Y., & Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. In *The 178th Meeting of the Acoustical Society of America*. San Diego, CA. Poster. https://doi.org/10.1121/1.5137431 → pages 41, 42, 76, 83, 86, 130

Lee, Y., & Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. In *The 179th Meeting of the Acoustical Society of America*. Acoustics Virtually Everywhere. Poster. https://doi.org/10.1121/1.5146847 → pages 41, 42, 46, 69, 76, 83, 130

Lein, T., Kupisch, T., & van de Weijer, J. (2016). Voice onset time and global foreign accent in German–French simultaneous bilinguals during adulthood. *International Journal of Bilingualism*, *20*(6), 732–749. https://doi.org/10.1177/1367006915589424 → page 92

Leung, M.-T., & Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, *6*(2), 305–325. https://doi.org/10.1075/ijcl.6.2.06leu → page 11

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science*, *10*(2), e1483. https://doi.org/10.1002/wcs.1483 → page 44

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. https://doi.org/10.1007/978-3-319-12619_7 → page 54

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279 → pages 5, 38

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, *5*(1), 91–107. https://doi.org/10.1146/annurev-linguistics-011516-033830 → page 10

Lieberman, P., & Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139165952 → page 103

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman, & C. N. Li (Eds.) *Language, Speech, and Mind: Studies in Honour of Victoria A. Fromkin*, (pp. 62–78). London: Routledge. → page 92

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830 → pages 101, 124, 125, 131

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28. https://doi.org/10.1177/002383096701000101 → page 102

Littell, P. (2010). Thank-you notes [Version 1.0: Agent focus]. http://totemfieldstoryboards.org/stories/thank_you_notes/ → page 21

Llompart, M., & Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in German vowels. *Journal of Phonetics*, *67*. https://doi.org/10.1016/j.wocn.2017.12.001 → page 88

Lloy, A., Johnson, K., & Babel, M. (2021). Examining the roles of language familiarity and bilingualism in talker recognition. In *The 13th International Symposium on Bilingualism*. Virtual. Poster. → pages 87, 138

Lloy, A., Johnson, K. A., & Babel, M. (2020). Bilingual talker identification with spontaneous speech in Cantonese and English: The role of language-specific knowledge. In *The 179th Meeting of the Acoustical Society of America*. Virtual. Poster. https://doi.org/10.1121/1.5147685 → pages 87, 138

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. https://doi.org/10.1177/002383098102400105 → pages 51, 133

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445 → page 69

Luke, K. K., & Wong, M. L. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series*, *25*, 312–333. https://www.jstor.org/stable/26455290 → page 11

Matthews, S., Yip, V., & Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. https://doi.org/10.4324/9780203835012 → pages x, 20, 48, 82, 101, 134

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner. Version 1.0.1. https://montrealcorpustools.github.io/Montreal-Forced-Aligner/ → page 28

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman and Hall/CRC, 2 ed. https://doi.org/10.1201/9780429029608 → pages 115, 117, 118

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9 → page 126

Ménard, L., Schwartz, J.-L., & Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, *50*(1), 14–28. https://doi.org/10.1016/j.specom.2007.06.004 → pages 99, 100

Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second Language Research*, *26*(1), 13–41. → page 40

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006 → page 91

Mielke, J., & Nielsen, K. (2018). Voice onset time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America*, *144*(4), 2166–2177. https://doi.org/10.1121/1.5059493 → page 101

Munson, B., & Babel, M. (2019). The phonetics of sex and gender. In W. F. Katz, & P. F. Assmann (Eds.) *The Routledge Handbook of Phonetics*. Routledge. https://doi.org/10.4324/9780429056253_19 → page 57

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & 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 47

Myers-Scotton, C. (2011). The matrix language frame model: Developments and responses. In *Codeswitching Worldwide*, vol. 126 of *Trends in Linguistics. Studies and Monographs*. De Gruyter Mouton. https://doi.org/10.1515/9783110808742.23 → page 55

