Chapter 3

The structure of acoustic voice variation in bilingual speech

3.1 Introduction

Voices provide a lot about the person talking, ranging from their current physical and emotional state to talker indexical features that help listeners identify who they are. In this context, voices can be described as auditory faces, in that they are uniquely individual, yet with broadly shared characteristics across the population (Belin et al., 2004). Voices are rich in information, and they convey all of this information alongside communicative content. Understanding the structure of a voice is no small feat, as is understanding how listeners use different dimensions within the voice in processing talker indexical, affective, social, and linguistic information. The difficulty here arises from the sheer variability across voices. While voices share attributes and relevant acoustic dimensions, much of the variation appears idiosyncratic (Lee et al., 2019). From the perspective of voice perception, the balance between shared and idiosyncratic makes sense. The shared dimensions allow listeners to perceive, classify, and understand new voices, while the idiosyncrasies enable identification and discrimination between voices. While this makes sense conceptually, understanding the structure of voice variation in speech production and its complement in listeners' ability to process that information remains an active area of research. While the focus of this chapter is acoustic voice variability,

the emphasis on describing and processing variation echoes one of the largest puzzles in phonetics: the "lack of invariance" problem (Liberman et al., 1967). That is, given the ubiquity of variation, how do perceivers efficiently extract relevant and important information from the communicative signal?

While variation is indeed wide-ranging, it remains far from random. Some of the most prevalent accounts of how individuals understand and process variation emphasize that variation in speech production is highly structured. This chapter looks at the structure of voices, and the following chapter examines structure for sound categories—both attempt to elucidate what exists in the signal for listeners to use. In the domain of voice quality, Jody Kreiman and colleagues have synthesized work from various areas and put forth a psychoacoustic model of voice quality (Kreiman et al., 2014). This model features a minimal set of acoustic dimensions necessary to encode (and thus reproduce) voice quality. While there are numerous dimensions in the model, extensive experimental work has validated the inclusion of each dimension (Kreiman et al., 2021, and references therein). As a result, Kreiman and colleagues argue that this set is both sufficient and necessary to capture a wide range of normal and disordered voices. This model includes acoustic dimensions that capture harmonic and inharmonic voice source, pitch, loudness, and vocal tract characteristics. While each dimension in the model could be considered independently by researchers, Kreiman and colleagues argue that these dimensions are more than the sum of their parts. The measures covary and conspire together to form a percept. While this model establishes a set of acoustic dimensions, it does not arbitrate between them in a way that establishes what matters for a given voice in a given language.

There is a large body of literature focused on understanding differences in variability across populations for a small set of acoustic measurements. Such studies typically compare summary statistics for fundamental frequency (F0) and a handful of spectral measures. This body of work will be summarized below in the context of crosslinguistic comparisons. Before summarizing this work, it is important to highlight that very little of it dives into the structure of voice variability, which is a relatively new area spearhead by Lee and colleagues (Lee et al., 2019; Lee and Kreiman, 2019, 2020). In this set of studies examining acoustic voice variation in different languages and speech styles, the authors leverage the psychoacoustic

model of voice quality (Kreiman et al., 2014) and adapt methods from the domain of face variability and perception (Burton et al., 2016). The driving question for Lee and colleagues is one of understanding what information exists in the signal and how it's structured. In many ways, this is the first step towards understanding which aspects of voice are available to listeners and thus useable in perceptual processes.

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

The takeaway from this work is that different voices share a handful of dimensions with one another and the group as a whole. Despite this shared structure, however, much of the way a voice varies is idiosyncratic. Typically shared dimensions were spectral shape and noise parameters in the higher frequencies, the fourth formant, and formant dispersion. The spectral measures are associated with vocal breathiness or brightness, and the formant-based measures with speaker identity and vocal tract size. Lee and Kreiman (2019) replicates this work with short samples of spontaneous speech from the same database, with the exception that fundamental frequency emerges as a shared relevant dimension. This arguably reflects the difference between read and spontaneous speech in English, with reading tending to be more monotone and spontaneous speech more affective. Lee and Kreiman (2020) replicates this work again with sentence reading in Seoul Korean, again finding minimal differences that are explained readily by typological differences from English. Unlike English, fundamental frequency and variability in the lower formants emerged as relevant dimensions in read Korean speech. The authors argue that this reflects phrasal intonation patterns that occur in reading.

