JASA ARTICLE

# The structure of acoustic voice variation in bilingual speech

Khia A. Johnson and Molly Babel[a)]
*Department of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada*

**ABSTRACT:**
When a bilingual switches languages, do they switch their voice? Using a conversational corpus of speech from early Cantonese-English bilinguals ($n = 34$), this paper examines the talker-specific acoustic signatures of bilingual voices. Following the psychoacoustic model of voice, 24 filter and source-based acoustic measurements are estimated. The analysis summarizes mean differences for these dimensions and identifies the underlying structure of each talker's voice across languages with principal component analyses. Canonical redundancy analyses demonstrate that while talkers vary in the degree to which they have the same voice across languages, all talkers show strong similarity with themselves, suggesting an individual's voice remains relatively constant across languages. Voice variability is sensitive to sample size, and we establish the required sample to settle on a consistent impression of one's voice. These results have implications for human and machine voice recognition for bilinguals and monolinguals and speak to the substance of voice prototypes. *© 2023 Acoustical Society of America*.

https://doi.org/10.1121/10.0019659

(Received 2 December 2022; revised 12 May 2023; accepted 23 May 2023; published online 7 June 2023)

[Editor: Jody Kreiman] Pages: 3221–3238

## I. INTRODUCTION

An individual's voice is a rich signal, carrying information about an array of biological, physiological, and psychological states while presenting linguistic and social meaning (Podesva and Callier, 2015). While many of these states are transient—for example, one's emotional state can vary across a conversation—an individual's vocal identity remains intact. Voices have been likened to auditory faces—they are uniquely individual yet share basic characteristics with the broader population (Belin *et al.*, 2004; Lee *et al.*, 2019). Where faces share an overall shape and composition of features (e.g., eyes, nose, etc.), voices share the acoustic consequences of similar vocal anatomy. At the same time, upon seeing a familiar face or hearing a familiar voice, one can often immediately identify the individual and ascertain their present state. In this way, both voices and faces signal identity along with aspects of the individual's physical and emotional state. Voices also simultaneously convey a communicative message. A perceiver is, thus, presented with a signal interwoven with talker-indexical, affective, social, and linguistic information. Bilinguals offer a unique angle on voice identity. While a bilingual's physiology does not change as they switch between their languages, their social personae may change, and the phonetic and phonological features of their languages may shift or alter the acoustic characteristics that present their identity. The goal of the current work is to provide a rigorous acoustic account of how the voice varies within and across a bilingual's two languages.

While the source-filter theory of speech production delineates the attributes that voices share (Fant, 1970), voices also vary in unique ways (Lee *et al.*, 2019; Lee and Kreiman, 2022). While variation is indeed wide-ranging, it remains far from random (e.g., Chodroff and Wilson, 2017; Johnson, 2021a; Lee *et al.*, 2019; Lee and Kreiman, 2022). Early work on voice variation focused on how articulatory settings correspond to voice quality (Laver, 1980; Pittam, 1987), while more recent accounts advocate for a psychoacoustically informed model of voice variation (Kreiman *et al.*, 2014). The rationale for this shift is motivated by the lack of a one-to-one mapping from articulation to perception via acoustics.

Voice, as a term, can refer to different aspects of speech, ranging from vocal fold behavior up to the holistic percept of an individual's voice. The latter may be more in line with a layperson's interpretation, which may be relevant for voice perception research. In its narrowest sense, voice has been defined as the behavior of the vocal folds—the glottal source—although Garellek (2019) notes that the acoustic and perceptual consequences of the glottal source cannot be entirely separated from supralaryngeal factors (i.e., the filter). Thus, contemporary models of voice quality account for voice more broadly, also capturing filter behavior via the resonant frequencies of the vocal tract.

As noted, the voice indexes many elements—stance, psychological states, physical characteristics, and identity (Podesva and Callier, 2015). Identity here includes the idea of "linguistic identity," which stems from scholarship summarizing how phonetic settings can vary across languages and dialects (see Mennen *et al.*, 2010; Pittam, 1987; Podesva and Callier, 2015). While acoustic and articulatory dimensions vary for linguistic reasons, the same set of dimensions can also vary for non-linguistic reasons. Speech acoustics, thus, index a multitude of talker traits and

[a)]Electronic mail: molly.babel@ubc.ca

attributes simultaneously. This observation is especially relevant in light of Kreiman and colleagues' argument that the perceptually validated set of dimensions in the psychoacoustic model of voice are more than the sum of their parts (Kreiman *et al.*, 2014; Kreiman *et al.*, 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 an individual's voice.

### A. Structure in voice quality variation

In summarizing the state of art on voice variation, we begin with contemporary scholarship that focuses on voices generally before turning to the bilingual-focused research, which historically examines fewer acoustic dimensions. Lee and colleagues (Lee *et al.*, 2019; Lee and Kreiman, 2020, 2022) use principal component analyses (PCA) on 26 acoustic dimensions to characterize the structure of voice (co)variation. Examining acoustic voice variation in different languages [American English (Lee and Kreiman, 2019, 2022) and Seoul Korean (Lee and Kreiman, 2020)] and speech styles (read and conversational speech), Lee and colleagues leverage the psychoacoustic model of voice quality (Kreiman *et al.*, 2014) and adapt methods from the domains of face variability and perception (Burton *et al.*, 2016).

To outline the structure of voice variability, Lee *et al.* (2019) used a series of PCAs to investigate how acoustic measurements pattern with one another. PCA is a dimensionality reduction technique—that is, it distills a large set of variables into components that reflect covarying bundles of variables. 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 first language (L1)] and sex (here, male or female). Crucially for the comparison with their later work on spontaneous speech (Lee and Kreiman, 2022), Lee *et al.* (2019) focused on sentence reading. Using a large set of talkers ($n = 201$) producing read sentences in English, Lee *et al.* (2019) demonstrated that voices share a basic structure. Shared structure is characterized by the same set of variables covarying and together accounting for comparable amounts of the overall variation in the PCAs. The most commonly shared component in Lee *et al.* (2019) consisted of spectral shape variables in the higher frequencies and spectral noise variables; these components accounted for approximately 20% of the overall variance. These variables are associated with vocal breathiness or brightness, although we add that while Kreiman *et al.* (2021) have perceptually validated the psychoacoustic model of voice quality, the exact correspondence between any of these psychoacoustic features and their percepts is not well understood. 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 the presence of this shared structure, however, Lee *et al.* (2019) argue that the rest of voice structure variation is largely idiosyncratic, although, given that PCA by definition eliminates some amount of idiosyncratic variation, this might be overestimating the presence of idiosyncratic vocal traits.

Lee and Kreiman (2022) replicated their work with short samples of spontaneous speech from the same database using a smaller, but still large subset of voices ($n = 99$). The results were similar, with the exception that fundamental frequency (F0) emerged as a shared relevant dimension. This result arguably reflects differences between read and spontaneous spoken English, with spontaneous speech exhibiting more affective qualities. In spontaneous speech, F0 varies along with the higher source spectral shape and noise parameters. Lee and Kreiman (2020) also replicated the basic tenets of Lee *et al.* (2019) again with sentence reading in Seoul Korean, finding some small differences that are readily explained by typological differences between Korean and English. Unlike in 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 read speech.

### B. Bilingual voice variation

Describing and analyzing acoustic voice variation in bilingual speech has motivation from both perception and production. Listeners are better at identification and discrimination when they have more familiarity with the language at hand, but performance in identification tasks tends to be above chance even for listeners who lack familiarity with the language (e.g., Orena *et al.*, 2019). Listener experience matters substantially less for discrimination (e.g., Perrachione *et al.*, 2019). In cases where listeners cannot rely on linguistic information to parse talker identity, they track 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 how listeners weight and prioritize information in the speech signal, as it delimits the hypothesis space. A focus on bilingual voices crucially allows for the decomposition of what is language-specific and what is individual-specific as the speaker and their vocal tract physiology remain intact across a bilingual's different languages.

Moreover, 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. Related to this is the fact that there are language-specific patterns in facial postures and dynamic deformation in speech. 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. Despite language-specific patterns in face movement during speech, it is still easy to identify a person from their face, regardless of the language being spoken.

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 across their studies, Lee and colleagues argue that both linguistic 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 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 that is held constant across languages is the biological part (i.e., anatomy and physiology). The same "hardware" can be used for drastically different ends. Given the multiple functions of the voice, voice variation across languages may indicate language-specific expression of talkers' social and cultural identities and not just language-specific settings for articulation (Loveday, 1981; Voigt *et al.*, 2016).

Cantonese-English bilinguals' spontaneous speech is the empirical focus of this work. We anticipate differences in voice variation across Cantonese and English due to phonetic and phonological differences between these languages. While all languages have consonants and vowels, they differ in distribution, articulation, and acoustics (e.g., Munson *et al.*, 2010). Cantonese and English differ in their consonant and vowel inventories, in addition to their suprasegmental and prosodic properties (Matthews *et al.*, 2013). A core difference between these languages is that Cantonese is a tone language and English is not. Cantonese has six lexical tones, which are often referred to by numbers 1–6: (1) high level, (2) high rising, (3) mid level, (4) low falling, (5) low rising, and (6) low level. Both segmental and suprasegmental differences in languages have implications for voice quality.