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, *43*(1-2), 65–84. http://digital.casalini.it/10.1400/190440 → pages 20, 23, 26

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. University of Adelaide, Adelaide, Australia. Version 0.5. http://ua.edu.au/ccs/teaching/lsr → page 61

Ng, M. L., Chen, Y., & 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. https://doi.org/10.1016/j.jvoice.2011.07.013 → pages 49, 50, 53, 61, 63, 67, 83, 130, 132

Ng, M. L., Hsueh, G., & 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. https://doi.org/10.3109/17549501003721080 → pages 49, 61, 83

Ng, R. W. M., Kwan, A. C., Lee, T., & Hain, T. (2017). ShefCE: A Cantonese-English bilingual speech corpus for pronunciation assessment. In

Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, (pp. 5825–5829). https://doi.org/10.1109/ICASSP.2017.7953273 → page 10

Nieuwenhuis, R., Manfred, t. G., & Pelzer, B. (2017). Weighted effect coding for observational data with wec. *The R Journal*, *9*(1), 477. https://doi.org/10.32614/rj-2017-017 → page 116

Olson, D. J. (2016). The role of code-switching and language context in bilingual phonetic transfer. *Journal of the International Phonetic Association*, *46*(3), 263–285. https://doi.org/10.1017/S0025100315000468 → pages 93, 94, 95

Ordin, M., & Mennen, I. (2017). Cross-linguistic differences in bilinguals' fundamental frequency ranges. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1493–1506. https://doi.org/10.1044/2016_JSLHR-S-16-0315 → page 52

Orena, A. J., Polka, L., & 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. https://doi.org/10.1121/1.5097735 → pages 4, 5, 44, 45, 100, 128, 134, 138, 139

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/icassp.2015.7178964 → page 11

Perrachione, T. K. (2018). Recognizing speakers across languages. In S. Frühholz, & P. Belin (Eds.) *The Oxford Handbook of Voice Perception*, (pp. 514–538). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.23 → page 44

Perrachione, T. K., Furbeck, K. T., & 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. https://doi.org/10.1121/1.5126697 → pages 44, 46, 53, 76, 128

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of

transcriber reliability. *Speech Communication*, *45*(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001 → pages 9, 10, 22

Pittam, J. (1987). The long-term spectral measurement of voice quality as a social and personality marker: A review. *Language and Speech*, *30*(1), 1–12. https://doi.org/10.1177/002383098703000101 → pages 39, 40

Podesva, R. J., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270 → pages 37, 38, 40

Polinsky, M. (2018). *Heritage Languages and their Speakers*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107252349 → page 135

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ → pages 61, 69, 106, 114

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75–86. https://doi.org/10.1037/a0027979 → page 128

Revelle, W. (2021). psych: Procedures for psychological, psychometric, and personality research. R package version 2.1.3. https://CRAN.R-project.org/package=psych → page 106

Ryabov, R., Malakh, M., Trachtenberg, M., Wohl, S., & Oliveira, G. (2016). Self-perceived and acoustic voice characteristics of Russian-English bilinguals. *Journal of Voice*, *30*(6), 772.e1–772.e8. https://doi.org/10.1016/j.jvoice.2015.11.009 → page 51

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. https://doi.org/10.1016/j.jml.2019.104070 → page 98

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. https://doi.org/10.1006/jpho.1997.0051 → pages 3, 94

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences*, vol. 3, (pp. 1846–1849). Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011 → page 56

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935345.013.72 → pages 89, 90

Simonet, M., & Amengual, M. (2019). Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, *24*(2), 208–221. https://doi.org/10.1177/1367006919826388 → pages 3, 22, 94, 95, 97

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, *3*(2), 621–640. https://doi.org/10.1111/j.1749-818X.2009.00125.x → page 55

Sjölander, K. (2004). The Snack Sound Toolkit. https://www.speech.kth.se/snack/ → page 57

