

UCLA

UCLA Previously Published Works

Title

Acoustic voice variation within and between speakers.

Permalink

<https://escholarship.org/uc/item/3hq140vh>

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

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

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

^{a)}yoongeonglee@ucla.edu

19 **I. INTRODUCTION**

20 What makes your voice yours? Individuals' voices, their "auditory faces" (Belin *et al.*,
21 2004), provide significant clues to personal identity along with information about talkers'
22 long-term physical, psychological, and social characteristics, based on the variability these
23 factors introduce into voice. Because even small changes in emotion, social context, and
24 physiologic state can cause significant variability in voice, no speaker ever says the same thing
25 in exactly the same way twice, whether quality is intentionally or incidentally manipulated
26 (see Kreiman and Sidtis 2011, for extended review). However, the extent and nature of
27 within-speaker variability in voice are unknown, despite the fact that the acoustic signal
28 serves as input to the perceptual system, which must be able to cope with this variability in
29 order to achieve a stable percept and/or recognition. Information about acoustic variability
30 is thus critical for formulating models of voice quality and talker recognition. This paper
31 describes a series of analyses exploring within- and between-talker acoustic variation and
32 the extent to which they conform to expectations derived from current models of voice
33 perception.

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

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

48 Facial recognition poses similar challenges to viewers, who must cope with changes in
49 lighting, expression, and orientation in order to identify or discriminate among faces (Hill
50 and Bruce, 1996; O’Toole *et al.*, 1998; Patterson and Baddeley, 1977). Because similarities
51 exist in voice and face processing (Stevenage *et al.*, 2018; Yovel and Belin, 2013), recent
52 findings from the face perception literature may provide insight into mechanisms for coping
53 with acoustic voice variability. In particular, facial identity learning improves when viewers
54 are exposed to highly but naturally varying sets of images of one person (for example, with
55 changes in orientation or emotion) during training (Kramer *et al.*, 2017; Murphy *et al.*,
56 2015; Ritchie and Burton, 2017). This suggests that variation in the same face provides
57 useful person-specific information and thus is important in identity learning and perception
58 (Burton, 2013; Burton *et al.*, 2016; Jenkins *et al.*, 2011). To our knowledge, no parallel
59 studies have appeared for voice learning, but some classic findings suggest acoustic variability
60 may also provide important information to listeners. These studies have reported that
61 increasing phonological length (i.e., the number of individual phonemes; Schweinberger *et al.*

62 1997) or acoustic duration (Bricker and Pruzansky, 1966; Cook and Wilding, 1997; Legge
63 *et al.*, 1984) of the voice samples leads to more accurate vocal identity processing, due to
64 the increased variety in speech sounds available in longer stimuli or the longer duration
65 (or both), which provide listeners with added articulatory and acoustic variability (cf. e.g.
66 *Lively et al.* 1993, for similar effects in learning phonological categories).

67 Taken together, these studies of faces and voices suggest that listeners need to learn
68 how a particular voice varies in order to recognize it accurately and efficiently. At first
69 glance, this claim appears consistent with prototype-based models of the cognitive and
70 neural processes underlying voice identity perception (Latinus and Belin, 2011a; Lavner
71 *et al.*, 2001; Papçun *et al.*, 1989; Yovel and Belin, 2013). In these accounts, listeners encode
72 and process voice identity in relation to a population prototype, which is a context-dependent
73 “average-sounding” voice, defined as a central tendency in a distribution of exemplars (Patel,
74 2008) that resides at the center of a multidimensional acoustical ‘voice space.’ Each voice is
75 further represented in terms of its deviations from that group prototype, stored as a unique
76 ‘reference pattern’ for that identity and passed on for further analysis (Latinus and Belin,
77 2011b; Papçun *et al.*, 1989). On further consideration, however, it becomes apparent that
78 these models are underspecified with respect to two important issues. First, the relationship
79 between between-talker variability in quality and the population prototype is unknown.
80 Although it is commonly assumed that prototypes are statistical averages derived from
81 multiple samples of a given talker’s voice (e.g., Latinus and Belin 2011a; Maguinness *et al.*
82 2018), to our knowledge no data exist about how much detail (and what kind of detail) about
83 quality is actually needed to specify the prototype, and how much is reserved as “deviations”

84 from the prototype. Second, the nature (or even the existence) of similar reference patterns
85 for individual talkers and the way in which within-talker variation affects formation of these
86 patterns have not to our knowledge been addressed, although such patterns would seem to be
87 essential for the formation of stable representations of voices and thus for voice recognition
88 ([Lavan *et al.*, 2018](#)).