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 a different mean F0 (e.g., Keating and Kuo, 2012), Altenberg and Ferrand (2006) addressed both whether different languages have different F0 baselines and whether F0 shifts when an individual changes languages. They found that Russian-English bilinguals exhibited differences in mean F0 across their two languages, and Cantonese-English bilinguals did not. Speakers did, however, produce a wider F0 range in Cantonese compared to their English. In a larger study of Cantonese-English bilinguals reading passages ($n = 40$), Ng *et al.* (2012) examined a variety of different voice measures with male and female talkers. 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 and 8 kHz) and lower spectral tilt (ratio of energy between 0 and 1 kHz and between 1 and 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 long-term average spectrum (LTAS) measure of the first spectral peak did not differ across languages, suggesting that vocal fold 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 diverges from the nonsignificant difference in Altenberg and Ferrand (2006). The mixed 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. Our focus on Cantonese-English bilinguals provides a small body of literature from which we can anticipate patterns.

We broaden our coverage of voice variation in bilinguals to other language comparisons to understand the ways in which bilingual voices have been shown to vary (or not). Lee and Sidtis (2017) compare F0, speech rate, and intensity in a small group of Mandarin-English bilinguals ($n = 11$) and Korean-English bilinguals ($n = 11$) across three different tasks. They report a higher mean F0 for Mandarin reading and all Korean styles (reading, picture description, and monologue) compared to English. Additionally, there were no differences in F0 variability across languages or tasks for the Mandarin-English bilinguals, but an increase in F0 variability in Korean monologue compared to English monologue. Last, while there were no interesting differences in intensity, the bilinguals spoke faster in Mandarin and Korean. This quantification was based on syllables per second, and, given increased syllable complexity in English compared to Mandarin and Korean, this increased syllables/

s rate in the non-English languages could be a reflection of there being fewer phones in a single syllable. 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, however, may be strongly associated with the type of bilinguals studied, and, to our knowledge, there is no principled reason that a language's having lexical tone would lead to a higher or lower F0 mean. Xue *et al.* (2002) found that Mandarin-English bilinguals produced lower F0 in Mandarin than English. This group differed from the participants in Lee and Sidtis (2017), as they are described as non-native English speakers. Producing higher F0 in a non-native language may reflect non-linguistic factors like stress or confidence (Järvinen *et al.*, 2013; Lee and Sidtis, 2017), although Yang *et al.* (2020) found no differences in F0 profiles across the languages of 12 Cantonese-dominant Cantonese-Mandarin bilinguals. Cheng (2020) finds that Korean has consistently higher F0 than English, regardless of whether speakers were early sequential or simultaneous bilinguals, and that differences in F0 range differ for cisgender males and females. Ryabov *et al.* (2016) looked at rate, duration, and F0 for Russian-English bilinguals and found no F0 differences but found that Russian had a faster speech rate. This result contradicts 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 conflicting results.

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, in this case, the crosslinguistic difference is likely to be sociocultural, 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, then 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 the social from the linguistic, it emphasizes that F0 can index social in addition to linguistic dimensions. 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. One can question whether linguistic and sociocultural factors are disentangleable in the first place.

Between-talker variability should perhaps be given more of a spotlight in this research domain. In work with speech rate, Bradlow *et al.* (2017) found talker and language differences. That is, some talkers are fast and others are slow, and some languages are faster while others are slower. Crucially, speech rate appeared to be a talker trait that stuck with the individual: 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. The work of Bradlow *et al.* (2017) highlights the utility of comparing within individuals and across languages.

### C. The present study

This paper integrates the crosslinguistic voice differences with the structure of acoustic voice variation to provide a more comprehensive picture of how voices vary across languages. With a corpus of Cantonese-English bilingual spontaneous speech (Johnson, 2021b), we describe the behavior of various spectral properties (e.g., Ng *et al.*, 2012) and also examine how acoustic variation is structured, following closely on the work of Lee *et al.* (2019) and Lee and Kreiman (2022). We build upon the methods of Lee and colleagues by extending the methods to bilingual speakers' speech in two languages, using longer samples, and assessing how large of a sample is necessary to characterize voice variation within and across talkers and languages. We also introduce a method to quantify structural similarity within and between individuals and languages.

## II. THE DATA

### A. The SpiCE corpus

The data used in this analysis come from the conversational interviews in the SpiCE corpus (Johnson, 2021b). The analysis uses the Cantonese and English interviews from the 34 early Cantonese-English bilinguals (17 self-identified female, 17 self-identified male) in the corpus. Participants ranged from 19 to 34 years of age. Interviews were conducted in English and Cantonese with a 24 year old Cantonese-English bilingual female. Interviews in each language were approximately 25 min and were in a casual interview format designed to elicit continuous speech from the interviewee. The language order was counterbalanced, and each interview was preceded by a sentence reading task and a short story narration in the target language for that part of the session.

The audio from these interviews is high-quality, with a 44.1 kHz sampling rate, 16-bit resolution, and minimal background noise. The analyses reported here used the channel that recorded the participants' speech. Code-switches are included in the analysis, as code-switches are representative of the particular talker's language behavior in a given language. Additionally, code-switching does not necessarily imply a categorical shift in language modes, and switches may be pronounced with matrix language phonology—that is, the base language of the sentence (e.g., Fricke *et al.*, 2016; 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 Hz 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. This method of identifying voiced portions of the speech signal captures vowels, approximants, and some voiced obstruents. Because /n/ and /l/ vary in their mid-frequency spectral properties (Garellek *et al.*, 2016) and Cantonese and English likely do not have comparable counts of these sounds (Cheng *et al.*, 2022; Soo *et al.*, 2021), force-aligned (McAuliffe *et al.*, 2017) intervals with /n/ and /l/ labels were removed after processing in VoiceSauce (Shue *et al.*, 2011).

## B. Acoustic measurements

All voiced segments, with the exception of intervals identified in a forced aligned transcription as containing /n/ or /l/, 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 [following the exclusionary criteria in Sec. II C: Pearson's $r = 0.94$, degrees of freedom (df) = 2 917 150, $p < 0.001$]. The choice of measurements in Lee *et al.* (2019) is based on Kreiman *et al.* (2014) psychoacoustic voice quality model, as well as the availability of algorithms in the software used to extract measurements. Measurements were output every 5 ms from voiced segments in VoiceSauce (Shue *et al.*, 2011).

The measurements made in this analysis are described below. Each measurement is bolded and followed by a short description. **F0** is associated with linguistic (e.g., lexical tone), prosodic, and talker characteristics and was measured in Hz using the STRAIGHT algorithm (Kawahara *et al.*, 2016). **F1, F2,** and **F3** contribute to linguistic contrasts—particularly for vowels and sonorant consonants—and 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**, measured in Hz and estimated along with F1–F3, is associated with talker characteristics, such as vocal tract length. **H1*–H2***, measured in dB, is the corrected (Iseli *et al.*, 2007) amplitude difference between the first two harmonics, characterizes source spectral shape, and is typically associated with phonation type, but can also be confounded by nasality (Chai and Garellek, 2022; Munson and Babel, 2019; Simpson, 2012). **H2*–H4*** is the corrected amplitude difference between the second and fourth harmonics and is the second of four measures capturing spectral shape—it is associated with phonation type and is measured in dB. If P0 aligns with H2, this measure, like H1*–H2*, may also be confounded with nasality (Chai and Garellek, 2022; Simpson, 2012). **H4*–H2kHz***, measured in dB, is the corrected amplitude difference between the fourth harmonic and the harmonic closest to 2000 Hz, is a third spectral shape measure and captures shape in a higher frequency range. **H2kHz*–H5kHz** is also measured in dB, and is the amplitude difference between the harmonics closest to 2000 Hz (corrected) and 5000 Hz (uncorrected). **CPP** is cepstral peak prominence and is measured in dB. CPP corresponds to the degree of harmonic regularity in voicing and is associated with non-modal phonation types. VoiceSauce computes CPP according to the algorithm in Hillenbrand *et al.* (1994), measuring 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.[1] **Energy**—root mean square (RMS) energy—is a measure of spectral noise in dB that reflects overall amplitude and is calculated over a window comprising five pitch periods. Energy is a perceptual correlate of volume or loudness. **SHR** is a (unitless) subharmonics-harmonics amplitude ratio; it is a measure of spectral noise associated with period-doubling or irregularities in phonation. VoiceSauce's implementation is based on the algorithm described in Sun (2002). All analyses and code for this project are available on the Open Science Framework (OSF) at https://osf.io/ybdkw/.

## C. Exclusionary criteria and post-processing

To eliminate impossible values, which are assumed to result from measurement error, 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 (s.d.s) 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 non-modal phonation, a zero for CPP likely reflects either a lack of voicing or an erroneous F0 measurement. Last, 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 or uncorrected. In aggregate, these filtering criteria led to the removal of 37% of the original set of observations. Both Energy and SHR were highly skewed in their distributions. Energy was log-scaled to address the skew, but log-transforming SHR did not attenuate the non-normalcy of the distribution and so was not logged. SHR is a ratio, where zero meaningfully indicates no subharmonics, which likely indicates breathier voice qualities. Because the abundance of 0 values is an issue for the PCA, values of zero were adjusted to 0.0001. This is ten times lower than the lowest reported SHR value from VoiceSauce (0.001).

