Title
Acoustic voice variation in spontaneous speech

Permalink
https://escholarship.org/uc/item/64v578cc

Journal
The Journal of the Acoustical Society of America, 151(5)

ISSN
0001-4966

Authors
Lee, Yoonjeong
Kreiman, Jody

Publication Date
2022-05-01

DOI
10.1121/10.0011471

Peer reviewed

JASA ARTICLE

Acoustic voice variation in spontaneous speech

Yoonjeong Lee[a] and Jody Kreiman[b]

Department of Head and Neck Surgery, David Geffen School of Medicine at UCLA, Los Angeles, California 90095-1794, USA

ABSTRACT:

This study replicates and extends the recent findings of Lee, Keating, and Kreiman [J. Acoust. Soc. Am. **146**(3), 1568–1579 (2019)] on acoustic voice variation in read speech, which showed remarkably similar acoustic voice spaces for groups of female and male talkers and the individual talkers within these groups. Principal component analysis was applied to acoustic indices of voice quality measured from phone conversations for 99/100 of the same talkers studied previously. The acoustic voice spaces derived from spontaneous speech are highly similar to those based on read speech, except that unlike read speech, variability in fundamental frequency accounted for significant acoustic variability. Implications of these findings for prototype models of speaker recognition and discrimination are considered. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0011471

(Received 22 February 2022; revised 4 May 2022; accepted 9 May 2022; published online 25 May 2022)

[Editor: Susanne Fuchs] Pages: 3462–3472

I. INTRODUCTION

It is a truism in phonetics that no one ever says the same thing twice in precisely the same way. An individual's voice is shaped in part by the varying emotional, social, physiological, and linguistic states that person experiences in various contexts, and substantial acoustic variability arises within an individual's voice as a result. This within-talker variability co-exists with between-talker variability, and both listeners and machine recognition systems face the significant challenge of attributing the differences they hear across utterances to a change in talkers or a change within a single talker (Afshan *et al.*, 2022; Lavan *et al.*, 2019a; Reich and Duke, 1979; Saslove and Yarmey, 1980; Wagner and Köster, 1999). This suggests that listeners need to learn how a given voice varies in order to recognize it accurately and efficiently and must have knowledge of how qualities are distributed across a population if they are to successfully discriminate among unknown talkers. It follows that models of speaker recognition and discrimination, whether by humans or by machines, must account for the ways in which this occurs.

Despite the importance of talker variability in formulating models of voice quality and talker recognition, until recently little research addressed within-speaker variability in voice, and it is not known what characteristics distinguish one voice from another or how (or how much) these characteristics vary across different communicative settings. Phonetic studies have shown more within-speaker acoustic variability in spontaneous speech than in laboratory speech (e.g., DiCanio *et al.*, 2015; see Wagner *et al.*, 2015, for review), while additional studies suggest that f_{o} is rather stable within talkers for speech versus non-speech vocalizations (e.g., Pisanski *et al.*, 2020). However, these studies typically examine only a small number of acoustic attributes (for example, f_{o}, duration, or formant frequencies), and to our knowledge have not considered inter-speaker variability in different contexts. In an initial examination of vocal variability more broadly construed, Lee *et al.* (2019) used multiple tokens of read speech from 100 talkers to examine the acoustic characteristics that varied both within and across individuals. Results showed that the acoustic spaces that characterize vocal variability are remarkably consistent, both within and across talkers. However, this consistency may have occurred at least partly due to the constrained nature of the underlying voice samples used in these analyses (repeated readings of a fixed set of sentences). The present study thus attempts to replicate these findings and to extend them to samples of spontaneous conversation from the same talkers, to determine whether the nature and extent of within- and/or between-talker variability depends on speaking style.

Current models of voice perception and recognition posit that listeners evaluate individual voices based on the relationship between a given token and an abstract prototypical voice, a context-dependent "average" reference token residing at the center of a multidimensional acoustic voice space (Latinus and Belin, 2011; Lavner *et al.*, 2001; Papcun *et al.*, 1989; Yovel and Belin, 2013). While this idea is widely accepted, these models do not specify what characterizes the prototype(s), how within-speaker variability affects voice perception and/or recognition, or how prototypes fit within the structure of the acoustic voice space for individuals and populations of talkers. As a result, we do not know how voices are encoded with respect to prototypes, either for individual speakers or for populations of speakers.

Starting with the set of perceptually valid acoustic measures of voice quality proposed by Kreiman *et al.* (2014)

[a] Also at: Department of Linguistics, University of Michigan, Ann Arbor, MI 48109-1220, USA. Electronic mail: yoonjeonglee@ucla.edu

[b] Also at: Department of Linguistics, University of California, Los Angeles, Los Angeles, CA 90095-1543, USA.

and Kreiman *et al.* (2021), Lee *et al.* (2019) conducted a series of principal component (PC) analyses to assess which acoustic parameters within this set indeed form dimensions of an acoustic space specific to an individual voice, as well as a global acoustic voice space for the population of talkers. Among the large array of vocal parameters available for each individual voice, a few parameters—which together quantified formant dispersion and the balance between high-frequency harmonic and inharmonic energy in the voice—consistently emerged in the first three PCs for all talkers, but most within-speaker acoustic variability in voice was idiosyncratic. Results further showed that the measures that varied in the speech of all individual talkers also characterized voice variation across talkers, suggesting that individual and population voice spaces have very similar acoustic structures.

The Lee *et al.* (2019) analyses used multiple sentence productions from 100 English speakers (50 female and 50 male speakers). Read sentences clearly do not represent the full range of acoustic variability that occurs within a talker in an average day's phonation. In this follow-up study, we examined how well findings from read speech generalize to spontaneous speech from the same set of talkers. Because the same parameters characterized acoustic variability for virtually every speaker in Lee *et al.* (2019), we hypothesized that these parameters would also emerge from parallel analyses of spontaneous speech for the same speakers, although the precise nature of the PCs may vary due to the greater acoustic variability usually observed for spontaneous utterances (Nakamura *et al.*, 2008; Lieberman *et al.*, 1985; Wagner *et al.*, 2015).

II. METHOD

A. Voice samples

This study used a set of recorded phone conversations from 99 speakers (49 females, 50 males; self-reported) out of the 100 originally analyzed in Lee *et al.* (2019). Voice samples were drawn from the University of California, Los Angeles Speaker Variability Database (Keating *et al.*, 2019; Keating *et al.*, 2021), which offers significant within- and between-speaker variability. All speakers were native speakers of English, with no known speech or hearing impairments, and were members of the UCLA community at the time of recording (Age range: F, 18–29; M, 18–26). Speakers were recorded in a sound-attenuated booth at a sampling rate of 22 kHz, using a Brüel & Kjær $\frac{1}{2}''$ microphone (Hottinger Brüel & Kjær A/S, Germany) (model 4193) firmly attached to a baseball cap worn by the speaker. All audio recordings in the database are accompanied by transcriptions in the form of Praat TextGrids. The recordings were first force-aligned, and the force-aligned segmentations were individually checked and manually corrected to provide precise alignments.