Conceptualizing what these dimensions mean and how to think about acoustic voice variability in this way is somewhat of a challenge, given the abstractness of these measurements. The domain of faces thus provides a useful analogy for

thinking about what shared structure looks like compared to idiosyncratic aspects of the structure. Burton et al. (2016) found that all faces share dimensions of variability related to angle (i.e., looking up, down, or to the side) as well as lighting. Idiosyncratic variation in structure arose from things like facial hairstyle, makeup, and expressions. As with the face literature, Lee and colleagues argue that the structure of voice spaces supports a prototype model of voice perception (Lavner et al., 2001), in which novel individual voices are perceived in relation to a speech community average.

In any case, Lee et al. (2019) argue that familiarity with a voice arises from learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions. And indeed, familiarity with a voice pays off—listeners are good at identifying familiar voices, but perform poorly on the same tasks with unfamiliar voices (Nygaard and Pisoni, 1998). The prototype model merely proposes a mechanism by which listeners learn a novel talker's voice. It's not just familiarity with a particular individual's voice that facilitates processing, it's familiarity on various levels, including language.

Summarize Baumann and Belin, 2010 paper (Baumann and Belin, 2008).

Section on LFE here — write during friday writing time That bilingual listeners are sensitive to this information signals its importance (Orena et al., 2019; Fricke et al., 2016). The following paragraphs summarize this work. Voice advantage (Levi, 2018). Talker change detection (Sharma et al., 2020) Talker identification generalizes, but not amazingly well Cite Perrachione JASA paper somewhere in here (Perrachione et al., 2019).

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

Describing and analyzing acoustic voice variation in bilingual speech has motivation in both perception and production. As apparent from the language familiarity

effect literature listeners are capable of learning and identifying voices in one language and then generalizing across languages. Listeners are better at identification and discrimination when they have more familiarity with the language, but performance on such tasks tends to be well above chance. In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic information in the voice. Understanding the structure of that variability brings us one step closer to understanding what listeners are using from the signal to process speech. On the production side of things, bilingual speech presents an ideal test case for the designation of voices as auditory faces. If the structure of variability from each of a bilingual's languages is matched, then voices can be straightforwardly thought of as auditory faces.

Additionally, understanding the structure of the same talker's voice in each language lends additional validation to the arguments made by Lee and Kreiman (2020) for the differences between English and Seoul Korean sentence reading, a cross-study comparison of different populations. Across each of their studies, Lee and colleagues argue that both language and biological factors contribute to the structure of voice variation. Bilingual speech, again, presents an ideal test ground for disentangling biological and linguistic factors from one another. It is important to note that this dichotomy is somewhat misleading. While there ultimately are biological constraints on a voice (e.g., vocal tract length, pathologies, etc.), individuals nonetheless exert remarkable and wide-ranging control over their voice space (), and are highly capable of manipulating factors that are not linguistically important but which signal social and contextual information. This applies both within languages (), as well as across languages in the case of bilinguals (Bullock and Toribio, 2009). Thus in the case of bilinguals, the only aspect we can be truly confident in being held constant across languages is the biological part. The same "hardware" can be used for drastically different ends.

In this chapter, I examine how voice varies across a bilingual's two languages. Some differences are expected. While all languages have consonants and vowels, they differ in distribution, articulation, and acoustics (e.g., Munson et al., 2010). Suprasegmental and prosodic properties also vary across languages. Languages can differ in terms of whether a suprasegmental dimension is exploited at all. For example, does a language encode lexical tone contrastively? Another way languages

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

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

For example, a small group of English-Cantonese bilinguals (n=9) in did not differ in mean fundamental frequency (F0), but exhibited greater variability in F0 (Altenberg and Ferrand, 2006).

