Title
Acoustic voice variation within and between speakers.

Journal
The Journal of the Acoustical Society of America, 146(3)

ISSN
0001-4966

Authors
Lee, Yoonjeong
Keating, Patricia
Kreiman, Jody

Publication Date
2019-09-01

DOI
10.1121/1.5125134

Peer reviewed

Acoustic voice variation within and between speakers

Yoonjeong Lee,[1, a)] Patricia Keating,[2] and Jody Kreiman[1, 2]

[1] *Department of Head and Neck Surgery, UCLA School of Medicine, Los Angeles, California 90095-1794, USA*

[2] *Department of Linguistics, University of California, Los Angeles, Los Angeles, California 90095-1543, USA*

Abstract

Little is known about the nature or extent of everyday variability in voice quality. This paper describes a series of principal component analyses to explore within- and between-talker acoustic variation and the extent to which they conform to expectations derived from current models of voice perception. Based on studies of faces and cognitive models of speaker recognition, we hypothesized that a few measures would be important across speakers, but that much of within-speaker variability would be idiosyncratic. Analyses used multiple sentence productions from fifty female and fifty male speakers of English, recorded over three days. Twenty-six acoustic variables from a psychoacoustic model of voice quality were measured every 5 ms on vowels and approximants. Across speakers the balance between higher harmonic amplitudes and inharmonic energy in the voice accounted for the most variance (females=20%, males=22%). Formant frequencies and their variability accounted for an additional 12% of variance across speakers. Remaining variance appeared largely idiosyncratic, suggesting that the speaker-specific voice space is different for different people. Results further showed that voice spaces for individuals and for the population of talkers have very similar acoustic structures. Implications for prototype models of voice perception and recognition are discussed.

Keywords: acoustic voice variation, within-speaker variability, between-speaker variability, prototype models of voice perception, speaker recognition, voice quality.

[a]yoonjeonglee@ucla.edu

I. INTRODUCTION

What makes your voice yours? Individuals' voices, their "auditory faces" (Belin *et al.*, 2004), provide significant clues to personal identity along with information about talkers' long-term physical, psychological, and social characteristics, based on the variability these factors introduce into voice. Because even small changes in emotion, social context, and physiologic state can cause significant variability in voice, no speaker ever says the same thing in exactly the same way twice, whether quality is intentionally or incidentally manipulated (see Kreiman and Sidtis 2011, for extended review). However, the extent and nature of within-speaker variability in voice are unknown, despite the fact that the acoustic signal serves as input to the perceptual system, which must be able to cope with this variability in order to achieve a stable percept and/or recognition. Information about acoustic variability is thus critical for formulating models of voice quality and talker recognition. This paper describes a series of analyses exploring within- and between-talker acoustic variation and the extent to which they conform to expectations derived from current models of voice perception.

Although listeners can cope to some extent with acoustic variability to establish stable identity percepts, across voices and listeners many studies have shown that within-speaker variability makes voice recognition and discrimination challenging tasks. In forensic contexts, for example, an earwitness's ability to identify a person from a voice lineup diminishes when vocal variability is introduced. Listeners often fail to reliably discriminate between talkers when exposed to voices disguised using falsetto, hyponasality, creaky voice, or whispering

(Hirson and Duckworth, 1993; LaRiviere, 1975; Reich and Duke, 1979; Reich *et al.*, 2005; Wagner and Köster, 1999); and changes in a speaker's emotional state substantially impair listeners' abilities to recognize Saslove and Yarmey 1980; cf. Read and Craik 1995) or discriminate among talkers (Lavan *et al.*, 2019). Within-talker variability can also interfere with a listener's ability to judge that samples come from the same (rather than different) talkers. In a "telling voices together" task, listeners frequently judged that exemplars from a single talker came from multiple speakers when samples were drawn from different speaking situations with varied interlocutors (Lavan *et al.*, 2018).

Facial recognition poses similar challenges to viewers, who must cope with changes in lighting, expression, and orientation in order to identify or discriminate among faces (Hill and Bruce, 1996; O'Toole *et al.*, 1998; Patterson and Baddeley, 1977). Because similarities exist in voice and face processing (Stevenage *et al.*, 2018; Yovel and Belin, 2013), recent findings from the face perception literature may provide insight into mechanisms for coping with acoustic voice variability. In particular, facial identity learning improves when viewers are exposed to highly but naturally varying sets of images of one person (for example, with changes in orientation or emotion) during training (Kramer *et al.*, 2017; Murphy *et al.*, 2015; Ritchie and Burton, 2017). This suggests that variation in the same face provides useful person-specific information and thus is important in identity learning and perception (Burton, 2013; Burton *et al.*, 2016; Jenkins *et al.*, 2011). To our knowledge, no parallel studies have appeared for voice learning, but some classic findings suggest acoustic variability may also provide important information to listeners. These studies have reported that increasing phonological length (i.e., the number of individual phonemes; Schweinberger *et al.*

1997) or acoustic duration (Bricker and Pruzansky, 1966; Cook and Wilding, 1997; Legge *et al.*, 1984) of the voice samples leads to more accurate vocal identity processing, due to the increased variety in speech sounds available in longer stimuli or the longer duration (or both), which provide listeners with added articulatory and acoustic variability (cf. e.g. Lively *et al.* 1993, for similar effects in learning phonological categories).

Taken together, these studies of faces and voices suggest that listeners need to learn how a particular voice varies in order to recognize it accurately and efficiently. At first glance, this claim appears consistent with prototype-based models of the cognitive and neural processes underlying voice identity perception (Latinus and Belin, 2011a; Lavner *et al.*, 2001; Papçun *et al.*, 1989; Yovel and Belin, 2013). In these accounts, listeners encode and process voice identity in relation to a population prototype, which is a context-dependent "average-sounding" voice, defined as a central tendency in a distribution of exemplars (Patel, 2008) that resides at the center of a multidimensional acoustical 'voice space.' Each voice is further represented in terms of its deviations from that group prototype, stored as a unique 'reference pattern' for that identity and passed on for further analysis (Latinus and Belin, 2011b; Papçun *et al.*, 1989). On further consideration, however, it becomes apparent that these models are underspecified with respect to two important issues. First, the relationship between between-talker variability in quality and the population prototype is unknown. Although it is commonly assumed that prototypes are statistical averages derived from multiple samples of a given talker's voice (e.g., Latinus and Belin 2011a; Maguinness *et al.* 2018), to our knowledge no data exist about how much detail (and what kind of detail) about quality is actually needed to specify the prototype, and how much is reserved as "deviations" from the prototype. Second, the nature (or even the existence) of similar reference patterns for individual talkers and the way in which within-talker variation affects formation of these patterns have not to our knowledge been addressed, although such patterns would seem to be essential for the formation of stable representations of voices and thus for voice recognition (Lavan *et al.*, 2018).

Existing cognitive and neuropsychological models of voice perception and recognition have not been fully exploited to generate clear hypotheses about the nature and extent of even between-talker acoustic variability in voice, which has been studied far more than within-talker variability. As discussed above, these models posit the existence of an acoustic voice space organized around a population prototype, so that voices are encoded and later recognized in terms of their distance from the prototype and the manner in which they deviate from this (presumed) population average. Because voice production and perception have co-evolved, it follows that if the perceptual models are correct, then there should be some acoustic features that consistently explain significant between-talker acoustic variance across all the talkers in a population. These features would characterize the central category member for the population of talkers, consistent with the existence of a perceptual space organized around a prototype, and would also specify the location of each voice in the space with respect to the prototype. Remaining differences between voices should be idiosyncratic, so that the features that differentiate pairs of talkers depend on the precise acoustic information involved in each comparison (e.g., Kreiman and Gerratt 1996). This would be consistent with what has been found for faces (Maguinness *et al.*, 2018; Stevenage *et al.*, 2018; Yovel and Belin, 2013), although we cannot assume that faces and voices are perceived in similar ways at all processing stages.