Among the speech tasks available in the database, recordings of informal telephone calls were used in this study. Speakers used their own cell phones to call and chat with a friend or family member for at least two minutes. Only the speaker's side of the conversation was recorded directly from the speaker's mouth, not *via* the telephone line.

B. Acoustic measurements and post-processing steps

Any non-speech items (sighs, laughs, coughs) were removed prior to acoustic analysis. In order to compare the acoustic variability of spontaneous speech to that reported for read speech (Lee *et al.*, 2019), the 26 acoustic variables previously measured were again obtained from all vowels and approximants (/l/, /r/, /j/, and /w/) in each recording (Table I). These variables form a psychoacoustic model of voice quality (Kreiman *et al.*, 2021), and as a set have been shown to adequately quantify the quality of virtually all samples of normal and disordered voice. In addition, to quantify time-varying changes in voice quality in continuous speech we calculated moving coefficients of variation ($moving\ CoV = moving\ \sigma / moving\ \mu$) (Kreiman *et al.*, 2003), using a smoothing window of 50 ms. Variables were then grouped into five categories: (i) f_o; (ii) formant frequencies (F_1, F_2, F_3, F_4) and formant dispersion (FD) (Fitch, 1997), calculated as the average frequency interval between immediately adjacent pairs of formants; (iii) spectral noise (cepstral peak prominence, CPP) (Hillenbrand *et al.*, 1994), the root mean square energy calculated over five pitch pulses (energy) and the amplitude ratio between subharmonics and harmonics (SHR) (Sun, 2002); (iv) harmonic source spectral shape (H1*–H2*, H2*–H4*, H4*–H2kHz*, H2kHz*–H5kHz); and (v) the coefficients of variation for all measures (CoVs) (Table I). All variables were measured automatically every 5 ms using VoiceSauce software (Shue *et al.*, 2009).

After removing data frames with spurious parameter values (e.g., impossible 0 s; data trimming removed less than 0.01% of the data), values of each acoustic 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 values across variables and talkers ranged from 0 to 1. Finally, moving coefficients of variation for all 13 variables were calculated for each complete conversation and separately for each sentence or full utterance in a conversation. These two different analysis scopes were initially examined separately because it was unclear which would be more appropriate for comparing acoustic spaces for spontaneous speech to those for read sentences (Lee *et al.*, 2019): Full conversations may better represent a speaker's detailed acoustic space, while individual sentences are a better match to the sentence stimuli used

TABLE I. Acoustic variables. Harmonic amplitudes marked with * have been corrected for the influence of formants (Hanson and Chuang, 1999; Iseli and Alwan, 2004).

Variable categories	Acoustic variables
Pitch	f_o
Formant frequencies	F_1, F_2, F_3, F_4, 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

in our previous study. Across speakers, these steps resulted in about 4×10^6 data frames (sentences, F: 2116k, M: 2021k; conversations, F: 2131k, M: 2039k).

C. Principal component analysis (PCA)

The procedure for PCA followed that described in full by Lee *et al.* (2019). By employing an oblique rotation (Cattell, 1978; Thurstone, 1947), the 26 variables—moving averages for 13 variables + moving coefficients of variation for the same 13 variables—were simultaneously entered into PCAs, which reduced the data into a smaller set of PCs (factorial solutions with eigenvalues greater than 1; Kaiser, 1960), each of which was formed by variables with loadings (weights) at or exceeding 0.32 (Tabachnick and Fidell, 2013). PCAs were conducted separately for each speaker (within-speaker analyses), and separate combined speaker analyses were also run for groups of the 50 male and 49 female speakers.

D. Degrees of acoustic variation in conversational versus read speech

Finally, to assess the extent to which talker variability depends on speaking style, we examined the small subsets of acoustic variables that differentiated the two speaking styles in the PCA solutions (see Secs. III C and III D). The normalized values of each of these measures were analyzed independently for each gender group using linear mixed effects models. All statistical analyses were made in R (R Core Team, 2021) using the lme4 package (Bates *et al.*, 2015). Style (Conversation vs Reading) was entered in the analysis as a fixed effect and random slopes were fit for speakers in each group (1+Style|Speaker). The random effects of speakers significantly improved model fit according to likelihood ratio tests (Baayen, 2008; $p < 0.05$ for all). *P*-values less than or equal to 0.05 were considered significant.

III. RESULTS

Analyses using sentences excerpted from conversations versus complete conversations produced highly similar acoustic spaces. The same variables largely emerged within the same PCs for each speaker and each speaker group, albeit with differences in weights for the individual variables in each PC. For this reason, we report only results for analyses of complete phone calls in this paper.

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

A total of 6–9 PCs emerged in analyses for individual speakers. For most speakers, seven (48/99 speakers) or eight (47/99 speakers) PCs were extracted. These components accounted for 64%–72% (F, mean = 68%; M, mean = 67%) of the cumulative acoustic variance. As the higher order PCs after PC6 accounted for very small amounts of acoustic variability, only the first six are reported in detail here.

We first counted the number of times each acoustic category appeared in a within-speaker solution, cumulated across speakers in each group (F: 49 speakers, M: 50 speakers). Figure 1 shows the distribution of variables with respect to weight in the first six components.

The first PC accounted for 20%–24% (mean = 21%) and 20%–29% (mean = 25%) of the variance for females and males, respectively. For both females and males, this PC represented the combination of variability in source spectral shape (with heaviest weights on H2kHz*–H5kHz CoV), in spectral noise (heaviest weight on CPP CoV), and in f_o CoV; (F, 44/49 speakers; M, 34/50 speakers; green, yellow, and red bars in the second and fourth columns of Fig. 1, respectively). For a subset of speakers, variability in lower formant frequencies—F_1 and/or F_2 CoV—also emerged in this PC, albeit with low weights (orange bars in the second and fourth columns of Fig. 1; F, 26/44 speakers; M, 9/34 speakers). For most of the remaining speakers (F, 4/49 speakers; M, 14/50 speakers), formant frequency CoV was the most representative variable in the first component, with highest weights on formant dispersion CoV and F_4 CoV (orange bars in the second and fourth figure panels). Last, acoustic variability for one female speaker and two male speakers was mostly related to source spectral shape.