Next, moving s.d.s were calculated for each of the 12 measures using a centered 50 ms window, such that each window includes approximately ten observations. The moving s.d.s 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 s.d.s, 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 Sec. III, and as a result, there should not be any undue effect on the outcome as the transformation from s.d. to coefficient of variation is a scaling transformation. The last round of exclusionary criteria uses these moving s.d.s. If an observation was missing a moving s.d. value, it was removed. Given the centered window, this means that observations falling less than 25 ms from a voicing boundary were not included. There were 24 total measures, with a measured value and a moving s.d. for each of the acoustic measurements listed above. These 24 measures were used in the analyses described in Sec. III. Across the 34 talkers, there were 2 917 152 observations after winnowing the data from an initial count of 6 387 510 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 Sec. III A and is used there, sample size may have an impact on the PCA-based analyses in Secs. III B and III C. 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 20 151 samples for each interview. This value was selected as it represents the interview with the fewest observations. 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. Thus, the sample size was controlled to better equate across talkers.

Following this last winnowing step, there were 1 370 268 total observations (34 talkers × 2 interviews × 20 151 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 5000 used in Lee *et al.* (2019).

## III. ANALYSIS AND RESULTS

### A. Analysis 1: Crosslinguistic comparison of acoustic measurements

#### *1. Methods and results*

We first present a crosslinguistic comparison for each talker and measure. Figure 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 comparisons have no bearing on how a given variable *varies*. Table I reports counts of talkers by effect size. Notably, across all talkers and variables, only 20.8% yielded non-trivial Cohen's *d* values, although all talkers had at least one non-trivial comparison. The distribution of these non-trivial counts by talker is depicted in Fig. 2. Additionally, Figs. 3 and 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.

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

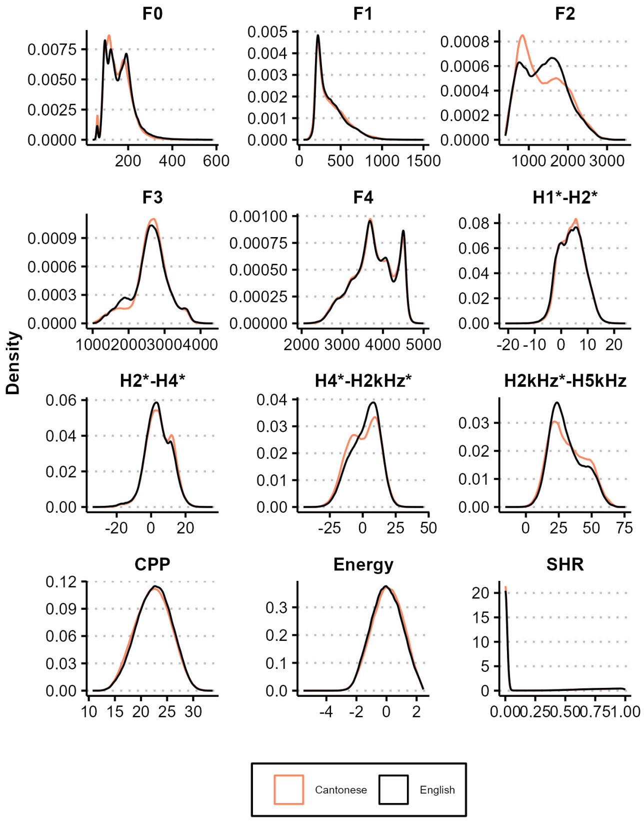

FIG. 1. (Color online) Each panel depicts a density plot that pools measurements from all talkers together to show the range of values for that measure. The *x* axes each have their own scale. Language is separated out by color.

TABLE I. 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<br>*0.0–0.2* | Small<br>*0.2–0.5* | Medium<br>*0.5–0.8* | Large<br>*>0.8* |
| F0 | 22 | 9 | 3 | — |
| F0 s.d. | 34 | — | — | — |
| F1 | 22 | 10 | 2 | — |
| F1 s.d. | 30 | 4 | — | — |
| F2 | 26 | 8 | — | — |
| F2 s.d. | 32 | 2 | — | — |
| F3 | 25 | 8 | 1 | — |
| F3 s.d. | 27 | 7 | — | — |
| F4 | 30 | 4 | — | — |
| F4 s.d. | 27 | 7 | — | — |
| H1*–H2* | 19 | 14 | 1 | — |
| H1*–H2* s.d. | 32 | 2 | — | — |
| H2*–H4* | 24 | 10 | — | — |
| H2*–H4* s.d. | 31 | 3 | — | — |
| H4*–H2K* | 25 | 8 | 1 | — |
| H4*–H2K* s.d. | 32 | 2 | — | — |
| H2K*–H5K | 23 | 10 | 1 | — |
| H2K*–H5K s.d. | 31 | 3 | — | — |
| CPP | 22 | 10 | 2 | — |
| CPP s.d. | 32 | 2 | — | — |
| Energy | 15 | 13 | 5 | 1 |
| Energy s.d. | 24 | 10 | — | — |
| SHR | 31 | 3 | — | — |
| SHR s.d. | 30 | 4 | — | — |

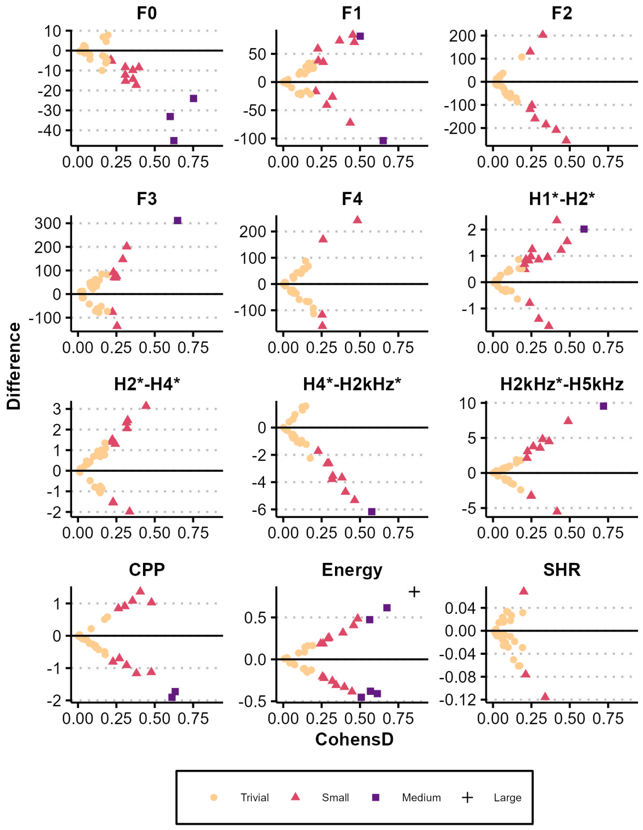

FIG. 3. (Color online) Each panel plots Cohen's *d* on the *x* axis (scales differ) and the difference between language means on the *y* axis. Positive values indicate a higher mean in Cantonese than English. The color reflects the levels of interpretation for Cohen's *d*. Each point represents a talker.

languages, then Cantonese had a lower mean F0 than English (12 of 34; female = 7), although most talkers did not exhibit a difference (22 of 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 (Altenberg and Ferrand, 2006; Ng *et al.*, 2012). This difference occurs at roughly similar rates for female and male talkers.

As for the two spectral shape measures with consistent patterns, H4*–H2kHz* was consistently lower in Cantonese when the comparison was not trivial ($n = 9$), although 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 (12 of 34), lower for a small number (3 of 34), but trivial for most (19 of 34). While based on different measures than Ng *et al.* (2012), the H1*–H2* results are consistent with previous findings that Cantonese tends to be breathier (or English creakier). However, other interpretations are possible: Cantonese could be breathier and English more modal. Cantonese can be more modal and English creakier. Cantonese could also be breathier and English creakier. One way to better understand these values is to consider H1*–H2* alongside CPP (Seyfarth and Garellek, 2018), but given that at the group level, the means ($M$) and ranges for CPP (Cantonese: $M = 22.26$, range $= 18.75, 25.08$; English $M = 22.43$, range $= 18.86, 25.42$) and H1*–H2* (Cantonese: $M = 4.07$, range $= –0.291, 10.28$; English $M = 3.73$, range $= –0.56, 10.2$) are so similar, the relationship between these measures at this coarse level of analysis does not offer a fruitful path to interpretation. 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.

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

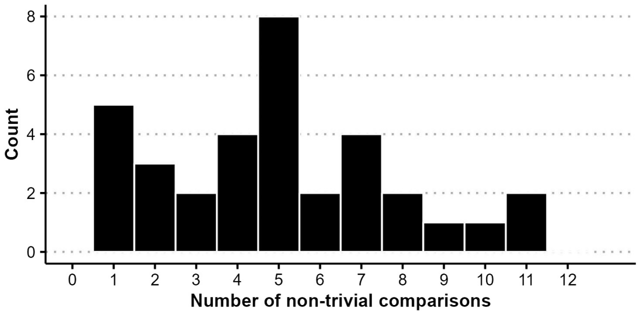

FIG. 2. A histogram summary of the number of non-trivial comparisons from Table I across the 34 talkers.