This was not the case in Ng et al. (2012), which examined voice differences with Cantonese-English bilinguals reading passages (n = 40). Based on Long-Term Average Spectral measures, females exhibited higher F0 in English than Cantonese, but males did not (Ng et al., 2012). In the same study, all participants had greater mean spectral energy values (mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese (Ng et al., 2012). Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English.

While the body of work in this area on Cantonese-English bilinguals remains small, there are useful insights from other language pairs.

Not all factors are necessarily easy to separate as linguistic or talker-instrinsic. For example... Speech rate shared across languages, for example

Together, these bodies of literature raise the question of whether bilingual talkers have the "same" voice in each of their languages. Using the corpus described in Chapter **??**, I look at spectral properties (Cheng, 2020; Altenberg and Ferrand,

2006; Ryabov et al., 2016; Ng et al., 2012), and also examine how acoustic variation is structured, following the work of Kreiman et al. (2014) and Lee et al. (2019). Describe and discuss psychoacoustic voice quality model

This work builds on Lee et al. (2019) in a handful of ways: it extends the methods to the case of bilinguals, considers longer samples, and addresses the role of sample duration both within and across talkers and languages.

3.2 Methods & Results

3.2.1 Data

The data used in this analysis come from the conversational interviews in the SpiCE corpus—both Cantonese and English are considered. As noted before, the 34 talkers studied here are all early Cantonese-English bilinguals. For additional information about the participant, please refer to sections **??** and **??** in the previous chapter.

While prior work uses short chunks of speech, the present analysis is focused on longer stretches of spontaneous speech. Additionally, there are practical reasons to exclude the sentence reading and storyboard tasks from this analysis. The sentence sets... Similarly, there are imbalances in the storyboard task. As talkers narrated the same story in both languages, they were often more confident the second time around.

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. As a reminder, both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. For the analysis in this chapter, the participant channel was extracted from the stereo recordings, including any code-switches they made during the interview. While it would be possible to exclude items not produced in the main interview language from the final sample using the time-aligned transcripts, this was not done. The driving reason for keeping code-switches in the analysis is that such code-switches are representative of the particular talker's language behavior. Further, just because someone switches lanuages, does not mean that they full switch. For example, individual words may be borrowed in and pronounced

with the phonology of the main language.

All voiced segments were identified with the *Point Process (periodic, cc)* and *To TextGrid (vuv)* Praat algorithms (Boersma and Weenink, 2021), implemented with the Parselmouth Python package (Jadoul et al., 2018). The pitch range settings used with *Point Process (periodic, cc)* were set to 100–500 Hz for female talkers, and to 75–300 for male talkers. While speech from the interviewer can occasionally be heard in the participant channel, it is quiet enough to have been largely ignored by the Praat algorithms. This method of identifying voiced portions of the speech signal captures vowels, approximants, and some voiced obstruents. This differs slightly from the methods described in Lee et al. (2019), the paper on which the methods of this chapter were modeled.

3.2.2 Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by Lee et al. (2019), with the exception of formant dispersion, which was excluded given its near perfect correlation with the measured value of F4. The choice of measurements in Lee et al. (2019) comes from the psychoacoustic voice quality model described in the introduction to this chapter (Kreiman et al., 2014). Measurements were made every 5 ms during voiced segments, as in Lee et al. (2019), using VoiceSauce (Shue et al., 2011). measurements were:

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 ALGORTHIM (), which is widely regarded to be more accurate that the alternative choices in VoiceSauce. It is one of the more widely studied variables on this list, as evidenced by the literature cited in the introduction (e.g.,)

F1, F2, and F3 The first three formant frequencies—also measured in Hertz—are typically discussed in relation to linguistic contrasts, particularly vowel and sonorant consonants.