89 Existing cognitive and neuropsychological models of voice perception and recognition
90 have not been fully exploited to generate clear hypotheses about the nature and extent
91 of even between-talker acoustic variability in voice, which has been studied far more than
92 within-talker variability. As discussed above, these models posit the existence of an acoustic
93 voice space organized around a population prototype, so that voices are encoded and later
94 recognized in terms of their distance from the prototype and the manner in which they
95 deviate from this (presumed) population average. Because voice production and perception
96 have co-evolved, it follows that if the perceptual models are correct, then there should be
97 some acoustic features that consistently explain significant between-talker acoustic variance
98 across all the talkers in a population. These features would characterize the central cat-
99 egory member for the population of talkers, consistent with the existence of a perceptual
100 space organized around a prototype, and would also specify the location of each voice in
101 the space with respect to the prototype. Remaining differences between voices should be
102 idiosyncratic, so that the features that differentiate pairs of talkers depend on the precise
103 acoustic information involved in each comparison (e.g., [Kreiman and Gerratt 1996](#)). This
104 would be consistent with what has been found for faces ([Maguinness *et al.*, 2018](#); [Stevenage](#)

105 *et al.*, 2018; Yovel and Belin, 2013), although we cannot assume that faces and voices are
106 perceived in similar ways at all processing stages.

107 Predictions are less clear for variation within a single talker across utterances, although
108 studies of variation in faces may again offer some clues. Principal component analyses exam-
109 ining how images of a face vary across different photographs of that person (Burton *et al.*,
110 2016) showed that the first few components (left-to-right head rotations, angle to camera,
111 the direction of lighting; and changes in expression like smiles, eye movements, mouth open-
112 ing, or lip rounding during speech) emerged consistently across individuals and accounted
113 for the most variance in different photos of the same person. Dimensions appearing in later
114 principal components (from the fourth onward) did not generalize well from one person to
115 another, so that some features were shared across faces, and some dimensions of variability
116 were idiosyncratic to specific faces. Given the many similarities between face and voice
117 processing in identity perception (see Yovel and Belin 2013, for review), this suggests that
118 voice spaces for individual talkers should be similarly structured. If “prototypes” for individ-
119 ual talkers are characterized by the same features across talkers, then these features would
120 naturally characterize a population prototype against which each individual voice could be
121 assessed.

122 Results from our preliminary studies (Keating and Kreiman, 2016; Kreiman *et al.*, 2017)
123 are also consistent with the hypothesis that voice spaces for individual talkers are struc-
124 tured similarly to population voice spaces. In those experiments, we used linear discrimi-
125 nant analyses to identify the acoustic features that maximally distinguished a large number
126 of individual voices. A small number of variables (F0, F4, the root mean square energy

127 calculated over five pitch pulses [energy], the relative amplitudes of the first and second
128 harmonics [H1-H2], and the amplitude ratio between subharmonics and harmonics [SHR])
129 proved important for distinguishing both male and female voices, but these accounted for
130 only about 50% of the acoustic variance in the data, the remaining variance being explained
131 by different variables depending on the particular voices being compared.

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

149 II. METHOD

150 A. Speakers and voice samples

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

163 The database provides significant within- and between-speaker variability. Speakers were
164 recorded on 3 different days and performed multiple speech tasks including reading, un-
165 scripted speech tasks, and a conversation. In order to control for variations due to differences
166 in phonemic content or emotional state across talkers, this initial study used recordings of 5
167 Harvard sentences ([IEEE Subcommittee 1969](#); Table I), read twice each day for a total of 6
168 repetitions per sentence over 3 recording sessions on different days. Variability reported in

169 this paper was calculated across sentence productions (different repetitions, sentences, and
 170 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.

171 B. Measurements and data processing

172 Acoustic measurements were made automatically every 5 ms on vowels and approximants
 173 (i.e., /l/, /r/, /w/) excerpted from each complete sentence, using VoiceSauce ([Shue et al.,](#)
 174 [2011](#)). Following the psychoacoustic model of voice quality described in [Kreiman et al.](#)
 175 [\(2014\)](#), acoustic parameters included fundamental frequency (F0); the first four formant
 176 frequencies (F1, F2, F3, F4), the relative amplitudes of the first and second harmonics
 177 (H1*-H2*) and the second and fourth harmonics (H2*-H4*); and the spectral slopes from
 178 the fourth harmonic to the harmonic nearest 2 kHz in frequency (H4*-H2kHz*) and from
 179 the harmonic nearest 2 kHz to the harmonic nearest 5 kHz in frequency (H2kHz*-H5kHz).
 180 Values of harmonics marked with ‘*’ were corrected for the influence of formants on harmonic
 181 amplitudes ([Hanson and Chuang, 1999](#); [Iseli and Alwan, 2004](#)). Our preliminary studies
 182 ([Keating and Kreiman, 2016](#); [Kreiman et al., 2017](#)) showed substantial correlations between

183 the relative amplitude of the cepstral peak prominence in relation to the expected amplitude
 184 as derived via linear regression (CPP; Hillenbrand *et al.* 1994) and the 4 measures of the
 185 shape of the inharmonic (noise) source spectrum included in the psychoacoustic model, so
 186 for simplicity CPP was used as the only measure of spectral noise and/or periodicity in these
 187 analyses.