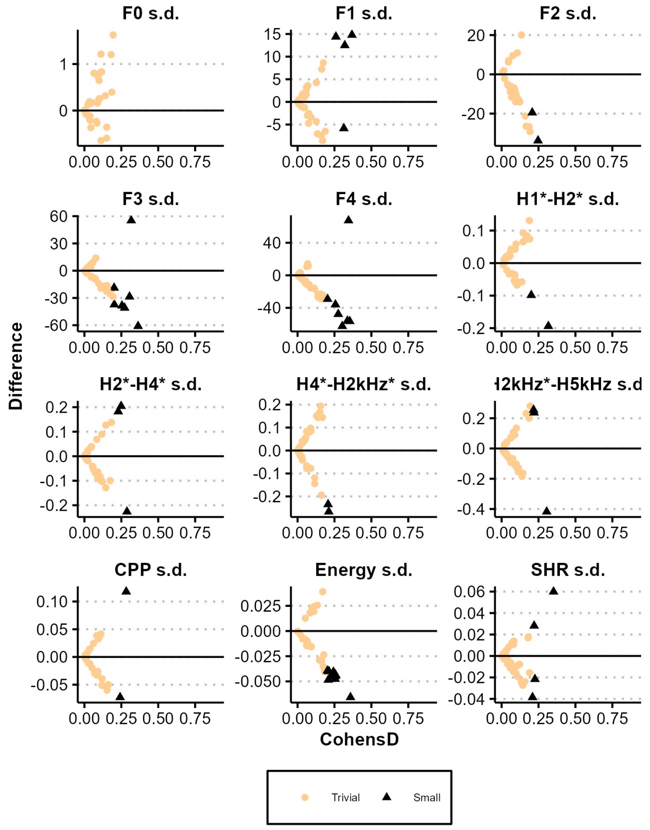

FIG. 4. (Color online) This figure uses the format of Fig. 3 but reports on the s.d. measures.

unlikely to vary across languages within the same talker given its association with vocal tract size and has relatively low counts of non-trivial differences across language.[2]

F2 exhibits variable behavior, showing the largest visible differences in Fig. 1. This visible group-level difference masks the fact that underlyingly, only some individuals show F2 differences by language. As shown in the Cohen's *d* results in Table I, eight individuals have non-trivial cross-language comparisons: Two were positive, and six were negative. 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 suggests 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 = 9, 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 (five talkers) and negative (seven talkers). 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 straightforward, however, as it is not immediately clear which type of non-modal phonation the measure entails. Given the H1*–H2* results, it suggests 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 [for example, see Seyfarth and Garellek (2018)]. In any case, listener assessments would ultimately help pinpoint how spectral shape and noise parameters map onto perceived voice quality.

### *2. Interim discussion*

Overall, while talkers show some clear across-language differences with crosstalker tendencies for lower mean F0 in Cantonese than English and phonation quality differences between the languages, these are far outnumbered by instances with no consistent differences or trivial differences. Together, this offers the initial conclusion that the acoustic variation in voice quality in Cantonese-English bilinguals is both subtle and individual in nature, potentially cueing social identity differences across languages, and not language-specific articulatory patterns. Alternatively, it may also indicate that there is not a veridical voice quality difference between Cantonese and English.

## B. Analysis 2: PCA

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, 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 cohere together and ultimately form a percept. While PCA does not shed light on perception, the signal-based account can be used to generate predictions about listener perception of voices.

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 here is on the internal structure of the components. That is, we delve into what makes up components for different talkers and whether an individual's voice structure varies (or not) across languages.

### *1. Methods*

We adapt methods from work on voices (Lee *et al.*, 2019; Lee and Kreiman, 2020) and faces (Burton *et al.*, 2016; Turk and Pentland, 1991). 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 Sec. III A 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; Kreiman *et al.*, 2021).

There are many different methods for setting the number of PCA components, 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.

### *2. Results and discussion*

The results and their discussion are presented jointly to facilitate understanding of the output of the PCAs.

The PCAs across both languages for all 34 talkers resulted in 10–14 components and accounted for 73.85%–81.95% of the total variation. Half of the talkers had the same number of components for each language (17 of 34), 16 talkers had a difference of one in the number components, and only one talker had a difference of two. Talkers had 3–10 identical component configurations across their languages ($M$ = 7.6)—that is, the same variables loaded on the components above the fair threshold (although loading values varied). These shared components represent 26.1%–83.3% of the total components for talkers ($M$ = 64.3%). The numbers comprising these summary statistics are provided in Table II. 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 and a component with just F2 and F3 are identified as different components in the crosslanguage comparison. This similarity will be taken into account in the next part of the analysis in Sec. III C.

TABLE II. 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 | 12 | 0.79 | 12 | 0.78 | 7 |
| VF19B | 12 | 0.79 | 12 | 0.79 | 8 |
| VF19C | 12 | 0.79 | 11 | 0.75 | 7 |
| VF19D | 12 | 0.77 | 12 | 0.77 | 10 |
| VF20A | 11 | 0.77 | 11 | 0.78 | 5 |
| VF20B | 13 | 0.81 | 12 | 0.80 | 8 |
| VF21A | 11 | 0.76 | 12 | 0.78 | 6 |
| VF21B | 12 | 0.78 | 12 | 0.78 | 6 |
| VF21C | 14 | 0.82 | 13 | 0.81 | 10 |
| VF21D | 12 | 0.79 | 12 | 0.79 | 10 |
| VF22A | 11 | 0.78 | 11 | 0.78 | 8 |
| VF23B | 12 | 0.79 | 12 | 0.79 | 9 |
| VF23C | 12 | 0.79 | 12 | 0.79 | 9 |
| VF26A | 11 | 0.74 | 12 | 0.76 | 6 |
| VF27A | 11 | 0.77 | 10 | 0.74 | 7 |
| VF32A | 12 | 0.78 | 12 | 0.77 | 8 |
| VF33B | 12 | 0.78 | 13 | 0.81 | 10 |
| VM19A | 11 | 0.76 | 12 | 0.79 | 9 |
| VM19B | 11 | 0.79 | 12 | 0.79 | 3 |
| VM19C | 11 | 0.77 | 11 | 0.76 | 7 |
| VM19D | 12 | 0.77 | 14 | 0.81 | 8 |
| VM20B | 12 | 0.80 | 11 | 0.76 | 7 |
| VM21A | 11 | 0.80 | 11 | 0.79 | 7 |
| VM21B | 11 | 0.78 | 11 | 0.78 | 8 |
| VM21C | 12 | 0.78 | 13 | 0.80 | 9 |
| VM21D | 12 | 0.78 | 11 | 0.75 | 6 |
| VM21E | 11 | 0.78 | 12 | 0.81 | 7 |
| VM22A | 12 | 0.77 | 12 | 0.79 | 9 |
| VM22B | 12 | 0.79 | 12 | 0.79 | 8 |
| VM23A | 12 | 0.80 | 12 | 0.78 | 7 |
| VM24A | 12 | 0.79 | 11 | 0.76 | 6 |
| VM25A | 12 | 0.78 | 12 | 0.78 | 9 |
| VM25B | 11 | 0.76 | 12 | 0.78 | 5 |
| VM34A | 11 | 0.78 | 12 | 0.80 | 9 |

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' 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 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 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.

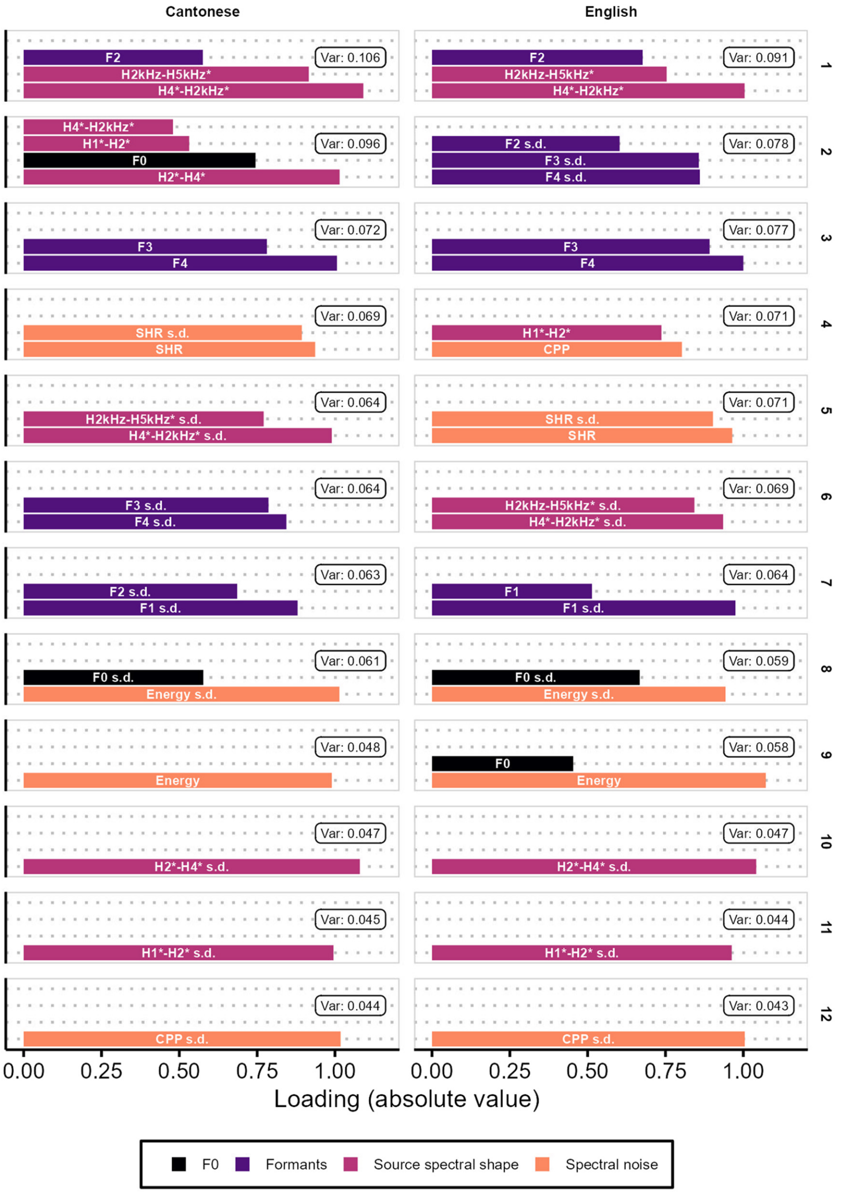

FIG. 5. (Color online) Components of the Cantonese and English PCAs for VF32A—a single talker from the corpus taken as an example. Loadings are represented by bar height and are labeled with the variable name; color represents conceptual groupings. The component's variance accounted for is superimposed.