F4 The fourth formant frequency is not typically discussed in linguistic contexts, and is instead associated with talker characteristics. In this light, it is not

particularly surpriseing that it was highly correlated with formant dispersion. Both measures reflect talker characteristics such as vocal tract length. F4 is measured in Hertz.

H1*–H2* The corrected amplitude difference between the first two harmonics is one of four primary measures used to characterize source spectral shape (Kreiman et al., 2014). It is associated with phonation type, and blah blah blah. The asterisks here and in the following spectral slope measures indicated that the value has been corrected (Iseli et al., 2007), in order to account for the impact of nearby formants on the amplitudes of harmonics. units

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

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

H2kHz*–H5kHz* The corrected amplitude difference between the harmonics closest to 2000 Hz and 5000 Hz is a measure of harmonic spectral slope that does not depend on. It captures the highest frequency band of the four shape measures, and it is associated with... units

CPP Cepstral Peak Prominence is a measure of the ratio between harmonic energy and spectral noise, and is associated with non-modal phonation types. As CPP is a ratio, it does not have units.

Energy Root Mean Square (RMS) **Energy** is a measure of spectral noise that reflects overall amplitude. units

SHR The subharmonics-harmonics amplitude ratio is a measure of spectral noise associated with period doubling or irregularities in phonation. While based on amplitude, this ratio is unitless.

The raw data output from VoiceSauce is available in the supplementary materials for this dissertations.

3.2.3 Exclusionary criteria and post-processing

Given the nature of the corpus and methods thus far, there is reason to suspect a sizable number of erroneous measurements. In an effort to filter these out prior to analysis, measurements were subjected to exclusionary criteria focused on identifying impossible values. Observations were excluded in cases where any of the following measurements had a value of zero: F0, F1, F2, F3, F4, CPP, or (uncorrected) H2kHz–H5kHz. 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. Filtering with CPP says... Only one of the uncorrected harmonic amplitude measures, as erroneous values tended to co-occur on the same observation, and the distribution of H2kHz–H5kHz did not span zero, with the exception of a spike of (erroneous) values equal to zero. This operationalization minimizes the removal of correctly measured zero values, which would have occurred with one of the other spectral shape parameters (corrected or uncorrected).

Moving standard deviations were calculated for each of the 12 measures using a centered 50 ms window, such that each window includes approximately ten observations. The moving standard deviations capture dynamic changes for each of the voice quality measures, which is important as they may better reflect what listeners attend to in talker identification and discrimination tasks (Lee et al., 2019). This analysis uses moving standard deviations, as opposed to the coefficients of variation used by Lee et al. (2019). This should not have any undue effect on the outcome, as all variables were scaled prior to inclusion in the principal components analysis described in the next section. The last round of exclusionary criteria uses these moving standard deviations. If an observation was missing a moving standard deviation value, it was removed. This means that obsevations falling extremely close to a voicing boundary were not included.

Overall, 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 are used in the analyses described in the following sections. Across the

34 talkers, there were 3,126,267 observations after winnowing the data. These observations were not evenly distributed across talkers and langauges. NEW not implemented yet: In order to control for the impact of passage length in the analysis, the number of samples for each talker was capped to include only the first 22,433 samples were considered for each interview. This value was selected as it represents the interview with the fewest number of samples across all talkers and languages. Following this last winnowing step, there were 1,525,444 total observations. While the winnowing process removed a lot of data, the number of samples here is still substantially larger than used in Lee et al. (2019), where the per-talker sample count was closer to 5,000. Further, a later section in this chapter directly tackles the issue of passage length, and (spoiler) indicates that 20,000 is sufficient for stabilization.