PC2 accounted for an average of 12% and 11% of acoustic variability, for female and male speakers, respectively (ranges: F = 10%–19%, M = 9%–14%). Across speakers, formant frequencies (F, 49/49 speakers; M, 50/50 speakers) emerged most frequently as the second PC (Fig. 1). However, sub-analyses revealed a difference in distribution patterns between female and male speaker groups. For most female speakers, formant dispersion, F_4 and F_3 were most important and consistently appeared together (F, 32/49 speakers). Among these variables, formant dispersion predominated (weights: FD > F_4 > F_3). For 12 other female speakers, F_2 and H4*–H2kHz* were the most important variables in the second component and always emerged together. For the remaining five speakers, the combination of F_1 and H2*–H4* was most important in this PC.

We observed the opposite pattern for male speakers, for most of whom the combination of F_2 and H4*–H2kHz* emerged as the most important variables for this PC (M, 33/50 speakers). Formant dispersion and higher formant frequencies explained the most variance for a smaller group of speakers (M, 16/50 speakers). Finally, one male speaker showed the combination of F_1 and H2*–H4* as the most heavily weighted variables for this PC.

Across speakers PC3 accounted for an additional 9% of acoustic variance (range = 7%–11%) and weighted mainly on formant frequencies (49/49 for females, 48/50 for males). The variables that emerged in this PC mirrored those in the second PC, but with complementary distributions for the two speaker groups. For most female speakers, weights were highest on either a combination of F_1 and H2*–H4* (23/49 speakers) or F_2 and H4*–H2kHz* (18/49 speakers), while for the remaining eight speakers higher formant frequencies and FD emerged as the most important variables within this PC. The largest group of male speakers (20/50

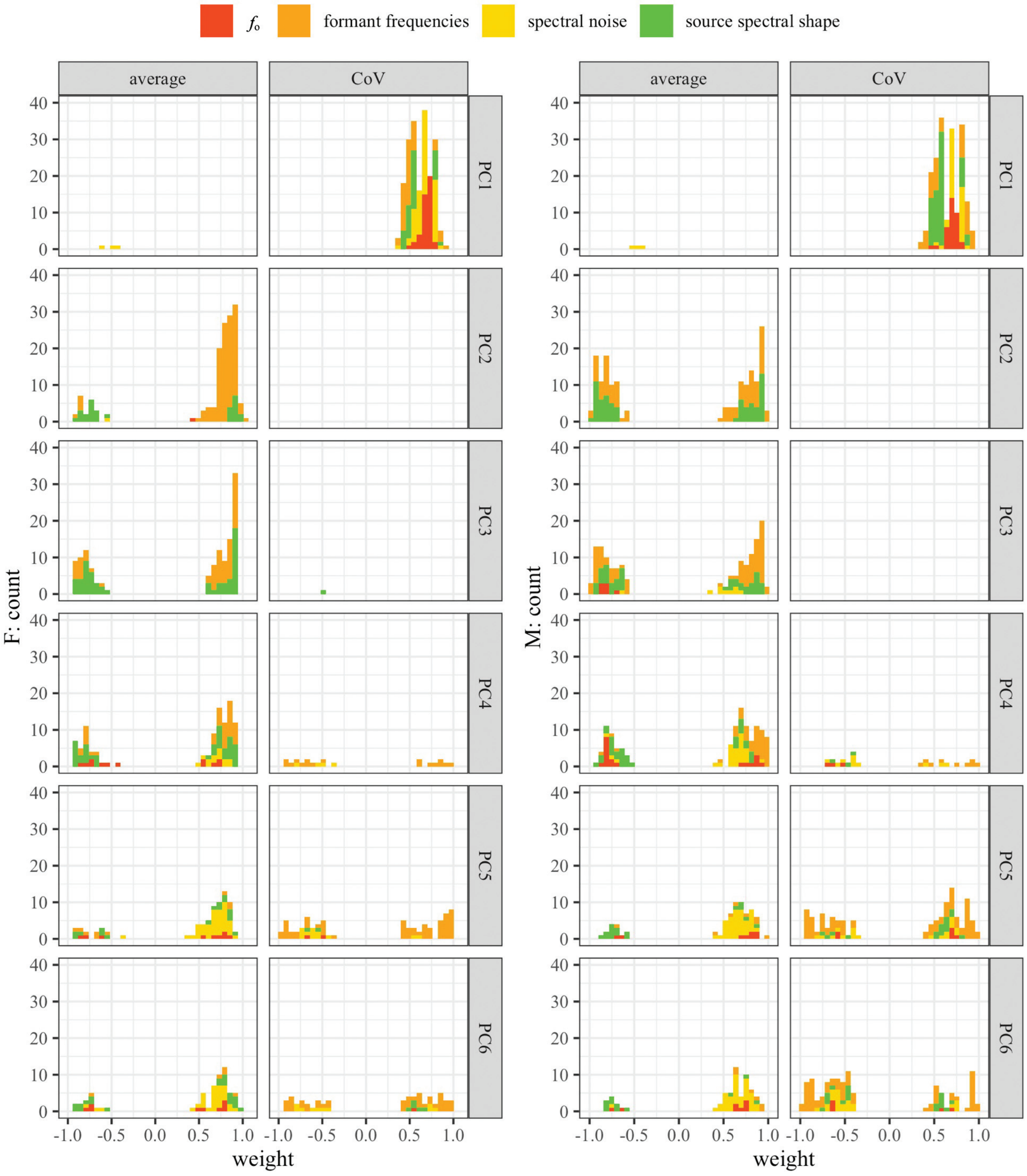

FIG. 1. Distribution of acoustic parameters plotted (stacked histogram) against the rotated component loadings (weight) for the first 6 PCs derived from measures of continuous speech. Left panel, female speakers. Right panel, male speakers. CoV, coefficient of variation.

speakers) showed higher formant frequencies and FD as the most important contributors to PC3. The combination of F_1 and H2*–H4* weighed most heavily for an additional 14/50 speakers. Seven speakers showed the combination of F_2 and H4*–H2kHz* as the most important variables, and another seven speakers showed f_o as the most important variable. Last, PC3 for two male speakers weighed primarily on H1*–H2*, Energy, and CPP.

PCs above the third combined to account for an average of 21% (female speakers) to 22% (male speakers) of the acoustic variance in the data, but in contrast to the first three PCs, this variance was largely idiosyncratic, and no particular acoustic category predominated (Fig. 1). For PC4–PC6, the distributions of the variables and 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. In other words, the component in which each variable appeared differed across individuals, ranging from PC4 to the last PC (PC6–PC9) across individuals, and no single component accounted for substantial variance.

