Chapter 3

The structure of acoustic voice variation in bilingual speech

3.1 Introduction

Voices provide a lot of information about the person talking, ranging from their current physical and emotional state to talker indexical features that help listeners identify who they are. In this context, voices can be described as auditory faces—they are uniquely individual yet share basic characteristics with the broader population (Belin et al., 2004). Voices convey this rich array of information along with the message being communicated. Understanding the structure of a voice is no small feat, as it means understanding how listeners leverage different vocal dimensions to process talker-indexical, affective, social, and linguistic information. The processing challenge arises from the sheer variability across voices.

Though voices share some attributes, they also vary in unique ways (Lee et al., 2019). From the perspective of voice perception, the balance between shared and idiosyncratic characteristics makes sense. The shared dimensions allow listeners to recognize the sound they hear as a voice. They also help listeners perceive, classify, and understand new voices. Idiosyncrasies, on the other hand, enable identification and discrimination between different voices. While this makes sense conceptually, understanding the structure of voice variation in speech production and its complement in listeners' ability to process that information remains an active area of

research. The focus of this chapter is acoustic voice variability, and the emphasis on describing and processing variation echoes one of the big puzzles in phonetics: the "lack of invariance" problem (Liberman et al., 1967). That is, given the ubiquity of variation, how do perceivers efficiently extract relevant and important information from the communicative signal? This chapter foregrounds the speech signal itself, asking what is available in the speech production for listeners to use.

While variation is indeed wide-ranging, it remains far from random. Some of the most prevalent accounts of how individuals understand and process variation emphasize its structure. While this chapter looks at the structure of voices and the following chapter examines sound category structure, both attempt to elucidate structure in the speech signal that may be beneficial for listeners in understanding and processing new talkers.

In the domain of voice, 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 one (Kreiman et al., 2021, and references therein). As a result, Kreiman and colleagues argue that this set of dimensions is both sufficient and necessary to capture a wide range of normal and disordered voices. The psychoacoustic model of voice quality includes acoustic dimensions that capture vocal tract anatomy, pitch, loudness, harmonic voice source, and inharmonic voice source characteristics. While each dimension in the model could be considered independently by researchers—some of which are studied in isolation—Kreiman and colleagues argue that these dimensions are more than the sum of their parts. The measures covary and conspire together to form a multidimensional percept of voice. While this model establishes a set of acoustic dimensions, it does not arbitrate between them in a way that establishes what matters for perceiving and processing a particular voice in a particular 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 is 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, Lee and colleagues leverage the psychoacoustic model of voice quality (Kreiman et al., 2014) and adapt methods from the domain of face variability and perception (Burton et al., 2016). Their driving question is one of understanding the structure of acoustic information in the speech signal. In many ways, this is the first step towards understanding which aspects of voice are available to listeners and thus useable in perceptual processes, particularly when coarse summary statistics do not indicate cross-talker differences.

To drill down into the structure of voice variability, Lee et al. (2019) use a series of principal components analyses 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. Commonly shared dimensions were spectral shape and noise parameters in the higher frequencies, the fourth formant, and formant dispersion. These spectral measures are associated with vocal breathiness or brightness. The formant-based measures are typically associated with speaker identity and vocal tract size. Lee and Kreiman (2019) replicates this work with short samples of spontaneous speech from the same database. The results were similar, with the exception that F0 emerges as a shared relevant dimension. This result arguably reflects the difference between reading and spontaneous spoken English, with reading tending to be more monotonous and spontaneous speech exhibiting more affective qualities. Lee and Kreiman (2020) replicates this work again with sentence reading in Seoul Korean, again finding minimal differences that are explained readily by typological differences from English. Unlike English,

F0 and variability in the lower formants emerged as relevant dimensions in read Korean speech. The authors argue that this reflects phrasal intonation patterns that occur in Korean reading.

Conceptualizing what these dimensions mean and how to think about acoustic voice variability in this way is challenging, as many of the acoustic dimensions considered do not map neatly onto a single percept. F0 is a straightforward example, given its clear relationship to pitch. Many of the spectral measures, both harmonic and noise-based, are much more challenging to interpret without considering multiple measures simultaneously. 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 things like lighting and viewing angle (i.e., looking up, down, or to the side). Idiosyncratic variation in face structure arose from things like facial hairstyle, makeup, and expressions. Burton et al. (2016) discuss their results as supporting a prototype model of faces. 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; Latinus and Belin, 2011), in which novel individual voices are perceived in the context of a speech community average, or prototype. Their results add to the argument that faces and voices share processing mechanisms (Yovel and Belin, 2013).

In any case, Lee et al. (2019) argue that familiarity with a voice arises from learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions. And indeed, familiarity with a voice pays off—listeners are good at identifying familiar voices but perform poorly on the same tasks with unfamiliar voices (Nygaard and Pisoni, 1998). The prototype model merely proposes a mechanism by which listeners learn a novel talker's voice.

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

2019). One of the biggest takeaway points from this literature is the Language Familiarity Effect (LFE), which encompasses a broad range of findings where listeners are better at identifying talkers in a familiar language (for a recent review, see Perrachione, 2018). Bilinguals are especially good at this kind of task and show evidence of generalizing across languages (Orena et al., 2019).

Very little of this work identifies what listeners use in the signal, and as such, claims about the relative importance of linguistic or talker-indexical information should be tempered. However, there are exceptions to this. For example, Perrachione et al. (2019) collected perceptual voice (dis)similarity ratings for Mandarin and English voices by Mandarin and English native listeners and report on the relationship between several acoustic measurements and rating data. Perrachione et al. (2019) found that when the talker was the same, regardless of the manipulations used in the study (language and time-reversal), all listeners rated stimuli pairs as highly similar. This result highlights that listeners are sensitive to low-level acoustic information present in voices, regardless of whether they know the language or understand the stimuli. Additionally, Perrachione et al. (2019) found that some acoustic measurements predict similarity ratings, while others do not. F0 was the most prominent measure, which is unsurprising given its salience, and how much the voice variability literature has focused on it (e.g., Keating and Kuo, 2012). Other measures predicting similarity were the harmonics-to-noise ratio and formant dispersion, which are associated with voice quality and vocal tract size, respectively. That listeners appear to use these measures is of direct relevance to the study presented in this chapter, and represents a point that will be returned to in this chapter's discussion.

In light of this 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 can affect voice quality. This chapter builds on Lee and colleagues' body of work by extending their methods to the case of spontaneous bilingual speech.

Describing and analyzing acoustic voice variation in bilingual speech has moti-

vation 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 (e.g., Orena et al., 2019). In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic information in the voice—or so the argument goes (Perrachione et al., 2019). Understanding the structure of that variability brings us one step closer to understanding what listeners are using from the signal to process speech, as it limits the hypothesis space. On the production side of things, bilingual speech presents an ideal test case for the argument that voices function like auditory faces. If the structure of variability from each of a bilingual's languages is matched, then voices can be straightforwardly thought of as auditory faces.

Additionally, examining 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. In comparing these studies, Lee and colleagues argue that both language and biological factors contribute to the structure of voice variation. Bilingual speech, again, presents an ideal test ground for disentangling biological and linguistic factors from one another. While common in the literature, this dichotomy is somewhat misleading. Voices ultimately have biological constraints, such as vocal tract lengths or pathologies. Yet at the same time, 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 across all aspects of an individual's linguistic repertoire (Bullock and Toribio, 2009; Wei, 2018). Thus in the case of bilinguals, the only aspect we can be truly confident in being held constant across languages is the biological part. The same "hardware" can be used for drastically different ends.

In this chapter, I examine how voice varies across a bilingual's two languages. Some differences are expected, despite the characterization of voices as auditory faces. While all languages have consonants and vowels, they differ in distribution, articulation, and acoustics (e.g., Munson et al., 2010). Suprasegmental and prosodic properties also vary. Languages differ in terms of whether a suprasegmental di-

mension is made use of 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 the languages considered are Cantonese (a language with lexical tone) and English (a language without lexical tone). Segmental and suprasegmental differences both have cascading effects on voice quality.

The following paragraphs detail comparisons that have been made between English and Cantonese in the literature thus far. As there is an additional body of work comparing English and Mandarin Chinese—typologically similar to Cantonese—comparisons between English and Mandarin are also summarized. While the most relevant comparisons for the present work are those made within 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 for a small set of spectral measurements. Results have been decidedly mixed.