3.2.4 Crosslinguistic comparison of acoustic measurements

Following from prior work, the first step in this analysis is a crosslinguistic comparison for each talker and measure. As discussed in the introduction to this chapter, there are some often (but not always) found differences for Cantonese and English. ADD 1-SENTENCE SUMMARY

Figure XX depicts the distribution of values for each of the measurments across languages. Given the highly skewed shape of

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

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

H1*–H2* was significantly higher in Cantonese for a relatively large subset of the talkers (13/34), lower for a small number (3/34), but trivial for most (18/34).

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

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

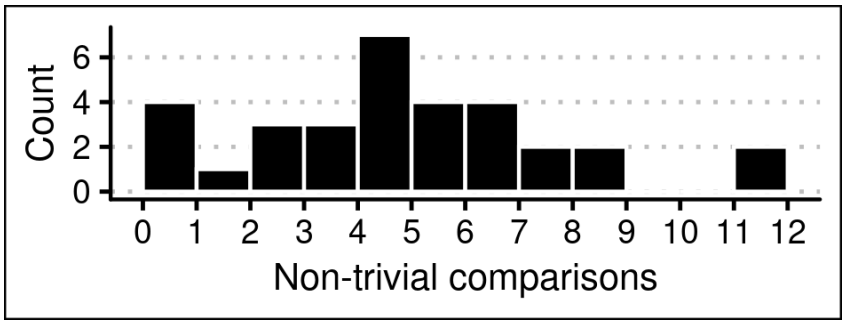

Figure 3.1: A summary of the number of non-trivial comparisons from Table 3.1 across the 34 talkers.

While based on a different measure than (Ng et al., 2012), this is consistent with the finding that Cantonese tends to be breathier, or English creakier—the current analysis does not distinguish between these interpretations.

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

3.2.5 Principal components analysis

Methods

Principal components analysis (PCA) is a dimensionality reduction technique appropriate for data that include a large number of (potentially) correlated variables. The distillation into components helps identify and facilitate describing the internal structure, in this case, of a voice. While the typical goal in analyses that use PCA is to identify a smaller number of components to use in modeling, the focus here is instead on understanding the internal structure. In this light, the components themselves will be examined.

I adapt methods from work on voices (Lee et al., 2019; Lee and Kreiman, 2020)

and faces (Burton et al., 2016; Turk and Pentland, 1991). The goal is to the capture similarities or differences in the structure of each talker's voice across languages. As such, I conducted PCAs separately for each talker-language pair, and compared the results of each talker's English and Cantonese PCAs. All 24 measures were normalized (z-scored) on by-PCA basis prior to the analysis. PCAs were implemented with the *parameters* package (Makowski et al., 2019) in R (R Core Team, 2020), using an oblique *promax* rotation to simplify the factor structure, as the measurements reported in the previous section were expected to be somewhat correlated given prior findings (Lee et al., 2019), and a broader undesttanding of how different acoustic measures align with one another (Kreiman et al., 2014, 2021).

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

3.2.6 PCA results

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

TABLE HERE

To assess whether talkers exhibit the same structure in voice variability across their languages, I first consider the patterns present across the different PCAs, as this provides context for understating what unique structural characteristics in talkers' voices looks like. To this end, I briefly summarize common patterns across PCA components, regardless of how much variance they account for, as the difference is often quite small. Figure 3.2 shows the first four components of a single

talker's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages. It also highlights the importance of not attributing to much value to the ordering of components, but rather to their composition and variance accounted for.