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

203 Data frames with missing or obviously erroneous parameter values (for example, impos-
 204 sible 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

205 variable were normalized with respect to the overall minimum and maximum values from
 206 the entire set of voice samples from males or females, as appropriate, so that all variables
 207 ranged from 0 to 1. Then, for each sentence production, a smoothing window of 50 ms
 208 (10 observations) was used to calculate moving averages and moving coefficients of varia-
 209 tion for the 13 variables during that sentence. Across speakers, the above winnowing and
 210 post-processing steps resulted in about 515k data frames (F: 266k, M: 249k).

211 C. Principal component analysis

212 In principal component analysis (PCA), variables that are correlated with one another
 213 but relatively independent of other subsets of variables are combined into components, with
 214 the goal of reducing a large number of variables into a smaller set which is thought to re-
 215 flect internal structures that have created the correlations among variables. As moderate
 216 correlations were expected between variables, we employed an oblique rotation to create the
 217 simplest possible factor structure for our data (Cattell, 1978; Thurstone, 1947). Analyses
 218 were conducted separately for each speaker (within-speaker analyses) and for the combined

219 male and female speakers as groups (combined speaker analyses). For within-speaker anal-
220 yses, PCA was performed separately on each individual talker’s acoustic measurement data
221 (26 variables: moving averages for 13 variables + moving coefficients of variation for the
222 same 13 variables) to reveal the dimensions of the acoustic variability space for that partic-
223 ular voice. For combined speaker analyses, PCA was performed separately on the acoustic
224 data (all 26 variables) from females and males, pooling the 50 speakers’ data in each analy-
225 sis. PCs were restricted to the resulting factorial solutions with eigenvalues greater than 1,
226 ensuring that each retained factor accounted for an interpretable amount of variance in the
227 data (Kaiser, 1960). Results were also visually confirmed with Scree plots (Cattell, 1966).
228 Following usual practice, variables with loadings (weights) at or exceeding 0.32 on a given
229 component were considered to form a principal component (Tabachnick and Fidell, 2013).

230 III. RESULTS

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

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

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

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

252 Sub-analyses of factors contributing to the first PC are shown in Figs. 2 and 3. For
 253 most speakers, PC1 represented the combination of **variability (measured by CoVs) in**
 254 **source spectral shape** (F: 41/50 speakers, M: 46/50 speakers) and in **spectral noise** (F:
 255 45/50 speakers, M: 47/50 speakers), which usually emerged together (F: 40/50 speakers, M:
 256 44/50 speakers) (Fig. 2). An additional analysis (Fig. 3) revealed that across speakers all
 257 4 CoV measures of source spectral variability (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-

