Chapter 3

The structure of acoustic voice variation

3.1 Introduction

Voices can tell you a lot about the person who is talking, and have been discussed as an "auditory face" (Belin et al., 2004). They simultaneously provide information about the talker's current physical and emotional state, as well as cues to who they are (Belin et al., 2004). Understanding the dimesions that conspire together to form a voice is no small feat. It is similarly challenging to understand which factors contribute to processing and talker identification from a listener's perspective. The reason these areas are difficult to understand and study is because voices vary drastically. While there are certainly shared attributes and acoustic dimensions comprising disparate voices, much of the variation appears idiosyncratic in nature (Lee et al., 2019). Conceptually, this makes sense from the perceptual end—if voices didn't have anything unique, how would we ever identify them? Yet understanding and processing variability remains one of the big puzzles of speech perception, the "lack of invariance" problem in certain circles.

There is good news in the domain of voice quality. Kreiman and colleagues have synthesized work in various areas and put forth proposals for how both ends of the process work. The psychoacoustic model of voice quality described in (Kreiman et al., 2014)—and later validated in (Kreiman et al., 2021)—argues for a set of

acoustic measurements that is sufficient and necessary to capture a wide range of normal and disorderd voice qualities. These measurements range from... to..., and while each individual item on the list could be examined in isolation, Kreiman and colleagues argue that they are more than the sum of their parts. The measures covary and conspire together in different ways that to form a percept. While this model establishes a set of measurements to sufficiently describe a voice, it does not arbitrate between the measures in a way that establishes what matters for a given voice in a given language.

Work towards understanding the structure of voice quality variation and what it means for perecptions in a variable world remains in its early stages, particularly in the case of voices. Lee et al. (2019) approach this topic by leveraging the psychoacoustic model of voice quality and adapting the methods using in the domain of face variability and perception (Burton et al., 2016). Lee and colleagues work is largely focused on examining acoustic data to get an idea of what information exists in the signal, before speculating as to whether or not a listener uses it. In their paper—the methods of which will be described in greater detail later on—they make a case for a prototype model of voice perception based on how individuals do and do not vary from the a community average of sorts. In effect, this mirrors the findings in the face domain.

The prototype model argument is perhaps easier to conceptualize with faces. Burton et al. (2016) find that faces share a set of relevant dimensions in how they vary. These are things like ..., that is ar eyou looking at a face straight on? In profile? Another shared dimension is lighting. These are things that all faces are subject to. idiosyncratic dimensions are things like expressions, makeup, or facial hair. There might be broad similarity across different subsets of individuals, but not all faces vary over time in that way. The analogs for voice are somewhat less readily available, but nonetheless relevant. Lee et al. (2019) find that....

Despite the high degree of variability, idiosyncrasies, and similarity or deviation from a prototype, listeners nonetheless use talker-specific information to recognize and discriminate voices. Listeners are good at identifying familiar voices, but perform poorly on the same tasks with unfamiliar voices (Nygaard and Pisoni, 1998). It was suggested by Lee et al. (2019) that familiarity with a voice largely comes from learning how that voice varies across time and space, whether within

an utterance or across environments, physical states, and emotions. evidence here from talker identification and discrimination work that there si both linguistic and crosslinguistic type stuff

Lee et al. (2019) restrict variability in their analysis by focusing on English speech produced by native English speakers from the UCLA Variability Speaker Database (Keating et al., 2019). They use....methods.

This work is groundbreaking in a lot of ways, even if it looks only at a particular dialect and register of American English spoken by monolinguals in southern California CHECK. It is comforting that subsequent work by Lee and colleagues has extended the basic conceptualization to connected speech (Lee and Kreiman, 2019) and to Seoul Korean (Lee and Kreiman, 2020). Across the original paper and follow up presentations, voice variability was largely accounted for by harmonic and noise spectral shape parameters associated with a voice's timbre (brightness, breathiness). In the extension to English connected speech, the behavior of fundamental frequency differed, with it accounting for more variation in spontaneous speech. This tracks from assumptions about read speech following more predictable (monotone) patterns, and spontaneous speech being more affective in nature. In the case of Seoul Korean sentence reading, fundamental frequency and lower formant frequencies emerged as important dimensions of variation. The authors argue that this reflects phrasal intonation patterns that occur in reading.