In a small study of Cantonese-English bilingual (n=9), Russain-English bilingual (n=9), and English monolingual (n=10) young women, Altenberg and Ferrand (2006) examined F0 patterns in conversational speech across the different languages and populations. As some languages reportedly have different mean F0 (e.g., Keating and Kuo, 2012), Altenberg and Ferrand (2006) focused on whether F0 shifts when an individual switches languages and whether different languages have different baselines. Ultimately, Russian-English bilinguals exhibited differences in mean F0, and Cantonese-English bilinguals did not. Though, they did produce a wider F0 range in Cantonese compared to their English. While the results in Altenberg and Ferrand (2006) ultimately paint a coarse picture of bilingual F0 production with a small sample size, they highlight an important point of departure—bilinguals can differ in F0 across languages.

In a larger study of Cantonese-English bilinguals reading passages (n=40), Ng et al. (2012) examined a variety of different voice measures with both male and female talkers. Results were based on Long-Term Average Spectral (LTAS) measures. Female talkers exhibited higher F0 in English than Cantonese, but males did not. In the same study, all participants had greater mean spectral energy values

(mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese (Ng et al., 2012). Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English. The LTAS measure of the first spectral peak did not differ across languages, suggesting that vocal stiffness remained consistent in the bilinguals' two languages.

Ng et al. (2010) examined F0 in spontaneous speech from 86 Cantonese-English bilingual children and found it to be lower in Cantonese compared to English. This corroborates Ng et al. (2012), and goes against the nonsignificant difference in (Altenberg and Ferrand, 2006). This mixed bag of results could ultimately be attributed to differences in sample sizes, the quantity of speech analyzed, or in language backgrounds of the bilinguals studied. While the picture regarding voice quality measures appears clearer and more consistent, those conclusions arise from a single study. In any case, these three studies offer reason to expect that Cantonese and English might differ in measures associated with pitch and phonation type.

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

non-native language arguably reflects factors like stress or confidence (Järvinen et al., 2013; Lee and Sidtis, 2017).

The speculation that higher F0 is a feature of tone languages does not align with the observation in Ng et al. (2012), who argued the opposite for Cantonese: that lower F0 could be accounted for by lexical tone. While the tone inventories for Cantonese and Mandarin have substantial differences, it seems clear that appealing to the presence or absence of lexical tone is too simplistic of an account. Alternatively—or perhaps, concurrently—talkers may be expressing different social and cultural identities in each of their languages (Loveday, 1981; Voigt et al., 2016). Regardless of whether language, experiential, or social factors drive differences across languages, this body of work highlights the importance of comparing within the same task.

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

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

Russian was faster. This result goes against the findings for the bilinguals studied in Altenberg and Ferrand (2006), where Russian exhibited consistently higher F0 than English. While higher F0 and slower speech rates can be characteristics of speech by non-native or non-dominant speakers (Järvinen et al., 2013), such an explanation cannot account for both outcomes.

Another example of less than clear-cut results comes from Ordin and Mennen (2017). They demonstrate differences in F0 range and level across languages for female Welsh-English bilinguals in a reading task, for whom Welsh had a higher and wider F0 range. This result did not hold for males from the same population, who varied more in their F0 level and range. The authors argue that the crosslinguistic difference is likely to be sociocultural in this case, as different patterns were observed for male and female speakers on a within-speaker basis. This gender difference means that the result is unlikely to be due to anatomical or purely linguistic reasons.

Considering these studies together, a few key observations are especially relevant to the present chapter. While studying bilingual talkers provides a clear path to disambiguating the role of anatomical differences in voices, it does not necessarily facilitate disentangling linguistic and sociocultural factors from one another. Most likely, both contribute simultaneously to the differences in voice patterns across languages. For example, there is clear evidence that Korean has a higher F0 than English, given results from two studies with different populations of bilinguals Cheng (2020); Lee and Sidtis (2017). Conversely, Ordin and Mennen (2017) show social rather than linguistic stratification.

This body of work mostly focuses on linguistic and social differences. While some of it dives into individual differences, between-talker variability should perhaps be given more of a spotlight. In work with speech rate, Bradlow et al. (2017) found that some talkers are fast and others are slow and that some languages are fast while others are slower. Crucially, these relationships held across talkers in various languages. That is, if someone was a fast talker in their dominant language, they were also a fast talker in their non-dominant language, and likewise for slow talkers. In this sense, both talker-indexical and linguistic (or sociocultural) factors contribute to speech rate behavior. It is not a particularly big leap to suggest that other speech signal variables might pattern in the same way. Adding to this picture

of variability across individuals, it is important to remember that bilinguals are sophisticated social actors and are fully capable of tailoring their speech behavior to a wide variety of contexts (Bullock and Toribio, 2009).

While this body of work highlights important points, it is limited by its laser focus on F0, with occasional forays into speech rate, intensity, and other spectral measures. The focus on F0 is not without reason—Perrachione et al. (2019) found it to be the most important perceptual dimension for voice similarity ratings. Yet at the same time, there is so much more to voice than pitch, particularly if the characterization of voices as auditory faces holds up.

This chapter brings together work describing crosslinguistic voice differences and work describing the structure of acoustic voice variation, to provide a more comprehensive picture of how voices vary across languages. Using the corpus introduced in **??**, I describe various spectral properties Ng et al. (e.g. 2012), and also examine how acoustic variation is structured, following the work of Kreiman, Lee, and colleagues (Kreiman et al., 2014; Lee et al., 2019). This chapter builds on Lee et al. (2019) in a handful of ways: it extends the methods to the case of bilinguals, considers longer samples, and addresses the role of sample duration both within and across talkers and languages. I also extend their methods by introducing a mechanism to assess structural similarity within and between individuals and languages.

3.2 Methods & Results

3.2.1 Data

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

While prior work by Lee and colleagues (e.g., Lee et al., 2019) uses relatively short chunks of speech, the present analysis is focused on longer stretches of spontaneous speech. While it would have been possible to include the sentence reading

and storyboard task recordings from each participant, there are practical reasons for excluding them from the analysis. The sentence sets were overall quite short and thus unlikely to be sufficiently representative on their own. Additionally, as many of the SpiCE talkers were not confident in their Cantonese reading, there was a wide range of familiarity with the materials represented. Some talkers knew all of the sentences, and others struggled with some of them. This renders the sentences less comparable to their English counterparts in the SpiCE corpus. There are also imbalances in the storyboard task. As talkers narrated the same story in both languages, they were often more confident the second time around. Excluding both of these tasks is motivated by prior work that highlights how confidence (Järvinen et al., 2013) and speaking style (Lee and Sidtis, 2017) impact voice quality.

As discussed in the previous chapter, the recordings are high-quality, with a 44.1 kHz sampling rate, 16-bit resolution, and minimal background noise. Recall that both the participant and interviewer wore head-mounted microphones connected to separate channels, and levels were adjusted to minimize speech from the other talker. For the analysis in this chapter, the participant channel was extracted from the stereo recordings, including any code-switches they made during the interview. While it would be possible to exclude items not produced in the primary language of the interview, this was not done. The driving reason for keeping code-switches in the analysis is that such code-switches are representative of the particular talker's language behavior. Further, just because someone switches languages, does not mean that they fully and immediately switch language modes (e.g., Fricke et al., 2016). For example, individual words may be borrowed and pronounced with the phonology of the interview's primary language (c.f.., the matrix language in code-switching Myers-Scotton, 2011).

All voiced segments were identified with the *Point Process (periodic, cc)* and *To TextGrid (vuv)* Praat algorithms (Boersma and Weenink, 2021), implemented with the Parselmouth Python package (Jadoul et al., 2018). The pitch range settings used with *Point Process (periodic, cc)* were 100–500 Hz for female talkers and 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 ignored by the Praat algorithms, and likely exerted little to no influence on the results. This method of identifying voiced portions of the speech signal captures vowels, approximants, and

some voiced obstruents. This result of this process differs slightly from the methods described in Lee et al. (2019), the paper on which the methods of this chapter were modeled. Lee et al. (2019) examined only vowels and approximants.