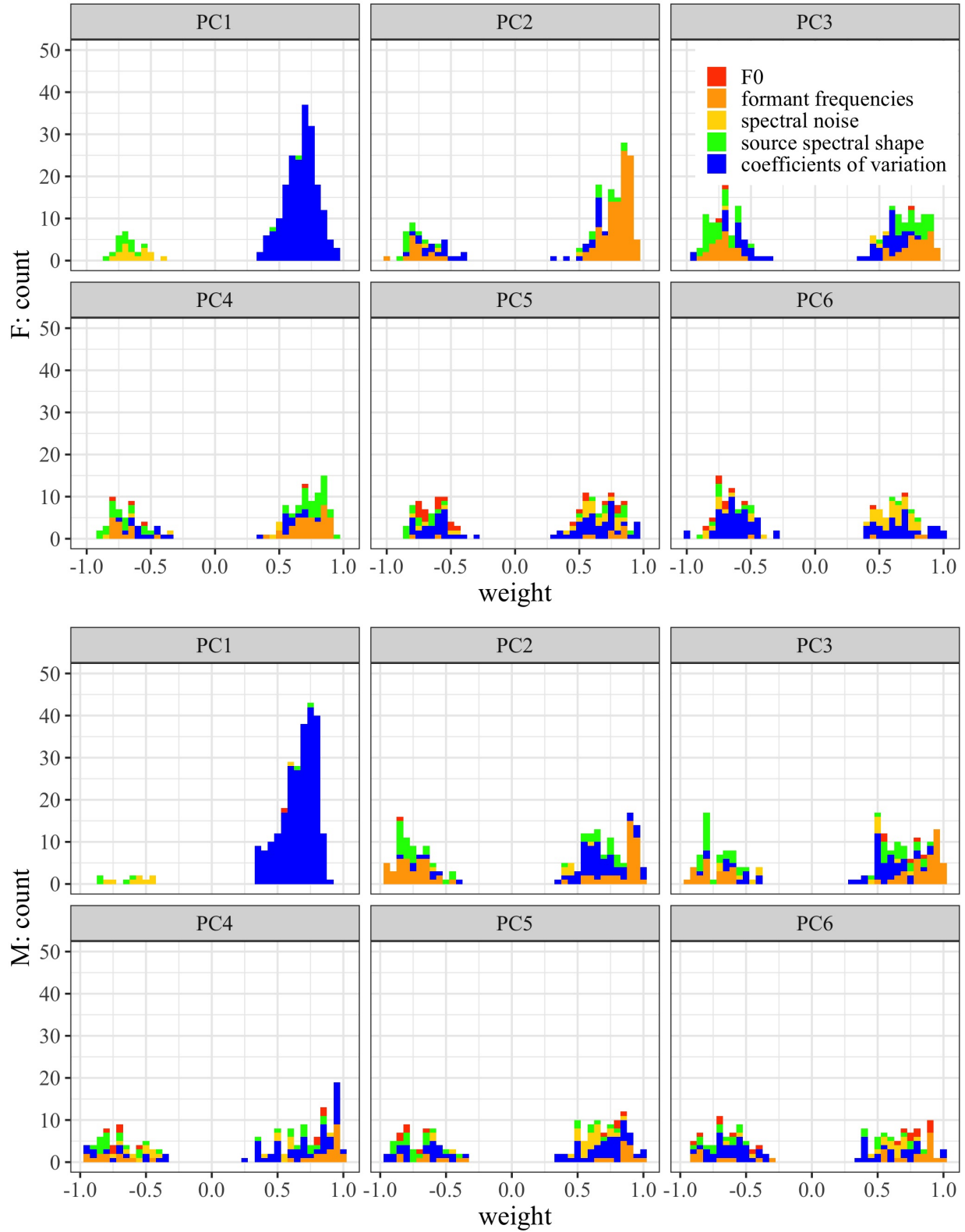


FIG. 1. Distribution of acoustic parameters plotted (stacked histogram) against the rotated component loadings (weight) for the first 6 PCs. Upper panel: female speakers. Lower panel: male speakers.

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

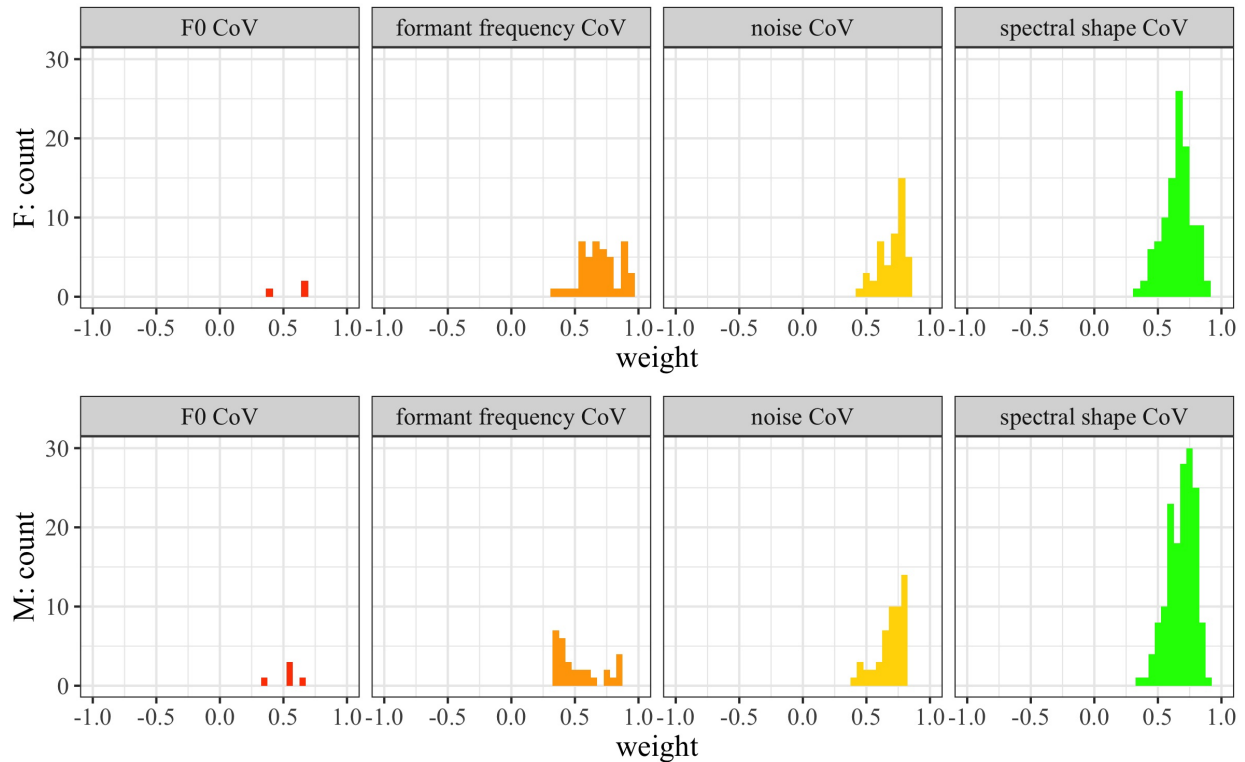


FIG. 2. (Color online) Distribution of variability parameters in PC1 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). ‘CoV’ = coefficient of variation.