Predictions are less clear for variation within a single talker across utterances, although studies of variation in faces may again offer some clues. Principal component analyses examining how images of a face vary across different photographs of that person (Burton *et al.*, 2016) showed that the first few components (left-to-right head rotations, angle to camera, the direction of lighting; and changes in expression like smiles, eye movements, mouth opening, or lip rounding during speech) emerged consistently across individuals and accounted for the most variance in different photos of the same person. Dimensions appearing in later principal components (from the fourth onward) did not generalize well from one person to another, so that some features were shared across faces, and some dimensions of variability were idiosyncratic to specific faces. Given the many similarities between face and voice processing in identity perception (see Yovel and Belin 2013, for review), this suggests that voice spaces for individual talkers should be similarly structured. If "prototypes" for individual talkers are characterized by the same features across talkers, then these features would naturally characterize a population prototype against which each individual voice could be assessed.

Results from our preliminary studies (Keating and Kreiman, 2016; Kreiman *et al.*, 2017) are also consistent with the hypothesis that voice spaces for individual talkers are structured similarly to population voice spaces. In those experiments, we used linear discriminant analyses to identify the acoustic features that maximally distinguished a large number of individual voices. A small number of variables (F0, F4, the root mean square energy calculated over five pitch pulses [energy], the relative amplitudes of the first and second harmonics [H1-H2], and the amplitude ratio between subharmonics and harmonics [SHR]) proved important for distinguishing both male and female voices, but these accounted for only about 50% of the acoustic variance in the data, the remaining variance being explained by different variables depending on the particular voices being compared.

In the present study we focused on the acoustic attributes that characterize different voice samples from individual talkers, as well as on the population of talkers as a whole. Following Burton *et al.* (2016), we used principal component analysis to assess voice variation both within and across speakers. The components that emerge from such analyses can be thought of as forming dimensions of an acoustic space specific to a given voice, in which that voice varies, in contrast to the discriminant analysis approach in our previous work. Based on Burton *et al.* (2016) and on prototype models of voice processing, we hypothesized that a few common acoustic dimensions would consistently emerge from analyses of individual speakers as explaining the most within-talker acoustic variability, but that much more of what characterizes vocal variability within a speaker would be idiosyncratic. Because voice quality is inherently dynamic, we tested the above hypothesis against multiple sentence productions from 100 native speakers of English, using a set of acoustic measures that combine to completely specify voice quality (Kreiman *et al.*, 2014). This approach contrasts with previous studies of vocal acoustic spaces (e.g., Baumann and Belin 2010; Murry and Singh 1980; Murry *et al.* 1978), which used limited sets of steady-state vowels. Finally, we compared the dimensions characterizing acoustic variability across speakers to those characterizing within-speaker acoustic variability, also in contrast to previous work.

II. METHOD

A. Speakers and voice samples

In this experiment, the voices of 50 female and 50 male speakers were drawn from the University of California, Los Angeles Speaker Variability Database (Keating *et al.*, 2019). All were native speakers of English, similar in age (F: 18-29, M: 18-26), with no known vocal disorder or speech complaints, and all were UCLA undergraduate students at the time of recording. As noted previously, virtually nothing is known about acoustic differences between different populations of speakers. For this reason, in this initial study we opted to control for possible systematic differences between populations by studying a homogeneous group, so that we would be able to unambiguously attribute acoustic differences to within- or between-speaker factors, without the added complication of differences between populations. Recordings were made in a sound-attenuated booth at a sampling rate of 22 kHz using a Bruel & Kjaer $\frac{1}{2}$" microphone (model 4193) securely attached to a baseball cap worn by the speaker.

The database provides significant within- and between-speaker variability. Speakers were recorded on 3 different days and performed multiple speech tasks including reading, unscripted speech tasks, and a conversation. In order to control for variations due to differences in phonemic content or emotional state across talkers, this initial study used recordings of 5 Harvard sentences (IEEE Subcommittee 1969; Table I), read twice each day for a total of 6 repetitions per sentence over 3 recording sessions on different days. Variability reported in this paper was calculated across sentence productions (different repetitions, sentences, and days), and its scope is limited to the reading task.

TABLE I. Reading materials.

Harvard sentences
A pot of tea helps to pass the evening.
The boy was there when the sun rose.
Kick the ball straight and follow through.
Help the woman get back to her feet.
The soft cushion broke the man's fall.

B. Measurements and data processing

Acoustic measurements were made automatically every 5 ms on vowels and approximants (i.e., /l/, /r/, /w/) excerpted from each complete sentence, using VoiceSauce (Shue *et al.*, 2011). Following the psychoacoustic model of voice quality described in Kreiman *et al.* (2014), acoustic parameters included fundamental frequency (F0); the first four formant frequencies (F1, F2, F3, F4), the relative amplitudes of the first and second harmonics (H1*-H2*) and the second and fourth harmonics (H2*-H4*); and the spectral slopes from the fourth harmonic to the harmonic nearest 2 kHz in frequency (H4*-H2kHz*) and from the harmonic nearest 2 kHz to the harmonic nearest 5 kHz in frequency (H2kHz*-H5kHz). Values of harmonics marked with '*' were corrected for the influence of formants on harmonic amplitudes (Hanson and Chuang, 1999; Iseli and Alwan, 2004). Our preliminary studies (Keating and Kreiman, 2016; Kreiman *et al.*, 2017) showed substantial correlations between the relative amplitude of the cepstral peak prominence in relation to the expected amplitude as derived via linear regression (CPP; Hillenbrand *et al.* 1994) and the 4 measures of the shape of the inharmonic (noise) source spectrum included in the psychoacoustic model, so for simplicity CPP was used as the only measure of spectral noise and/or periodicity in these analyses.

Several additional modifications were made to adapt the model to automatic measurement of continuous speech. Formant dispersion (FD, often associated with vocal tract length [Fitch 1997]) was calculated as the average difference in frequency between each adjacent pair of formants (cf. Pisanski *et al.* 2014 for related measures). Amplitude was measured as the root mean square energy calculated over five pitch pulses (energy). Period doubling, which is not included in the original psychoacoustic model but is common in the speech of UCLA students, was measured as the amplitude ratio between subharmonics and harmonics (SHR; Sun 2002). Finally, dynamic changes in voice quality were quantified using moving coefficients of variation ($moving\ CoV = \frac{moving\ \sigma}{moving\ \mu}$) for each parameter. In choosing this measure, we assumed that listeners do not generally rely on exact pitch and amplitude contours or on the precise timing of changes in spectral shape when telling speakers apart, although such details can be salient when discriminating among speech tokens from a single speaker. This approach has the added advantage that quantifying the amount of variability is straightforward, whereas there is no obvious way to quantify and objectively compare exact patterns of acoustic variation. Table II provides a complete list of variables.

Data frames with missing or obviously erroneous parameter values (for example, impossible 0 values) were removed. Next, for each speaker, the obtained values of each acoustic

TABLE II. Acoustic variables.

Variable categories	**Acoustic variables**
pitch	F0
formant frequencies	F1, F2, F3, F4, FD
harmonic source spectral shape	H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz
inharmonic source/spectral noise	CPP, energy, SHR
variability	coefficients of variation for all acoustic measures

variable were normalized with respect to the overall minimum and maximum values from the entire set of voice samples from males or females, as appropriate, so that all variables ranged from 0 to 1. Then, for each sentence production, a smoothing window of 50 ms (10 observations) was used to calculate moving averages and moving coefficients of variation for the 13 variables during that sentence. Across speakers, the above winnowing and post-processing steps resulted in about 515k data frames (F: 266k, M: 249k).

C. Principal component analysis

In principal component analysis (PCA), variables that are correlated with one another but relatively independent of other subsets of variables are combined into components, with the goal of reducing a large number of variables into a smaller set which is thought to reflect internal structures that have created the correlations among variables. As moderate correlations were expected between variables, we employed an oblique rotation to create the simplest possible factor structure for our data (Cattell, 1978; Thurstone, 1947). Analyses were conducted separately for each speaker (within-speaker analyses) and for the combined male and female speakers as groups (combined speaker analyses). For within-speaker analyses, PCA was performed separately on each individual talker's acoustic measurement data (26 variables: moving averages for 13 variables + moving coefficients of variation for the same 13 variables) to reveal the dimensions of the acoustic variability space for that particular voice. For combined speaker analyses, PCA was performed separately on the acoustic data (all 26 variables) from females and males, pooling the 50 speakers' data in each analysis. PCs were restricted to the resulting factorial solutions with eigenvalues greater than 1, ensuring that each retained factor accounted for an interpretable amount of variance in the data (Kaiser, 1960). Results were also visually confirmed with Scree plots (Cattell, 1966). Following usual practice, variables with loadings (weights) at or exceeding 0.32 on a given component were considered to form a principal component (Tabachnick and Fidell, 2013).