In sum, variability (measured by coefficients of variation) in f_o, source spectral shape, and spectral noise (especially in H2kHz*–H5kHz* and CPP) accounted for the most acoustic variability within conversational speech from individual talkers. Across talkers, the next most frequently emerging variables were means for formant dispersion and the combination of F_2 and H4*–H2kHz*. Additionally, F_1 and H2*–H4* were important for some speakers. The first three PCs were largely shared across voices and together accounted for most of the explained variance in the underlying acoustic data (40%–50% total). The remaining PCs differed widely across voices and cumulatively accounted for about 21% of the explained variance (17%–30% total).

B. Between-speaker group PCA: Population voice spaces

A second set of PCAs examined the structure of the acoustic space for the combined groups of female and male speakers. Nine PCs were extracted for the female speaker group, and seven PCs were extracted for the male speaker group, accounting for 71% of the cumulative variance for female speakers and 64% for male speakers. Details of the analyses are included in the Appendix. Patterns of acoustic variability in these multi-talker spaces largely mirrored the patterns found within speakers (Figs. 2 and 3). The first component was composed of variability (measured by CoVs) in source spectral noise, spectral shape, and f_o, accounting for 22% and 24% of variance across females and males, respectively. The individual components of this PC were similarly weighted (except for F_1 CoV for the female group).

The second component accounted for 11% of acoustic variance in female voices and corresponded to formant frequencies (F_4, FD, F_3). For male speakers, spectral slope in the higher frequencies (H4*–H2kHz*, H2kHz*–H5kHz) and F_2 accounted for 11% of variance in the combined acoustic data. For the third component, an additional 9% of the variance was accounted for by spectral shape in the lower frequencies (H2*–H4*) and F_1 for female speakers; formant frequencies (F_4, FD, F_3) accounted for 9% of the variance for male speakers.

C. Differences between acoustic voice spaces in conversation and reading

Table II shows all the variables that emerged in the first three PCs from the group solutions in conversation and reading. While the earlier PCs were very similarly structured for the two speaking styles across speakers, some additional measures of variability emerged only in conversations (shown in bold in the table). For female speakers, four additional CoV measures (H1*–H2* CoV, f_o CoV, energy CoV, and F_1 CoV) emerged in the first PC for conversational speech. For male speakers, energy CoV, f_o CoV, and F_2 CoV measures additionally emerged in PC1 and PC2 in conversational speech. Full details of PCAs for reading can be found in Lee *et al.* (2019).

D. Speaking style effect on acoustic variability in the voice

Overall, measured variables varied much more for conversational speech than for read sentences. Figure 3 shows the effects of speaking styles on variables that emerged only in conversation from the PC analysis above (shown in bold in Table II). Variable ranges (whiskers) and quartile ranges (boxes) were wider, and more outliers (circles) were observed.

For female speakers, styles differed significantly for all variables except for f_o ($\chi^2[1] = 2.94$, $p = 0.086$). Energy was greater in conversation than in reading (median: 0.0074 vs 0.0069, $\chi^2[1] = 4.49$, $p < 0.05$). F_1 and H1*–H2* varied more in conversation than in reading (for F_1, standard deviation (SD): 0.16 vs 0.14, $\chi^2[1] = 24.4$, $p < 0.0001$; for H1*–H2*, SD: 0.097 vs 0.063, $\chi^2[1] = 74.82$, $p < 0.0001$), as indicated by the wider boxes (in red) in Fig. 4.

For male speakers, F_2 was higher for conversational speech than for reading (median: 0.39 vs 0.34; $\chi^2[1] = 90.85$, $p < 0.0001$). While energy and f_o did not differ significantly between the two speaking styles (energy: $\chi^2[1] = 0.45$, $p = 0.5$; f_o: $\chi^2[1] = 0.25$, $p = 0.62$), these measures varied more in conversations compared to read speech as shown in outlier distributions in Fig. 4.[1]

IV. DISCUSSION AND CONCLUSION

This study employed PCA to acoustic measures of unscripted phone conversations to determine how the acoustic voice spaces derived from spontaneous speech compared to those for sentences read by the same talkers (Lee *et al.*, 2019). Results from spontaneous speech largely replicated those from sentence reading. In both cases, the most acoustic variance within talkers, regardless of gender, was accounted for by a set of variability measures. Variability in higher-frequency harmonic and inharmonic energy, which often covary (Kreiman and Gerratt, 2012) and are often associated with the degree of perceived breathiness or brightness in the voice (Samlan *et al.*, 2013), was consistently associated with the first component, and formant dispersion was similarly associated with the second component. We note that these acoustic variables are associated with important biological and social traits across many species, including sex, body size, arousal, and dominance (Anikin, 2020; Congdon *et al.*, 2019; Fitch, 1997). Their repeated emergence from our analyses is consistent with an evolutionary basis for these aspects of vocal variability (see Lee *et al.*, 2021, or Lee and Kreiman, 2022 for more discussion of the biological versus social origins of systematic vocal variability).

One notable difference between these results and our previous findings was the emergence of variability in f_o from analyses of spontaneous speech, but not from read speech. This finding is not especially surprising, and probably reflects the fact that sentence reading tends to be highly stylized, with similar patterns and amounts of variability across speakers, while in free conversation the amount and kind of variability present can alter from talker to talker and