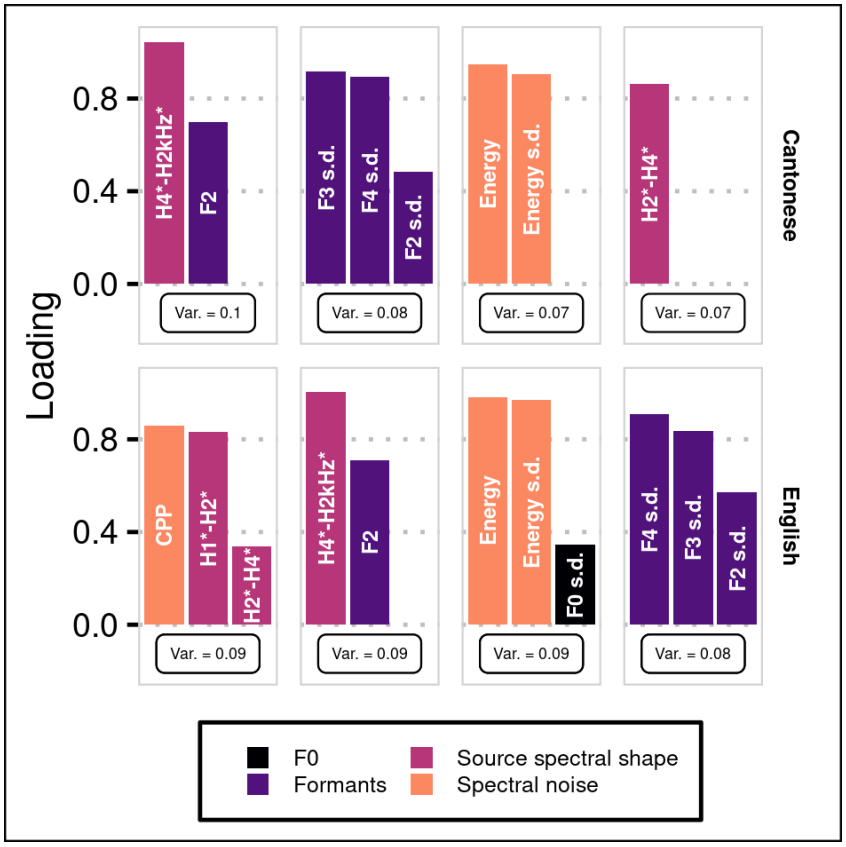

Figure 3.2: In the first four components of a talker's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

Broadly speaking, there were a lot of similarities in component composition across both talkers and languages, with the eight most commonly occurring components summarized in Table 3.2. For context, recall that PCAs had anywhere from 10–15 components total. These eight components consisted of source spec-

tral shape, spectral noise, as well as formant variables. On the other hand, F0 co-occurred with a wide variety of variables (often Energy), but in a manner that was less consistent across talkers. There were additional components (not reported here) that were shared by less than half of talkers. In summary, despite the greater amount of shared structure across PCAs than found in Lee et al. (2019), there is still ample room for idiosyncratic variation, both in terms of which variables co-occur, as well as in how much variance different components account for.

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

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

3.2.7 Canonical redundancy analysis

Methods

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

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

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

Results

Redundancy indices for within-talker comparisons ranged from 0.82 to 0.99, (*Mdn* = 0.93, *M* = 0.92, *SD* = 0.04), and are displayed in Figure 3.3, with the two redundancy indices for a given pair plotted against one another. Comparisons across talkers within-language (range: 0.63–0.98, *Mdn* = 0.84, *M* = 0.84, *SD* = 0.6) and