III. RESULTS

Although all 26 acoustic variables were entered simultaneously into the analyses, for brevity and clarity results are first described with respect to 5 categories, following Kreiman *et al.* (2019): i) F0; ii) formant frequencies (F1, F2, F3, F4, FD); iii) harmonic source spectral shape (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz); iv) spectral noise (CPP plus energy and SHR); and v) the coefficients of variation for all measures (CoVs) (Table II). Detailed analyses follow these summary descriptions. We first present results from within-speaker PCA analyses, followed by analyses of the combined male and female speakers.

A. Within-speaker PCAs: Common dimensions and speaker-specific patterns

Analyses for individual speakers resulted in 6-9 principal components (PCs) having eigenvalues greater than 1. Most speakers showed 7 (31/100 speakers) or 8 (59/100 speakers) extracted PCs. These components accounted for 65%-74% (M=69%) of the cumulative acoustic variance for individual female speakers and 62%-73% (M=68%) for individual male speakers (see Appendix A for details). While all individual PCs were included in subsequent analyses, because the higher order PCs accounted for very small amounts of acoustic variability (Appendix A), only the first 6 are reported in detail.

We first counted the number of times each acoustic category appeared in a within-speaker solution, cumulated across the 50 speakers in each group. Fig. 1 shows the distribution of variables with respect to weight in the first six components. The first component accounted for 17%-23% (M=20%) and 20%-25% (M=22%) of the variance for females and males, respectively. For both females and males, the combined coefficients of variation emerged most frequently in PC1 across individual speakers (blue bars in Fig. 1).

Sub-analyses of factors contributing to the first PC are shown in Figs. 2 and 3. For most speakers, PC1 represented the combination of **variability (measured by CoVs) in source spectral shape** (F: 41/50 speakers, M: 46/50 speakers) and in **spectral noise** (F: 45/50 speakers, M: 47/50 speakers), which usually emerged together (F: 40/50 speakers, M: 44/50 speakers) (Fig. 2). An additional analysis (Fig. 3) revealed that across speakers all 4 CoV measures of source spectral variability (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-

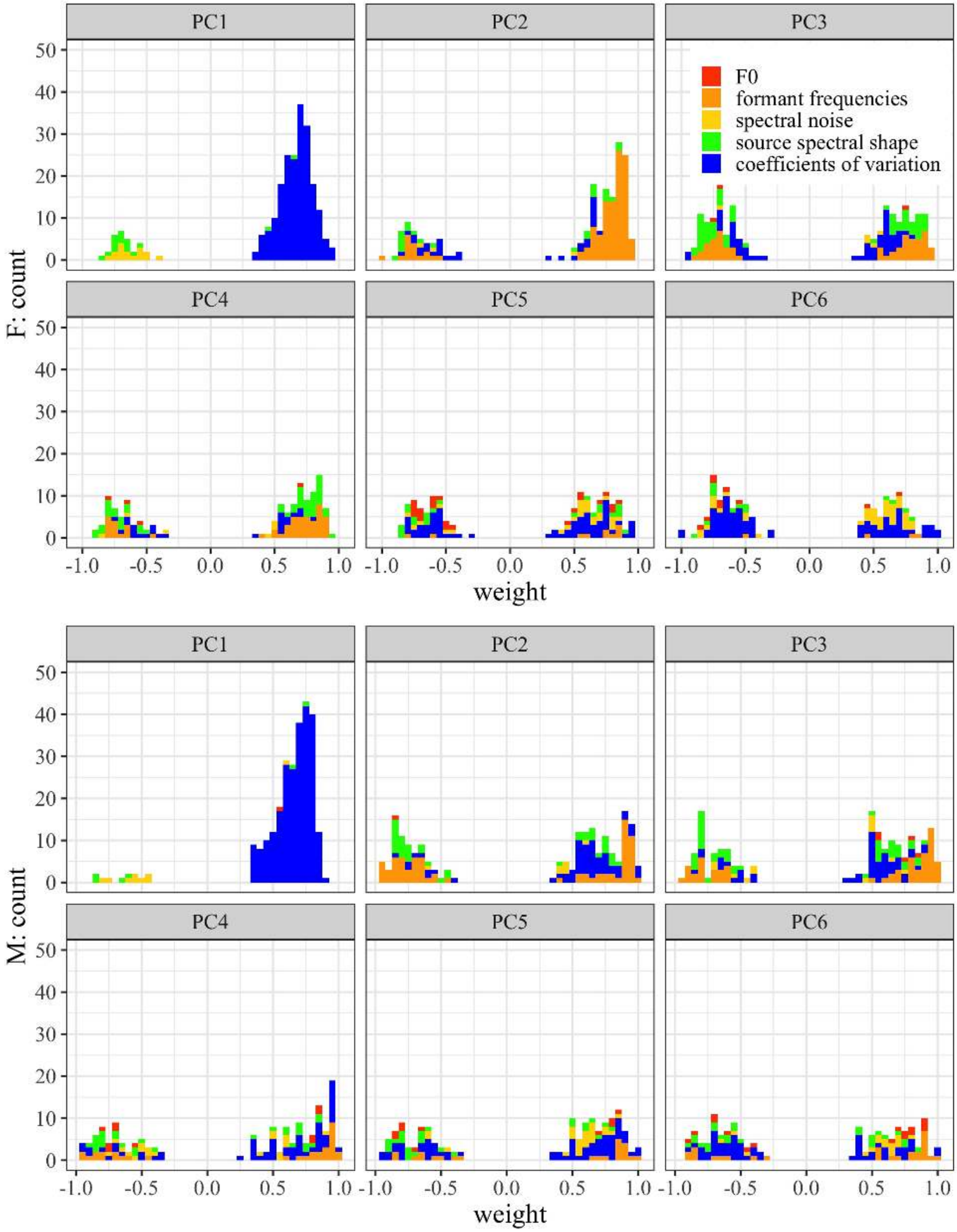

FIG. 1. Distribution of acoustic parameters plotted (stacked histogram) against the rotated component loadings (weight) for the first 6 PCs. Upper panel: female speakers. Lower panel: male speakers.

H5kHz) emerged in the first component, but **H2kHz*-H5kHz** predominated; spectral noise variability was mostly related to **coefficients of variation for CPP**.