3.2.2 Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by Lee et al. (2019), except formant dispersion, which was excluded given its near-perfect correlation with the measured value of F4. The choice of measurements in Lee et al. (2019) is based on the psychoacoustic voice quality model described in the introduction to this chapter (Kreiman et al., 2014), as well as the availability of algorithms in the software used to extract measurements. Measurements were made every 5 ms during voiced segments in VoiceSauce Version 1.28? (Shue et al., 2011). The measurements are described below. Note that the shorthand name for each measurement is presented in boldface, and will be used throughout the rest of the chapter.

F0 Fundamental frequency is a correlate of pitch and is associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics. F0 was measured in Hertz using the STRAIGHT algorithm (Kawahara et al., 2016), which is regarded to be more accurate than other options in VoiceSauce. It is one of the more widely studied variables on this list, as evidenced by the literature cited in the introduction.

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

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

H1*–H2* The corrected amplitude difference between the first two harmonics is one of four primary measures used to characterize source spectral shape—also called spectral tilt—in the psychoacoustic model of voice quality (Kreiman et al., 2014). It is typically associated with phonation type but can be confounded by nasality (Garellek, 2019; Munson and Babel, 2019). The asterisks here—and in the following spectral shape measures—indicate that the value has been corrected (Iseli et al., 2007), to account for the amplifying impact of nearby formants on the amplitudes of harmonics. This allows for different vowels and other voiced segments to be compared with one another. This amplitude difference is measured in dB. Note that this measure—along with the following three spectral shape measures—depends on an accurate F0 measurement.

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

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

H2kHz*–H5kHz* The amplitude difference between the harmonics closest to 2000 Hz (corrected) and 5000 Hz (uncorrected) is a measure of harmonic spectral shape that does not depend on F0. The amplitude of the harmonic nearest 5000 Hz is not corrected by VoiceSauce, given inaccuracies in the correction algorithm at higher amplitudes. It captures the highest frequency band of the four shape measures, reflects phonation type and is measured in dB.

CPP Cepstral Peak Prominence measures the degree of harmonic regularity in voicing, and as such, it is associated with non-modal phonation types. VoiceSauce computes CPP according to the algorithm in Hillenbrand et al. (1994). Specifically, CPP measures the difference between the amplitude of the peak

in a cepstrum and the value at the same quefrency on the regression line for that cepstrum. It is measured in dB.

Energy Root Mean Square (RMS) Energy is a measure of spectral noise that reflects overall amplitude and is calculated over a window comprising five pitch periods. Energy is measured in dB.

SHR The subharmonics-harmonics amplitude ratio is a measure of spectral noise associated with period-doubling or irregularities in phonation. VoiceSauce's implementation is based on the algorithm described in Sun (2002). While based on amplitude, this ratio is unitless.

The raw VoiceSauce output used in this chapter is available in a repository on the Open Science Framework, in the data subfolder at https://osf.io/9ptk4/. The analysis code used for the following sections is available on GitHub, at https://github.com/khiajohnson/dissertation. Note that the diss repo is currently private!

3.2.3 Exclusionary criteria and post-processing

Given the nature of the corpus and the level of automation in the methods thus far, there is reason to expect a sizable number of erroneous measurements. To filter these out before analysis, measurements were subjected to exclusionary criteria focused on identifying impossible values. Observations were excluded in cases where any of the following measurements had a value of zero: F0, F1, F2, F3, F4, CPP, or uncorrected H5kHz. Observations were also excluded if Energy was more than three standard deviations above the mean. This may exclude some valid measurements but removes the long right tail of likely erroneous measures, as humans can only produce speech so loud.

Filtering based on F0 and the four formant frequencies reflects the observation that zero measurements are not possible for voiced portions of the speech signal. The interpretation for zero in CPP would indicate there is no cepstral peak, that is, no regularity in the voicing. In this sense, a zero for CPP likely also reflects either a lack of voicing or an erroneous F0 measurement. Lastly, only the uncorrected spectral measure for H5kHz was used in filtering, as erroneous values tended to co-occur on the same observation. The distribution of H5kHz did not span zero, except

for a spike of erroneous values equal to zero. This operationalization minimizes the removal of correctly measured zero values, which occurred with all of the other spectral shape parameters, whether corrected or uncorrected.

Moving standard deviations were calculated for each of the 12 measures using a centered 50 ms window, such that each window includes approximately ten observations. The moving standard deviations capture dynamic changes for each of the voice quality measures, which is important, as they may better reflect what listeners attend to in talker identification and discrimination tasks (Lee et al., 2019). This analysis uses moving standard deviations, as opposed to the coefficients of variation used by Lee et al. (2019). This should not have any undue effect on the outcome, as all variables were scaled before inclusion in the principal components analysis described in the next section. The last round of exclusionary criteria uses these moving standard deviations. If an observation was missing a moving standard deviation value, it was removed. Given the centered window, this means that observations falling less than 25 ms away from a voicing boundary were not included.

There were 24 total measures, with a measured value and a moving standard deviation for each of the acoustic measurements listed above. These 24 measures were used in the analyses described in the following sections. Across the 34 talkers, there were 3,071,736 observations after winnowing the data from an initial count of 6,560,403 observations. These observations were not evenly distributed across talkers and languages. While this full set of observations is perfectly valid for the crosslinguistic comparison in Section 3.2.4 and is used there, sample size is likely to have an impact on the principal components based analysis in Sections 3.2.5 and 3.2.6.

To control for the impact of sample size in that part of the analysis, the number of samples for each talker was capped to include only the first 20,124 samples for each interview. This value was selected as it represents the interview with the fewest observations. Put simply, differences in sample size reflect the variability in how much different individuals in the corpus talked. Those who produced longer passages of speech ultimately had more observations of voiced speech. Passage length was expected to impact the analysis, given how much affect and style can vary within a single conversation. Over time, individuals cover more of their range of variation, and as such, a regression to the mean is expected over time. To level

the playing field in this first analysis, the sample size was controlled. At the end of this chapter, in Section 3.2.7, a follow-up analysis validates this assumption. To preview those results, 20,000 samples appear sufficient for capturing the range of variability in acoustic voice variation.

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

3.2.4 Crosslinguistic comparison of acoustic measurements

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

For each acoustic measurement and talker, I conducted a Student's *t*-test and calculated Cohen's *d* using the *lsr* package (Navarro, 2015) in R (R Core Team, 2020); this provides a high-level assessment of whether variable means differed across the two languages. These 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, though most talkers (32/34) had at least one non-trivial comparison. The distribution of these counts is depicted in Figure 3.2.

For the non-trivial comparisons, there were consistent patterns across languages for a handful of the variables, including F0, H4*–H2kHz, and to a lesser extent, H1*–H2**. If there was a non-trivial difference in F0 across languages, then Can-