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 4 in Cantonese and component 5 in English. Both components comprise SHR s.d. and SHR exclusively and account for 6.9% and 7.1% 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 Cantonese component 6 (F3 s.d. and F4 s.d.) consists of a subset of the variables in her English component 2 (F2 s.d., F4 s.d., and F3 s.d.). Last, sometimes variables just pattern differently—in Cantonese, F0 patterns with H4*–H2kHz*, H1*–H2*, and H2*–H4* in component 2, while in English, F0 patterns in component 9 with Energy. While an in-depth analysis of each component of each PCA is beyond the scope and goals of this paper, examining VF32A's components in this way highlights the importance of not attributing too much value to the ordering of 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. We summarize the components that were present in every talker's PCA and describe the composition of others that occurred frequently or in notable combinations. There were 140 unique components across all voices and languages, relatively evenly split across Cantonese ($n = 73$) and English ($n = 67$).

The most commonly shared component accounting for the most variation across talkers had a structure of H2kHz*–H5kHz s.d. and H4*–H2kHz* s.d. All 34 talkers had these measures patterning together in both Cantonese and English, although it did not account for a particularly large proportion of variance. All talkers also had CPP s.d. as a single component in English, and most did in Cantonese as well (n = 32). SHR and SHR s.d. went together for the vast majority of talkers (Cantonese = 31, English = 31). Similarly, like the most common component involving H2kHz*–H5kHz s.d. and H4*–H2kHz* s.d., these were composed of CPP s.d., SHR, and SHR s.d. and accounted for relatively small amounts of variance. However, it is of note that these most common component structures are formed by variables associated with phonation type and source qualities.

Another very frequent component included phonation variables and filter qualities: F2, H2K*–H5K, and H4*–2K* loaded on a component for most talkers in both languages (Cantonese = 32, English = 28). This component account accounted for a relatively high proportion of variance. While a concise summarization of what this component means is challenging, it includes both higher spectral shape parameters and up to three formants (F2, F3, H2K*–H5K, and H4*–2K* for English = 1; F2, F3, F4, H2K*–H5K, and H4*–2K* for English = 5 and Cantonese = 2). These variables are typically associated with phonation type from mid-frequency measures and vowel quality (or other aspects of the filter), respectively. 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)]. The higher formants F3 and F4 also patterned together for most talkers (Cantonese = 27, English = 26).

H2*–H4* s.d. most commonly occurred alone (Cantonese = 17, English = 17) or in combination with H1*–H2* s.d. (Cantonese = 10, English = 17). H2*–H4* s.d. also occurred along with H1*–H2* s.d. and CPP (Cantonese = 2), F1 s.d. (Cantonese = 1), or F1 (Cantonese = 1). H2*–H4* was also in a component with F1 (Cantonese = 13, English = 3). These components, F1 withstanding, reflect variability in non-modal phonation quality and the timbre of the voice—often described as brightness in Lee *et al.* (2019).

Formant s.d. parameters often co-occurred. In both languages, this component typically consisted of F2 s.d., F3 s.d., and F4 s.d. (Cantonese = 17, English = 19), although many cases excluded F2 s.d. and only contained F3 s.d. and F4 s.d. (Cantonese = 10, English = 13). 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 s.d.s 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.

CPP co-occurs with source components, notably with F0, F0 s.d., Energy, Energy s.d., SHR, and most of the harmonic-based measures of phonation quality. CPP s.d., however, only occurs with itself, and does so for nearly all of the talkers in Cantonese and all in English (Cantonese = 32, English = 34). CPP s.d. co-occurs with other components in only two other instances for two talkers. These patterns 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 and other source components 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, SHR by itself for a single talker per language and SHR s.d., for a single talker in Cantonese. The pair was sometimes accompanied by H1*–H2* (Cantonese = 2, English = 1) or F0 (English = 1). SHR s.d. was present in components with H1*–H2* s.d. and CPP (Cantonese = 1) and F0, H1*–H2*, and H2*–H4* (English = 1). SHR is associated with period-doubling and irregularities in phonation. SHR and SHR s.d. co-occurring so often (and so rarely with other variables) suggests that SHR and its variability together form a meaningful dimension in voice quality.

While this covers many of the variables that went into the PCAs, F0 and F0 s.d. are notably sparse in the above paragraphs. F0 s.d. was fairly consistent in emerging with Energy s.d. (Cantonese = 17, English = 18) and in a small number of other component configurations that included CPP, H1*–H2* s.d., and F1 s.d. F0 did not occur frequently on its own (Cantonese = 4, English = 4) and was more often accompanied with Energy (Cantonese = 13, English 7). Beyond the combination of F0 and Energy, F0 appeared in 21 different component configurations, but none of these occurred more than five times. Across these different components, F0 was accompanied by all kinds of variables: F0 s.d., H1*–H2*, H1*–H2* s.d., H2*–H4*, F1, CPP, Energy, Energy s.d., SHR, and SHR s.d. Some of these combinations only occurred once. The lack of consistency in F0 across talkers is notable for a few reasons. First, in Lee and colleagues' work, F0 emerged as an important feature of acoustic voice variation structure in English spontaneous speech (Lee and Kreiman, 2022) 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). Second, F0 plays a major role in prior work on voice production and perception, given its salience as an acoustic dimension (Perrachione *et al.*, 2019). While neither F0 nor F0 s.d. featured dominantly in their own component, 26% of all components included F0 or F0 s.d., and it may be these varied combinations of F0 with other variables (e.g., F0 and H2*–H4* for four English voices, F0 and H1*–H2* for five English voices, F0 s.d. and CPP for four Cantonese voices) that provide listeners with a unique talker signature to latch on to in voice perception.

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. 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 on OSF.

In summary, this PCA analysis found a greater amount of component structure overlap than was reported in similar voice analyses (Lee *et al.*, 2019; Lee and Kreiman, 2022). 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 18.1% to 26.2%, the unaccounted for variability may also be idiosyncratic in nature.

## C. Analysis 3: Canonical redundancy analysis

While interpretation of the PCAs allows for the characterization of acoustic voice structure, it does not provide a quantitative comparison of voices. The goal of the next analysis 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* (CCA), which provides a metric of redundancy (i.e., overlap) between the two PCAs compared, resulting in a metric that is easy to interpret.

### *1. Methods*

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 CCA (Jolliffe, 2002; Stewart and Love, 1968). 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 loading 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.

### *2. Results*

Redundancy indices for within-talker comparisons ranged from 0.79 to 0.97 [median ($Mdn$) = 0.91, $M$ = 0.91, s.d. = 0.04] and are displayed in Fig. 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 0.95 ($Mdn$ = 0.81, $M$ = 0.81, s.d. = 0.5). Comparisons across both talkers and languages ranged from 0.65 to 0.95 ($Mdn$ = 0.81, $M$ = 0.81, s.d. = 0.5).

Within-talker values were confirmed to be higher than across-talker comparisons, per a Welch's $t$-test [$t$(70.352) = –18.68, $p < 0.001$, $d$ = 1.8]—this result indicates that regardless of language, talkers are more similar to themselves than talkers are to each other.

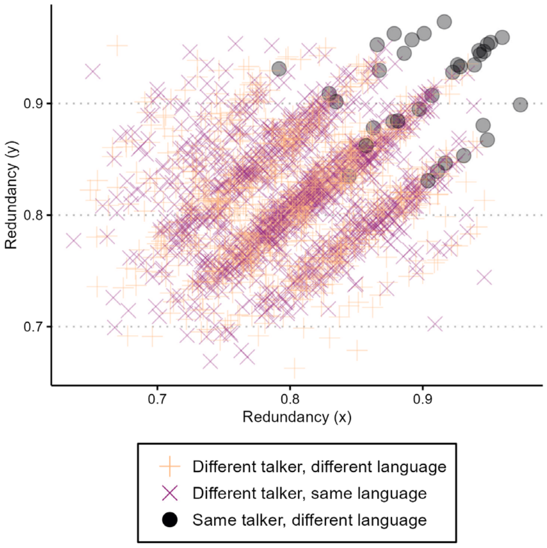

FIG. 6. (Color online) This plot depicts the relationship between the two redundancy indices for three different types of comparisons. Across-talker comparisons represented by orange "+" (different language) and pink "×" (same language) overlap in their entirety. Within-talker comparisons are represented by the black circles and are clearly clustered at the top right.

