Chapter 3

The structure of acoustic voice variation in bilingual speech

3.1 Introduction

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

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

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

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

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

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

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

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

3.1.1 Voice and voice quality

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

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

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

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

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

3.1.2 Structure in voice quality variation

There is a large body of literature focused on understanding differences in variability across populations for a small set of these acoustic measurements. Such studies typically compare summary statistics for F0 and a handful of spectral measures. This body of work is summarized in Section 3.1.4 in the context of crosslinguistic

comparisons. Before summarizing this work, it is important to highlight that very little of it dives into the *structure* of voice variability, which is a relatively new area spearheaded by Lee and colleagues (Lee et al., 2019; Lee 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. As noted in Section 3.1.1, acoustic dimensions do not behave in isolation; rather, they pattern together in complex ways. Part of what is novel about Lee et al.'s approach is that it gets at how these variables behave in relation to one another—that is, how covariation is structured. In many ways, this is the first step towards understanding which aspects of voice are available to listeners and thus useable in perceptual processes, particularly when coarse summary statistics (i.e., means, ranges, and standard deviations) do not indicate cross-talker differences.

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

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

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

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

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

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

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

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

In any case, Lee et al. (2019) argue that familiarity with a voice arises from

learning how that voice varies across time and space, whether within an utterance or across environments, physical states, and emotions. This familiarity could easily be characterized in terms of the extent and manner that a voice deviates from a prototype.

3.1.3 Voice perception

The literature on voice perception has approached the question of what listeners use in voice identification, discrimination, and learning through the lens of familiarity (Levi, 2019; Perrachione, 2018). This body of experimental work pairs different combinations of listeners, talkers, languages, and stimuli manipulations to probe how listeners identify and discriminate among talkers. While identification and discrimination are often talked about in conjunction with one another, the processes are likely supported by different perceptual mechanisms (Perrachione et al., 2019). One of the key findings from this literature is the Language Familiarity Effect (LFE), which encompasses a broad range of findings where listeners are better at identifying talkers in a familiar language (for a recent review, see Perrachione, 2018). Bilinguals are especially good at this kind of task and show evidence of generalizing across languages they know (Orena et al., 2019).

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

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

3.1.4 Bilingual voices

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

Describing and analyzing acoustic voice variation in bilingual speech has motivation in both perception and production. As apparent from the LFE literature, listeners are capable of learning and identifying voices in one language and then generalizing across languages. Listeners are better at identification and discrimination when they have more familiarity with the language, but performance on such tasks tends to be above chance even for listeners who lack familiarity with the language (e.g., Orena et al., 2019). Knowledge of the language used in the experiment lends the greatest advantage; and, knowledge of a related language also provides a benefit (Zarate et al., 2015). Presumably, listeners in the latter situation can extract some degree of linguistic information given overlap in the sound structure of the two languages. In cases where listeners cannot rely on linguistic information, they must be tracking non-linguistic acoustic/auditory information in

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

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

Additionally, examining the structure of the same talker's voice in each language lends additional validation to the arguments made by Lee 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, the language versus biology dichotomy is somewhat misleading. Voices ultimately have biological constraints due to physical and physiological limitations (e.g., vocal tract length, vocal fold mass) or pathologies. Yet, at the same time, individuals exert remarkable and wide-ranging control over their voice space and are highly capable of manipulating factors that are not linguistically important but which signal social and contextual information. This applies across all aspects of an individual's linguistic repertoire

(Bullock 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.

English and Cantonese

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

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

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

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

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

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

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

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

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

In a small study of Cantonese-English bilingual (n=9), Russian-English bilingual (n=9), and English monolingual (n=10) young women, Altenberg 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 across their two languages, and Cantonese-English bilinguals did not. Though, they did produce a wider F0 range in Cantonese compared to their English. While the results in Altenberg 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 lower F0 in Cantonese than English, but males did not. In the same study, all participants had greater mean spectral energy values (mean amplitude of energy between 0–8 kHz) and lower spectral tilt (ratio of energy between 0–1 kHz and 1–5 kHz) in Cantonese (Ng et al., 2012). Respectively, these findings suggest a greater degree of laryngeal tension and breathier voice quality in Cantonese compared to English. The LTAS measure of the first spectral peak did not differ across languages, suggesting that vocal stiffness remained consistent in the bilinguals' two languages.

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

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

English and other languages

The authors of these studies speculate that Cantonese's status as a tone language may account for some of these differences compared to English. It is important to emphasize that this explanation is pure speculation. In this light, it is also relevant to consider the larger body of research comparing voice quality for Mandarin and English. Lee 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 lower F0 in Mandarin than English. 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 a simple

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

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

While the studies reviewed thus far provide a mixed picture of voice differences across language pairs, there is a strong focus on F0. Both the F0-centricity and variable outcomes are apparent in work on other language pairs as well. For example, Cheng (2020) finds that Korean has consistently higher F0 than English, regardless of whether they were early sequential or simultaneous bilinguals, and that differences in F0 range differed for cisgender males and females. This result builds on the findings for Korean-English bilinguals (Lee 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. Ordin and Mennen (2017) argue that if a difference in F0 stemmed purely from language differences, that males and females would both show the pattern. Because this is not the case, they argue that the result is unlikely to be due to anatomical or purely linguistic reasons. While this argument does not necessarily disentangle social from linguistic, it emphasizes the need to consider social dimensions.

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

This body of work mostly focuses on linguistic and social differences. While

some of it dives into individual differences, between-talker variability should 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).

3.1.5 The present study

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

This chapter brings together work describing crosslinguistic voice differences and the structure of acoustic voice variation to provide a more comprehensive picture of how voices vary across languages. Using the corpus introduced in Chapter 2, this chapter describes the behavior of various spectral properties (e.g. Ng et al., 2012), and also examines how acoustic variation is structured, following the work of Lee et al. (2019). This chapter builds on Lee et al. (2019) in a handful of ways: it extends the methods to the case of bilinguals, considers longer samples, and addresses the role of sample size both within and across talkers and languages. It also extends their methods by introducing a mechanism to assess structural similarity within and between individuals and languages.

3.2 Methods & results

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

3.2.1 Data

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

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

Cantonese reading, there was a wide range of familiarity with the materials represented. Some talkers knew all of the sentences, and others struggled with some of them. This variability renders the sentences less comparable to their English counterparts in the SpiCE corpus. There are also imbalances in the storyboard task. As talkers narrated the same story in both languages, they were often more confident the second time around. Excluding both of these tasks is motivated by prior work that highlights how confidence (Järvinen et al., 2013) and speaking style (Lee 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., 2016b). For example, individual words may be borrowed and pronounced with the phonology of the interview's primary language (cf. the matrix language in code-switching Myers-Scotton, 2011).

All voiced segments were identified with the *Point Process (periodic, cc)* and *To TextGrid (vuv)* Praat algorithms (Boersma 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. These settings reflect a balance between known differences between male and female pitch (Simpson, 2009) and the wide range of F0 variability in spontaneous speech while guarding against the pitch estimation issues of doubling and halving. While speech from the interviewer can occasion-

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

3.2.2 Acoustic measurements

All voiced segments were subjected to the same set of acoustic measurements of voice quality made by Lee et al. (2019), except formant dispersion, which was excluded given its very strong correlation with the measured value of F4 in this chapter (following the exclusionary criteria in Section 3.2.3: Pearson's $r = 0.94$, df $= 3071734$, $p < 0.001$). The choice of measurements in Lee et al. (2019) is based on Kreiman et al.'s (2014) psychoacoustic voice quality model, as well as the availability of algorithms in the software used to extract measurements. Measurements were made every 5 ms during voiced segments in VoiceSauce (Shue et al., 2011).[2] The measurements are described below. Note that the shorthand name for each measurement is presented in boldface and will be used throughout the rest of the chapter.

F0 Fundamental frequency is a correlate of pitch and is associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics. F0 was measured in Hertz using the STRAIGHT algorithm (Kawahara et al., 2016). It is one of the more widely studied variables on this list, as evidenced by the literature cited in Section 3.1.4.

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

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

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

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

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

H1*–H2* The corrected amplitude difference between the first two harmonics is one of four primary measures used to characterize source spectral shape—also called spectral tilt—in the psychoacoustic model of voice quality (Kreiman et al., 2014). It is typically associated with phonation type, such that lower values fall on the creakier end of the spectrum and higher values on the breathier end of the spectrum (Garellek, 2019). Interpretation is less than straightforward, however, as determining where on the spectrum from creaky to breathy a particular observation falls depends on a measure of spectral noise (e.g., CPP below; Garellek, 2019). Additionally, H1*–H2* can be confounded by nasality (Munson 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. Like H1*–

H2*, it is associated with phonation type and is measured in dB.