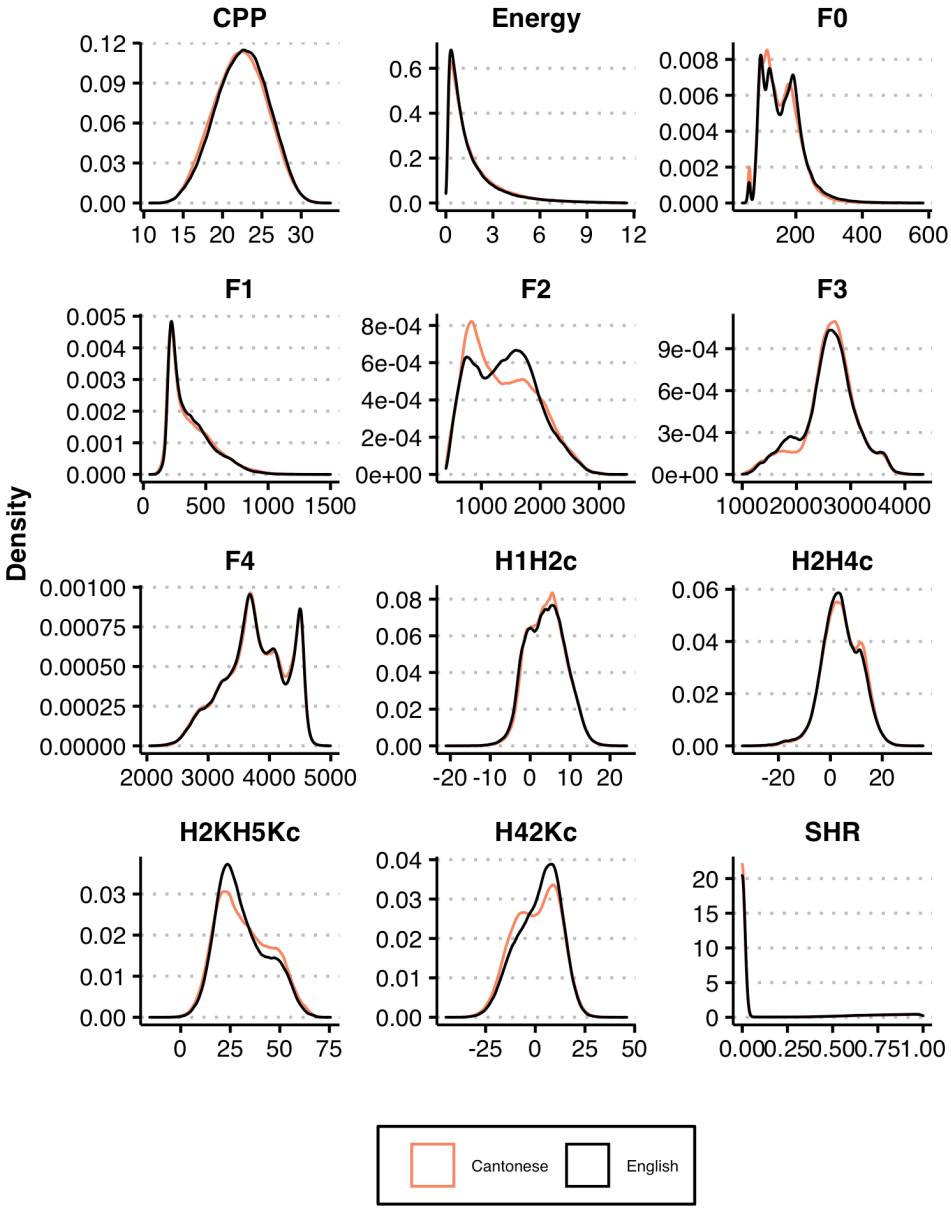

Figure 3.1: Each panel depicts a density plot that pool measurments from all talkers together to show the range of values for that measure. The x-axes each have their own scale. Language is separated out by color.

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

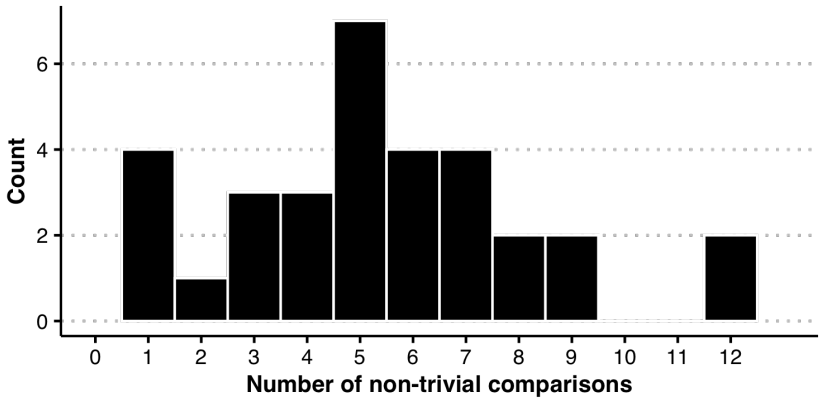

Figure 3.2: A histogram summary of the number of non-trivial comparisons from Table 3.1 across the 34 talkers.

tonese 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.

As for the two spectral shape measures, H4*–H2kHz was consistently lower in Cantonese when the comparison was not trivial (n=9), though most talkers did not exhibit a difference on this measure. H1*–H2* was significantly higher in Cantonese for a relatively large subset of the talkers (13/34), lower for a small number (3/34), but trivial for most (18/34). While based on 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.

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

Other measures, such as Energy, have a high number of nontrivial comparisons

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

CPP also exhibits a split between positive (n=6) and negative (7). Higher CPP values are associated with both breathy or creaky non-modal phonation types. In this sense, a positive difference would indicate that Cantonese was more non-modal, while a negative difference would indicate that English was more non-modal. Interpreting CPP is not so straightforward, however, as it is not immediately clear which type of non-modal phonation the measure entails. Given the results of H1*–H2**, it seems clear that knowing where on the creaky-modal-breathy spectrum a given speaker falls is pertinent to interpreting this measure. CPP would likely corroborate that outcome. How much more interpretation should I add in here?

Overall, while talkers show some clear across-language differences, these are far outnumbered by instances with no difference. The variability observed here fits in with the variable outcomes of previous work but does not necessarily fall neatly along the lines prior work would suggest that male and female talkers fall along.

3.2.5 Principal components analysis

Methods

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

Often, the goal of PCA is to take a large number of dimensions and extract a much smaller set to use for some additional purpose (e.g., linear regression). The

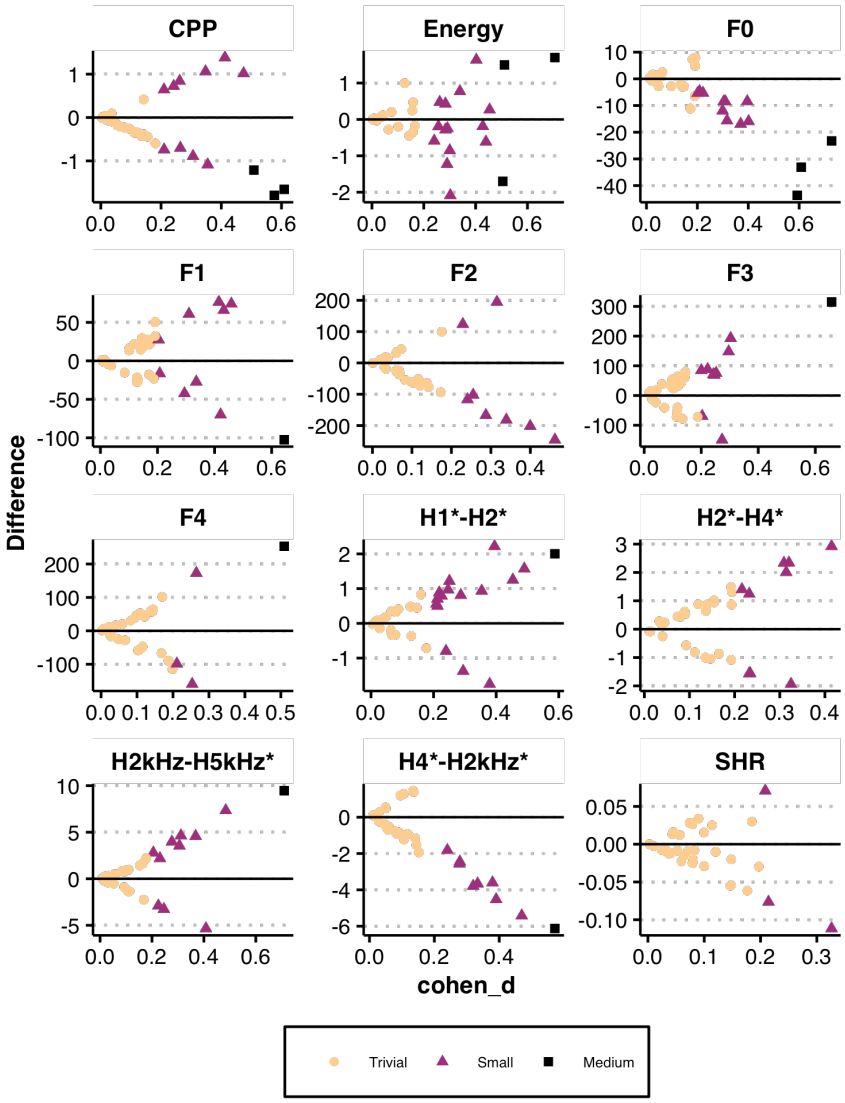

Figure 3.3: Each panel plots Cohen's *d* on the x-axis, and the difference between means from the t-tests on the y-axis. Positive values indicate a higher mean in Cantonese than English. The color reflect the levels of interpreations for Cohen's *d*.