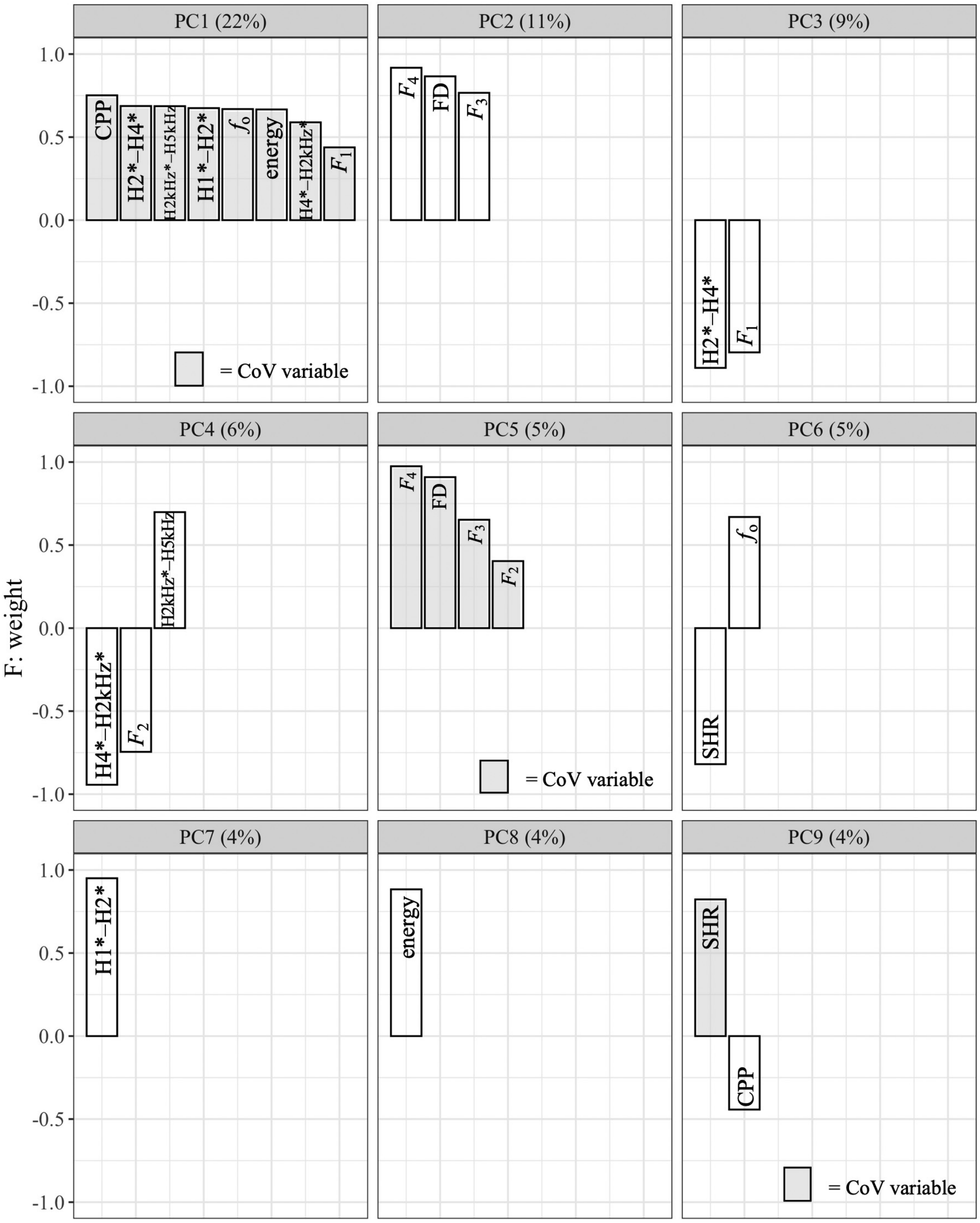

FIG. 2. Acoustic parameters emerging in nine PCs for the female speaker group. Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. Shaded bars represent CoV variables. (N%), variance explained. CoV, coefficient of variation.

from moment to moment within talkers (Lavan *et al.*, 2019b; Lieberman *et al.*, 1985; Nakamura *et al.*, 2008). The varied length of the utterances in telephone conversation (whether shorter or longer) compared to the fixed length of the utterances in sentence reading might be another factor introducing more variance in measured variables. However, variability in spectral shape and in noise emerged from both read and spontaneous speech, indicating that although the nature of variability may differ across speaking styles, it is variability in general that best distinguishes voices.

Again paralleling results from read speech, the next most important measures distinguishing talkers were formant dispersion (the average interval between formant frequencies) and higher formant frequencies, which are considered relatively independent of vowel quality (Fant, 1960) but are often associated with speaker identity (e.g., Fitch, 1997; Pisanski *et al.*, 2014; Ives *et al.*, 2005; Smith *et al.*, 2005). These body-size related parameters emerged in the second and third PCs, for female speakers and male speakers, respectively. Variables related to vowel quality (i.e., lower formant frequencies and their nearby harmonics) were also important for many speakers as these variables emerged in the first three PCs (predominantly in PC2 and PC3). Across speaking styles, F_2 was important for many speakers (see Table II). Additionally, the