While this body of work compares outcomes for English speakers and Korean speakers, the groups considered are necessarily distinct populations. Lee and Kreiman (2020) argue that there are both language and biologial factors contribute to the structure of voice variation. Bilingual speech would be an ideal testing ground for examining this distinction. The (likely unintended) dichotomy of language and biology is misleading. While there are ultimately vocal tract and other biological constraints on a voice, individuals show remarkable control over the space (), and are 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 differnt ends.

In this chapter, I examine how voice varies across a bilingual's two languages. Some differences are expected. Languages differ in terms of their consonant and vowel inventories, which affect the spectral properties of a language. While all languages have consonants and vowels, they differ with respect to distribution, articulation, and acoustics (e.g., Munson et al., 2010). For example...phrasal accent stuff in korean, danish vowels vs. 3-vowel system, slavic consonants vs hawaiian

Some prior work has looked at various vocal features across languages, both in the case of bilinguals and comparing separate monolingual populations. While this work uses rather broad strokes to compare crosslinguistic behavior—things like means and standard deviations, as opposed to a deep dive into the structure of variability—it captures some interesting and useful generalizations, suggesting that while some aspects of voice variability differ for linguistic reasons, other talker-indexical features remain constant across languages, and still others can be influenced by both linguistic and non-linguistic factors.

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 language 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? Expand this

Within an individual bilingual, the acoustic variability within each language can also be related to the social identities a talker adopts within each language (see discussion in Cheng (2020)). Further, bilinguals are sophisticated social actors, and can conciously leverage their use of language to convey different aspects of identity... add more here

In an effort to understand what aspects of an individual's voice vary across languages and what are more or less fixed talker-specific attributes, researchers have compared particular spectral properties of bilingual speech, both within and across talkers. Results have been decidedly mixed (Cheng, 2020; Altenberg and Ferrand, 2006; Ryabov et al., 2016). Big picture summary, and flesh out each of the following in greater detail.

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 similar language pairs. In a study of Mandarin and English...

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

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. Talker identification generalizes, but not amazingly well

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

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 informa-

tion 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, particuarly 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 associ-

ated 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 only the first XXX 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 XXX total observations.

3.2.4 Principal components analysis

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. I adapt methods from work on voices (Lee et al., 2019) and faces (Burton et al., 2016; Turk and Pentland, 1991). The goal is to the capture similarities or differences for 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 for 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 (Lee et al., 2019).

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); however, all loadings were retained for the canonical redundancy analysis described in the next section.

3.2.5 Canonical redundancy analysis

In order to assess whether variation in a talker's voice is structurally similar across both languages, I compare the PCA output from English and Cantonese by calculating redundancy indices from 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. 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).

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.

3.3 Results

3.3.1 Crosslinguistic comparison of acoustic measurements

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). 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

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

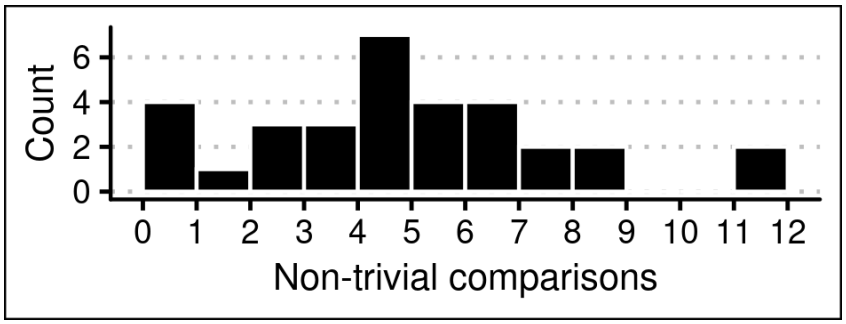

Figure 3.1: A summary of the number of non-trivial comparisons from Table 3.1 across the 34 talkers.

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.

3.3.2 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. 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.