H4*–H2kHz* The corrected amplitude difference between the fourth harmonic and the harmonic closest to 2,000 Hz is the third spectral shape measure. Unlike the previous two, one of the harmonics depends on F0, while the other does not. It captures shape in a higher frequency range and is also associated with phonation in a similar manner to H1*–H2*. Like the other spectral shape measures, it is in dB.

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

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

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

SHR The subharmonics-harmonics amplitude ratio is a measure of spectral noise associated with period-doubling or irregularities in phonation. VoiceSauce's

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

implementation is based on the algorithm described in Sun (2002). While based on amplitude, this ratio is unitless.

The raw VoiceSauce output used in this chapter is available in a repository on the Open Science Framework, in the data subfolder at https://osf.io/9ptk4/. The analysis code used for the following sections is available on GitHub, at https://github.com/khiajohnson/dissertation.[4]

3.2.3 Exclusionary criteria and post-processing

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

Filtering based on F0 and the four formant frequencies reflects the observation that zero measurements are not possible for voiced portions of the speech signal. The interpretation for zero in CPP would indicate there is no cepstral peak, that is, no regularity in the voicing. As nonzero values for CPP reflect a range of modal and nonmodal phonation, a zero for CPP likely reflects either a lack of voicing or an erroneous F0 measurement. Lastly, only the spectral measure for H5kHz was used in filtering (uncorrected, and not the difference used in the analysis), as erroneous values tended to co-occur on the same observation. The distribution of H5kHz did not span zero, except for a spike of erroneous values equal to zero. This operationalization minimizes the removal of correctly measured zero values, which occurred with all of the other spectral shape parameters, whether corrected

[4]Note that this repository is currently private.

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

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

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

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

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

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

3.2.4 Crosslinguistic comparison of acoustic measurements

Following prior work, the first step in this analysis is a crosslinguistic comparison for each talker and measure. As discussed in the introduction to this chapter, there are some commonly found—though inconsistent—differences between Cantonese and English. Prior work has found that speakers sometimes produce lower and more variable F0 in Cantonese (Altenberg 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, Cohen's *d* was calculated using the *lsr* package (Navarro, 2015) in R (R Core Team, 2020); this provides a high-level assessment of whether variable means differed across the two languages. These

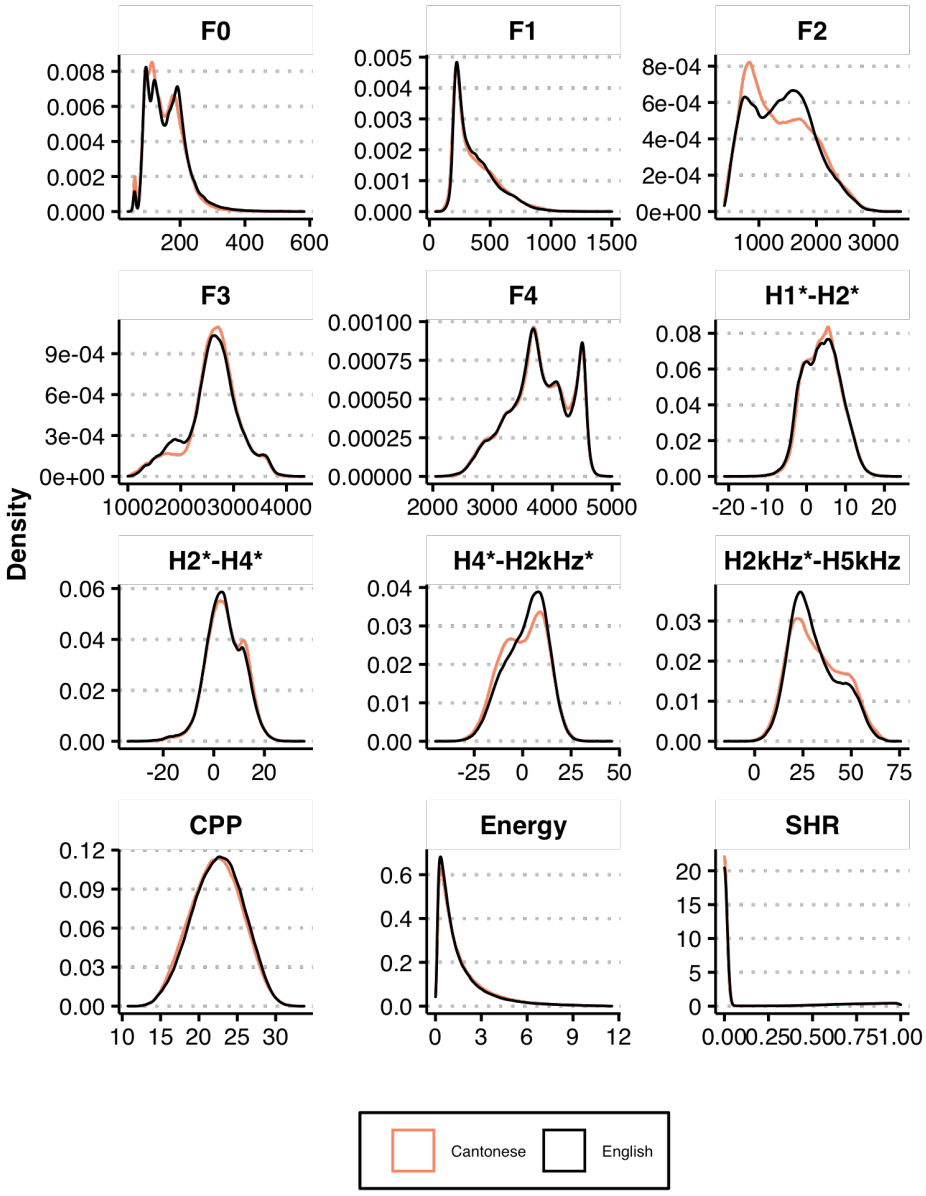

Figure 3.1: Each panel depicts a density plot that pools measurments from all talkers together to show the range of values for that measure. The x-axes each have their own scale. Language is separated out by color.

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

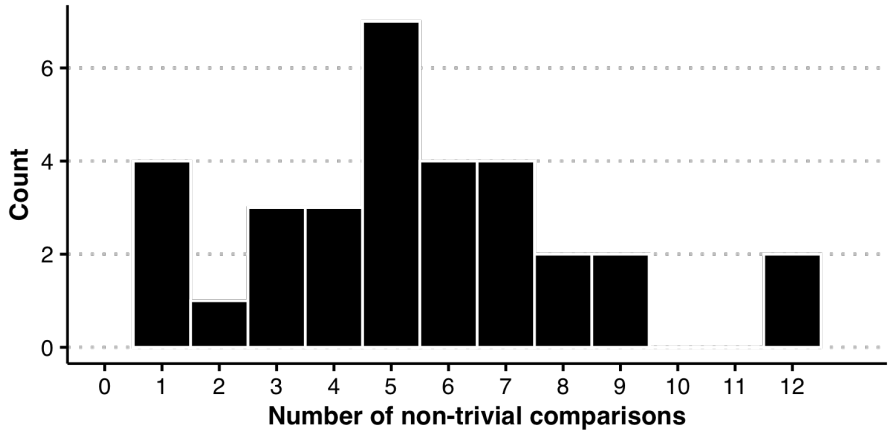

Figure 3.2: A histogram summary of the number of non-trivial comparisons from Table 3.3 across the 34 talkers.