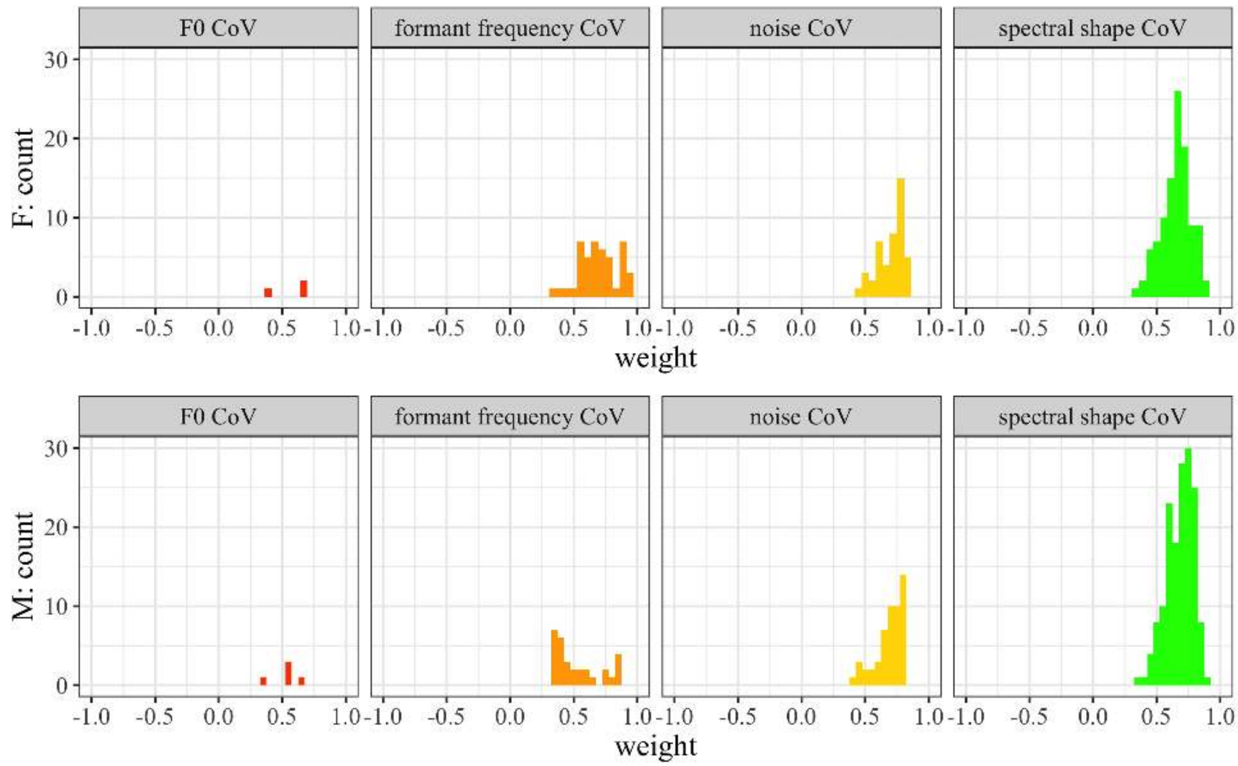

FIG. 2. (Color online) Distribution of variability parameters in PC1 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). 'CoV' = coefficient of variation.

For most of the remaining speakers (F: 10/50 speakers, M: 4/50 speakers), formant frequency CoV was the most representative variable in the first component. Lastly, two male speakers showed source spectral shape alone as the primary variable associated with this PC.

PC2 accounted for an average of 12% of acoustic variability, for both male and female speakers (ranges: females = 10%-16%; males = 10%-14%.). For both females and males, **formant frequencies (F: 50/50 speakers, M: 41/50 speakers) and/or their CoVs**

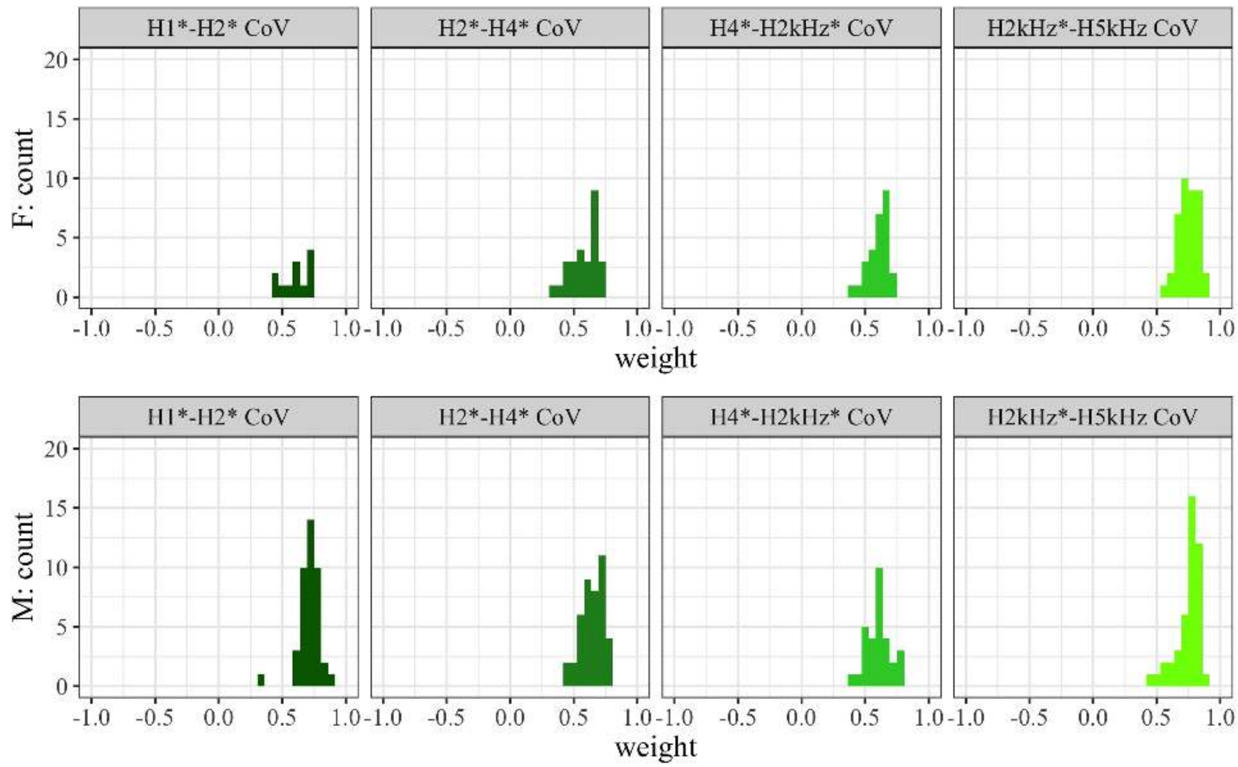

FIG. 3. (Color online) Distribution of spectral source variability parameters in PC1 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). ‘CoV’ = coefficient of variation.

(F: 21/50 speakers, M: 30/50 speakers) emerged most frequently as the second PC (Fig. 1). Sub-analyses are shown in Fig. 4; bars in this figure include both formant frequencies and coefficients of variation for each formant. **Formant dispersion** (F: 37/50 speakers, M: 28/50 speakers) and **F4** (F: 35/50 speakers, M: 28/50 speakers) appeared most important and frequently appeared together as a pair across speakers.

PC3-PC6 combined to account for an average across voices of 29% (females) and 28% (males) of the acoustic variance in the data (see also Appendix A), but in contrast to the first two PCs, this variance was largely idiosyncratic, and no particular acoustic category predominated (Fig. 1). For PC3-PC6, the distributions of the five variable categories and

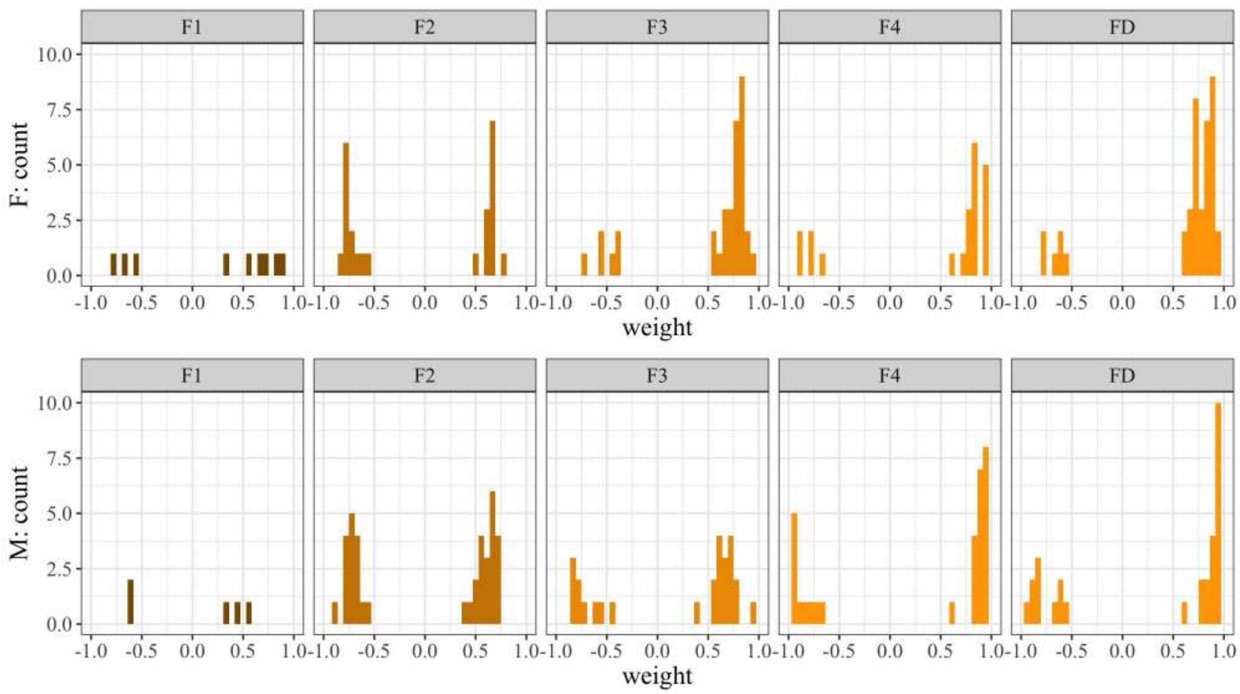

FIG. 4. (Color online) Distribution of formant frequency parameters in PC2 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). Each figure reflects values derived from both moving averages and moving coefficients of variation for each formant frequency measure. 'FD' = formant dispersion.