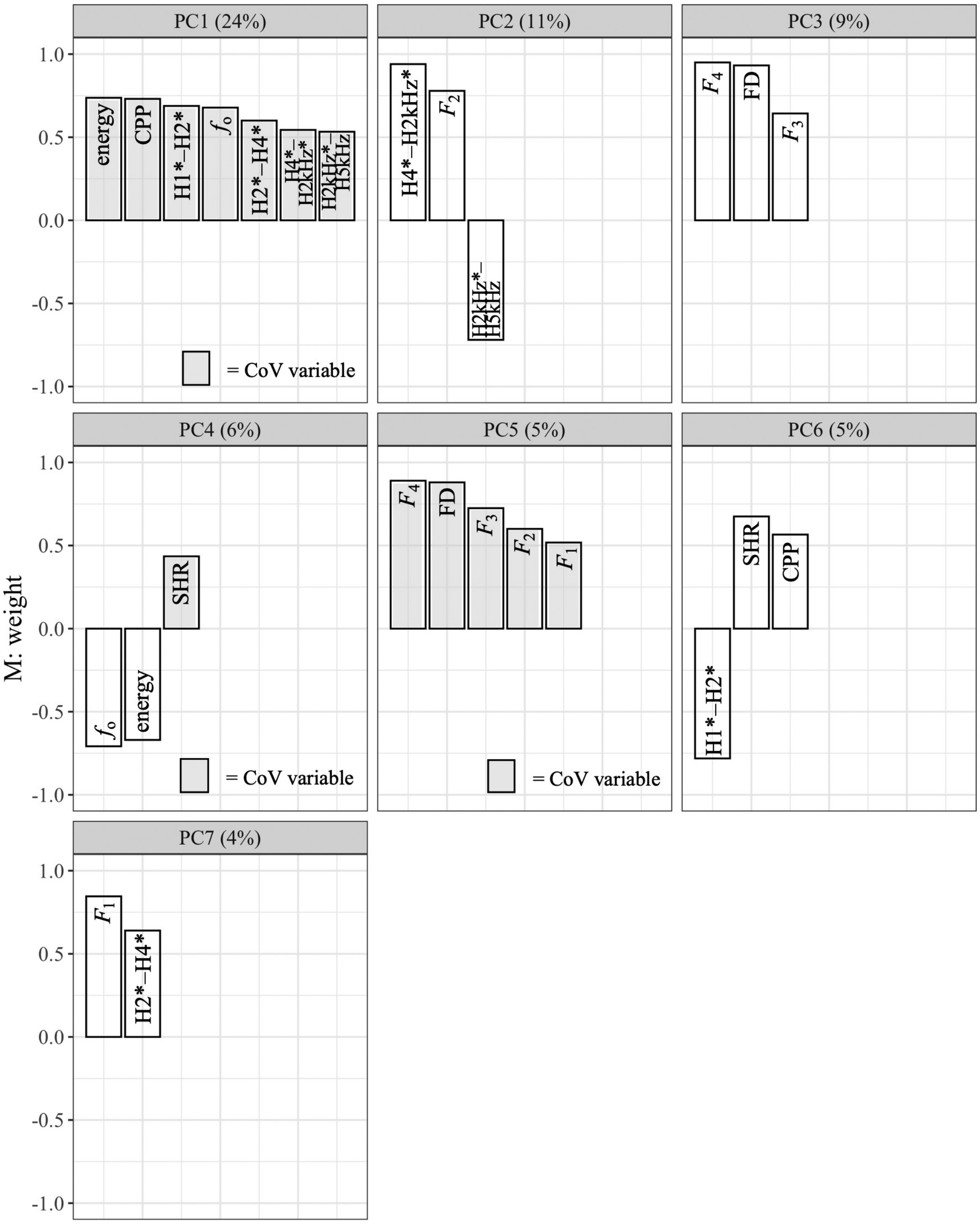

FIG. 3. Acoustic parameters emerging in seven PCs for the male speaker group. Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. Shaded bars represent CoV variables. (N%), variance explained. CoV, coefficient of variation.

vowel productions in spontaneous speech varied even more for the female speakers as indicated by both F_1 and F_2 emerging early (in PC3 and PC4) for female speakers. The more varied vowel productions observed in conversations than in read sentences may arise from the difference in the segmental context causing varied productions of each vowel category (due to coarticulation between the vowel and the neighboring segments). Finally, in both sets of analyses, higher PCs representing detailed patterns of acoustic variability in the talker-specific voice spaces were idiosyncratic.

If acoustic voice spaces are in fact characterized by a few shared dimensions and a mass of idiosyncratic detail, what are the implications for voice discrimination and recognition? The fact that acoustic spaces for individual talkers are both low dimensional and highly similar suggests that listeners could evaluate voice quality using a "quick and dirty" algorithm to locate a given voice in the population space. This first step may make it easy to compare one talker's voice to another's, and may reflect listeners' lifelong experience with this simple voice space. Indeed, listeners are excellent at "telling voices apart" (Lavan *et al.*, 2019b). Logically, the next step in processing voice quality is then to address the idiosyncratic details that characterize an individual talker. However, building mental models of individual

TABLE II. Acoustic parameters emerging in the first three PCs in different speaking styles (conversation vs reading) for the female speaker group (left two columns) and for the male speaker group (right two columns). Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. Variables that emerged *only* in conversation are **in bold**. Gray cells for PC3 of the female speaker group indicate parameters that actually emerged in PC4 in different speaking styles. CoV, coefficient of variation.

PC	Conversation (female group)	Reading (female group)	Conversation (male group)	Reading (male group)
1	CPP CoV H2*–H4* CoV H2kHz*–H5kHz CoV **H1*–H2* CoV** **f_o CoV** **energy CoV** H4*–H2kHz* CoV **F_1 CoV**	H2kHz*–H5kHz CoV CPP CoV H4*–H2kHz* CoV H2*–H4* CoV	**energy CoV** CPP CoV H1*–H2* CoV **f_o CoV** H2*–H4* CoV H2kHz*–H5kHz CoV H4*–H2kHz* CoV	H2kHz*–H5kHz CoV CPP CoV H1*–H2* CoV H2*–H4* CoV H4*–H2kHz* CoV
2	F_4 FD F_3	F_4 FD F_3	H4*–H2kHz* F_2 H2kHz*–H5kHz **F_2 CoV**	H4*–H2kHz* F_2 H2kHz*–H5kHz
3	H2*–H4* F_1	H4*–H2kHz* F_2 H2kHz*–H5kHz F_2 CoV	F_4 FD F_3	F_4 FD F_3

voice spaces requires substantial experience with specific talkers. Lacking wide familiarity with how a person's voice varies across changing contexts, it is challenging to recognize that two voice samples come from the same talker ("telling voices together"; Johnson *et al.*, 2020; Lavan *et al.*, 2019b). Consistent with this view, recent evoked potential data (Plant-Hébert *et al.*, 2021) showed that responses to familiar and unfamiliar voices vary with time window as well as familiarity status, implying a two-step process beginning with the evaluation of shared features, and moving on to incorporation of idiosyncratic features as needed for familiar voices.

The increased variability in conversations for the same voices leads to the prediction that it may be harder to tell speakers together *and* apart from these utterances. Consistent with this prediction, Afshan *et al.* (2022) reported that listeners had greater difficulty in discriminating a subset of these voices from conversation than from read sentences, and when utterances were produced in different styles than when styles were matched. Moderate speaking style variability especially made the "telling voices together" task harder than the "telling voices apart" task. They further reported that the listeners attended to speaker-specific idiosyncrasies when "telling speakers together," and that they "tell speakers apart" based on their relative positions within a shared acoustic space. It remains to be explored whether the stylistically different samples of the same voices are close together in individual and group voice spaces. Analysis of which specific voices should be easier to tell together/tell apart for both kinds of speaking styles based on our acoustic results is under way.

One possible limitation of this study is the relative homogeneity of the speakers with respect to age, health, native language, and speaking style, which may have contributed to the consistency in voice spaces. Results from ongoing analyses of recordings from speakers of Seoul Korean and Hmong (Lee *et al.*, 2021; Lee and Kreiman, 2022) argue against this possibility. The same few PCs emerged first from these new analyses of different speakers as well, along with additional PCs that reflected the status of f_o and phonation in the phonology of the language. Ongoing work with additional languages and speakers (including those with vocal pathology; Lee and Kreiman, 2021) will continue to address this issue. Other sources of acoustic variability, including emotion, are represented in our spontaneous speech samples by virtue of the topics speakers chose but were not explicitly manipulated. Given the robustness of our findings, we again predict that these factors will add acoustic variability, but that the same PCs will emerge first from analyses of emotional speech. This hypothesis remains to be tested.