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

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

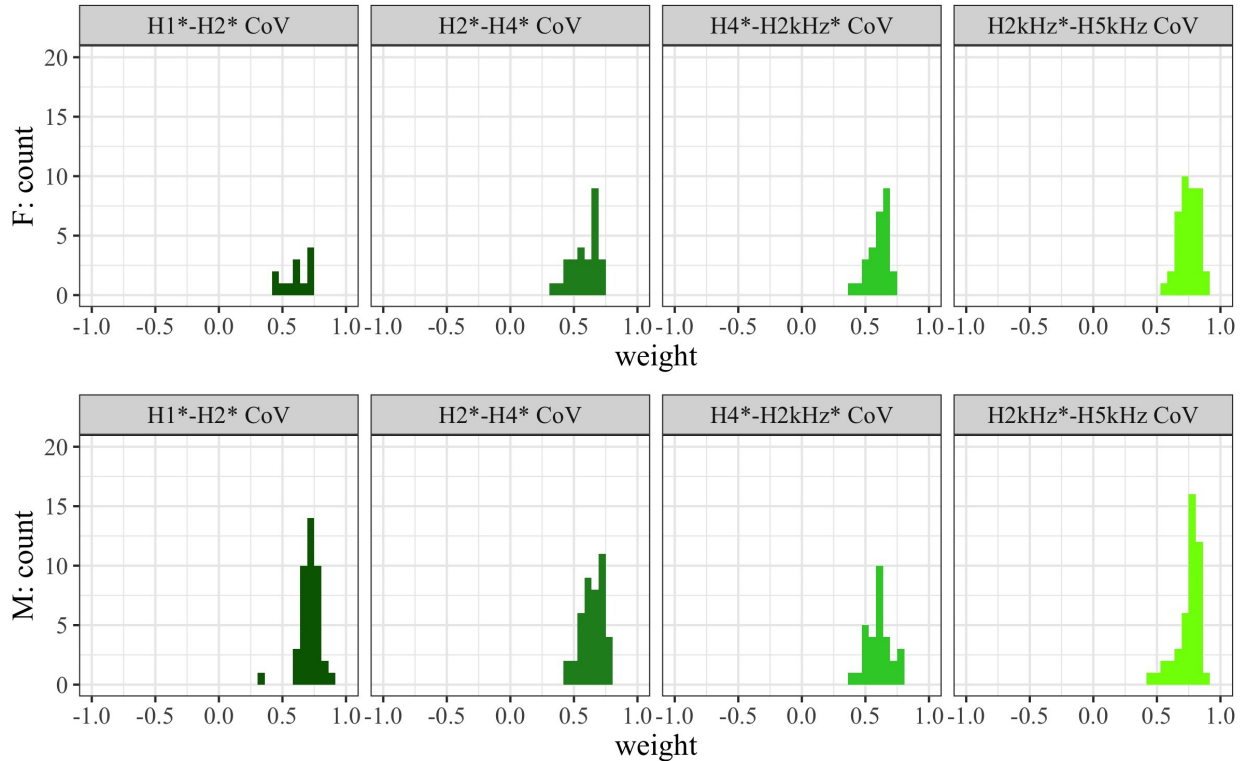


FIG. 3. (Color online) Distribution of spectral source variability parameters in PC1 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). ‘CoV’ = coefficient of variation.