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

As for the two spectral shape measures, H4*–H2kHz* was consistently lower

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

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

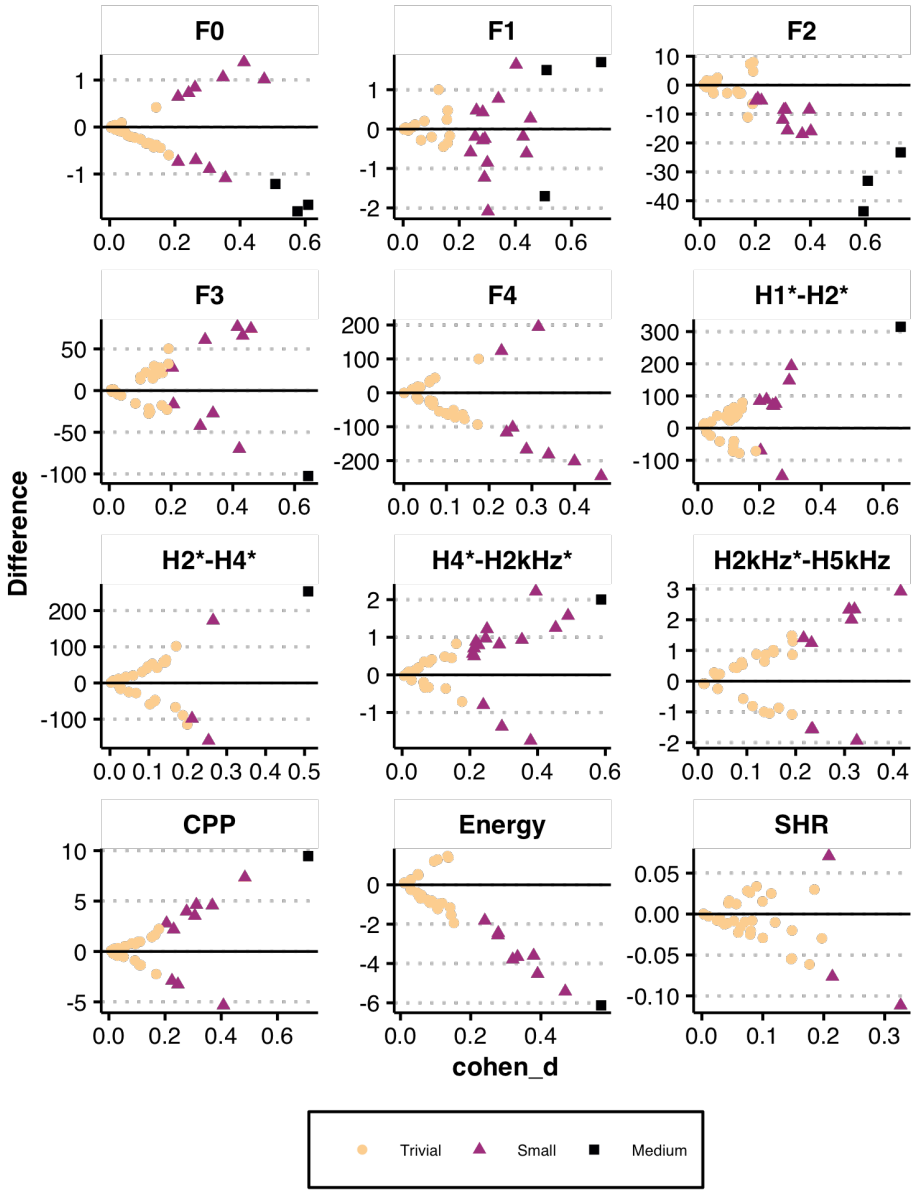

Figure 3.3: Each panel plots Cohen's *d* on the x-axis (scales differ) and the difference between language means on the y-axis. Positive values indicate a higher mean in Cantonese than English. The color reflects the levels of interpreation for Cohen's *d*. Each point represents a talker.

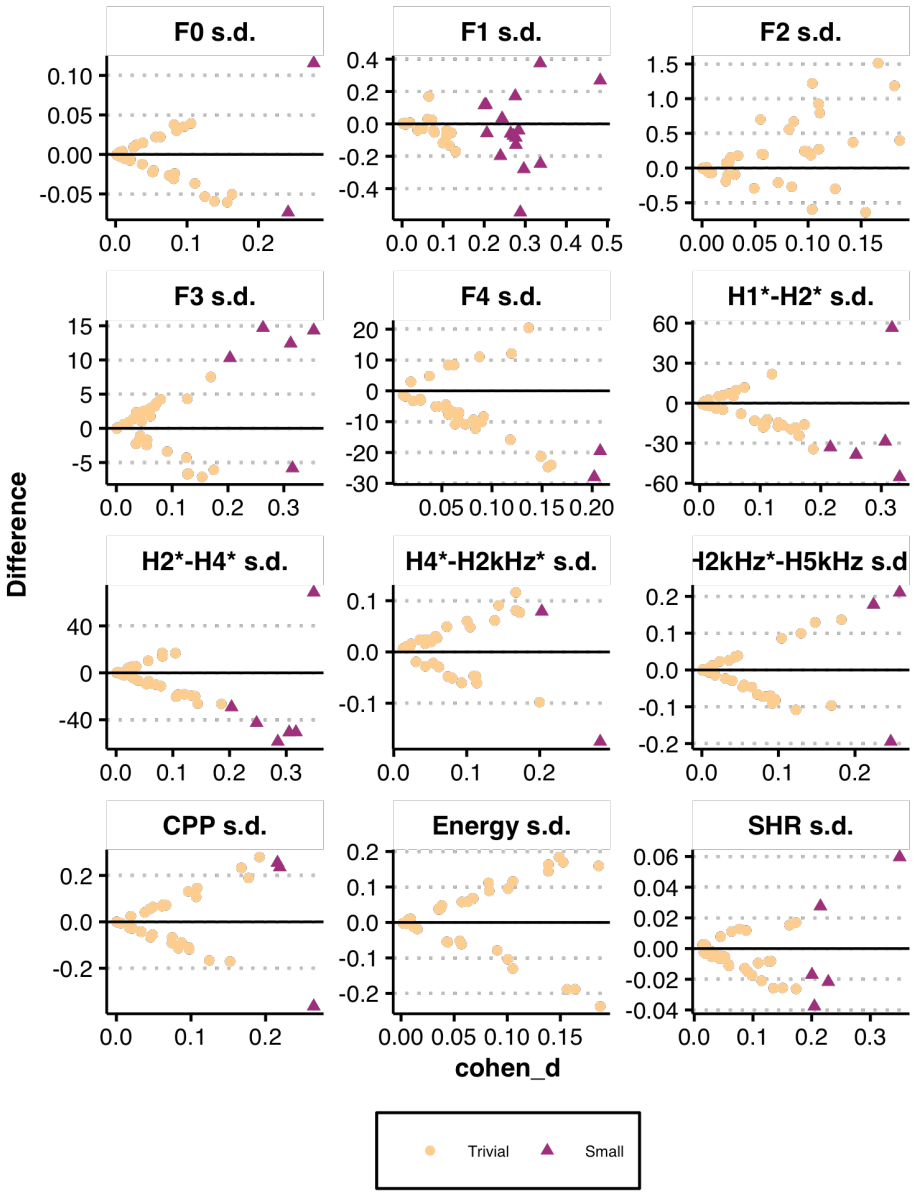

Figure 3.4: This figure uses the format of 3.3, but reports on the standard deviation measures.

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 different measures than (Ng et al., 2012), the H1*–H2* results are consistent with the finding that Cantonese tends to be breathier (or English creakier)—the current analysis does not distinguish between these interpretations. The H4*–H2kHz* results are not consistent with Ng et al. (2012), yet for both spectral shape measures, it is important to reiterate that they are difficult to interpret on their own. The interpretation here should be taken with a grain of salt.

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

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

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

which likely varies by individual (Järvinen et al., 2013).

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

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

3.2.5 Principal components analysis

Methods

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

does not shed light on perception. Rather, it offers a signal-based account that can be used to generate perception prediction and interpret the results of perception research, as outlined in Chapter 1.

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

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

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

Additionally, across each of the components, only loadings with an absolute value of 0.45 or higher were interpreted. While Lee et al. (2019) use a threshold of 0.32, Tabachnick 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 greater than the fair threshold are interpreted. This methodological decision facilitates interpreting meaningful loadings on components.

Results

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

To assess whether talkers exhibit the same structure in voice variability across their languages, patterns present across the different PCAs are considered. This provides context for understanding what unique structural characteristics in talkers'

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

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

voices look like. To this end, this section briefly summarizes 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. Figure 3.5 can be interpreted as follows. The left column visualizes the VF32A's Cantonese PCA, and the right column English. Each panel depicts a single component, and the components are numbered along the right in order by the amount of variance accounted for in the PCA.