their weights overlapped highly, for both male and female speakers, reflecting differences across voices in the amount of variance explained by each measure. As shown in Fig. 1, most of the variables are approximately evenly distributed across PCs, with the exception of F0 (red bars), which emerged only sporadically. In other words, the component in which each variable appeared differed across individuals, ranging from PC3 to PC6(~9) across individuals; and no single component accounted for substantial variance.

Notably, F0 and/or its CoV only emerged in the first two components for 4/100 speakers (2 female and 2 male). Among those 4 speakers, only one (male) speaker showed F0 as the most weighted variable within the PC (red bar in PC1, Fig. 1, bottom panel).

Interim summary and discussion

To summarize, variability (measured by coefficients of variation) in source spectral shape and spectral noise, especially in H2kHz*-H5kHz and CPP, accounted for the most acoustic variability within individual speakers. Across speakers, the next most frequently emerging variables were means and variability for formant dispersion and F4. The first two PCs were largely shared across voices, and together accounted for slightly more than half of the explained variance in the underlying acoustic data (32%-34% total). The remaining PCs differed widely across voices, and cumulatively accounted for slightly less than half of the explained variance (28%-29% total).

The general picture that emerges from these results is one of surprisingly similar acoustic organization across talkers. This pattern of a common core of variables shared by virtually all voices, accompanied by unique deviations from that central pattern, is consistent with what might be required as input to a recognition/perception system organized around prototypes, and suggests that such a model applies to between-talker variability as well as to within-talker acoustic variability. The analyses in the next section test this hypothesis.

B. Between-speaker group PCA: "General" voice spaces

As described above, a second set of PCAs examined the structure of the acoustic space for the combined groups of female and male speakers. Eight PCs were extracted for both speaker groups, accounting for 67% of the cumulative variance for female speakers and 66% for male speakers. Not surprisingly, given how consistent results were across individual speakers, patterns of acoustic variability in these multi-talker spaces largely mirrored the patterns found within speakers. Fig. 5 shows the group results, and details of the analyses are included in Appendix B. The first PC weighted most heavily on **variability (measured by CoVs) in source spectral shape and spectral noise**, accounting for 18% and 20% of variance across females and males, respectively. As in the within-speaker analyses, **coefficients of variation for H2kHz*-H5kHz and CPP** were the most important components of this PC.

The second component accounted for 11% of acoustic variance in female voices and corresponded to **formant frequencies (F4, FD, F3)**. For males, **spectral slope in the higher frequencies (H4*-H2kHz*, H2kHz*-H5kHz) and F2** accounted for 10% of variance in the combined acoustic data. The opposite was observed for the third component: an additional 10% of the variance was accounted for by spectral shape in the higher frequencies and F2 for females; formant frequencies accounted for 9% of the variance in male voices. F0 only emerged in later components (PC5 for females, PC4 for males) with noise and spectral shape variables, and accounted for very little variance in the data (6% for females, 7% for males). CoVs for F0 and noise measures emerged in PC6 for female speakers and PC7 for male speakers and accounted for 5% of acoustic variance across speaker groups.

IV. DISCUSSION

Acoustic variability is a key factor in models of voice perception and speaker identification, because perceptual processes must cope with variable input in order to achieve perceptual constancy. Using principal component analysis (PCA), this study identified voice quality measures that accounted for perceptually-relevant acoustic variance both within individual

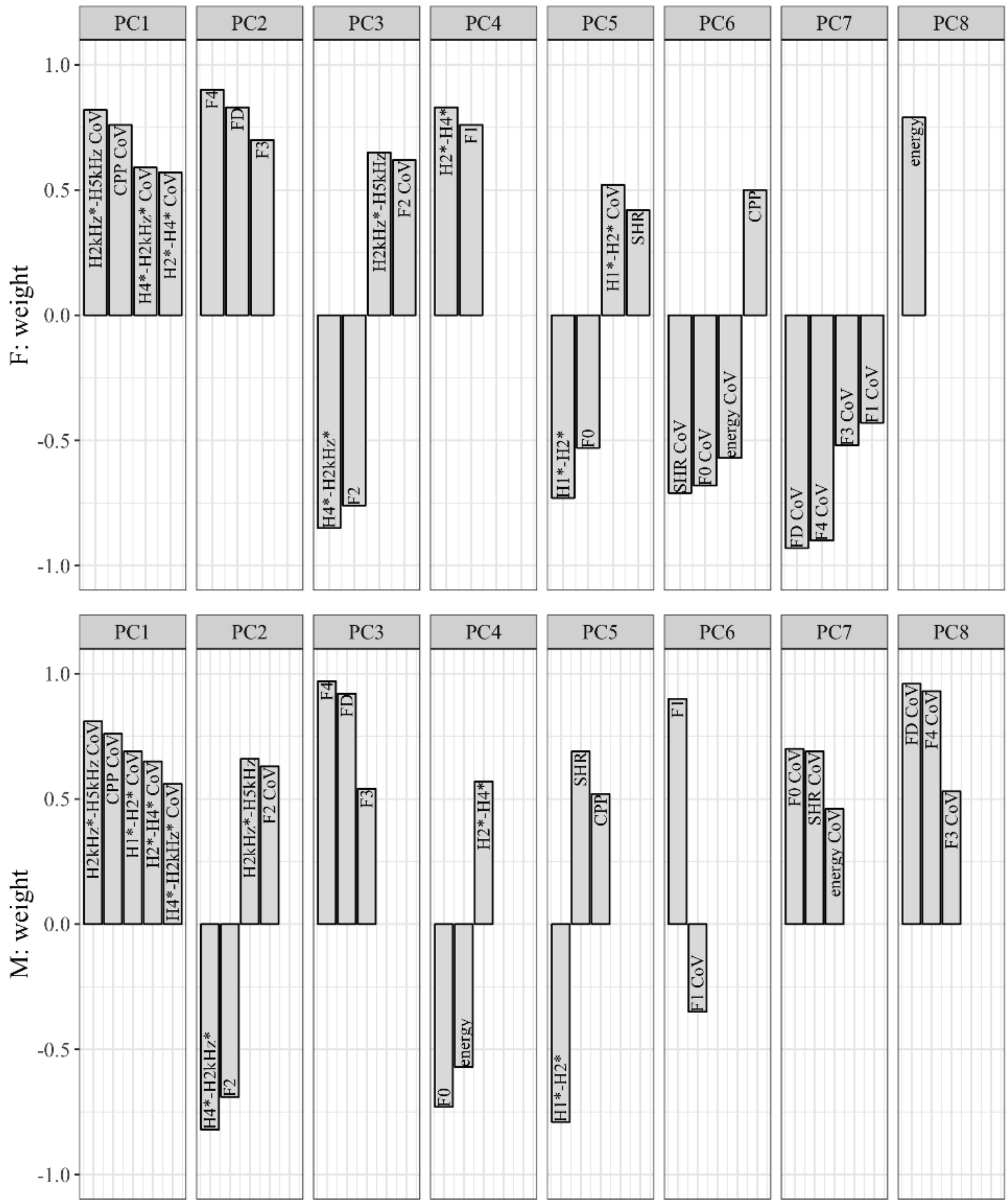

FIG. 5. Acoustic parameters emerging in 8 PCs for female speaker group (upper panel) and male speaker group (bottom panel). Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. See also Appendices B 1 and B 2 for variance accounted for by each PC. ‘CoV’ = coefficient of variation.

speakers and for pooled groups of speakers. Unlike previous studies of vocal variation, which typically use sustained vowels produced in isolation by relatively small numbers of talkers, this study included multiple complete sentences from large numbers of female and male talkers, and thus reflected vocal variation within and across utterances and multiple recording sessions.