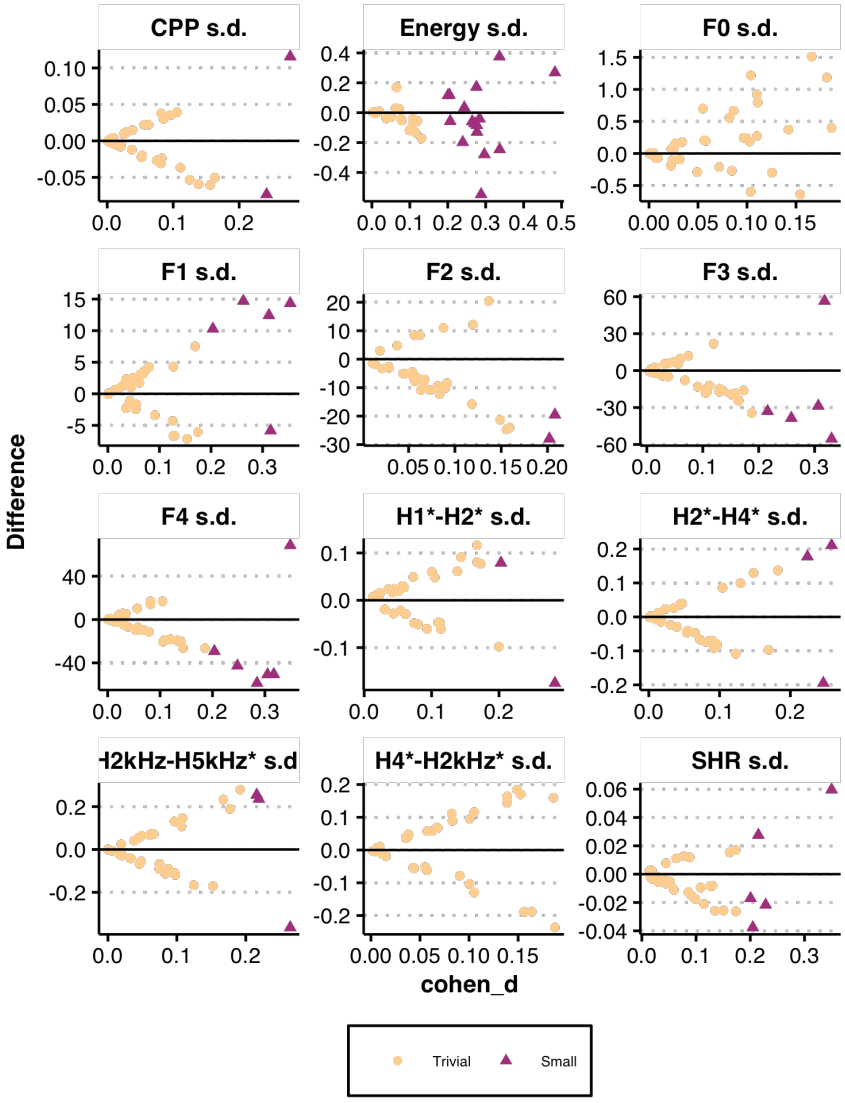

Figure 3.4: This figure is a continuation of 3.3.

focus in this chapter is on the internal structure of the components. That is, I examine what makes up components for different talkers and whether an individual's voice structure varies (or not) across languages.

I adapt methods from work on voices (Lee et al., 2019; Lee and Kreiman, 2020) and faces (Burton et al., 2016; Turk and Pentland, 1991). The goal of this analysis is to 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 standardized on a by-PCA basis before 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 understanding 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 (2002) recommended adjustment to the Kaiser-Guttman rule. This rule was used in place of a more sophisticated test (e.g., broken sticks), as it is not detrimental to this exploratory analysis to err on the side of including marginal components. Additionally, across each of the components, only loadings with an absolute value of 0.45 or higher were interpreted (Lee et al., 2019; Tabachnick and Fidell, 2013). While Lee et al. (2019) use a threshold of 0.32, Tabachnick and Fidell (2013) note that higher loadings indicate that a particular variable is a better measure of the component, with 0.32 corresponding to poor (but still interpretable) overlap between the variable and the component. The guidelines in Tabachnick and Fidell (2013) indicate that loadings of 0.45 correspond to fair, 0.55 to good, 0.63 to very good, and 0.71 and above to excellent. Given the large number of components and loadings in this analysis, only loadings over the fair threshold are interpreted.

Results

The PCAs across both languages for all 34 talkers resulted in 10–14 components and accounted for 74.9–82.7% of the total variation. Half of the talkers had the

same number of components for each language (17 of 34). Of the remainder, 16 of the 34 talkers had a difference of one in the number components, and only one had a difference of two. Talkers had 4–11 identical component configurations across their languages (M=7.82). These shared components represent 33.3%–91.7% of the total components for talkers (M=66.7%). The numbers comprising these summary statistics are provided in Table 3.2. While this already indicates a substantial amount of shared lower-dimensional structure across languages, it likely underestimates the actual shared structure. The reason is that similarity of component structure is not taken into account (i.e., a component of F2, F3, and F4 versus a component with just F2 and F3), as is the case in Section 3.2.6.

To assess whether talkers exhibit the same structure in voice variability across their languages, I first consider the patterns present across the different PCAs. 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 3.5 shows all of the components of participant VF32A's Cantonese and English PCAs, illustrating some examples of how components can vary (or not) across languages. It also highlights the importance of not attributing too much value to the ordering of components, but rather to their composition and variance accounted for.

Broadly, there were many similarities in component composition across talkers and languages. The following paragraphs summarize the components that were present in every PCA, regardless of talker or language. The shared component accounting for the most variation across talkers had a core structure consisting of F2 and H4*-H2kHz*. These usually went along with H2kHz*-H5kHz* (Cantonese = 34, English = 31), and occasionally with F3 and F4 (Cantonese = 3, English = 3). In a similar vein, all talkers had a component consisting of H4*-H2kHz* s.d. and H2kHz*–H5kHz* s.d., though it accounted for a smaller proportion of the total variation. Should I describe what these mean here? Or in the discussion?

In the case of the moving standard deviation parameters, there were a few common configurations. Formant s.d. parameters often co-occurred. In both languages, the component typically consisted of F3 s.d. and F4 s.d. (Cantonese = 32, English = 26), though a subset of these cases also included F2 s.d. (Cantonese = 6, En-

Table 3.2: The number of components and the variance accounted for is listed for each PCA. The last column indicates the number of identical components across langauges.