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

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

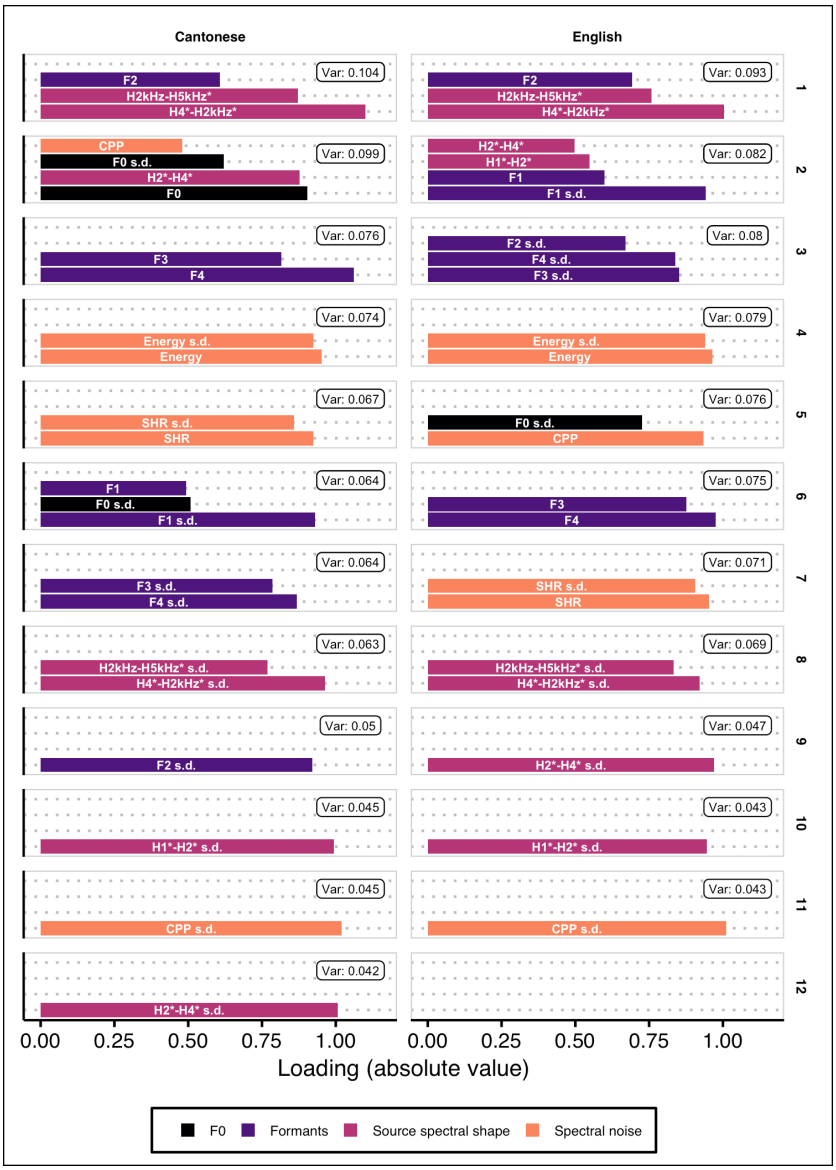

Figure 3.5: In this depiction of the components of the Cantonese and English PCAs for VF32A—a single talker from the corpus taken as an example. Loadings are represented by bar height and are labelled with the variable name; color represents conceptual groupings. The component's variance accounted for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

components. Instead, it is more appropriate to attend to component composition and the variance accounted for by different components.

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

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

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

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

While the formant and spectral shape moving standard deviations often exhibited these common patterns, variables in these categories were just as likely to

pattern in more idiosyncratic ways, loading alongside each other, F0, formants, and spectral measures. This kind of variability is not readily summarizable.

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

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

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

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

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

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

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

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

of the voices they measure. Considering that the total variance unaccounted for by the PCAs ranges from 17.3%–25.1%, the unaccounted for variability may also be idiosyncratic in nature.

3.2.6 Canonical redundancy analysis

Methods

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

To assess whether variation in a talker's voice is structurally similar across both languages, PCA output from both languages is compared by calculating redundancy indices in a canonical correlation analysis (CCA: Stewart 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)—this result indicates that regardless of language, talkers are more similar to themselves than talkers are to each other. A second Welch's t-test testing the same versus different language for the across-talker comparisons did not find a difference between those groups ($t(4485.9) = -1.53$, $p = 0.13$, d = 0.05). This result demonstrates that language is not a delineating factor, or at the very least, the role of language is eclipsed by the role of talker. This interpretation makes sense, given the high degree of within-talker similarity demonstrated in the first Welch's t-test.

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

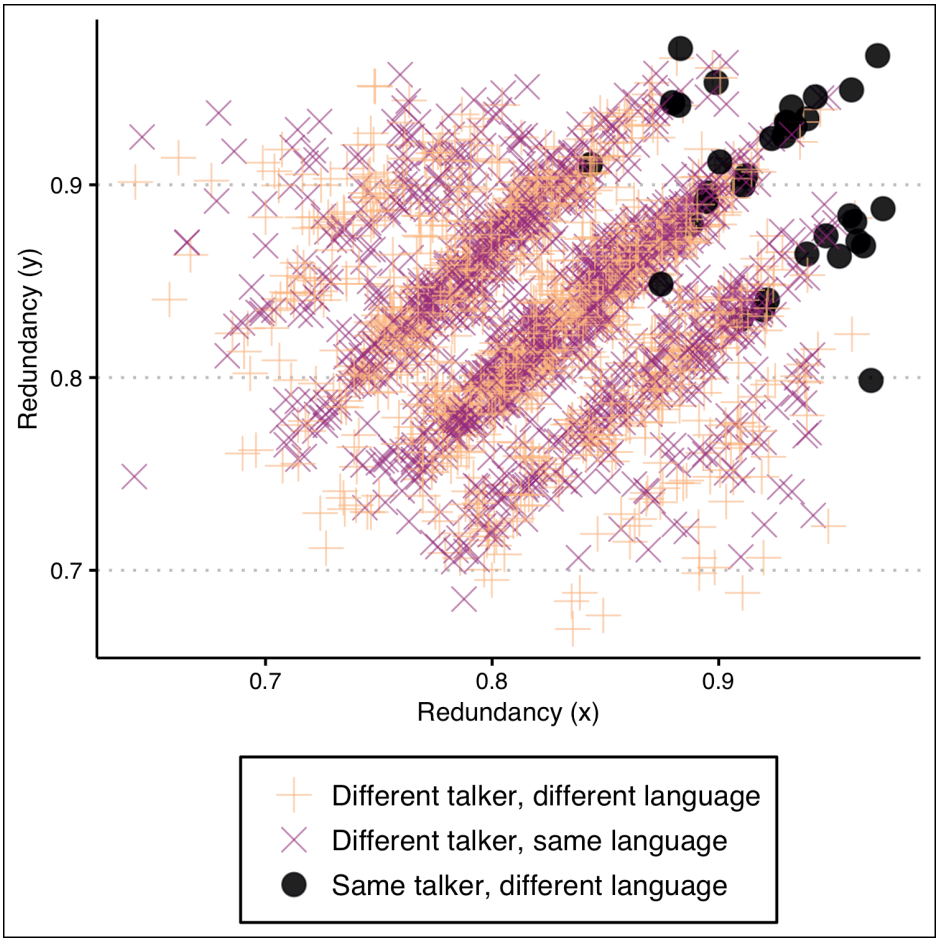

Figure 3.6: This plot depicts the relationship between the two redundancy indices for three different types of comparisons. Across-talker comparisons represented by orange "+" (different language) and pink "x" (same language) overlap in their entirety. Within-talker comparisons are represented by the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

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.

3.2.7 Passage length analysis

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

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

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

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

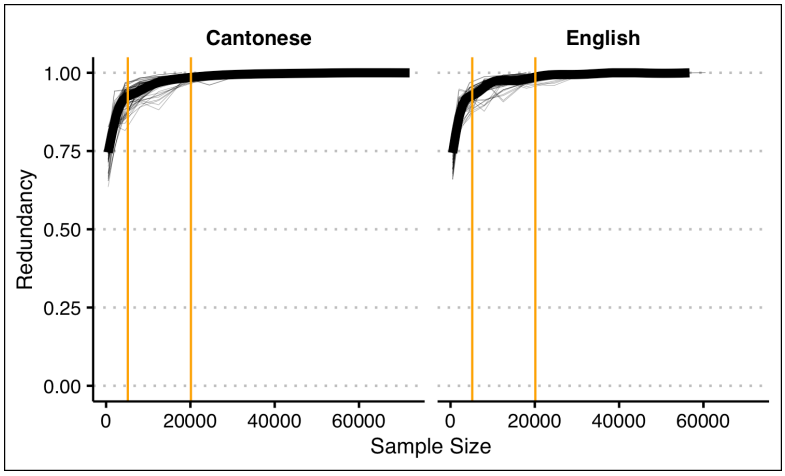

Figure 3.7: Passage length redundancy indices are plotted against the sample size of the smaller PCA. Smoothed curves show a rapid increase in redundancy followed by a levelling off between the vertical orange lines, which represent the sample sizes used in prior work (x = 5,000) and the present study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

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

3.3 Discussion and conclusion

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

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

at perceptual tasks in both domains.