As hypothesized, results of analyses of within-speaker acoustic variability paralleled findings for individual faces (Burton *et al.*, 2016), in that a small number of components emerged consistently across talkers. For both females and males, variability in higher-frequency harmonic and inharmonic energy (often associated with the degree of perceived breathiness or brightness; Samlan *et al.* 2013) combined to account for the most variance within talkers. These two measures generally emerged as a pair within the same PC, consistent with the manner in which they covary in controlling the perceived levels of noise in a voice (Kreiman and Gerratt, 2012). The second PC was consistently associated with higher formant frequencies and with the average interval between formant frequencies (i.e., formant dispersion). These measures have been associated with speaker identity (e.g., Ives *et al.* 2005; Smith *et al.* 2005) and with vocal tract length and perception of speaker size (Fitch, 1997; Pisanski *et al.*, 2014), but appear to be relatively independent of vowel quality (Fant, 1960).

However, an equal amount of within-talker acoustic variability was in fact specific to individual voices. The talker-specific dimensionality of the derived voice spaces differed across different talkers, and different measures, different combinations of measures, or different orderings of the same sets of measures emerged in PCs after the first two. This suggests that each individual “auditory face” is indeed unique, allowing for the formation of person-specific patterns/representations for a particular voice.

Similar dimensions also emerged in the first three components from group PCAs combining the 50 male and 50 female speakers into separate group analyses, with the balance between higher-frequency harmonic and inharmonic energy again accounting for the most variability. Frequencies of higher formants, formant dispersion, and mid-frequency measures (near the F2 range) emerged in the second and third components, with only differences in order of emergence across groups. As with analyses of individual voices, later components included very different measures across the two groups. Although this finding may appear trivial given the homogeneity of the individual results, in fact there is no a priori reason why individual solutions should coincide as they did, and no a priori reason why individual and group acoustic spaces should be so similar. However, prototype models seemingly require that acoustic spaces for individual talkers and population spaces be structured similarly, so that listeners can evaluate the location of each voice with respect to the population prototype. This result thus provides strong evidence consistent with such models.

Two limitations of this work must be noted. First, acoustic measures were based on read speech, not on spontaneous vocalization or conversation. This has the advantage of controlling for variations due to differences in phonemic content or emotional state across talkers, while still sampling variability across utterances and recording sessions within talkers, but clearly does not represent the full range of acoustic variability that occurs within a talker in an average day’s phonation. The UCLA Speaker Variability Database (Keating *et al.*, 2019) also includes a recording of an unscripted telephone conversation for each talker, and analyses are underway to determine how well the present findings extend to more natural utterances. Second, the sample of speakers studied was restricted with respect to speakers' ages (a limitation of the database) and native languages (a design decision). For this initial study, we view both of these limitations as necessary: No information is available about differences in acoustic variability across different populations of speakers, and even speculation is lacking with regard to how many and what kinds of populations exist, so no basis exists for distinguishing variability within a population from variability across populations. The methods presented here offer a means of investigating this question, which will be important for further development of models of voice perception. Similarly, the manner (if any) in which within- and between-speaker acoustic variability interact with linguistic factors such as tone and phonemic voice quality differences remains unknown, again making it desirable to control this factor in the present study. A systematic investigation of the interactions among these factors is also underway.

The fact that F0 did not emerge early among the principal components extracted for either the within-speaker or group analyses is counter-intuitive, given how important F0 is to many aspects of voice perception (e.g., Baumann and Belin 2010; Kreiman *et al.* 1992; Murry and Singh 1980; Murry *et al.* 1978; Walden *et al.* 1978; see Kreiman and Sidtis 2011, for review). The lack of a major F0 component in our results may be an artefact of our normalization technique, which was based on acoustic ranges but did not take into account differences in perceptual sensitivity to different variables. However, we note that previous studies reporting an F0 factor have used similar normalization procedures and steady-state vowels (e.g., Baumann and Belin 2010). We additionally note that F0 may vary in limited ways during reading, reducing its contributions to both within- and between-speaker acoustic differences. However, F0 did emerge as important for discriminating among voices for both females and males in our previous studies using linear discriminant analysis (LDA) and the same voice stimuli (Keating and Kreiman, 2016; Kreiman *et al.*, 2017), making it unlikely that our results are due to the use of read speech in this study. (Future studies using spontaneous speech will test this possibility directly.) Finally, LDA and PCA differ in the nature of the questions they ask: LDA provides insight into the variables that maximally separate stimuli, while PCA can reveal the structure of the acoustic space in which the stimuli vary, somewhat analogous to "telling voices apart" versus "telling voices together" (Lavan *et al.*, 2018). These different emphases may partially explain differences in the importance of F0 across experiments. In any event, this apparent discrepancy between acoustic structure and perceptual data requires further consideration.

These results have implications for current prototype-based models of voice processing (Kreiman and Sidtis, 2011; Lavner *et al.*, 2001; Yovel and Belin, 2013), which as previously noted are underspecified with respect to within-person variability in voice. Perceptual processes must be adapted to the acoustic input they receive, so understanding the structure of acoustic voice spaces can provide insight into why and how voice perception functions as it does. Converging evidence from different scientific disciplines has shown that assessing who is speaking utilizes both featural and pattern recognition strategies. Perceiving unfamiliar voices requires both reference to a population prototype and evaluation of the manner in which the voice deviates from that prototype, while familiar voices are recognized using holistic pattern recognition processes (Schweinberger *et al.* 1997; Van Lancker *et al.*

1985; see Kreiman and Sidtis 2011, for review). Our results suggest that reference patterns for individual speakers are mainly computed over the balance of higher-frequency harmonic versus inharmonic energy in the voice and over formant dispersion, and are located in a group voice space with similar structure. However, this shared structure accounts for only a fraction of either within- or between-speaker acoustic variability, with most variability being idiosyncratic. Thus, it may be misleading to think of prototypes as "average tokens" computed across complete acoustic signals. Our results suggest instead that they are specified by a very small number of acoustic attributes.

These results further suggest that for unfamiliar voices, "deviations from the prototype" include two different kinds of variability: differences within talkers from their own prototype, and deviations of representations for individual speakers from a group prototype. Listeners who are unfamiliar with the voices should be adept at assessing the second kind of variability ("telling voices apart;" Lavan *et al.* 2018), given that the same acoustic features appear to characterize both group and individual prototypes. However, listeners who are unfamiliar with a talker's voice should have difficulty in associating different tokens of a single talker's voice with each other ("telling voices together;" Lavan *et al.* 2018), given their unfamiliarity with the specific idiosyncrasies that characterize that talker's overall acoustic variability. The present data allow us to make specific acoustic-based predictions about which voice samples from different talkers will be confused and which samples from the same talker will fail to be correctly recognized as coming from the same talker. These predictions will be explored in our ongoing work.

Finally, these results suggest that learning to recognize a voice involves learning the specific manner(s) in which that voice varies around its prototype—in other words, variability in voice may be essential to learning, in the same way that it is essential for learning faces (Kramer *et al.*, 2017; Ritchie and Burton, 2017) and categories of any kind. Previous studies have suggested that familiar voices are unique patterns, such that a given feature may be essential for recognizing one voice, but irrelevant for another (Lattner *et al.*, 2005; Schweinberger, 2001; Van Lancker *et al.*, 1985). The present data are consistent with this view; but familiarity with a voice involves much more than knowledge of acoustic variability. Mental representations of familiar voices are linked to faces (e.g., Schweinberger 2013), and hearing a familiar voice activates a plethora of personal information about the speaker, possibly organized in "person identity nodes" (see Kreiman and Sidtis 2011, section 6.6, and Barton and Corrow 2016, for review). Thus, the manner in which voices become familiar, and even what familiarity entails, remain unknown, although the present data shed some light on possible mechanisms of acoustic learning.