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

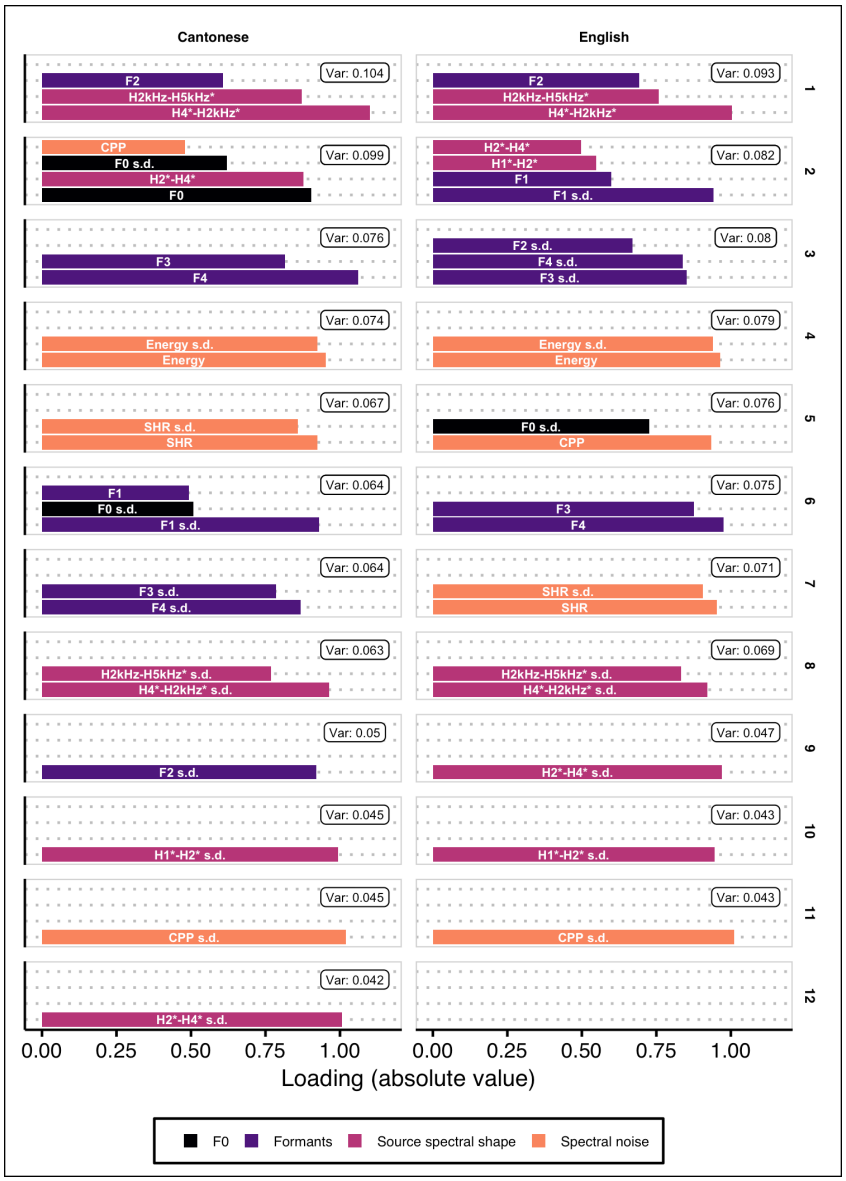

Figure 3.5: In this depiction of the components of VF32A's Cantonese and English PCAs, loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings; and, the component's variance is superimposed.

glish = 10). In the case of spectral shape, the variable for H2*–H4* s.d. commonly occurred alone (Cantonese = 18, English = 18) or in combination with H1*–H2* s.d. (Cantonese = 13, English = 14). While the formant and spectral shape moving standard deviations often exhibited these common patterns, variables in these categories were just as likely to pattern in more idiosyncratic ways, loading alongside each other, F0, formants, and spectral measures. This kind of variability is not readily summarizable.

The spectral noise parameters had a relatively consistent component structure across talkers and languages. Energy and Energy s.d. consistently loaded on the same component and were sometimes accompanied by F0 (Cantonese = 6, English = 2) and F0 s.d. (Cantonese = 1). CPP s.d. occurred consistently on its own component for all English PCAs, and 31 of the Cantonese PCAs. In the remaining three Cantonese PCAs, CPP s.d. was accompanied by CPP (n=1) or H1*-H2* s.d. (n=2). CPP patterned less consistently but was most often accompanied by F0 s.d. (Cantonese = 19, English = 14). SHR and SHR s.d. exclusively loaded together for 31 talkers in each language and SHR by itself for a single talker per language. The pair was sometimes accompanied by H1*-H2* (Cantonese = 2, English = 2), H2*-H4* (English = 1), or F0 (English = 3).

While this covers many of the variables that went into the PCAs, F0 is notably sparse in the above paragraphs. While F0 s.d. was fairly consistent in emerging either with CPP (Cantonese = 21, English = 17) or alone (Cantonese = 9, English = 10), the same cannot be said for F0. No particular component structure with F0 occurred more than six times, and across the wide range of configurations, F0 was accompanied by all kinds of variables: F0 s.d., H1*–H2*, H1*–H2* s.d., H2*–H4*, F1 s.d., F4 s.d., CPP, Energy, Energy s.d., and SHR, SHR s.d. The lack of consistency in F0 across talkers is notable for a few reasons. First, F0 plays a major role in prior work on voice production and perception, given its salience as an acoustic dimension (Perrachione et al., 2019). A second reason for it being notable comes from Lee and colleagues' work, where F0 emerged as an important feature of acoustic voice variation structure in English spontaneous speech (Lee and Kreiman, 2019) and Korean sentence reading (Lee and Kreiman, 2020), but not for English sentence reading (Lee et al., 2019).

On the whole, variables emerged on a single component. That is, very few

variables had complex loading structures. Across talkers, only three had complex loading structures for H2*–H4* in each language. F0 and F0 s.d. participated in complex loadings for a single English talker, and twice in the Cantonese PCAs. The remaining variables that participated in complex loading structures only occurred in one or two PCAs across all talkers and languages. This means that for a given PCA, the interpretation of components is reasonably straightforward, even if drawing generalizations over the full group is not.

There were additional components (not reported here) that were shared by less than half of the talkers. A full list of component configurations, along with the number of occurrences and range of variation accounted for is provided in the supplementary materials. Well... it will be! Also, would a table in this section help with interpretation? Or is prose plus supplementary materials good enough?

In summary, this PCA analysis found a greater amount of component structure overlap than was reported in Lee et al. (2019). At the same time, idiosyncratic variation was still readily apparent in the PCAs, both in how variables co-occur, as well as in how much variance is accounted for by the different components. Additionally, it is important to remember that these PCAs represent the lower dimensional structure of the voices they measure. Considering that the total variance unaccounted for by the PCAs ranges from 17.3%–25.1%, this unaccounted for variability may also be idiosyncratic in nature.

3.2.6 Canonical redundancy analysis

Methods

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 relate to one another. The two sets of variables are transformed such that the correlation between the rotated versions is maximized. This is useful here, as a talker may have similar components in their English PCA and Cantonese PCA, but these components might not necessarily be in the same order, even if they account for

comparable amounts of variance.

Redundancy is a relatively simple way to characterize the relationship between the loadings matrices of two PCAs—the two sets of variables under consideration here. For example, the two redundancy indices represent the amount of variation in a talker's Cantonese PCA output that can be accounted for via canonical variates by their English PCA output and vice versa. Notably, the two redundancy indices are not symmetrical (Stewart 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. The PCA with more components will likely account for more of the variation in a PCA with fewer components than the reverse.

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

Results

Redundancy indices for within-talker comparisons ranged from 0.80 to 0.97, (Mdn = 0.92, M = 0.91, SD = 0.04) and are displayed in Figure 3.6, with the two redundancy indices for a given pairwise comparison plotted against one another. Comparisons across talkers within-language ranged from 0.64 to 0.96 (Mdn = 0.83, M = 0.83, SD = 0.5). Comparisons across both talkers and languages ranged from 0.64 to 0.97 (Mdn = 0.83, M = 0.83, SD = 0.5). Within-talker values were confirmed to be higher than across-talker comparisons, per a Welch's t-test ($t(70.93) = -17.35$, $p < 0.001$, d = 1.77). A second Welch's t-test testing the same versus different language for the across talker comparisons did not find a difference between those groups ($t(4485.9) = -1.53$, $p = 0.13$, d = 0.05).

While the across-talker comparisons were generally lower than the within-talker ones, the redundancy indices are overall still relatively high. The high values are not unexpected. As PCA is a dimensionality reduction technique, the discarded com-

ponents almost certainly contain idiosyncratic variation. Moreover, and following from Section 3.2.5, there were a substantial number of commonly occurring patterns across talkers and languages. Together, this supports the conceptualization of a voice space comprising a shared structure—as in the case of the prototype account—where voices can only deviate from one another so much.