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

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

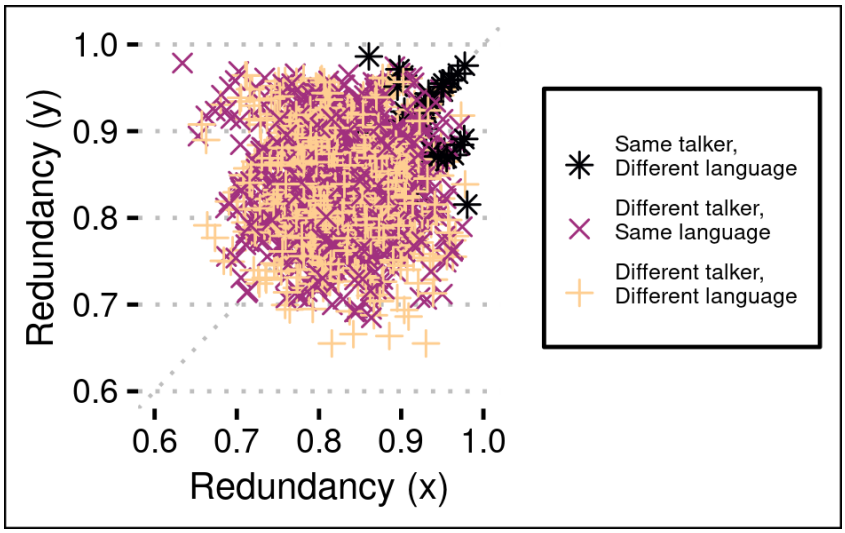

Figure 3.3: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are clustered at the top right.

3.3 Discussion and conclusion

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

an apt comparison.

The comparison of these 34 Cantonese-English bilinguals' voices across languages suggest more similarity for an individual across languages than found within a more tightly controlled group of monolingual English speakers (Lee et al., 2019)—several analysis decisions may have contributed to this. I compared similar components independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that any comparisons made are among like items. Any downside to this methodological decision is mitigated by the fact that most components made relatively small contributions, accounting for 4.2–10.3% (95% highest density interval) of the PCA's total variance.

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

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

Bibliography

Altenberg, E. P. and Ferrand, C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, 20(1):89–96. → pages 6, 13

Baumann, O. and Belin, P. (2008). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, 74(1):110–120. → page 4

Belin, P., Fecteau, S., and Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135. → page 1

Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Version 6.1.38. → page 8

Bullock, B. E. and Toribio, A. J. (2009). Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, L., Winford, D., and deBot, K., editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam. → page 5

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., and Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1):202–223. → pages 3, 4, 14

Cheng, A. (2020). Cross-linguistic F0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, 147(2):EL67–EL73. → page 6

Fricke, M., Kroll, J. F., and Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, 89:110–137. → page 4

Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America*, 121(4):2283–2295. → page 9

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 8

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → pages 14, 17

Keating, P., Kreiman, J., and Alwan, A. (2019). A new speech database for within- and between-speaker variability. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, pages 736–739, Melbourne, Australia. → page 3

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, 1(1):e009. → pages 2, 3, 7, 8, 9, 14

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, 149(1):457–465. → pages 2, 14

Lavner, Y., Rosenhouse, J., and Gath, I. (2001). The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology*, 4(1):63–74. → page 4

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3):1568–1579. → pages 1, 2, 3, 4, 7, 8, 10, 11, 13, 14, 16, 19

Lee, Y. and Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. → pages 2, 3

Lee, Y. and Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. 10.1121/1.5146847. → pages 2, 3, 5, 13

Levi, S. V. (2018). Another bilingual advantage? Perception of talker-voice information. *Bilingualism: Language and Cognition*, 21(3):523–536. → page 4

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461. → page 2

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 14

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → page 18

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

Ng, M. L., Chen, Y., and Chan, E. Y. (2012). Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—a long-term average spectral analysis. *Journal of Voice*, 26(4):e171–e176. → pages 6, 7, 13

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376. → page 4

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

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, 146(5):3384–3399. → page 4

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → page 14

Ryabov, R., Malakh, M., Trachtenberg, M., Wohl, S., and Oliveira, G. (2016). Self-perceived and acoustic voice characteristics of Russian-English bilinguals. *Journal of Voice*, 30(6):772.e1 – 772.e8. → page 7

Sharma, N. K., Krishnamohan, V., Ganapathy, S., Gangopadhayay, A., and Fink, L. (2020). Acoustic and linguistic features influence talker change detection. *The Journal of the Acoustical Society of America*, 148(5):EL414–EL419. Publisher: Acoustical Society of America. → page 4

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences*, volume 3, pages 1846–1849, Hong Kong. → page 8

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163. → page 17

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 14

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Sco. Press. → page 14