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf → pages 25, 26

Sóskuthy, M., & Stuart-Smith, J. (2020). Voice quality and coda /r/ in Glasgow English in the early 20th century. *Language Variation and Change*, *32*(2), 133–157. https://doi.org/10.1017/S0954394520000071 → page 74

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*(2), 218–231. https://doi.org/10.3758/BF03193744 → pages 46, 82

Stan Development Team (2021). *Stan Modeling Language Users Guide and Reference Manual*. https://mc-stan.org → page 114

Statistics Canada (2017). Proportion of mother tongue responses for various regions in Canada, 2016 Census. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm → pages 13, 137

Stewart, D., & Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, *70*(3, pt.1), 160–163. https://doi.org/10.1037/h0026143 → page 77

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3-4), 505–549. https://doi.org/10.1515/lp-2015-0015 → pages 101, 109, 125

Sun, J. (2020). jieba. Version 0.42.1. https://github.com/fxsjy/jieba → page 28

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (pp. I–333–I–336). https://doi.org/10.1109/ICASSP.2002.5743722 → page 59

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, *9*(1), 97–114. https://doi.org/10.1017/S1366728905002403 → pages 92, 93, 94, 95

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 ed. → page 70

Tanner, J., Sonderegger, M., Stuart-Smith, J., & Fruehwald, J. (2020). Toward "English" phonetics: Variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00038 → page 6

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. http://d-scholarship.pitt.edu/35721/ → pages 20, 29

Tsui, R. K.-Y., Tong, X., & Chan, C. S. K. (2019). Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, *40*(1), 29–58. https://doi.org/10.1017/S0142716418000449 → pages 96, 97

Turk, M., & Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1991.139758 → page 69

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008 → pages 114, 117

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1. https://mc-stan.org/loo/ → page 118

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4 → page 118

Voigt, R., Jurafsky, D., & Sumner, M. (2016). Between- and within-speaker effects of bilingualism on F0 variation. In *Proceedings of Interspeech 2016*, (pp. 1122–1126). San Francisco, CA. https://doi.org/10.21437/Interspeech.2016-1506 → pages 51, 133

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9–30. https://doi.org/10.1093/applin/amx039 → page 47

Wilson, C., & Mihalicek, V. (2011). *Language Files: Materials for an Introduction to Language and Linguistics*. Columbus, UH: Ohio State University Press. https://linguistics.osu.edu/research/pubs/lang-files → pages x, 48, 82

Winterstein, G., Tang, C., & Lai, R. (2020). CantoMap: A Hong Kong Cantonese MapTask corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 2906–2913). Marseille, France. https://aclanthology.org/2020.lrec-1.355 → page 11

Wong, W. Y. P. (2006). *Syllable Fusion in Hong Kong Cantonese Connected Speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. http://rave.ohiolink.edu/etdc/view?acc_num=osu1143227948 → pages 26, 28

Xue, S. A., Hagstrom, F., & 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. https://doi.org/10.1179/136132802805576544 → page 50

Yang, J. (2019). Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, (p. 0267658319851820). https://doi.org/10.1177/0267658319851820 → pages 96, 101

Yang, Y., Chen, S., & Chen, X. (2020). F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Proceedings of Interspeech 2020*, (pp. 4163–4167). https://doi.org/10.21437/Interspeech.2020-2549 → page 51

Yau, M. (2019). PyJyutping. https://github.com/MacroYau/PyJyutping → page 11

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150. https://doi.org/10.1146/annurev-linguistics-011516-033815 → page 127

Yu, H. (2013). Mountains of gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, (pp. 108–121). Routledge. https://doi.org/10.4324/9780203100387.ch7 → page 13

Yuan, J., Ryant, N., & Liberman, M. (2014). Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 2539–2543). https://doi.org/10.1109/ICASSP.2014.6854058 → page 29

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 11475. https://doi.org/10.1038/srep11475 → page 45