

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

© Khia Anne Johnson, 2021

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

Crosslinguistic 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

Abstract	iii
Lay Summary	v
Preface	vi
Table of Contents	vii
List of Tables	x
List of Figures	xii
Acknowledgments	xvi
1 Introduction	1
1.1 Bilingualism	2
1.2 Processing bilingual talkers	3
1.3 Variability in conversational speech	5
1.4 Thesis goals and research questions	6
2 The SpiCE Corpus	7
2.1 Introduction	7
2.2 Corpus design and creation	11
2.2.1 Recruitment	12

2.2.2	Participants	12
2.2.3	Recording setup	19
2.2.4	Recording procedure	19
2.3	Annotation	25
2.3.1	Cloud speech-to-text	25
2.3.2	Orthographic transcription hand-correction	25
2.3.3	Forced alignment	28
2.4	Descriptive statistics	29
2.4.1	Cantonese interviews	30
2.4.2	English interviews	32
2.5	SpiCE corpus release	33
2.6	Discussion and conclusion	34
3	The Structure of Acoustic Voice Variation in Bilingual Speech . . .	37
3.1	Introduction	37
3.1.1	Voice and voice quality	39
3.1.2	Structure in voice quality variation	40
3.1.3	Voice perception	44
3.1.4	Bilingual voices	45
3.1.5	The present study	53
3.2	Methods and results	54
3.2.1	Data	54
3.2.2	Acoustic measurements	56
3.2.3	Exclusionary criteria and post-processing	59
3.2.4	Crosslinguistic comparison of acoustic measurements . . .	61
3.2.5	Principal components analysis	68
3.2.6	Canonical redundancy analysis	77
3.2.7	Passage length analysis	80
3.3	Discussion and conclusion	82

4	The Structure of Voice Onset Time Variation in Bilingual Long-lag	
	Stops	88
4.1	Introduction	88
4.1.1	Identifying “links” across bilinguals’ languages	89
4.1.2	Crosslinguistic influence and representation	91
4.1.3	Adapting the uniformity framework	98
4.1.4	Long-lag stops in Cantonese and English	101
4.2	Methods	102
4.2.1	Corpus	102
4.2.2	Segmentation and measurement	102
4.3	Analysis and results	104
4.3.1	Ordinal relationships	105
4.3.2	Pairwise correlations	106
4.3.3	Linear mixed-effects model	113
4.4	Discussion	124
5	Discussion and Conclusion	129
5.1	Recap	130
5.2	General discussion	132
5.2.1	Talker-indexical and linguistic influences	132
5.2.2	Shared structure and consequences for perception	133
5.3	Limitations	136
5.4	Current and future directions	137
5.5	Conclusion	139
	Bibliography	141

List of Tables

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.	14
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.	22
Table 2.3	All Cantonese sentences are widely-known imperatives associated with Chinese New Year.	23
Table 3.1	The Cantonese segmental inventory as described by Matthews et al. (2013). Note that Cantonese vowels combine into many different diphthongs.	48
Table 3.2	The English segmental inventory as described by Wilson & Michalick (2011), with [ʔ ɾ w] excluded. Note that some English vowels combine into diphthongs.	48
Table 3.3	This table reports counts of Cohen’s <i>d</i> 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.	64

Table 3.4	The number of components, variance accounted for, and number of identical components across languages for each PCA. . .	71
Table 4.1	The number of stop tokens (overall and range across talkers) and word types for each language and sound category.	104
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.	106
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 <i>r</i> , and the Holm-adjusted <i>p</i> -value given 15 comparisons.	110
Table 4.4	Population parameter summary.	119
Table 4.5	Group parameter variability summary.	121

List of Figures

Figure 2.1	This four panel bar chart summarizes where the SpiCE participants lived during different portions of their lives.	15
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.	16
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.	17
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.	18
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 file-name, (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.	24

Figure 2.6	This screenshot from Praat shows what the final transcript looks like for a small portion of a Cantonese interview.	30
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.	31
Figure 2.8	The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.	33
Figure 2.9	The distribution of log word frequency for English and Cantonese words in the Cantonese interviews.	34
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.	35
Figure 3.1	Each panel depicts a density plot that pools measurements 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.	62
Figure 3.2	A histogram summary of the number of non-trivial comparisons from Table 3.3 across the 34 talkers.	63
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 interpretation for Cohen’s d . Each point represents a talker.	65
Figure 3.4	This figure uses the format of 3.3, but reports on the standard deviation measures.	66

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.	73
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.	79
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 ($x = 20,124$). . .	81
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.	85
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.	107
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.	108

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. <i>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.</i> . . .	111
Figure 4.4	Correlations for the across-language comparisons of raw mean VOT are depicted 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. . .	112
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.	120
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 <i>brms</i> ’ conditional effects function. Notably, the error overlaps almost completely for /p/, but not at all for /t/ and /k/.	121
Figure 4.7	This figure depicts the posterior distributions for the standard deviation of each of the grouping parameters, both intercepts and slopes.	122
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.	123