It is reassuring that the results from the first two approaches used here reflect prior findings. For example, when there was a difference for measures like F0 or H1*–H2*, it tended to mirror expectations from the literature that Cantonese tends to have lower pitch and breathier voice quality than English (Ng et al., 2012, 2010). At the same time, most talkers did not exhibit a meaningful difference, validating prior work that found no differences (Altenberg 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 emerged for the SpiCE talkers. Respectively, they are associated with (i) perceived breathiness or brightness, (ii) vocal tract size or speaker identity, and (iii) a combination of phonation type and vocal tract configuration—perhaps reflecting shared linguistic variation. Much like Lee et al. (2019), the key shared dimensions relate to the timbre, identity, and vocal tract size.

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

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

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

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

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

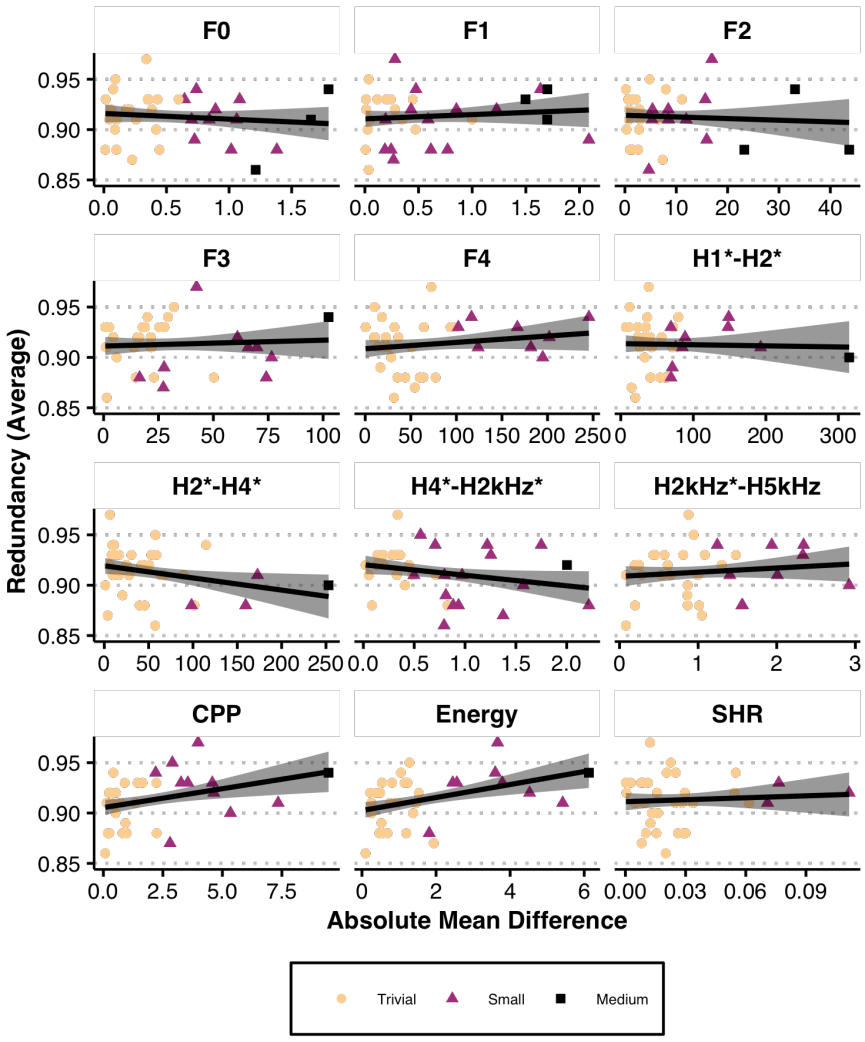

Figure 3.8: The average redundancy value for each talker is plotted against the absolute value of the difference of means across languages for that talker. Color and shape indicate the size of Cohens' *d*. The superimposed regression line summarizes the relationship between these values.

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

While methodological choices may account for some part of these results, the data differences between the current chapter and previous studies are also pertinent. This chapter uses substantially longer passages than the short samples in Lee et al. (2019). Larger speech samples allow for a stable underlying structure to emerge. Smaller samples, conversely, may reflect more ephemeral variation in a talker's voice, and thus not be representative of the talker's full range. The passage length analysis in this chapter shows that the number of samples needed for stabilization is substantially larger than the 5,000 samples used in Lee et al. (2019). This does not necessarily discount their work, however, as the current chapter uses spontaneous speech, which is arguably more variable than read speech.[5] It's plausible that an analysis of sentence reading would not need as much data to cover talkers' range of variability in reading aloud. The body of literature in the introduction establishes differences in voice quality across speaking styles (e.g., Lee 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 variability and structure of talkers' voices maps onto listeners' organization of a voice space for use in talker recognition and discrimination. Turning to listener and behavioral data will help in deciphering what is meaningful variation within

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

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

Bibliography

Afouras, Triantafyllos, Chung, Joon Son, and Zisserman, Andrew. Now you're speaking my language: Visual language identification. In *Proceedings of Interspeech 2020*, pages 2402–2406, 2020. doi:10.21437/Interspeech.2020-2921. → pages 46, 82

Alderete, John, Chan, Queenie, and Yeung, H. Henny. Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, 191:103952, 2019. doi:10.1016/j.cognition.2019.04.021. → page 11

Altenberg, Evelyn P. and Ferrand, Carole T. Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, 20(1):89–96, 2006. doi:10.1016/j.jvoice.2005.01.005. → pages 49, 52, 61, 63, 83

Amengual, Mark. Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, 23(5):954–970, 2017. doi:10.1177/1367006917741364. → pages 3, 13

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

Antoniou, Mark, Best, Catherine T., Tyler, Michael D., and Kroos, Christian. Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, 38(4):640–653, 2010. doi:10.1016/j.wocn.2010.09.005. → page 93

Antoniou, Mark, Best, Catherine T., Tyler, Michael D., and Kroos, Christian. Inter-language interference in VOT production by L2-dominant bilinguals:

Asymmetries in phonetic code-switching. *Journal of Phonetics*, 39(4): 558–570, 2011. doi:10.1016/j.wocn.2011.03.001. → page 93

Ardila, Rosana, Branson, Megan, Davis, Kelly, Kohler, Michael, Meyer, Josh, Henretty, Michael, Morais, Reuben, Saunders, Lindsay, Tyers, Francis, and Weber, Gregor. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, 2020. https://www.aclweb.org/anthology/2020.lrec-1.520. → page 11

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

Balukas, Colleen and Koops, Christian. Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, 19 (4):423–443, 2015. doi:10.1177/1367006913516035. → page 93

Barr, Dale J., Levy, Roger, Scheepers, Christoph, and Tily, Harry J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013. doi:10.1016/j.jml.2012.11.001. → page 112

Bates, Douglas, Kliegl, Reinhold, Vasishth, Shravan, and Baayen, Harald. Parsimonious mixed models. *ArXiv Preprints*, pages 1–21, May 2018. http://arxiv.org/abs/1506.04967. → page 112

Bauer, Robert S. and Benedict, Paul K. *Modern Cantonese Phonology*. De Gruyter Mouton, Berlin, 1997. doi:10.1515/9783110823707. → page 99

Belin, Pascal, Fecteau, Shirley, and Bédard, Catherine. Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3): 129–135, 2004. doi:10.1016/j.tics.2004.01.008. → page 37

Boersma, Paul and Weenink, David. Praat: Doing phonetics by computer, 2021. http://www.praat.org/. Version 6.1.38. → page 55

Bolton, Kingsley, Bacon-Shone, John, and Lee, Siu-lun. Societal multilingualism in Hong Kong. In *Multilingual Global Cities*, pages 160–184. Routledge, 2020. doi:10.4324/9780429463860-12. → page 15

Bradlow, Ann R, Ackerman, Lauren, Burchfield, L Ann, Hesterberg, Lisa, Luque, Jenna, and Mok, Kelsey. Language- and talker-dependent variation in global features of native and non-native speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 356–359, Hong Kong, 2011. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Bradlow/Bradlow.pdf. → pages 10, 98