V. CONCLUSION

Principal component analysis identified measures that characterize variability in voice quality within and between speakers and provided evidence for how voice spaces—individually and generally—may be formulated with reference to acoustic attributes. Among the large array of vocal parameters available for each individual voice, a few components (the balance between high-frequency harmonic and inharmonic energy in the voice, and formant dispersion) emerged consistently across talkers, but most within-speaker acoustic variability in voice was idiosyncratic. Results further showed that the measures that were frequently shared by individual talkers also characterized voice variation across talkers, suggesting that individual and "general" voice spaces have very similar acoustic structures. This aligns well with the input seemingly required by prototype models of voice recognition. Our results have implications for unfamiliar voice perception and processing, specifically providing evidence for the nature of reference patterns and deviations from "average-sounding" across voices, in individual and universal voice spaces. Going forward, it will be essential to consider how listeners organize these identified measures of within-person variability into a personal identity and how that relates to perceived differences between talkers.

ACKNOWLEDGMENTS

This work was supported by NIH grant DC01797 and NSF grant IIS-1704167. Preliminary analyses of these data appeared in *Proceedings of the 19th International Congress of Phonetic Sciences*. We thank Abeer Alwan, Zhaoyan Zhang, Bruce Gerratt, and three anonymous reviewers for valuable comments. We also thank Meng Yang for help with VoiceSauce analyses. The UCLA Speaker Variability Database is freely available by request to the second or third author. A text file including results of the acoustic analyses can be downloaded from XX.

APPENDIX A: AVERAGE PERCENTAGE OF ACOUSTIC VARIANCE EXPLAINED BY EACH PC AS A FUNCTION OF THE NUMBER OF PCS, FOR FEMALE AND MALE SPEAKERS. NUMBERS IN PARENTHESES INDICATE THE NUMBER OF SPEAKERS FOR WHOM THAT NUMBER OF PCS WAS EXTRACTED.

PC	**9 PCs** **(F: 8/50, M: 1/50)**	**8 PCs** **(F: 29/50, M: 30/50)**	**7 PCs** **(F: 13/50, M: 18/50)**	**6 PCs** **(F: 0/50, M: 1/50)**
1	F: 19% (17%-21%), M: 21%	F: 20% (18%-23%), M: 22% (20%-25%)	F: 20% (18%-23%), M: 22% (20%-25%)	F: N/A, M: 22%
2	F: 12% (10%-13%), M: 10%	F: 12% (11%-16%), M: 12% (10%-14%)	F: 13% (11%-14%), M: 12% (10%-13%)	F: N/A, M: 13%
3	F: 10% (8%-11%), M: 9%	F: 10% (9%-11%), M: 10% (8%-11%)	F: 10% (8%-11%), M: 10% (9%-12%)	F: N/A, M: 10%
4	F: 8% (7%-8%), M: 7%	F: 8% (7%-9%), M: 7% (6%-9%)	F: 8% (7%-9%), M: 7% (6%-9%)	F: N/A, M: 7%
5	F: 6% (5%-6%), M: 7%	F: 6% (5%-7%), M: 6% (5%-7%)	F: 6% (5%-7%), M: 6% (5%-7%)	F: N/A, M: 6%
6	F: 5% (5%), M: 4%	F: 5% (5%-6%), M: 5% (5%-6%)	F: 5% (5%-6%), M: 5% (4%-6%)	F: N/A, M: 5%
7	F: 5% (4%-5%), M: 5%	F: 4% (4%-5%) M: 4% (4%-5%)	F: 4% (4%-5%), M: 4% (4%-5%)	
8	F: 4% (4%-5%), M: 4%	F: 4% (4%), M: 4% (4%)		
9	F: 4% (4%), M: 4%			
Total	**F: 73% (71%-74%),** **M: 71%**	**F: 69% (68%-72%),** **M: 70% (67%-73%)**	**F: 66% (65%-68%),** **M: 66% (65%-68%)**	**F: N/A,** **M: 63%**

APPENDIX B: PCA PATTERN MATRICES FOR FEMALE (1) AND MALE (2) SPEAKER GROUP ANALYSES.

1. PCA pattern matrix for female speaker group analysis. 'CoV' = coefficient of variation.

PC	Variable group	Variables	Weight	Variance explained
1	spectral shape variability	H2kHz*-H5kHz CoV	0.82	18%
	noise variability	CPP CoV	0.76	
	spectral shape variability	H4*-H2kHz* CoV	0.59	
		H2*-H4* CoV	0.57	
2	formant frequencies	F4	0.90	11%
		FD	0.83	
		F3	0.70	
3	spectral shape	H4*-H2kHz*	-0.85	10%
	formant frequencies	F2	-0.76	
	spectral shape	H2kHz*-H5kHz	0.65	
	formant frequency variability	F2 CoV	0.62	
4	spectral shape	H2*-H4*	0.83	8%
	formant frequency	F1	0.76	
5	spectral shape	H1*-H2*	-0.73	6%
	F0	F0	-0.53	
	spectral shape variability	H1*-H2* CoV	0.52	
	noise	SHR	0.42	
6	noise variability	SHR CoV	-0.71	5%
	F0 variability	F0 CoV	-0.68	
	noise variability	energy CoV	-0.57	
	noise	CPP	0.50	
7	formant frequency variability	FD CoV	-0.93	5%
		F4 CoV	-0.90	
		F3 CoV	-0.52	
		F1 CoV	-0.43	
8	noise	energy	0.79	4%

2. PCA pattern matrix for male speaker group analysis. 'CoV' = coefficient of variation.

PC	Variable group	Variables	Weight	Variance explained
1	spectral shape variability	H2kHz*-H5kHz CoV	0.81	20%
	noise variability	CPP CoV	0.76	
	spectral shape variability	H1*-H2* CoV	0.69	
		H2*-H4* CoV	0.65	
		H4*-H2kHz* CoV	0.56	
2	spectral shape	H4*-H2kHz*	-0.82	10%
	formant frequencies	F2	-0.69	
	spectral shape	H2kHz*-H5kHz	0.66	
	formant frequency variability	F2 CoV	0.63	
3	formant frequencies	F4	0.97	9%
		FD	0.92	
		F3	0.54	
4	F0	F0	-0.73	7%
	noise	energy	-0.57	
	spectral shape	H2*-H4*	0.57	
5	spectral shape	H1*-H2*	-0.79	6%
	noise	SHR	0.69	
		CPP	0.52	
6	formant frequencies	F1	0.90	5%
	formant frequency variability	F1 CoV	-0.35	
7	F0 variability	F0 CoV	0.70	5%
	noise variability	SHR CoV	0.69	
		energy CoV	0.46	
8	formant frequency variability	FD CoV	0.96	4%
		F4 CoV	0.93	
		F3 CoV	0.53	

References

Barton, J. J., and Corrow, S. L. (**2016**). “Recognizing and identifying people: A neuropsychological review,” Cortex **75**, 132–150.

Baumann, O., and Belin, P. (**2010**). “Perceptual scaling of voice identity: Common dimensions for different vowels and speakers,” Psychol. Res. **74**(1), 110–120.

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

Bricker, P. D., and Pruzansky, S. (**1966**). “Effects of stimulus content and duration on talker identification,” J. Acoust. Soc. Am. **40**(6), 1441–1449.

Burton, A. M. (**2013**). “Why has research in face recognition progressed so slowly? The importance of variability,” Q. J. Exp. Psychol. **66**(8), 1467–1485.

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

Cattell, R. B. (**1966**). “The scree test for the number of factors,” Multivar. Behav. Res. **1**(2), 245–276.

Cattell, R. B. (**1978**). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences* (Springer, Boston).

Cook, S., and Wilding, J. (**1997**). “Earwitness testimony: Never mind the variety, hear the length,” Appl. Cogn. Psychol. **11**(2), 95–111.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (The Hague: Mouton & Co).

Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in Rhesus Macaques," J. Acoust. Soc. Am. **102**(2), 1213–1222.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**(2), 1064–1077.

Hill, H., and Bruce, V. (**1996**). "Effects of lighting on the perception of facial surfaces," J. Exp. Psychol. Hum. Percept. Perform. **22**(4), 986–1004.

Hillenbrand, J., Cleveland, R., and Erickson, R. (**1994**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Lang. Hear. Res. **37**, 769–778.