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

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

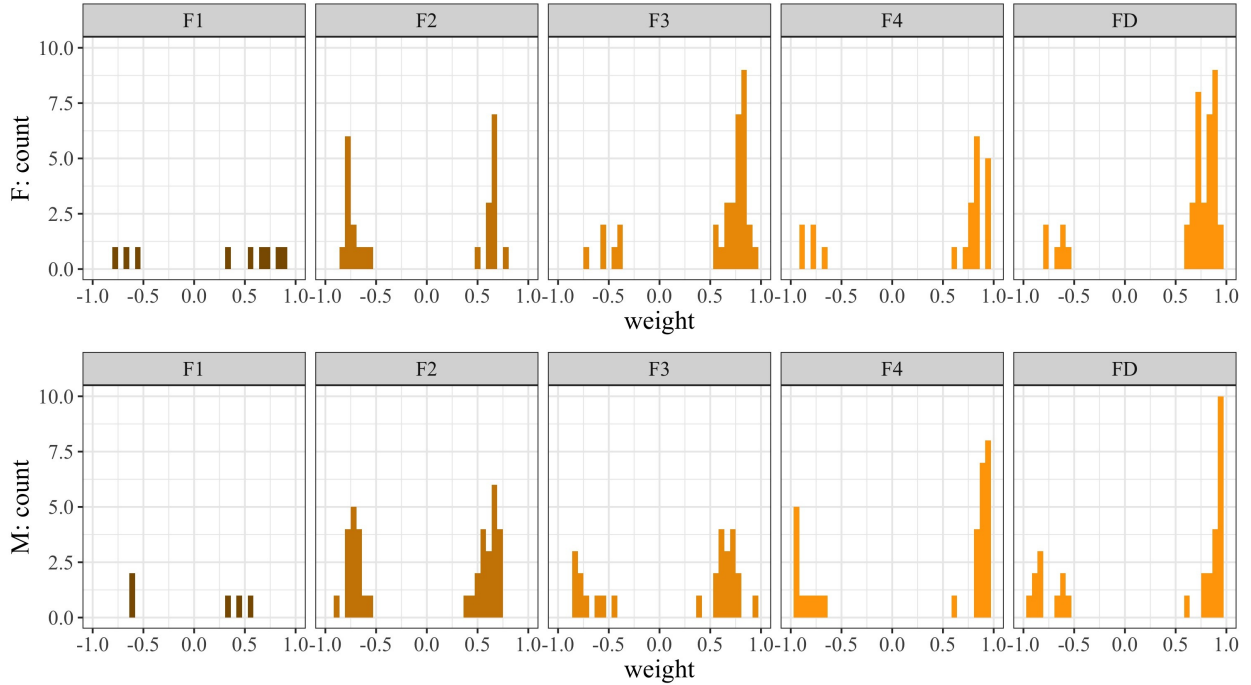


FIG. 4. (Color online) Distribution of formant frequency parameters in PC2 plotted against the rotated component loadings (weight) for female speakers (upper panel) and male speakers (bottom panel). Each figure reflects values derived from both moving averages and moving coefficients of variation for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

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

282 Notably, F0 and/or its CoV only emerged in the first two components for 4/100 speakers
 283 (2 female and 2 male). Among those 4 speakers, only one (male) speaker showed F0 as the
 284 most weighted variable within the PC (red bar in PC1, Fig. 1, bottom panel).

285 *Interim summary and discussion*

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

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

300 **B. Between-speaker group PCA: “General” voice spaces**

301 As described above, a second set of PCAs examined the structure of the acoustic space for
302 the combined groups of female and male speakers. Eight PCs were extracted for both speaker
303 groups, accounting for 67% of the cumulative variance for female speakers and 66% for male
304 speakers. Not surprisingly, given how consistent results were across individual speakers,
305 patterns of acoustic variability in these multi-talker spaces largely mirrored the patterns

306 found within speakers. Fig. 5 shows the group results, and details of the analyses are included
 307 in Appendix B. The first PC weighted most heavily on **variability (measured by CoVs)**
 308 **in source spectral shape and spectral noise**, accounting for 18% and 20% of variance
 309 across females and males, respectively. As in the within-speaker analyses, **coefficients of**
 310 **variation for H2kHz*-H5kHz and CPP** were the most important components of this
 311 PC.