Bradlow, Ann R., Kim, Midam, and Blasingame, Michael. 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, 2017. doi:10.1121/1.4976044. → pages 53, 107, 121, 129

Brehm, Laurel and Alday, Phillip M. A decade of mixed models: It's past time to set your contrasts. *OSF Preprints*, July 2021. https://osf.io/3tgq6/. → page 113

Brown, Esther L. and Amengual, Mark. Fine-grained and probabilistic cross-linguistic influence in the pronunciation of cognates: Evidence from corpus-based spontaneous conversation and experimentally elicited data. *Studies in Hispanic and Lusophone Linguistics*, 8(1):59–83, 2015. doi:10.1515/shll-2015-0003. → page 93

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

Bruggeman, Laurence and Cutler, Anne. No L1 privilege in talker adaptation. *Bilingualism: Language and Cognition*, pages 1–13, 2019. doi:10.1017/S1366728919000646. → page 131

Bullock, Barbara E. and Toribio, Almeida Jacqueline. Trying to hit a moving target: On the sociophonetics of code-switching. In Isurin, Ludmila, Winford, Donald, and deBot, Kees, editors, *Studies in Bilingualism*, volume 41, pages 189–206. John Benjamins Publishing Company, Amsterdam, 2009. doi:10.1075/sibil.41.12bul. → pages 4, 47, 53, 92, 93, 94, 129, 130

Burkner, Paul-Christian. brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi:10.18637/jss.v080.i01. → page 111

Burton, A. Mike, Kramer, Robin S. S., Ritchie, Kay L., and Jenkins, Rob. Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1):202–223, 2016. doi:10.1111/cogs.12231. → pages 41, 43, 69

Casillas, Joseph V. Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages*, 6(1):9, 2021. doi:10.3390/languages6010009. → pages 91, 92, 94, 133

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

Chan, Alice Y. W. and Li, David C. S. English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems. *Language, Culture and Curriculum*, 13(1):67–85, 2000. doi:10.1080/07908310008666590. → page 99

Chan, Leighanne, Johnson, Khia A., and Babel, Molly. Lexically-guided perceptual learning in early Cantonese-English bilinguals. *The Journal of the Acoustical Society of America*, 147(3):EL277–EL282, 2020. doi:10.1121/10.0000942. → pages 3, 131

Chang, Charles B. Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity. In Raimy, Eric and Cairns, Charles E., editors, *The Segment in Phonetics and Phonology*, pages 199–217. John Wiley & Sons, Inc., Chichester, UK, 1 edition, 2015. doi:10.1002/9781118555491.ch9. → page 90

Cheng, Andrew. Cross-linguistic F0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, 147(2): EL67–EL73, 2020. doi:10.1121/10.0000498. → pages 51, 52

Cho, Taehong and Ladefoged, Peter. Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, 27(2):207–229, 1999. doi:10.1006/jpho.1999.0094. → pages 102, 106, 118, 125

Chodroff, Eleanor. *Structured variation in obstruent production and perception*. PhD dissertation, Johns Hopkins University, Baltimore, MD, 2017. https://jscholarship.library.jhu.edu/handle/1774.2/44696. → page 98

Chodroff, Eleanor and Baese-Berk, Melissa. Constraints on variability in the voice onset time of L2 English stop consonants. In Calhoun, Sasha, Escudero, Paola, Tabain, Marija, and Warren, Paul, editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 661–665, Melbourne, Austrailia, 2019. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_710.pdf. → pages 98, 100, 103, 106, 120, 125, 130

Chodroff, Eleanor and Wilson, Colin. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47, 2017. doi:10.1016/j.wocn.2017.01.001. → pages 39, 97, 98, 99, 100, 101, 102, 106, 107, 109, 110, 113, 116, 118, 120, 125, 133

Chodroff, Eleanor and Wilson, Colin. Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4(s2), 2018. doi:10.1515/lingvan_2017_0047. → page 100

Chodroff, Eleanor, Golden, Alessandra, and Wilson, Colin. Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, 145(1): EL109–EL115, 2019. doi:10.1121/1.5088035. → page 110

Clumeck, Harold, Barton, David, Macken, Marlys A., and Huntington, Dorothy A. The aspiration contrast in Cantonese word-initial stops: Data from children and adults. *Journal of Chinese Linguistics*, 9(2):210–225, 1981. https://www.jstor.org/stable/23753507. → pages 99, 120

Deuchar, Margaret, Davies, Peredur, Herring, Jon Russell, Parafita Couto, M. Carmen, and Carter, Diana. Building bilingual corpora. In Thomas, Enlli M. and Mennen, Ineke, editors, *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters, 2014. doi:10.21832/9781783091713_008. → pages 9, 34

Ethnologue. Chinese, Yue. In Eberhard, David Mm, Simons, Gary F., and Fennig, Charles D., editors, *Ethnologue: Languages of the World*. SIL

International, Dallas, TX, 24 edition, 2021. http://www.ethnologue.com. Online version. → page 11

Faytak, Matthew Donald. *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. Doctoral dissertation, University of California, Berkeley, 2018. https://escholarship.org/uc/item/0jr0010h. → pages 97, 98, 99

Flege, James Emil and Bohn, Ocke-Schwen. The revised speech learning model (SLM-r). In Wayland, Ratree, editor, *Second Language Speech Learning: Theoretical and Empirical Progress*, pages 3–83. Cambridge University Press, 2021. doi:10.1017/9781108886901.002. → pages 1, 3, 88, 89, 90, 91, 92, 96, 99

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

Fricke, Melinda, Kroll, Judith F., and Dussias, Paola E. Phonetic variation in bilingual speech: A lens for studying the production-comprehension link. *Journal of Memory and Language*, 89:110–137, 2016b. doi:10.1016/j.jml.2015.10.001. → pages 55, 92, 93, 94, 132, 133

Fricke, Melinda, Zirnstein, Megan, Navarro-Torres, Christian, and Kroll, Judith F. Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, 22(1):200–207, 2019. doi:10.1017/S1366728918000482. → pages 3, 97

Gahl, Susanne, Yao, Yao, and Johnson, Keith. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806, 2012. doi:10.1016/j.jml.2011.11.006. → pages 8, 120

Garellek, Marc. The phonetics of voice. In Katz, William F. and Assmann, Peter F., editors, *The Routledge Handbook of Phonetics*. Routledge, 2019. doi:10.4324/9780429056253_5. → pages 39, 40, 42, 57, 84

Gelman, Andrew, Simpson, Daniel, and Betancourt, Michael. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017. doi:10.3390/e19100555. → pages 111, 114

Gertken, Libby M., Amengual, Mark, and Birdsong, David. Assessing language dominance with the Bilingual Language Profile. In Leclercq, Pascale, Edmonds, Amanda, and Hilton, Heather, editors, *Measuring L2 proficiency: Perspectives from SLA*, pages 208–225. Multilingual Matters, Bristol, UK, 2014. doi:10.21832/9781783092291-014. → page 2

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

Goldrick, Matthew, Runnqvist, Elin, and Costa, Albert. Language switching makes pronunciation less nativelike. *Psychological Science*, 25(4):1031–1036, 2014. doi:10.1177/0956797613520014. → pages 92, 93, 94

Google. Cloud speech-to-text, 2019. https://cloud.google.com/speech-to-text/. v1. → pages 11, 25

Grieve, Jack. Observation, experimentation, and replication in linguistics. *Linguistics*, 0, 2021. doi:10.1515/ling-2021-0094. → page 10

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

Grosjean, François. An attempt to isolate, and then differentiate, transfer and interference. *International Journal of Bilingualism*, 16(1):11–21, 2011. doi:10.1177/1367006911403210. → pages 93, 95, 96

Guion, Susan G. The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, 60(2):98–128, 2003. doi:10.1159/000071449. → page 91

Haines, Nathaniel, Kvam, Peter D., Irving, Louis H., Smith, Colin, Beauchaine, Theodore P., Pitt, Mark A., Ahn, Woo-Young, and Turner, Brandon M. Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. *PsyArXiv Preprints*, August 2020. doi:10.31234/osf.io/xr7y3. → page 110

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

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

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

Iseli, Markus, Shue, Yen-Liang, and Alwan, Abeer. 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, 2007. doi:10.1121/1.2697522. → page 57

Jadoul, Yannick, Thompson, Bill, and de Boer, Bart. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. doi:10.1016/j.wocn.2018.07.001. → page 55