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 [$t(4484) = -0.98$, $p = 0.33$, $d = 0.03$]. 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 Sec. III B, there were a substantial number of commonly occurring patterns across talkers and languages. Together, this supports the conceptualization of a voice space comprising a shared structure—as in the case of the prototype account (Kreiman and Sidtis, 2011; Lavner *et al.*, 2001; Lee *et al.*, 2019; Lee and Kreiman, 2022)—where voices can only deviate from one another so much.

### D. Analysis 4: Passage length analysis

There are two goals in the passage length analysis. The first is to determine how much of a voice sample is necessary to identify a stable description of voice variability for a talker. The second is a *post hoc* confirmation that the choice of samples in Sec. III B is sufficient.

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, 12 500, 18 000, 24 500, 32 000, 40 500, 50 000, and 60 500 observations. Each PCA 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 20 151 to 68 312, there were 6–11 total PCAs (and, thus, comparisons) per interview, depending on its maximum possible passage length. While these sample 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 ten, the lowest number found in Sec. III B. 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 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 5000 represents the average sample size from Lee *et al.* (2019). The vertical line at 20 151 represents the sample size used in Sec. III B. 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.

It is apparent from this visualization that the sample size used for PCAs in Sec. III B 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 Sec. III B 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) and Lee and Kreiman (2022) sufficiently captured the range and structure of talker variability. As a reviewer points out, however, the elbow in Fig. 7 starts around the 5000 sample mark. This may denote 5000 samples as a minimum threshold. As redundancy approaches 1, it is unclear where meaningful differences in characterizing the voice space exist in these values.

## IV. GENERAL DISCUSSION

How does a bilingual's voice vary across their two languages? To answer this question, we explored spectral properties and structural similarities in Cantonese-English bilinguals' spontaneous speech. The analyses cover three different exploratory approaches to the question of understanding crosslinguistic (dis)similarity in bilingual voices.

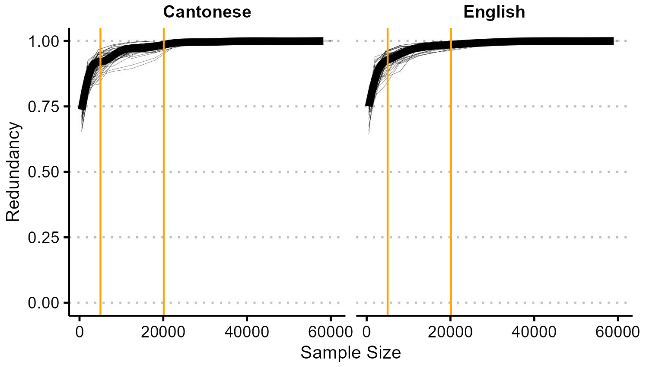

FIG. 7. (Color online) Passage length redundancy indices are plotted against the sample size of the smaller PCA. Smoothed curves show a rapid increase in redundancy followed by a leveling off between the vertical orange lines, which represent the sample sizes used in prior work ($x = 5000$) and the present study ($x = 20\,151$).

The first analysis is a coarse approach, comparing overall distributions using 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 s.d.s to describe how voices differ (or not). This analysis method suggests that *some* voices vary across languages and that there are acoustic dimensions that are more likely to differ across Cantonese and English within a talker. F0, in particular, was relatively consistent; all non-trivial Cohen's *d* values were in the direction of Cantonese having lower F0 than English. Several acoustic measurements intended to capture phonation quality were different across languages but in inconsistent directions across Cantonese and English. These diverging patterns suggest that the locus of the acoustic variation is likely not due to the languages themselves but may be related to the social identities individuals adopt in their two languages, structured by social factors not considered in our approach.

The second analysis builds upon Lee and colleagues' foundational work (Lee *et al.*, 2019; Lee and Kreiman, 2022) on voice variation using PCAs, extending it to the case of bilingual speech. We find more component structure overlap than their work; however, there was ample talker-specific idiosyncratic variation in the PCAs (between 18% and 26% of the variance). We introduce canonical redundancy as a metric for objectively assessing crosslinguistic similarity from the output of two PCAs. These methods are then extended to determine how much speech is needed to characterize an individual's voice and validate our methodological choices. In future research, we will validate whether these canonical redundancy measures correlate with listeners' perceptual organization of the voices within and across languages.

The results suggest that Cantonese-English bilinguals exhibit similar spectral properties and similar lower-dimensional structure in their acoustic voice variation in their two languages. This similarity is most apparent on a within-talker basis but still present across talkers and languages, despite substantial segmental and suprasegmental differences between English and Cantonese (Matthews *et al.*, 2013). It may be the case that some language combinations or bilingual individuals will exhibit more dissimilarity than observed here. On the other hand, the strong theme of within-talker voice similarity may not be surprising as any bilingual individual is still constrained to using the same vocal anatomy to articulate their oral languages. Bilinguals appear to have the same "voice" in each of the two languages, supporting 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 (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.

The language familiarity effect (LFE) in voice perception warrants our attention, as its existence suggests not all listeners can equivalently exploit the available acoustic information that signals talker identity. The LFE is a term that characterizes the observation that listeners who know the language being uttered have a performance advantage in voice line-ups (Goggin *et al.*, 1991; Hollien *et al.*, 1982; Johnson *et al.*, 2011; Thompson, 1987) and voice identification tasks (Bregman and Creel, 2014; Nygaard and Pisoni, 1998; Perrachione and Wong, 2007). The LFE appears to be gradient, increasing in its strength with increased proficiency in the target language (Bregman and Creel, 2014; Orena *et al.*, 2015; Xie and Myers, 2015). Phonological knowledge about a language appears to be a key to the LFE (Johnson *et al.*, 2011; Perrachione *et al.*, 2015; Perrachione *et al.*, 2019), although note that this phonological knowledge need not be particularly sophisticated because 7–8 month old infants have an advantage (Johnson *et al.*, 2011), as do anglophones in Montreal, Quebec when tested in French, who regularly *hear* French, although they are not competent in it (Orena *et al.*, 2019). If bilingual voices share so much low-dimensional voice structure, as we suggest, why cannot listeners globally take advantage of this information? We offer two possible explanations. One possibility is that the linguistic information essentially distracts listeners from the lower-dimensional structure. The second is simply that learning a talker's vocal identity is about learning *how* their voice can vary. It may be that listeners need sufficient exposure to a voice—more than what is feasibly granted in an experimental setting—to delimit a voice's range of variation. Indeed, being highly familiar does appear to confer a voice particular privileges (e.g., being selectively ignored or attended; Johnsrude *et al.*, 2013). Note too that the LFE appears stronger in tasks that require "telling voices together" compared to those that require "telling voices apart," which are conceptualizations of, broadly, talker categorization and discrimination tasks, respectively. It seems as though language experience matters substantially less in discrimination tasks, where listeners may be able to take advantage of low-dimensional voice structure to cue talker differences (Lavan *et al.*, 2019a; Park *et al.*, 2018; Perrachione *et al.*, 2019).

Returning to the results at hand, our findings from the first two analyses reflect prior research. 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 a different phonation quality than English (Ng *et al.*, 2012; Ng *et al.*, 2010). Previous work described Cantonese voice quality as breathier, but it may be the case that Cantonese is better described as more modal and English creakier. 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, potentially due to the larger sample sizes used in our analysis, these component structures also emerged for the Cantonese-English bilinguals. 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 and prior work (Lee *et al.*, 2019; Lee and Kreiman, 2020) supports the prototype model in voice (Kreiman and Sidtis, 2011; Latinus and Belin, 2011; Lavner *et al.*, 2001). Within this model, a prototype is typically thought of as a speech community average, although, as suggested by Lee and colleagues, the prototype could be the shared voice structure and not an average, although there is evidence that an average is used by listeners (Lavan *et al.*, 2019b). That there are similarities across disparate populations and languages (e.g., Lee and Kreiman, 2020) suggests that prototypes may extend beyond tightly defined speech communities.

The PCA analysis adds additional commonly occurring components to the mix, suggesting that there 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 F0 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 what 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 additional set of shared components serves to flesh out the structure of what a prototypical voice might look like.

This high degree of similarity does not preclude cross-linguistic differences on a within-talker basis but rather suggests that such differences occur on a more global level. This is apparent in Fig. 8, which depicts the relationship between within-talker, across-language redundancy (averaged) from Sec. III C and the difference between the mean values for each of the acoustic measurements in Sec. III A. 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 between 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 how voices vary.

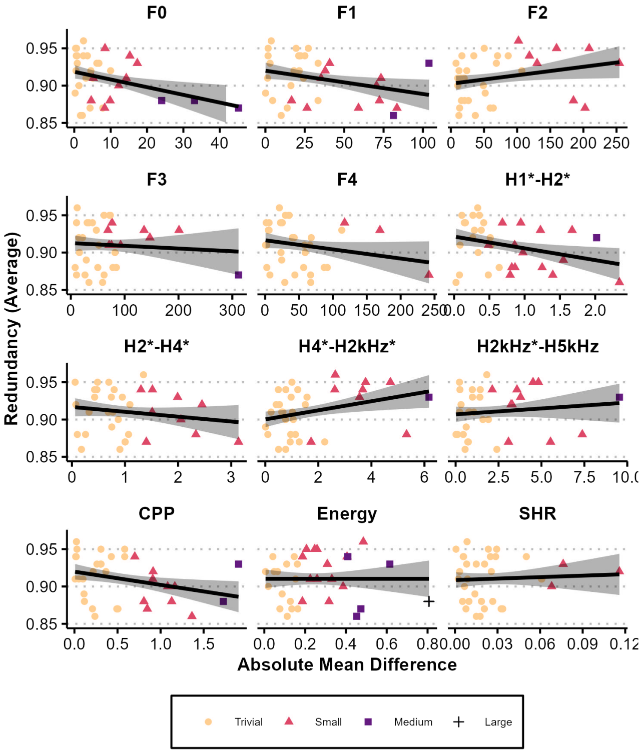

FIG. 8. (Color online) The average redundancy value for each talker is plotted against the absolute value of the difference of means across languages for that talker. Color and shape indicate the size of Cohen's *d*. The superimposed regression line summarizes the relationship between these values.