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

322 IV. DISCUSSION

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

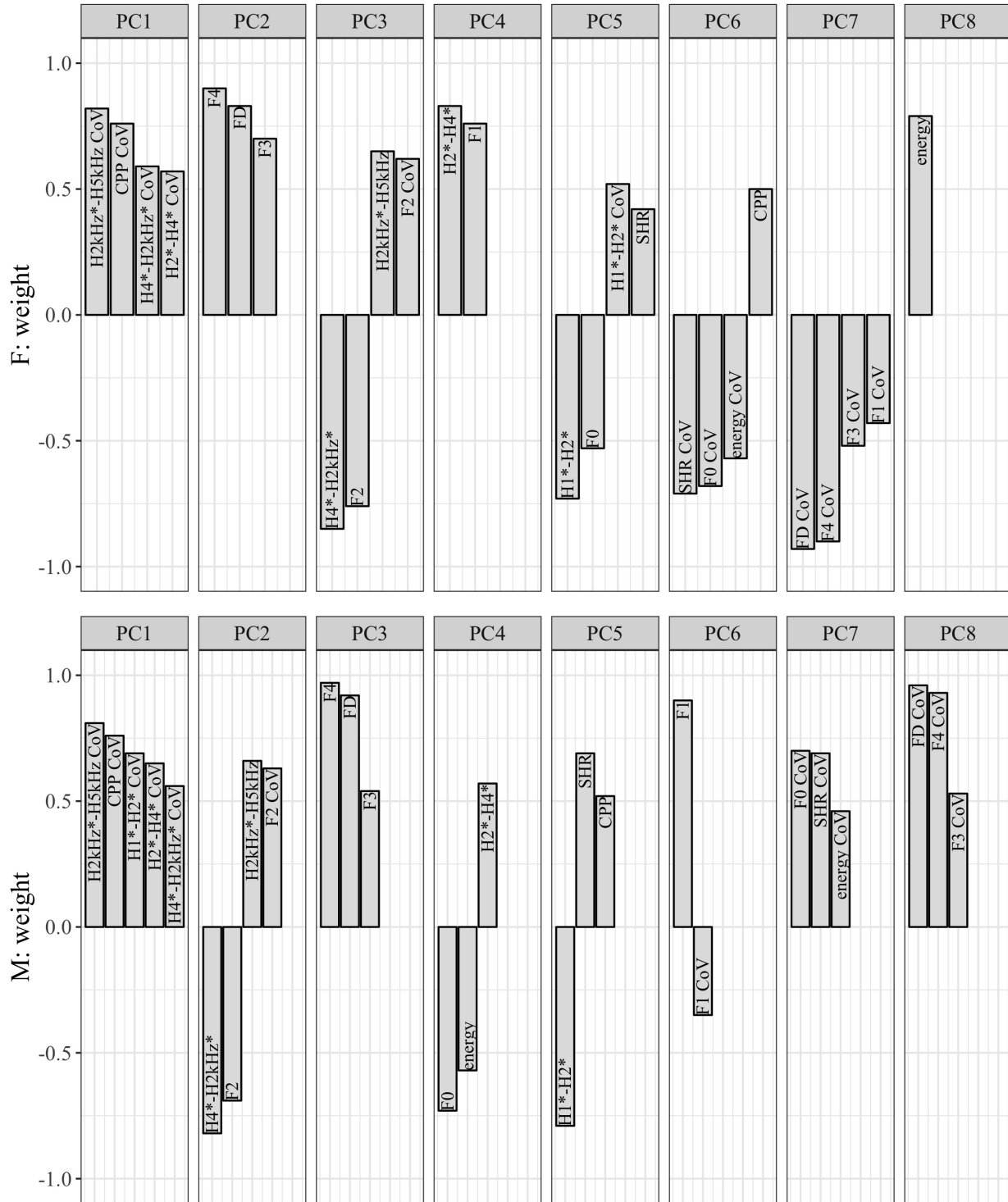


FIG. 5. Acoustic parameters emerging in 8 PCs for female speaker group (upper panel) and male speaker group (bottom panel). Variables within each PC are ordered from the highest absolute value of rotated component loadings (weight) to the lowest value. See also Appendices B1 and B2 for variance accounted for by each PC. ‘CoV’ = coefficient of variation.

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

332 As hypothesized, results of analyses of within-speaker acoustic variability paralleled find-
333 ings for individual faces (Burton *et al.*, 2016), in that a small number of components emerged
334 consistently across talkers. For both females and males, variability in higher-frequency har-
335 monic and inharmonic energy (often associated with the degree of perceived breathiness or
336 brightness; Samlan *et al.* 2013) combined to account for the most variance within talkers.
337 These two measures generally emerged as a pair within the same PC, consistent with the
338 manner in which they covary in controlling the perceived levels of noise in a voice (Kreiman
339 and Gerratt, 2012). The second PC was consistently associated with higher formant fre-
340 quencies and with the average interval between formant frequencies (i.e., formant dispersion).
341 These measures have been associated with speaker identity (e.g., Ives *et al.* 2005; Smith *et al.*
342 2005) and with vocal tract length and perception of speaker size (Fitch, 1997; Pisanski *et al.*,
343 2014), but appear to be relatively independent of vowel quality (Fant, 1960).

344 However, an equal amount of within-talker acoustic variability was in fact specific to in-
345 dividual voices. The talker-specific dimensionality of the derived voice spaces differed across
346 different talkers, and different measures, different combinations of measures, or different or-
347 derings of the same sets of measures emerged in PCs after the first two. This suggests that

348 each individual “auditory face” is indeed unique, allowing for the formation of person-specific
349 patterns/representations for a particular voice.

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

363 Two limitations of this work must be noted. First, acoustic measures were based on read
364 speech, not on spontaneous vocalization or conversation. This has the advantage of control-
365 ling for variations due to differences in phonemic content or emotional state across talkers,
366 while still sampling variability across utterances and recording sessions within talkers, but
367 clearly does not represent the full range of acoustic variability that occurs within a talker
368 in an average day’s phonation. The UCLA Speaker Variability Database ([Keating *et al.*,
369 2019](#)) also includes a recording of an unscripted telephone conversation for each talker, and

370 analyses are underway to determine how well the present findings extend to more natural
371 utterances. Second, the sample of speakers studied was restricted with respect to speakers'
372 ages (a limitation of the database) and native languages (a design decision). For this initial
373 study, we view both of these limitations as necessary: No information is available about dif-
374 ferences in acoustic variability across different populations of speakers, and even speculation
375 is lacking with regard to how many and what kinds of populations exist, so no basis exists
376 for distinguishing variability within a population from variability across populations. The
377 methods presented here offer a means of investigating this question, which will be important
378 for further development of models of voice perception. Similarly, the manner (if any) in
379 which within- and between-speaker acoustic variability interact with linguistic factors such
380 as tone and phonemic voice quality differences remains unknown, again making it desirable
381 to control this factor in the present study. A systematic investigation of the interactions
382 among these factors is also underway.