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-

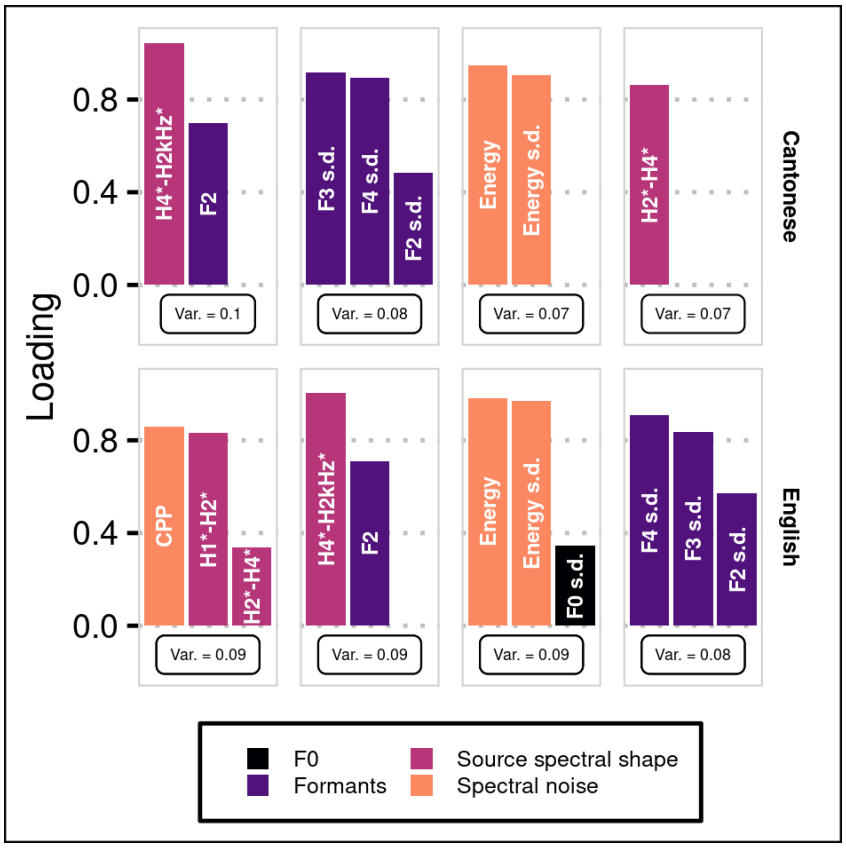

Figure 3.2: In the first four components of a talker's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

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

3.3.3 Within-talker analysis

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). 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 3.3.2, there were a substantial number of commonly occurring patterns across talkers and languages.

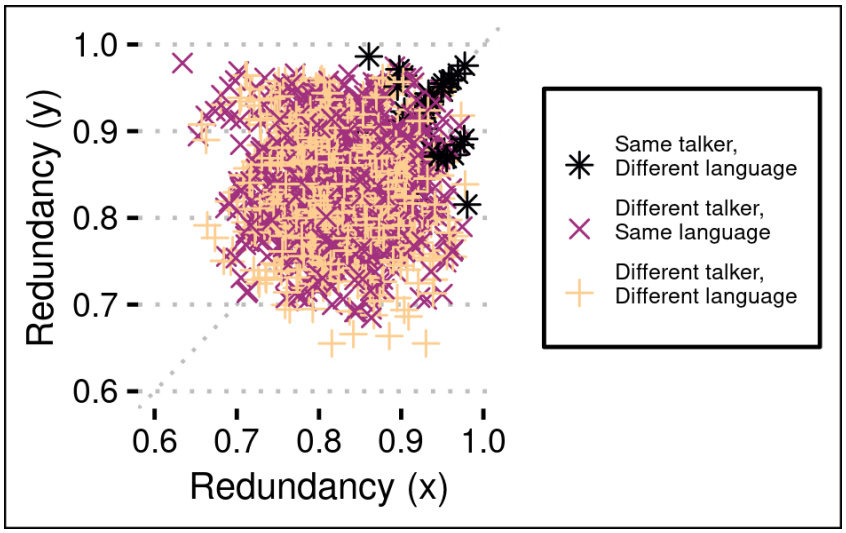

Figure 3.3: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are clustered at the top right.

3.4 Discussion and conclusion

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

The comparison of these 34 Cantonese-English bilinguals' voices across languages suggest more similarity for an individual across languages than found within

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

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

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

Bibliography

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 4, 5, 13

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 6

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 3

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 2, 10

Cheng, A. (2020). Cross-linguistic F0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, 147(2):EL67–EL73. → pages 4, 5

Fricke, M., Kroll, J. F., and Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, 89:110–137. → page 5

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 7

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 6

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → page 10

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 1, 5, 7

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. → page 1

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, 5, 6, 7, 9, 10, 14, 17

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

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

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 10

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → page 16

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 4

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 5, 11, 13

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376. → page 2

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 5

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → page 10

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. → pages 4, 5

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 7

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163. → pages 10, 11

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 10

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 10