Hirson, A., and Duckworth, M. (**1993**). "Glottal fry and voice disguise: a case study in forensic phonetics," J. Biomed. Eng. **15**(3), 193–200.

IEEE Subcommittee (**1969**). "IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements," IEEE Trans. Signal Process. **17**, 227–246.

Iseli, M., and Alwan, A. (**2004**). "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proceedings of IEEE ICASSP*, Vol. 1, pp. 10–13.

Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**(6), 3816–3822.

Jenkins, R., White, D., Van Montfort, X., and Burton, A. M. (**2011**). "Variability in photos of the same face," Cognition **121**(3), 313–323.

Kaiser, H. F. (**1960**). “The applications of electronic computer to factor analysis,” Educ. Psychol. Meas. **20**(1), 141–151.

Keating, P., and Kreiman, J. (**2016**). “Acoustic similarities among female voices,” J. Acoust. Soc. Am. **140**, 3393.

Keating, P. A., Kreiman, J., and Alwan, A. (**2019**). “A new speech database for within- and between-speaker variability,” in *Proceedings of the ICPhS XIX*.

Kramer, R. S., Jenkins, R., Young, A. W., and Burton, A. M. (**2017**). “Natural variability is essential to learning new faces,” Vis. Cogn. **25**(4-6), 470–476.

Kreiman, J., Auszmann, A., and Gerratt, B. (**2019**). “What does it mean for a voice to sound 'normal'?,” in *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, edited by M. Barkat-Defradas, B. Weiss, J. Trouvain, and J. Ohala (Springer, Singapore).

Kreiman, J., Gerratt, B., Garellek, M., Samlan, R., and Zhang, Z. (**2014**). “Toward a unified theory of voice production and perception,” Loquens **1**(1), 1–9.

Kreiman, J., Gerratt, B., Precoda, K., and Berke, G. (**1992**). “Individual differences in voice quality perception,” J. Speech Lang. Hear. Res. **35**(3), 512–520.

Kreiman, J., and Gerratt, B. R. (**1996**). “The perceptual structure of pathologic voice quality,” J. Acoust. Soc. Am. **100**(3), 1787–1795.

Kreiman, J., and Gerratt, B. R. (**2012**). “Perceptual interaction of the harmonic source and noise in voice,” J. Acoust. Soc. Am. **131**(1), 492–500.

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

Kreiman, J. E., Keating, P., and Vesselinova, N. (**2017**). “Acoustic similarities among voices. Part 2: Male speakers,” **142**, 2519.

LaRiviere, C. (**1975**). “Contributions of fundamental frequency and formant frequencies to speaker identification,” Phonetica **31**(3-4), 185–197.

Latinus, M., and Belin, P. (**2011**a). “Anti-voice adaptation suggests prototype-based coding of voice identity,” Front. Psychol. **2**(175), 1–12.

Latinus, M., and Belin, P. (**2011**b). “Primer: Human voice perception,” Curr. Biol. **21**(4), R143–R145.

Lattner, S., Meyer, M. E., and Friederici, A. D. (**2005**). “Voice perception: Sex, pitch, and the right hemisphere,” Hum. Brain Mapp. **24**(1), 11–20.

Lavan, N., Burston, L. F., and Garrido, L. (**2018**). “How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices,” Br. J. Psychol. **110**, S76–S93.

Lavan, N., Burston, L. F. K., Ladwa, P., Merriman, S. E., and Knight, S. (**2019**). “Breaking voice identity perception : Expressive voices are more confusable for listeners,” Q. J. Exp. Psychol. 1–9.

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.

Legge, G. E., Grosmann, C., and Pieper, C. M. (**1984**). “Learning unfamiliar voices,” J. Exp. Psychol. Learn. Mem. Cogn. **10**(2), 298–303.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). “Training japanese listeners to identify english /r/ and /l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. **94**, 1242–1255.

Maguinness, C., Roswandowitz, C., and von Kriegstein, K. (**2018**). "Understanding the mechanisms of familiar voice-identity recognition in the human brain," Neuropsychologia **116**, 179–193.

Murphy, J., Ipser, A., Gaigg, S. B., and Cook, R. (**2015**). "Exemplar variance supports robust learning of facial identity," J. Exp. Psychol. Hum. Percept. Perform. **41**(3), 577–581.

Murry, T., and Singh, S. (**1980**). "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am. **68**, 1294–1300.

Murry, T., Singh, S., and Sargent, M. (**1978**). "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am. **64**, 81–87.

O'Toole, A. J., Edelman, S., and Bülthoff, H. H. (**1998**). "Stimulus-specific effects in face recognition over changes in viewpoint," Vis. Res. **38**(15-16), 2351–2363.

Papçun, G., Kreiman, J., and Davis, A. (**1989**). "Long-term memory for unfamiliar voices," J. Acoust. Soc. Am. **85**(2), 913–925.

Patel, A. D. (**2008**). "Music and the brain: Three links to language," in *The Oxford Handbook of Music Psychology*, edited by S. Hallam, I. Cross, and M. Thaut, 1st ed. (Oxford Univ. Press, Oxford), pp. 208–216.

Patterson, K. E., and Baddeley, A. D. (**1977**). "When face recognition fails," J. Exp. Psychol. Learn. Mem. Cogn. **3**(4), 406–417.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., and Feinberg, D. R. (**2014**). "Vocal indicators of body size in men and women: a meta-analysis," Animal Behav. **95**, 89–99.

Read, D., and Craik, F. I. (**1995**). "Earwitness identification: Some influences on voice recognition," J. Exp. Psychol. Appl. **1**(1), 6–18.

Reich, A. R., and Duke, J. E. (**1979**). "Effects of selected vocal disguises upon speaker identification by listening," J. Acoust. Soc. Am. **66**(4), 1023–1028.

Reich, A. R., Moll, K. L., and Curtis, J. F. (**2005**). "Effects of selected vocal disguises upon spectrographic speaker identification," J. Acoust. Soc. Am. **60**(4), 919–925.

Ritchie, K. L., and Burton, A. M. (**2017**). "Learning faces from variability," Q. J. Exp. Psychol. **70**(5), 897–905.

Samlan, R. A., Story, B. H., and Bunton, K. (**2013**). "Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling," J. Speech Lang. Hear. Res. **56**(4), 1209–1223.

Saslove, H., and Yarmey, A. D. (**1980**). "Long-term auditory memory: Speaker identification," J. Appl. Psychol. **65**(1), 111–116.

Schweinberger, S. R. (**2001**). "Human brain potential correlates of voice priming and voice recognition," Neuropsychologia **39**(9), 921–936.

Schweinberger, S. R. (**2013**). "Audiovisual integration in speaker identification," in *Integrating Face and Voice in Person Perception*, edited by P. Belin, S. Campanella, and T. Ethofer (Springer Science + Business Media, New York), pp. 119–134.

Schweinberger, S. R., Herholz, A., and Sommer, W. (**1997**). "Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues," J. Speech Lang. Hear. Res. **40**(2), 453–463.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "VoiceSauce: A program for voice analysis," in *Proceedings of the ICPhS XVII*, pp. 1846–1849.

Smith, D., Patterson, R., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**(1), 305–318.

Stevenage, S. V., Neil, G. J., Parsons, B., and Humphreys, A. (**2018**). "A sound effect: Exploration of the distinctiveness advantage in voice recognition," Appl. Cogn. Psychol. **32**(5), 526–536.

Sun, X. (**2002**). "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in *Proceedings of IEEE ICASSP*, Vol. 1, pp. 333–336.

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

Thurstone, L. L. (**1947**). *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind* (University of Chicago Press, Chicago).

Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices," J. Phon. **13**, 19–38.

Wagner, I., and Köster, O. (**1999**). "Perceptual recognition of familiar voices using falsetto as a type of voice disguise," in *Proceedings of the ICPhS XI*, pp. 1381–1384.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M. (**1978**). "Correlates of psychological dimensions in talker similarity," J. Speech Lang. Hear. Res. **21**(2), 265–275.

Yovel, G., and Belin, P. (**2013**). "A unified coding strategy for processing faces and voices," Trends Cogn. Sci. **17**(6), 263–271.