Taken together, our results have implications for prototype-based models of voice perception (Kreiman and Sidtis, 2011; Lavner *et al.*, 2001; cf. Yovel and Belin, 2013; see Lee *et al.*, 2019 for detailed discussion). In our earlier paper (Lee *et al.*, 2019) we argued that a common set of variables shared by virtually all talkers, accompanied by unique deviations from that central pattern, is consistent with what might be required as input to a recognition system organized around prototypes, accounting for both between and within talker acoustic variability. However, authors are often quite vague with respect to what they mean by "prototype" in this context. Definitions typically suggest some kind of average token in a space, but it generally remains unclear whether that space is perceptual or acoustic or something else, and what its dimensions might be is typically also unstated (see Yovel and Belin, 2013). If we define a voice space in terms of acoustic variables, with dimensions that correspond to the first three PCs describing voice variability, our data suggest

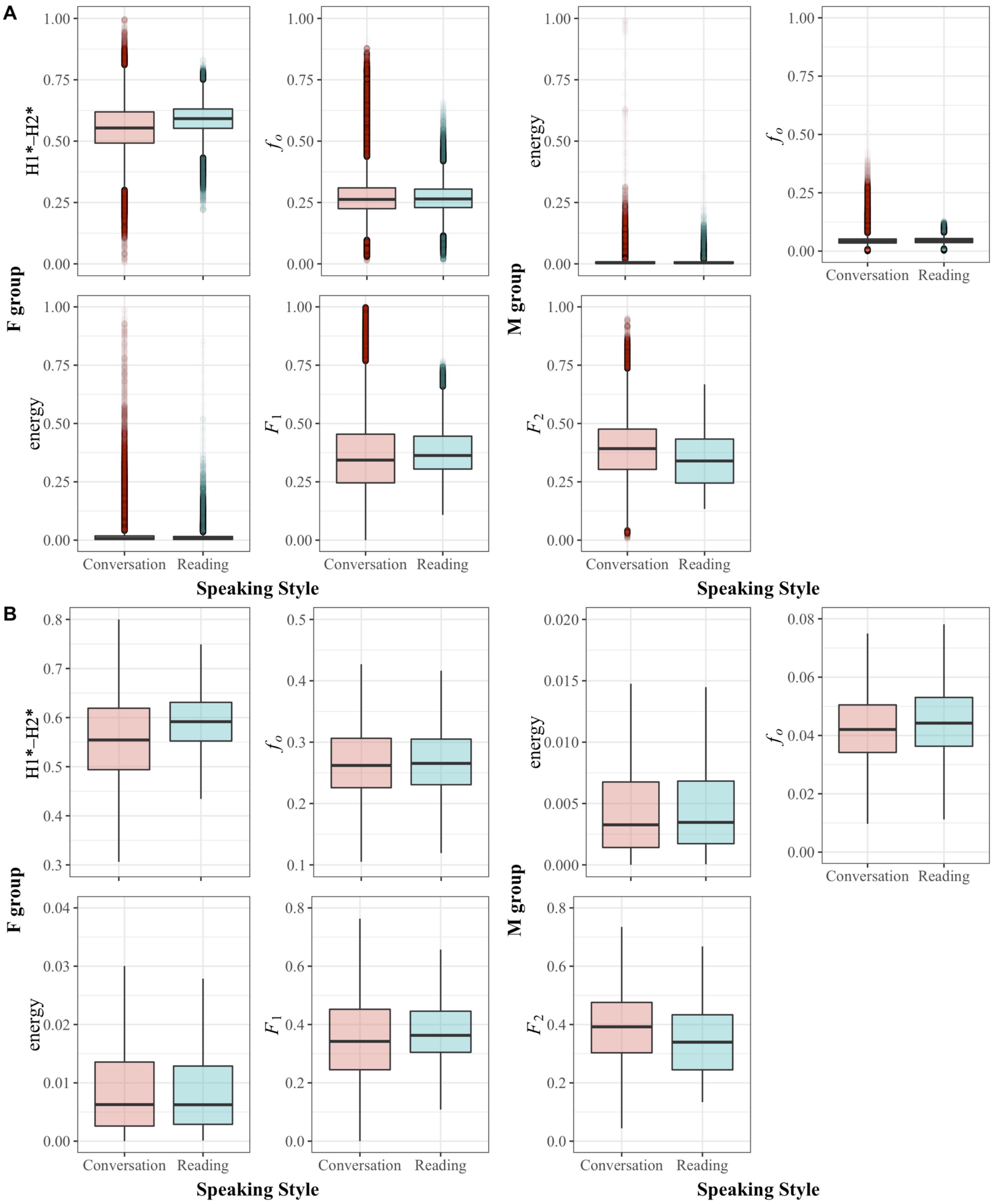

FIG. 4. (Color online) Effects of speaking style (conversation vs reading) on normalized measured variables for female speaker group (left panel) and male speaker group (right panel). The entire data range including outliers is shown in (A), and the range excluding outliers is shown in (B). The horizontal line within each box represents the median value.

two ways to map prototypes into these acoustic voice spaces. One (the more traditional version) would be to define a specific token as the average of those acoustic parameters. As an average, this token would sit in the center of the acoustic space defined by the dimensions that emerged as PCs in our analyses. However, if the center of the space is defined by the shared dimensions, and if the space characterizes all or almost all voices, there may be no need to hypothesize the existence of a token that sits in that position–the token is already specified by the space, and is (hypothetically) part of what listeners know by virtue of their experience with voices. In other words, if individual and group voice spaces are similarly structured with respect to a very small set of acoustic attributes, it is possible that prototypes might not be "average tokens" computed across complete [sets of] acoustic signals, because "average" is defined as the center of the shared acoustic space, rather than by the acoustic mean of many tokens from a single talker or a population of talkers. In this way, voice "prototypes" may be artifacts of a shared voice space, and the separate concept of a prototype may not be needed to explain listeners' behavior. More information about what listeners actually know about voice acoustics is needed to shed light on the relationship between voice waveforms and perceptual processes.

ACKNOWLEDGMENTS

This work was supported by NIH grant DC01797 and NSF grant IIS-1704167. NIH grant DC011300 and NSF grant SPRF-2105410 supported Y.L. for manuscript preparation and revision. We thank Associate Editor Dr. Susanne Fuchs, Dr. Kasia Pisanski, and two anonymous reviewers for their insightful comments on an earlier version of this paper. We also thank Alysa Wang and Nora Alwash for help with data coding.

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

(1) PCA pattern matrix for female speaker group analysis.