Such high similarity in the PCAs was not entirely expected, given the results of Lee *et al.* (2019) and Lee and Kreiman (2022), where a handful of shared components were evident but were complemented by numerous idiosyncratic components. 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 crucially ensures that comparisons are made among like items. Any downside to this methodological decision is mitigated by the fact that most components made relatively small contributions in how much of the overall variance they accounted for (see Table II).

While methodological choices may account for some part of these results, the data subject to the analysis are also relevant. Simply, more data are used in our analysis, and larger speech samples allow for a more 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 Sec. III D shows that the number of samples needed for full stabilization is substantially larger than the 5000 samples used in prior work. This does not necessarily discount Lee and colleagues' work, however, as our use of spontaneous interview speech, as opposed to Lee and colleagues' read speech (Lee *et al.*, 2019) and telephone speech (Lee and Kreiman, 2022), is likely more variable. Lee and Kreiman (2022) examined spontaneous speech collected from short phone calls that lasted minimally 2 min. By this estimation, the sample size was likely on the lower side, compared to the 20–25 min interviews in the SpiCE corpus. However, it is not possible to make a direct comparison without knowing the number of samples. Moreover, the elbow in the curve is around 5000 samples, which is an indication that such a value may be a minimum threshold that is, indeed, sufficient to accurately acoustically characterize a voice within a given speech style. The methods presented here offer a tool for researchers to assess whether the quantity of speech is representative of an individual.

The time sample needed to provide a stable acoustic characterization of a voice may not be what a listener needs to make a reliable judgment about a voice. Listeners can make quick and accurate judgments about social identity from a single syllable (e.g., Purnell *et al.*, 1999). Developing an understanding of how listeners categorize and organize the rich social and linguistic signals in human voices is clearly a work-in-progress. This work contributes to this effort. Ultimately, an empirically and theoretically grounded understanding of the acoustic structure of voices has relevance for one's dynamic identity construction and how voices map onto listeners' organization of a voice space for use in talker recognition and discrimination. This latter connection is our primary interest. Turning to listener and behavioral data will help decipher what is meaningful variation within a voice from the low-level noise that cannot be attributed to a particular vocal signature. Verification from listener performance will help adjudicate which analytical choices present an acoustic voice space that matches listener organization. For example, it may be the case that a smaller sample of a voice is sufficient for the less challenging task of "telling apart." Short samples may illustrate that two voices vary enough along some acoustic-auditory dimensions that they likely belong to different individuals. Listeners may need experience with substantially longer samples to accurately do the more challenging "telling together" task.

While the specifics of the results differ, our conclusions align with Lee *et al.* (2019) and Lee and Kreiman (2022), who posit that the structure of voice spaces supports a prototype model of voice perception (Latinus and Belin, 2011; Latinus *et al.*, 2013; Lavner *et al.*, 2001) in which novel individual voices are perceived in the context of one or more prototypes housed in listeners' memory. Lavner *et al.* (2001) 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" (p. 64). 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). 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. Recently, Lee *et al.* (2019) and Lee and Kreiman (2022) suggest that the common structure they (and we) identify across voices suggests an alternative understanding of the voice prototype. The prototype might not be an average voice, but rather a distillation of the voice space that is shared across voices. That is, the shared acoustic structure identified here and in Lee *et al.* (2019) and Lee and Kreiman (2022) may serve as the prototype. Considering the prototype in this way is a rather stark turn from some previous work. Latinus *et al.* (2013), for example, hypothesize that voices are organized around F0, formant dispersion, and harmonics-to-noise ratio (HNR) on a gender-specific basis. And, indeed, they find neurological evidence supporting their claim. In bringing the psychoacoustic voice model (Kreiman *et al.*, 2014) to the study of voice organization, we bring a broader range of acoustic-auditory measures that are known to characterize voices and be important to listeners. Our results do, indeed, suggest that measures of F0, higher formants (which strongly correlate with measures of formant dispersion), and several measures of phonation type (related to HNR) are important dimensions for acoustically characterizing voices. Future research is necessary to more directly connect our work to listeners' voice organization.

Bilinguals offer a crucial angle on voice and cognitive organization of voices. As we show, an individual's voice shares considerable amounts of structure across languages. These results suggest that while a language's structure determines what spectral and temporal dimensions will be used to cue linguistic meaning, an individual's vocal physiology, anatomy, and social persona will limit the range of variation. These results have implications for bilingual voice recognition for humans and machines, and they suggest that voice prototypes might not be language-specific. We, of course, only examine a single bilingual combination—Cantonese and English. Future research may demonstrate that properties of the languages under comparison—for example, their typological similarity, phonological overlap, and the accompanying sociocultural space—or the type of bilinguals compared—for example, early versus late, from a community with heavy code-switching or highly structured diglossia—may determine the degree to which bilinguals exhibit one or more voices.

## ACKNOWLEDGMENTS

This research was supported by an award to K. A. J. from the University of British Columbia Public Scholars Initiative and by Social Sciences and Humanities Research Council of Canada (SSHRC) and Natural Sciences and Engineering Research Council of Canada (NSERC) awards to M.B. This work has benefited from feedback from Rob Fuhrman and Rachel Soo.

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

[2] The correlation between F4 for talkers across languages is very high [$t(32) = 11.6, p < 0.001, r = 0.90$].

Afouras, T., Chung, J. S., and Zisserman, A. (**2020**). "Now you're speaking my language: Visual language identification," in *Proceedings of Interspeech 2020*, October 25–29, Shanghai, China, pp. 2402–2406.

Altenberg, E. P., and Ferrand, C. T. (**2006**). "Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women," J. Voice **20**(1), 89–96.

Belin, P., Fecteau, S., and Bédard, C. (**2004**). "Thinking the voice: Neural correlates of voice perception," Trends Cogn. Sci. **8**(3), 129–135.

Boersma, P., and Weenink, D. (**2021**). "Praat: Doing phonetics by computer (version 6.1.38) [computer program]," http://www.praat.org/ (Last viewed January 2, 2021).

Bradlow, A. R., Kim, M., and Blasingame, M. (**2017**). "Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate," J. Acoust. Soc. Am. **141**(2), 886–899.

Bregman, M. R., and Creel, S. C. (**2014**). "Gradient language dominance affects talker learning," Cognition **130**(1), 85–95.

Bullock, B. E., and Toribio, A. J. (**2009**). "Trying to hit a moving target: On the sociophonetics of code-switching," in *Studies in Bilingualism*, edited by L. Isurin, D. Winford, and K. deBot (John Benjamins, Amsterdam), pp. 189–206.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., and Jenkins, R. (**2016**). "Identity from variation: Representations of faces derived from multiple instances," Cogn. Sci. **40**(1), 202–223.

Chai, Y., and Garellek, M. (**2022**). "On H1–H2 as an acoustic measure of linguistic phonation type," J. Acoust. Soc. Am. **152**(3), 1856–1870.

Cheng, A. (**2020**). "Cross-linguistic F0 differences in bilingual speakers of English and Korean," J. Acoust. Soc. Am. **147**(2), EL67–EL73.

Cheng, L. S., Babel, M., and Yao, Y. (**2022**). "Production and perception across three Hong Kong Cantonese consonant mergers: Community- and individual-level perspectives," Lab. Phonol. **13**(1), 14.

Chodroff, E., and Wilson, C. (**2017**). "Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English," J. Phon. **61**, 30–47.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton de Gruyter, Berlin).

Fricke, M., Kroll, J. F., and Dussias, P. E. (**2016**). "Phonetic variation in bilingual speech: A lens for studying the production-comprehension link," J. Mem. Lang. **89**, 110–137.

Garellek, M. (**2019**). "The phonetics of voice," in *The Routledge Handbook of Phonetics*, edited by W. F. Katz and P. F. Assmann (Routledge, Abingdon, UK).

Garellek, M., Ritchart, A., and Kuang, J. (**2016**). "Breathy voice during nasality: A cross-linguistic study," J. Phon. **59**, 110–121.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cognit. **19**(5), 448–458.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Hear. Res. **37**(4), 769–778.

Hollien, H., Majewski, W., and Doherty, E. T. (**1982**). "Perceptual identification of voices under normal, stress and disguise speaking conditions," J. Phon. **10**(2), 139–148.

Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**(4), 2283–2295.

Jadoul, Y., Thompson, B., and de Boer, B. (**2018**). "Introducing Parselmouth: A Python interface to Praat," J. Phon. **71**, 1–15.

Järvinen, K., Laukkanen, A.-M., and Aaltonen, O. (**2013**). "Speaking a foreign language and its effect on F0," Logoped. Phoniatr. Vocol. **38**(2), 47–51.

Johnson, E. K., Westrek, E., Nazzi, T., and Cutler, A. (**2011**). "Infant ability to tell voices apart rests on language experience," Dev. Sci. **14**(5), 1002–1011.