Järvinen, Kati, Laukkanen, Anne-Maria, and Aaltonen, Olli. Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*, 38(2):47–51, 2013. ISSN 1401-5439. doi:10.3109/14015439.2012.687764. https://doi.org/10.3109/14015439.2012.687764. → pages 50, 52, 55, 68

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

Johnson, Khia A. Probabilistic reduction in Spanish-English bilingual speech. In Calhoun, Sasha, Escudero, Paola, Tabain, Marija, and Warren, Paul, editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1263–1267, Melbourne, Australia, 2019. → page 120

Johnson, Khia A. SpiCE: Speech in Cantonese and English, 2021. https://doi.org/10.5683/SP2/MJOXP3. V1. → pages 6, 8, 124

Johnson, Khia A. and Babel, Molly. Language contact within the speaker: Phonetic variation and crosslinguistic influence. *OSF Preprints*, 2021a. doi:10.31219/osf.io/jhsfc. https://osf.io/jhsfc/. → pages 94, 129

Johnson, Khia A. and Babel, Molly. Language contact within the speaker: Phonetic variation and crosslinguistic influence. Technical report, OSF Preprints, 2021b. https://osf.io/jhsfc/. → page 8

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

Ju, Min and Luce, Paul A. Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science*, 15(5):314–318, 2004. doi:10.1111/j.0956-7976.2004.00675.x. → pages 4, 130

Kawahara, Hideki, Agiomyrgiannakis, Yannis, and Zen, Heiga. 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, 2016. doi:10.21437/SSW.2016-36. → page 56

Keating, Patricia and Kuo, Grace. Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, 132(2):1050–1060, 2012. doi:10.1121/1.4730893. → pages 45, 49, 50, 126

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

Keshet, J., Sonderegger, M., and Knowles, T. AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction, 2014. https://github.com/mlml/autovot/. Version 0.94. → page 101

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

Kreiman, Jody, Gerratt, Bruce R., Garellek, Marc, Samlan, Robin, and Zhang, Zhaoyan. Toward a unified theory of voice production and perception. *Loquens*, 1(1):e009, 2014. doi:10.3989/loquens.2014.009. → pages 39, 40, 41, 56, 57, 69

Kreiman, Jody, Lee, Yoonjeong, Garellek, Marc, Samlan, Robin, and Gerratt, Bruce R. Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, 149(1):457–465, 2021. doi:10.1121/10.0003331. → pages 40, 69

Kruschke, John K. Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3):299–312, 2011. doi:10.1177/1745691611406925. → pages 111, 112

Labov, William, Ash, Sharon, and Boberg, Charles. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton, 2008. doi:10.1515/9783110167467. → page 67

Latinus, Marianne and Belin, Pascal. Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2:175, 2011. doi:10.3389/fpsyg.2011.00175. → pages 43, 83

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

Lavner, Yizhar, Rosenhouse, Judith, and Gath, Isak. The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, 4(1):63–74, 2001. doi:10.1023/A:1009656816383. → pages 43, 83

Lee, Binna and Sidtis, Diana Van Lancker. The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, 20(3):174–185, 2017. doi:10.1080/2050571x.2016.1273572. → pages 50, 51, 52, 55, 86, 126

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

Lee, Yoonjeong and Kreiman, Jody. Within- and between-speaker acoustic variability: Spontaneous versus read speech. In *The 178th Meeting of the Acoustical Society of America*, San Diego, CA, 2019. doi:10.1121/1.5137431. Poster. → pages 41, 42, 76, 83, 86, 124

Lee, Yoonjeong and Kreiman, Jody. Language effects on acoustic voice variation within and between talkers. In *The 179th Meeting of the Acoustical Society of*

America, Acoustics Virtually Everywhere, 2020. doi:10.1121/1.5146847. Poster. → pages 41, 42, 46, 69, 76, 83, 124

Lee, Yoonjeong, Keating, Patricia, and Kreiman, Jody. Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3):1568–1579, 2019. doi:10.1121/1.5125134. → pages 38, 39, 41, 42, 43, 45, 53, 54, 56, 60, 61, 69, 70, 72, 74, 76, 81, 82, 83, 84, 86, 124

Lein, Tatjana, Kupisch, Tanja, and van de Weijer, Joost. Voice onset time and global foreign accent in German–French simultaneous bilinguals during adulthood. *International Journal of Bilingualism*, 20(6):732–749, 2016. doi:10.1177/1367006915589424. → page 91

Leung, Man-Tak and Law, Sam-Po. HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6(2):305–325, 2001. doi:10.1075/ijcl.6.2.06leu. → page 11

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

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

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

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

Lieberman, Philip and Blumstein, Sheila E. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge, 1988. doi:10.1017/CBO9781139165952. → page 101

Lindblom, Björn and Maddieson, Ian. Phonetic universals in consonant systems. In Hyman, Larry M. and Li, Charles N., editors, *Language, Speech, and Mind: Studies in Honour of Victoria A. Fromkin*, pages 62–78. Routledge, London, 1988. → page 91

Lisker, Leigh and Abramson, Arthur S. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422, 1964. doi:10.1080/00437956.1964.11659830. → pages 99, 118, 120, 125

Lisker, Leigh and Abramson, Arthur S. Some effects of context on voice onset time in english stops. *Language and Speech*, 10(1):1–28, 1967. doi:10.1177/002383096701000101. → page 100

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

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

Lloy, Angelina, Johnson, Khia A., and Babel, Molly. Bilingual talker identification with spontaneous speech in Cantonese and English: The role of language-specific knowledge. In *The 179th Meeting of the Acoustical Society of America*, Virtual, 2020. doi:10.1121/1.5147685. Poster. → pages 87, 132

Lloy, Angelina, Johnson, Khia, and Babel, Molly. Examining the roles of language familiarity and bilingualism in talker recognition. In *The 13th International Symposium on Bilingualism*, Virtual, 2021. Poster. → pages 87, 132

Loveday, Leo. 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, 1981. doi:10.1177/002383098102400105. → pages 51, 127

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

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

Matthews, Stephen, Yip, Virginia, and Yip, Virginia. *Cantonese: A Comprehensive Grammar*. Routledge, 2013. doi:10.4324/9780203835012. → pages 20, 48, 82, 99, 128

McAuliffe, Michael, Socolof, Michaela, Stengel-Eskin, Elias, Mihuc, Sarah, Wagner, Michael, and Sonderegger, Morgan. Montreal Forced Aligner, 2017. https://montrealcorpustools.github.io/Montreal-Forced-Aligner/. Version 1.0.1. → page 28

McElreath, Richard. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, 2 edition, 2020. doi:10.1201/9780429029608. → pages 112, 113, 114

McMurray, Bob, Tanenhaus, Michael K., and Aslin, Richard N. Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2): B33–B42, 2002. doi:10.1016/S0010-0277(02)00157-9. → page 121

Ménard, Lucie, Schwartz, Jean-Luc, and Aubin, Jérôme. Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50(1):14–28, 2008. doi:10.1016/j.specom.2007.06.004. → pages 97, 98

Mennen, Ineke, Scobbie, James M, de Leeuw, Esther, Schaeffler, Sonja, and Schaeffler, Felix. Measuring language-specific phonetic settings. *Second Language Research*, 26(1):13–41, 2010. → page 40

Mielke, Jeff. A phonetically based metric of sound similarity. *Lingua*, 122(2): 145–163, 2012. doi:10.1016/j.lingua.2011.04.006. → page 90

Mielke, Jeff and Nielsen, Kuniko. Voice onset time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America*, 144(4):2166–2177, 2018. doi:10.1121/1.5059493. → page 99

Munson, Benjamin and Babel, Molly. The phonetics of sex and gender. In Katz, William F. and Assmann, Peter F., editors, *The Routledge Handbook of Phonetics*. Routledge, 2019. doi:10.4324/9780429056253_19. → page 57

Munson, Benjamin, Edwards, Jan, Schellinger, Sarah K, Beckman, Mary E, and Meyer, Marie K. Deconstructing phonetic transcription: Covert contrast,

perceptual bias, and an extraterrestrial view of vox humana. *Clinical linguistics & phonetics*, 24(4-5):245–260, 2010. → page 47

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

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

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

Ng, Manwa L., Hsueh, Gigi, and Sam Leung, Cheung-Shing. Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, 12 (3):230–236, 2010. doi:10.3109/17549501003721080. → pages 49, 61, 83