PC (variance explained)	Variables	Weight
1 (22%)	CPP CoV	0.75
	H2*–H4*CoV	0.69
	H2kHz*–H5kHz CoV	0.69
	H1*–H2* CoV	0.67
	f_o CoV	0.67
	energy CoV	0.67
	H4*–H2kHz* CoV	0.59
	F_1 CoV	0.44
2 (11%)	F_4	0.92
	FD	0.87
	F_3	0.77
3 (9%)	H2*–H4*	−0.89
	F_1	−0.80
4 (6%)	H4*–H2kHz*	−0.94
	F_2	−0.75
	H2kHz*–H5kHz	0.70
5 (5%)	F_4 CoV	0.97
	FD CoV	0.91
	F_3 CoV	0.65
	F_2 CoV	0.40
6 (5%)	SHR	−0.82
	f_o	0.67
7 (4%)	H1*–H2*	0.95
8 (4%)	energy	0.88
9 (4%)	SHR CoV	0.82
	CPP	−0.44

(2) PCA pattern matrix for male speaker group analysis.

PC (variance explained)	Variables	Weight
1 (24%)	Energy CoV	0.74
	CPP CoV	0.73
	H1*–H2* CoV	0.69
	f_o CoV	0.68
	H2*–H4*CoV	0.60
	H2kHz*–H5kHz CoV	0.54
	H4*–H2kHz* CoV	0.53
2 (11%)	H4*–H2kHz*	0.94
	F_2 CoV	0.78
	H2kHz*–H5kHz	−0.72
3 (9%)	F_4	0.95
	FD	0.93
	F_3	0.64
4 (6%)	f_o	−0.71
	energy	−0.67
	SHR CoV	0.43
5 (5%)	F_4 CoV	0.89
	FD CoV	0.88
	F_3 CoV	0.72
	F_2 CoV	0.60
	F_1 CoV	0.52
6 (5%)	H1*–H2*	−0.78
	SHR	0.67
	CPP	0.57
7 (4%)	F_1	0.85
	H2*–H4*	0.64

[1] In addition, to assess effects of the large difference in sample size across speaking styles (almost ten times larger for the phone call speech data than the sentence reading data), an additional post-hoc analysis was conducted across different subsets of data. For each speaker, we extracted several consecutive chunks of spontaneous speech from the beginning, the middle, and the end of a conversation that matched the size of the read speech for comparisons. Across different subsets of data, patterns conformed to the main finding.

Afshan, A., Kreiman, J., and Alwan, A. (**2022**). "Speaker discrimination performance for 'easy' versus 'hard' voices in style-matched and -mismatched speech," J. Acoust. Soc. Am. **151**(2), 1393–1403.

Anikin, A. (**2020**). "A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations," Phonetica **77**(5), 327–349.

Baayen, R. H. (**2008**). *Analyzing Linguistic Data: A Practical Introduction to Statistics* (Cambridge University Press, Cambridge, UK).

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.

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

Congdon, J. V., Hahn, A. H., Filippi, P., Campbell, K. A., Hoang, J., Scully, E. N., Bowling, D. L., Reber, S. A., and Sturdy, C. B. (**2019**). "Hear them roar: A comparison of black-capped chickadee (*Poecile atricapillus*) and human (*Homo sapiens*) perception of arousal in vocalizations across all classes of terrestrial vertebrates," J. Comp. Psychol. **133**(4), 520–541.

DiCanio, C., Nam, H., Amith, J. D., García, R. C., and Whalen, D. H. (**2015**). "Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec," J. Phon. **48**, 45–59.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, the Netherlands).

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.

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.

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 International Conference on Acoustics, Speech, and Signal Processing*, May 17–21, Quebec, Canada, 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.

Johnson, K. A., Babel, M., and Fuhrman, R. A. (**2020**). "Bilingual acoustic voice variation is similarly structured across languages," in *Proceedings of Interspeech 2020*, October 25–29, Shanghai, China, pp. 2387–2391.

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

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. 737–739.

Keating, P., Kreiman, J., Alwan, A., Chong, A., and Lee, Y. (**2021**). "UCLA speaker variability database," *LDC2021S09* (Linguistic Data Consortium, Philadelphia, PA).

Kreiman, J., Auszmann, A., and Gerratt, B. R. (**2021**). "What does it mean for a voice to sound "normal?," in *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, edited by B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. Ohala (Springer, Singapore), pp. 83–100.

Kreiman, J., Gabelman, B., and Gerratt, B. R. (**2003**). "Perception of vocal tremor," J. Speech. Lang. Hear. Res. **46**(1), 203–214.

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., 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**, 457–465.

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

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

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

Lavan, N., Burston, L. F. K., Ladwa, P., Merriman, S. E., and Knight, S. (**2019a**). "Breaking voice identity perception: Expressive voices are more confusable for listeners," Q. J. Exp. Psychol. **72**(9), 2240–2248.

Lavan, N., Burton, A. M., Scott, S. K., and McGettigan, C. (**2019b**). "Flexible voices: Identity perception from variable vocal signals," Psychon. Bull. Rev. **26**(1), 90–102.

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, Y., Garellek, M., Esposito, C., and Kreiman, J. (**2021**). "A cross-linguistic investigation of acoustic voice spaces," J. Acoust. Soc. Am. **150**, A191.

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

Lee, Y., and Kreiman, J. (**2021**). "Acoustic spaces for normal and pathological voices," J. Acoust. Soc. Am. **150**, A191.

Lee, Y., and Kreiman, J. (**2022**). "Linguistic and personal influences on speaker variability," J. Acoust. Soc. Am. **151**, A62.

Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (**1985**). "Measures of the sentence intonation of read and spontaneous speech in American English," J. Acoust. Soc. Am. **77**(2), 649–657.

Nakamura, M., Iwano, K., and Furui, S. (**2008**). "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," Comput. Speech Lang. **22**(2), 171–184.

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

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," Anim. Behav. **95**, 89–99.

Pisanski, K., Raine, J., and Reby, D. (**2020**). "Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries," R. Soc. Open Sci. **7**(2), 191642.

Plant-Hébert, J., Boucher, V. J., and Jemel, B. (**2021**). "The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification," PLoS One **16**, e0250214.

R Core Team (**2021**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (Last viewed April 2022).

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.

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.

Shue, Y.-L., Keating, P., and Vicenik, C. (**2009**). "VOICESAUCE: A program for voice analysis," J. Acoust. Soc. Am. **126**(4), 2221.

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.

Sun, X. (**2002**). "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 13–17, Orlando, FL, pp. 333–336.

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

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

Wagner, I., and Köster, O. (**1999**). "Perceptual recognition of familiar voices using falsetto as a type of voice disguise," in *Proceedings of the 14th International Congress of Phonetic Sciences*, August 1–7, San Francisco, CA, pp. 1381–1384.

Wagner, P., Trouvain, J., and Zimmerer, F. (**2015**). "In defense of stylistic diversity in speech research," J. Phon. **48**, 1–12.

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