Johnson, K. A. (**2021a**). "Leveraging the uniformity framework to examine crosslinguistic similarity for long-lag stops in spontaneous Cantonese-English bilingual speech," in *Proceedings of Interspeech 2021*, Brno, Czech Republic, August 30–September 3, pp. 2671–2675.

Johnson, K. A. (**2021b**). "SpiCE: Speech in Cantonese and English," https://doi.org/10.5683/SP2/MJOXP3 (Last viewed May 20, 2021).

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004.

Jolliffe, I. T. (**2002**). *Principal Component Analysis*, 2nd ed. (Springer-Verlag, New York).

Kawahara, H., Agiomyrgiannakis, Y., and Zen, H. (**2016**). "Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, September 13–15, Sunnyvale, CA, pp. 221–228.

Keating, P., Kreiman, J., and Alwan, A. (**2019**). "A new speech database for within- and between-speaker variability," in *Proceedings of the 19th International Congress of Phonetic Sciences*, August 5–9, Melbourne, Australia, pp. 736–739.

Keating, P., and Kuo, G. (**2012**). "Comparison of speaking fundamental frequency in English and Mandarin," J. Acoust. Soc. Am. **132**(2), 1050–1060.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (**2014**). "Toward a unified theory of voice production and perception," Loquens **1**(1), e009.

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (**2021**). "Validating a psychoacoustic model of voice quality," J. Acoust. Soc. Am. **149**(1), 457–465.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, Hoboken, NJ).

Latinus, M., and Belin, P. (**2011**). "Anti-voice adaptation suggests prototype-based coding of voice identity," Front. Psychol. **2**, 175.

Latinus, M., McAleer, P., Bestelmeyer, P., and Belin, P. (**2013**). "Norm-based coding of voice identity in human auditory cortex," Curr. Biol. **23**(12), 1075–1080.

Lavan, N., Burston, L. F. K., and Garrido, L. (**2019a**). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**(3), 576–593.

Lavan, N., Knight, S., and McGettigan, C. (**2019b**). "Listeners form average-based representations of individual voice identities," Nat. Commun. **10**(1), 2404.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University, New York).

Lavner, Y., Rosenhouse, J., and Gath, I. (**2001**). "The prototype model in speaker identification by human listeners," Int. J. Speech Technol. **4**(1), 63–74.

Lee, B., and Sidtis, D. V. L. (**2017**). "The bilingual voice: Vocal characteristics when speaking two languages across speech tasks," Speech Lang. Hear. **20**(3), 174–185.

Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**(3), 1568–1579.

Lee, Y., and Kreiman, J. (**2019**). "Within- and between-speaker acoustic variability: Spontaneous versus read speech," J. Acoust. Soc. Am. **146**, 3011.

Lee, Y., and Kreiman, J. (**2020**). "Language effects on acoustic voice variation within and between talkers," J. Acoust. Soc. Am. **148**, 2473.

Lee, Y., and Kreiman, J. (**2022**). "Acoustic voice variation in spontaneous speech," J. Acoust. Soc. Am. **151**(5), 3462–3472.

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

Lüdecke, D., Ben-Shachar, M. S., Patil, I., and Makowski, D. (**2020**). "Extracting, computing and exploring the parameters of statistical models using R," J. Open Source Softw. **5**(53), 2445.

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

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal forced aligner (version 1.0.1)," https://montrealcorpustools.github.io/Montreal-Forced-Aligner/ (Last viewed October 1, 2020).

Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., and Schaeffler, F. (**2010**). "Measuring language-specific phonetic settings," Second Lang. Res. **26**(1), 13–41.

Munson, B., and Babel, M. (**2019**). “The phonetics of sex and gender,” in *The Routledge Handbook of Phonetics*, edited by W. F. Katz and P. F. Assmann (Routledge, London).

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K. (**2010**). “Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of vox humana,” Clin. Linguist. Phon. **24**(4–5), 245–260.

Myers-Scotton, C. (**2011**). “The matrix language frame model: Developments and responses,” in *Codeswitching Worldwide* (Mouton De Gruyter, Berlin).

Navarro, D. (**2015**). “Learning statistics with R: A tutorial for psychology students and other beginners (version 0.6),” https://learningstatisticswithr.com (Last viewed October 1, 2020).

Ng, M. L., Chen, Y., and Chan, E. Y. (**2012**). “Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—A long-term average spectral analysis,” J. Voice **26**(4), e171–e176.

Ng, M. L., Hsueh, G., and Sam Leung, C.-S. (**2010**). “Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children,” Int. J. Speech Lang. Pathol. **12**(3), 230–236.

Nygaard, L. C., and Pisoni, D. B. (**1998**). “Talker-specific learning in speech perception,” Percept. Psychophys. **60**(3), 355–376.

Ordin, M., and Mennen, I. (**2017**). “Cross-linguistic differences in bilinguals’ fundamental frequency ranges,” J. Speech Lang. Hear. Res. **60**(6), 1493–1506.

Orena, A. J., Polka, L., and Theodore, R. M. (**2019**). “Identifying bilingual talkers after a language switch: Language experience matters,” J. Acoust. Soc. Am. **145**(4), EL303–EL309.

Orena, A. J., Theodore, R. M., and Polka, L. (**2015**). “Language exposure facilitates talker learning prior to language comprehension, even in adults,” Cognition **143**, 36–40.

Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. A., and Alwan, A. (**2018**). “Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles,” J. Acoust. Soc. Am. **144**(1), 375–386.

Perrachione, T., Dougherty, S., McLaughlin, D., and Lember, R. (**2015**). “The effects of speech perception and speech comprehension on talker identification,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK.

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). “Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices,” J. Acoust. Soc. Am. **146**(5), 3384–3399.

Perrachione, T. K., and Wong, P. C. (**2007**). “Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex,” Neuropsychologia **45**(8), 1899–1910.

Pittam, J. (**1987**). “The long-term spectral measurement of voice quality as a social and personality marker: A review,” Lang. Speech **30**(1), 1–12.

Podesva, R. J., and Callier, P. (**2015**). “Voice quality and identity,” Annu. Rev. Appl. Linguist. **35**, 173–194.

Purnell, T., Idsardi, W., and Baugh, J. (**1999**). “Perceptual and phonetic experiments on American English dialect identification,” J. Lang. Soc. Psychol. **18**(1), 10–30.

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

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

Seyfarth, S., and Garellek, M. (**2018**). “Plosive voicing acoustics and voice quality in Yerevan Armenian,” J. Phon. **71**, 425–450.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). “VoiceSauce: A program for voice analysis,” in *Proceedings of the 17th International Congress of Phonetic Sciences*, August 17–21, Hong Kong, Vol. 3, pp. 1846–1849.

Simpson, A. P. (**2009**). “Phonetic differences between male and female speech,” Lang. Linguist. Compass **3**(2), 621–640.

Simpson, A. P. (**2012**). “The first and second harmonics should not be used to measure breathiness in male and female voices,” J. Phon. **40**(3), 477–490.

Sjölander, K. (**2004**). “The Snack Sound Toolkit,” https://www.speech.kth.se/snack/ (Last viewed June 1, 2023).

Soo, R., Johnson, K. A., and Babel, M. (**2021**). “Sound change in spontaneous bilingual speech: A corpus study on the Cantonese n-l merger in Cantonese-English bilinguals,” in *Proceedings of Interspeech 2021*, Brno, Czech Republic, August 30–September 3, pp. 421–425.

Sóskuthy, M., and Stuart-Smith, J. (**2020**). “Voice quality and coda /r/ in Glasgow English in the early 20th century,” Lang. Var. Change **32**(2), 133–157.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., and Werker, J. F. (**2007**). “Discriminating languages by speech-reading,” Percept. Psychophys. **69**(2), 218–231.

Stewart, D., and Love, W. (**1968**). “A general canonical correlation index,” Psychol. Bull. **70**(3, pt.1), 160–163.

Sun, X. (**2002**). “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*, May 13–17, Orlando, FL, Vol. 1, pp. I–333–I–336.

Tabachnick, B. G., and Fidell, L. S. (**2013**). *Using Multivariate Statistics*, 6th ed. (Pearson, London).

Thompson, C. P. (**1987**). “A language effect in voice identification,” Appl. Cogn. Psychol. **1**(2), 121–131.

Turk, M., and Pentland, A. (**1991**). “Face recognition using eigenfaces,” in *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 3–6, Maui, HI.

Voigt, R., Jurafsky, D., and Sumner, M. (**2016**). “Between- and within-speaker effects of bilingualism on F0 variation,” in *Proceedings of Interspeech 2016*, September 8–12, San Francisco, CA, pp. 1122–1126.

Wei, L. (**2018**). “Translanguaging as a practical theory of language,” Appl. Linguist. **39**(1), 9–30.

Xie, X., and Myers, E. (**2015**). “The impact of musical training and tone language experience on talker identification,” J. Acoust. Soc. Am. **137**(1), 419–432.

Xue, S. A., Hagstrom, F., and Hao, J. (**2002**). “Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: A functional system approach,” Asia Pac. J. Speech Lang. Hear. **7**(1), 55–62.

Yang, Y., Chen, S., and Chen, X. (**2020**). “F0 patterns in Mandarin statements of Mandarin and Cantonese speakers,” in *Proceedings of Interspeech 2020*, October 25–29, Shanghai, China, pp. 4163–4167.