Ng, Manwa L, Chen, Yang, and Chan, Ellen YK. 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, 2012. doi:10.1016/j.jvoice.2011.07.013. → pages 49, 50, 53, 61, 63, 67, 83, 124, 126

Ng, Raymond W. M., Kwan, Alvin C.M., Lee, Tan, and Hain, Thomas. ShefCE: A Cantonese-English bilingual speech corpus for pronunciation assessment. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5825–5829, 2017. doi:10.1109/ICASSP.2017.7953273. → page 10

Nieuwenhuis, Rense, Manfred, te Grotenhuis, and Pelzer, Ben. Weighted effect coding for observational data with wec. *The R Journal*, 9(1):477, 2017. doi:10.32614/rj-2017-017. → page 113

Olson, Daniel J. The role of code-switching and language context in bilingual phonetic transfer. *Journal of the International Phonetic Association*, 46(3): 263–285, 2016. doi:10.1017/S0025100315000468. → pages 92, 93, 94

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

Orena, Adriel John, Polka, Linda, and Theodore, Rachel M. Identifying bilingual talkers after a language switch: Language experience matters. *The Journal of the Acoustical Society of America*, 145(4):EL303–EL309, 2019. doi:10.1121/1.5097735. → pages 4, 5, 44, 45, 122, 128, 132, 133

Panayotov, Vassil, Chen, Guoguo, Povey, Daniel, and Khudanpur, Sanjeev. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015. doi:10.1109/icassp.2015.7178964. → page 11

Perrachione, Tyler K. Recognizing speakers across languages. In Frühholz, Sascha and Belin, Pascal, editors, *The Oxford Handbook of Voice Perception*, pages 514–538. Oxford University Press, 2018. doi:10.1093/oxfordhb/9780198743187.013.23. → page 44

Perrachione, Tyler K., Furbeck, Kristina T., and Thurston, Emily J. Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, 146(5):3384–3399, 2019. doi:10.1121/1.5126697. → pages 44, 46, 53, 76

Pitt, Mark A., Johnson, Keith, Hume, Elizabeth, Kiesling, Scott, and Raymond, William. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005. doi:10.1016/j.specom.2004.09.001. → pages 9, 10, 22

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

Podesva, Robert J. and Callier, Patrick. Voice quality and identity. *Annual Review of Applied Linguistics*, 35:173–194, 2015. doi:10.1017/S0267190514000270. → pages 37, 38, 40

Polinsky, Maria. *Heritage Languages and their Speakers*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge, 2018. doi:10.1017/9781107252349. → page 129

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. http://www.R-project.org/. → pages 61, 69, 106, 111

Reinisch, Eva, Weber, Andrea, and Mitterer, Holger. Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1):75–86, 2013. doi:10.1037/a0027979. → page 122

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

Ryabov, Rashel, Malakh, Marcella, Trachtenberg, Malka, Wohl, Sherrie, and Oliveira, Gisele. Self-perceived and acoustic voice characteristics of Russian-English bilinguals. *Journal of Voice*, 30(6):772.e1–772.e8, 2016. doi:10.1016/j.jvoice.2015.11.009. → page 51

Samuel, Arthur G. Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111:104070, 2020. doi:10.1016/j.jml.2019.104070. → page 97

Sancier, Michele L. and Fowler, Carol A. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25(4):421–436, 1997. doi:10.1006/jpho.1997.0051. → pages 3, 93

Shue, Yen-Liang, Keating, Patricia, Vicenik, Chad, and Yu, Kristine. VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences*, volume 3, pages 1846–1849, Hong Kong, 2011. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011. → page 56

Simonet, Miguel. The phonetics and phonology of bilingualism. In *Oxford Handbooks Online*. Oxford University Press, 2016. doi:10.1093/oxfordhb/9780199935345.013.72. → page 89

Simonet, Miquel and Amengual, Mark. Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, 24(2):208–221, 2019. doi:10.1177/1367006919826388. → pages 3, 22, 93, 94, 96

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

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

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

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

Soto-Faraco, Salvador, Navarra, Jordi, Weikum, Whitney M., Vouloumanos, Athena, Sebastián-Gallés, Núria, and Werker, Janet F. Discriminating languages by speech-reading. *Perception & Psychophysics*, 69(2):218–231, 2007. doi:10.3758/BF03193744. → pages 46, 82

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*, 2021. https://mc-stan.org. → page 111

Statistics Canada. Proportion of mother tongue responses for various regions in Canada, 2016 Census, 2017. https://www12.statcan.gc.ca/census-recensement /2016/dp-pd/dv-vd/lang/index-eng.cfm. → pages 13, 131

Stewart, Douglas and Love, William. A general canonical correlation index. *Psychological Bulletin*, 70(3, pt.1):160–163, 1968. doi:10.1037/h0026143. → page 77

Stuart-Smith, Jane, Sonderegger, Morgan, Rathcke, Tamara, and Macdonald, Rachel. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6(3-4):505–549, 2015. doi:10.1515/lp-2015-0015. → pages 99, 107, 120

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

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

Sundara, Megha, Polka, Linda, and Baum, Shari. Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition*, 9(1): 97–114, 2006. doi:10.1017/S1366728905002403. → pages 91, 92, 93, 94

Tabachnick, Barbara G. and Fidell, Linda S. *Using Multivariate Statistics*. Pearson Education, Inc., 6 edition, 2013. → page 70

Tanner, James, Sonderegger, Morgan, Stuart-Smith, Jane, and Fruehwald, Josef. Toward "English" phonetics: Variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, 3, 2020. doi:10.3389/frai.2020.00038. → page 6

Tse, Holman. *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA, 2019. http://d-scholarship.pitt.edu/35721/. → pages 20, 29

Tsui, Rachel Ka-Ying, Tong, Xiuli, and Chan, Chuck Siu Ki. Impact of language dominance on phonetic transfer in Cantonese–English bilingual language switching. *Applied Psycholinguistics*, 40(1):29–58, 2019. doi:10.1017/S0142716418000449. → page 95

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

Vasishth, Shravan, Nicenboim, Bruno, Beckman, Mary E., Li, Fangfang, and Kong, Eun Jong. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71:147–161, 2018. doi:10.1016/j.wocn.2018.07.008. → pages 111, 114

Voigt, Rob, Jurafsky, Dan, and Sumner, Meghan. Between- and within-speaker effects of bilingualism on f0 variation. In *Proceedings of Interspeech 2016*, pages 1122–1126, San Francisco, CA, 2016. doi:10.21437/Interspeech.2016-1506. → pages 51, 127

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

Wilson, C and Mihalicek, V. *Language files: Materials for an introduction to language and linguistics*. Ohio State University Press, Columbus, UH, 2011. https://linguistics.osu.edu/research/pubs/lang-files. → pages 48, 82

Winterstein, Grégoire, Tang, Carmen, and Lai, Regine. CantoMap: A Hong Kong Cantonese MapTask corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2906–2913, Marseille, France, 2020. https://aclanthology.org/2020.lrec-1.355. → page 11

Wong, Wai Yi Peggy. *Syllable fusion in Hong Kong Cantonese connected speech*. Doctoral dissertation, The Ohio State University, Columbus, OH, 2006. http://rave.ohiolink.edu/etdc/view?acc_num=osu1143227948. → pages 26, 28

Xue, Steve An, Hagstrom, Fran, and Hao, JianPing. 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, 2002. doi:10.1179/136132802805576544. → page 50

Yang, Jing. Comparison of VOTs in Mandarin–English bilingual children and corresponding monolingual children and adults. *Second Language Research*, page 0267658319851820, 2019. doi:10.1177/0267658319851820. → pages 95, 99

Yang, Yike, Chen, Si, and Chen, Xi. F0 patterns in Mandarin statements of Mandarin and Cantonese speakers. In *Proceedings of Interspeech 2020*, pages 4163–4167, 2020. doi:10.21437/Interspeech.2020-2549. → page 51

Yau, Macro. PyJyutping, 2019. https://github.com/MacroYau/PyJyutping. → page 11

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

Yuan, Jiahong, Ryant, Neville, and Liberman, Mark. Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone. In *Proceedings of the 2014 IEEE International Conference on Acoustics,*

Speech and Signal Processing, pages 2539–2543, 2014. doi:10.1109/ICASSP.2014.6854058. → page 29

Zarate, Jean Mary, Tian, Xing, Woods, Kevin J. P., and Poeppel, David. Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, 5(1):11475, 2015. doi:10.1038/srep11475. → page 45