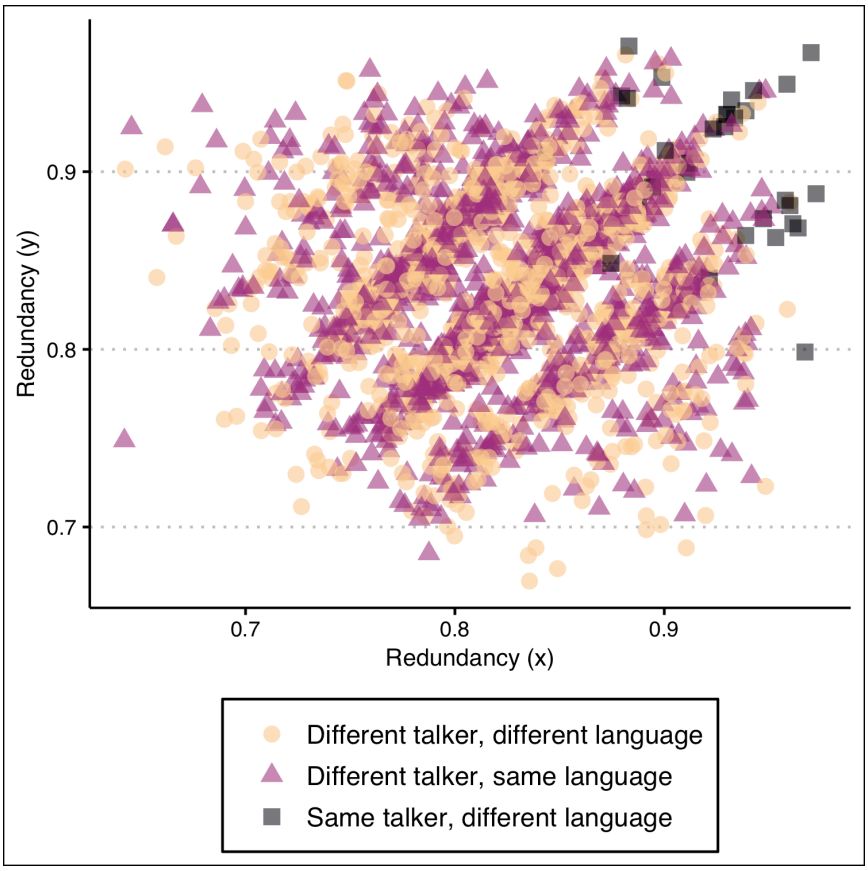

Figure 3.6: The relationship between the two redundancy indices for three different types of comparisons. Within-talker comparisons are represented by the black squares and are clearly clustered at the top right.

3.2.7 Passage length analysis

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

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

Figure 3.7 plots polynomial smooths for each interview, with superimposed mean smooths. The x-axis represents the sample size of the shorter passage length in the comparison. The y-axis represents an average of the two redundancy indices. The vertical line at 5000 represents the average sample size from Lee et al. (2019). The vertical line at 20124 represents the sample size used in Sections 3.2.5 and 3.2.5. While there are some gains in sample sizes above the second vertical line, they are comparatively small. It is readily apparent from this plot that the sample size used for PCAs in this chapter was sufficient to capture most of the range of talkers' within-interview variability. As the leveling-off point likely varies across speech styles, it is not immediately apparent whether the sample size in Lee et al.

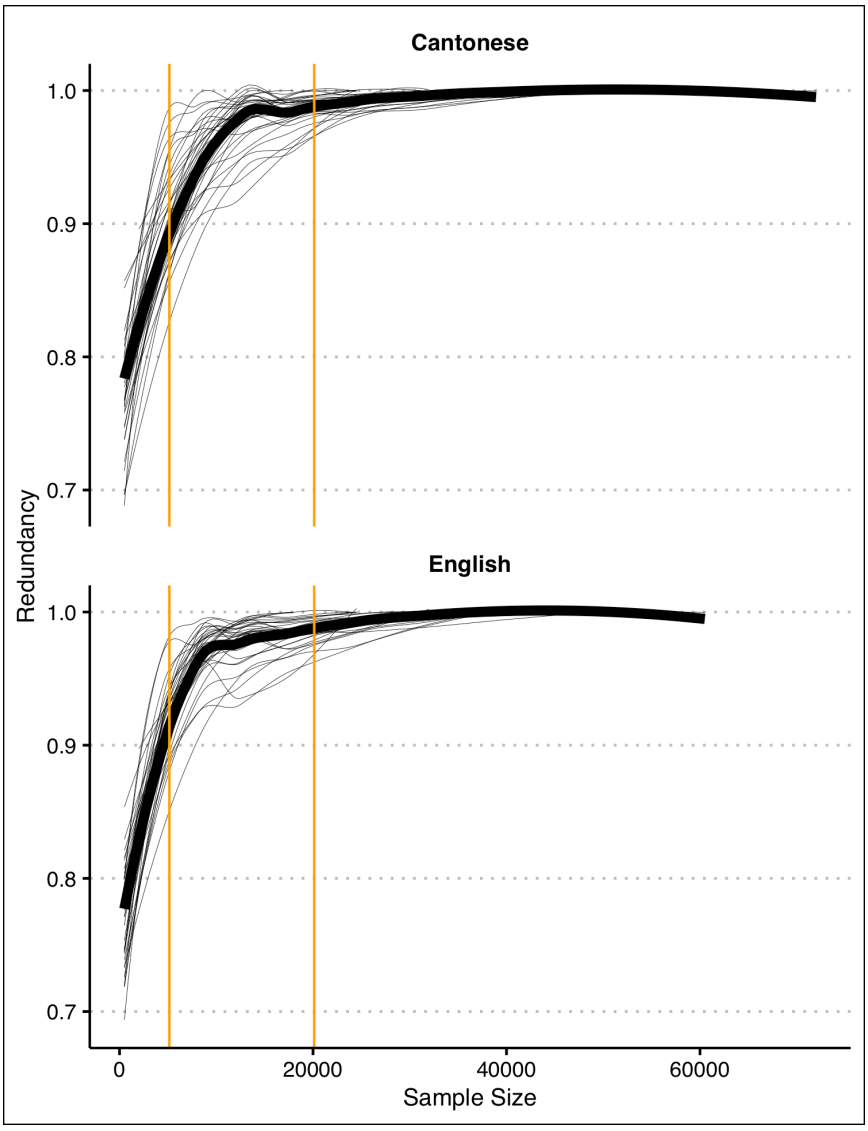

Figure 3.7: Passage length redundancy indices are plotted against the sample size of the smaller PCA. Smoothed curves show a rapid increase in redundancy followed by a levelling off between the vertical orange lines, which represent the sample sizes used in prior work (x = 5000) and the present study (x = 20124).

(2019) sufficiently captured the range of talker variability and thus may not adequately capture the structure of their variability.

3.3 Discussion and conclusion

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

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

uals and languages in question to excel at perceptual tasks in both domains.

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

In the PCAs, similarity to prior work emerges in the structure of various components, including the ones that account for the most variability. Lee et al. (2019) report that three of the largest components captured lower-dimensional structure for (i) higher harmonic spectral shape variation, (ii) higher formants, and (iii) a combination of lower spectral shape with the lower formants. While the amount of overall variance accounted for differs here, these component structures also occurred for the SpiCE talkers. Respectively, they are associated with (i) perceived breathiness or brightness, (ii) vocal tract size or speaker identity, and (iii) a combination of phonation type and vocal tract configuration—perhaps reflecting shared linguistic variation. The cross-study overlap in component structure adds credibility to the idea of a prototype model in voice (Lavner et al., 2001; Latinus and Belin, 2011). Much like Lee et al. (2019), the key shared dimensions relate to the timbre, identity, and vocal tract configuration.

This high degree of similarity does not preclude crosslinguistic differences on a within-talker basis but rather suggests that such differences occur on a more global level. This is apparent in Figure 3.8, which depicts the relationship between talkers' average redundancy from Section 3.2.6 and the difference between the mean values for each of the acoustic measurements in Section 3.2.4. If there were clear relationships between large crosslinguistic differences and redundancy, the regression lines should be strongly negative—this does not seem to be the case.

Such high similarity in the PCAs was not entirely expected, given the results of Lee et al. (2019), where a handful of shared components were evident but were complemented by numerous idiosyncratic components. At face value, the results in this chapter suggest that a heterogeneous bilingual population has more across-

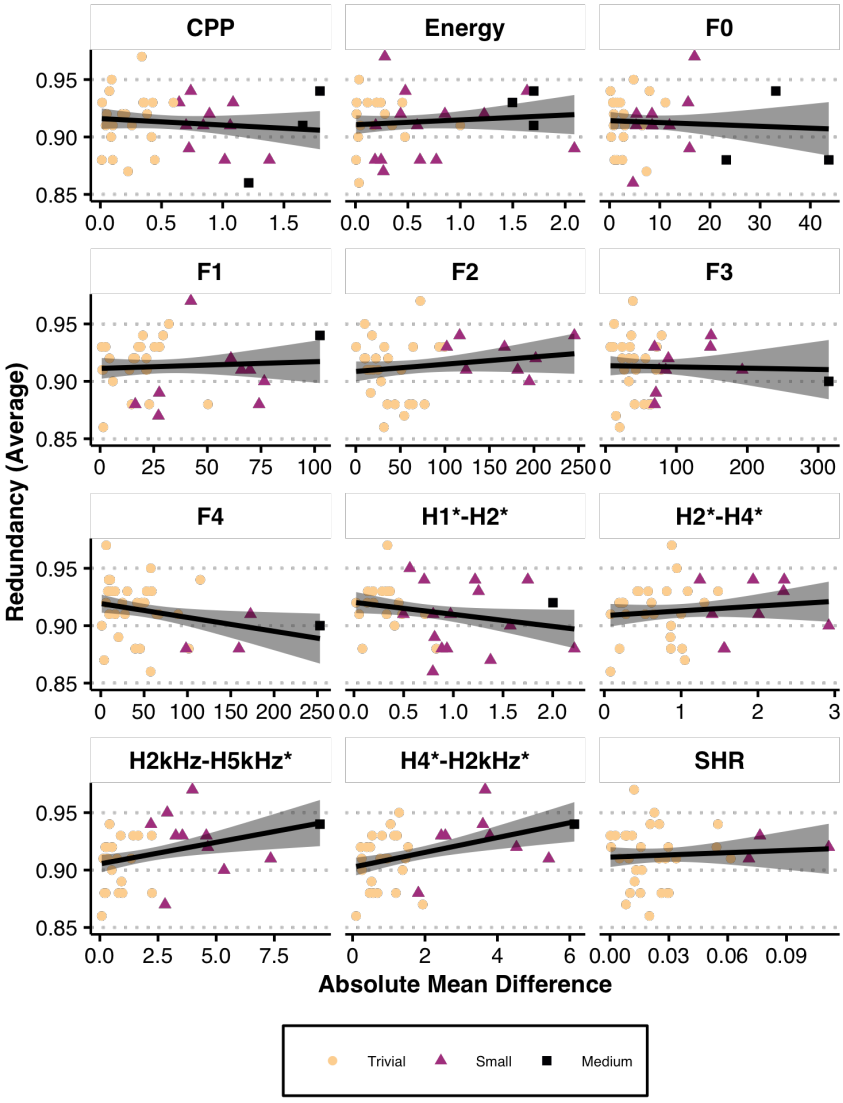

Figure 3.8: This plot depicts the relationship between the absolute value of the difference of means from the t-tests, plotted against the average redundancy value for the talker. Color and shape indicate whether the Cohens' *d* of the t-test was trivial, small, or medium. The superimposed regression line summarizes the relationship between these values.

talker similarity than a tightly controlled group of monolingual English speakers. Several analysis decisions may have contributed to this apparent difference. I compared similar components independent of order, which ignores the fact that similar components may account for different amounts of variance, but ensures that comparisons are made among like items. Any downside to this methodological decision is mitigated by the fact that most components made relatively small contributions in how much of the overall variance they accounted for (see Table 3.2). As such, I predict that increased across-talker similarity would be found in a reanalysis of the UCLA Speaker Variability Database (Keating et al., 2019) using the adapted methods of this chapter.

While methodological choices may account for some part of these results, the data differences between the current chapter and previous studies are also pertinent. This chapter uses substantially longer passages than the short samples in Lee et al. (2019). Larger speech samples clearly allow for a stable underlying structure to emerge. Smaller samples, conversely, may reflect more ephemeral variation in a talker's voice, and thus not be representative of the talker's full range. The passage length analysis in this chapter shows that the number of samples needed for stabilization is substantially larger than the 5000 samples used in Lee et al. (2019). This does not necessarily discount their work, however, as the current chapter uses spontaneous speech, which is arguably more variable than read speech.[1] It's plausible that an analysis of sentence reading would not need as much data to cover talkers' range of variability in reading aloud. The body of literature in the introduction establishes differences in voice quality across speaking styles (e.g., Lee and Sidtis, 2017). As such, the threshold suggested here may only be appropriate for the speaking style of peer-to-peer conversational interviews. In any case, the methods presented here offer a tool for researchers to use in assessing whether their sample size is representative of a larger whole. Understanding how this interacts with speaking style is left for future directions.

Ultimately, the goal of this line of research is to understand how the acoustic

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

variability and structure of talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination. Turning to listener and behavioral data will help in deciphering what is meaningful variation within a voice from low-level noise that cannot be attributed to a particular vocal signature. Verification from listener performance will help adjudicate which statistical choices present an acoustic voice space that matches listener organization. The results of this chapter set up predictions for that work. These predictions will be revisited in general discussion.

Bibliography

Afouras, T., Chung, J. S., and Zisserman, A. (2020). Now you' re speaking my language: Visual language identification. In *Proceedings of Interspeech 2020*, pages 2402–2406. ISCA. → page 34

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 7, 8, 10, 17, 20, 35

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 12

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts l2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899. → page 10

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. → pages 6, 11

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, 24

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 9, 10

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 12

Garellek, M. (2019). The phonetics of voice. In Katz, W. F. and Assmann, P. F., editors, *The Routledge Handbook of Phonetics*. Routledge. → page 14

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of speech and hearing research*, 37(4):769–778. → page 14

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 14

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15. → page 12

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, 2 edition. → pages 24, 29

Järvinen, K., Laukkanen, A.-M., and Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, 38(2):47–51. → pages 9, 10, 12, 21

Kawahara, H., Agiomyrgiannakis, Y., and Zen, H. (2016). Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pages 221–228. → page 13

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. → pages 3, 37

Keating, P. and Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, 132(2):1050–1060. → pages 5, 7, 8

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, 11, 13, 14, 24

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, 24

Latinus, M. and Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2:175. → pages 4, 35

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. → pages 4, 35

Lee, B. and Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, 20(3):174–185. → pages 8, 9, 10, 12, 37

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, 3, 4, 5, 11, 13, 16, 17, 24, 28, 29, 32, 34, 35, 37

Lee, Y. and Kreiman, J. (2019). Within- and between-speaker acoustic variability: Spontaneous versus read speech. → pages 3, 28

Lee, Y. and Kreiman, J. (2020). Language effects on acoustic voice variation within and between talkers. 10.1121/1.5146847. → pages 3, 6, 24, 28

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science*, 10(2):e1483. → page 4

Liang, S. (2015). *Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography*. Springer International Publishing. → page 11

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

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, 24(1):71–89. → page 9

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). Describe and understand your model's parameters. R package. → page 24

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → page 34

Munson, B. and Babel, M. (2019). The phonetics of sex and gender. In Katz, W. F. and Assmann, P. F., editors, *The Routledge Handbook of Phonetics*. Routledge. → page 14

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 6

Myers-Scotton, C. (2011). The matrix language frame model: Developments and responses. In *Codeswitching Worldwide*, volume 126 of *Trends in Linguistics. Studies and Monographs*. De Gruyter Mouton. → page 12

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. University of Adelaide, Adelaide, Australia. R package version 0.5. → page 17

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 7, 8, 9, 11, 17, 20, 35

Ng, M. L., Hsueh, G., and Sam Leung, C.-S. (2010). Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, 12(3):230–236. → pages 8, 17, 35

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

Ordin, M. and Mennen, I. (2017). Cross-linguistic differences in bilinguals' fundamental frequency ranges. *Journal of Speech, Language, and Hearing Research*, 60(6):1493–1506. → page 10

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. → pages 5, 6

Perrachione, T. K. (2018). Recognizing speakers across languages. In Frühholz, S. and Belin, P., editors, *The Oxford Handbook of Voice Perception*, pages 514–538. Oxford University Press. → pages 4, 5

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, 146(5):3384–3399. → pages 4, 5, 6, 11, 28

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. → pages 17, 24

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 9

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 13

Sjölander, K. (2004). The Snack Sound Toolkit. → page 13

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., and Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, 69(2):218–231. → page 34

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163. → pages 29, 30

Sun, X. (2002). Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–333–I–336. → page 15

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition. → page 24

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 24

Voigt, R., Jurafsky, D., and Sumner, M. (2016). Between- and within-speaker effects of bilingualism on f0 variation. In *Proceedings of Interspeech 2016*, pages 1122–1126, San Francisco, CA. → page 9

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, 39(1):9–30. → page 6

Xue, S. A., Hagstrom, F., and Hao, J. (2002). Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: a functional system approach. *Asia Pacific Journal of Speech, Language and Hearing*, 7(1):55–62. → page 8

Yang, Y., Chen, S., and Chen, X. (2020). F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Proceedings of Interspeech 2020*, pages 4163–4167. ISCA. → page 9

Yovel, G. and Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, 17(6):263–271. → page 4