383 The fact that F0 did not emerge early among the principal components extracted for
384 either the within-speaker or group analyses is counter-intuitive, given how important F0 is
385 to many aspects of voice perception (e.g., [Baumann and Belin 2010](#); [Kreiman *et al.* 1992](#);
386 [Murry and Singh 1980](#); [Murry *et al.* 1978](#); [Walden *et al.* 1978](#); see [Kreiman and Sidtis 2011](#),
387 for review). The lack of a major F0 component in our results may be an artefact of our
388 normalization technique, which was based on acoustic ranges but did not take into account
389 differences in perceptual sensitivity to different variables. However, we note that previous
390 studies reporting an F0 factor have used similar normalization procedures and steady-state
391 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.*

414 1985; see [Kreiman and Sidtis 2011](#), for review). Our results suggest that reference patterns
415 for individual speakers are mainly computed over the balance of higher-frequency harmonic
416 versus inharmonic energy in the voice and over formant dispersion, and are located in a
417 group voice space with similar structure. However, this shared structure accounts for only a
418 fraction of either within- or between-speaker acoustic variability, with most variability being
419 idiosyncratic. Thus, it may be misleading to think of prototypes as “average tokens” com-
420 puted across complete acoustic signals. Our results suggest instead that they are specified
421 by a very small number of acoustic attributes.

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

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

449 V. CONCLUSION

450 Principal component analysis identified measures that characterize variability in voice
451 quality within and between speakers and provided evidence for how voice spaces—individually
452 and generally—may be formulated with reference to acoustic attributes. Among the large
453 array of vocal parameters available for each individual voice, a few components (the bal-
454 ance between high-frequency harmonic and inharmonic energy in the voice, and formant
455 dispersion) emerged consistently across talkers, but most within-speaker acoustic variability

456 in voice was idiosyncratic. Results further showed that the measures that were frequently
457 shared by individual talkers also characterized voice variation across talkers, suggesting that
458 individual and “general” voice spaces have very similar acoustic structures. This aligns well
459 with the input seemingly required by prototype models of voice recognition. Our results have
460 implications for unfamiliar voice perception and processing, specifically providing evidence
461 for the nature of reference patterns and deviations from “average-sounding” across voices,
462 in individual and universal voice spaces. Going forward, it will be essential to consider
463 how listeners organize these identified measures of within-person variability into a personal
464 identity and how that relates to perceived differences between talkers.

465 **ACKNOWLEDGMENTS**

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

473 **APPENDIX A: AVERAGE PERCENTAGE OF ACOUSTIC VARIANCE EXPLAINED**
 474 **BY EACH PC AS A FUNCTION OF THE NUMBER OF PCS, FOR**
 475 **FEMALE AND MALE SPEAKERS. NUMBERS IN PARENTHE-**
 476 **SES INDICATE THE NUMBER OF SPEAKERS FOR WHOM**
 477 **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%

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

480 **1. PCA pattern matrix for female speaker group analysis. ‘CoV’ = coefficient of**
 481 **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%

482 **2. PCA pattern matrix for male speaker group analysis. ‘CoV’ = coefficient of**
 483 **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	

484 **References**

485

- 486 Barton, J. J., and Corrow, S. L. (2016). “Recognizing and identifying people: A neuropsy-
487 chological review,” *Cortex* **75**, 132–150.
- 488 Baumann, O., and Belin, P. (2010). “Perceptual scaling of voice identity: Common dimen-
489 sions for different vowels and speakers,” *Psychol. Res.* **74**(1), 110–120.
- 490 Belin, P., Fecteau, S., and Bédard, C. (2004). “Thinking the voice: Neural correlates of
491 voice perception,” *Trends Cogn. Sci.* **8**(3), 129–135.
- 492 Bricker, P. D., and Pruzansky, S. (1966). “Effects of stimulus content and duration on talker
493 identification,” *J. Acoust. Soc. Am.* **40**(6), 1441–1449.
- 494 Burton, A. M. (2013). “Why has research in face recognition progressed so slowly? The
495 importance of variability,” *Q. J. Exp. Psychol.* **66**(8), 1467–1485.
- 496 Burton, A. M., Kramer, R. S., Ritchie, K. L., and Jenkins, R. (2016). “Identity from
497 variation: Representations of faces derived from multiple instances,” *Cogn. Sci.* **40**(1),
498 202–223.
- 499 Cattell, R. B. (1966). “The scree test for the number of factors,” *Multivar. Behav. Res.*
500 **1**(2), 245–276.
- 501 Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*
502 (Springer, Boston).
- 503 Cook, S., and Wilding, J. (1997). “Earwitness testimony: Never mind the variety, hear the
504 length,” *Appl. Cogn. Psychol.* **11**(2), 95–111.
- 505 Fant, G. (1960). *Acoustic Theory of Speech Production* (The Hague: Mouton & Co).

- 506 Fitch, W. T. (1997). “Vocal tract length and formant frequency dispersion correlate with
507 body size in Rhesus Macaques,” *J. Acoust. Soc. Am.* **102**(2), 1213–1222.
- 508 Hanson, H. M., and Chuang, E. S. (1999). “Glottal characteristics of male speakers: Acous-
509 tic correlates and comparison with female data,” *J. Acoust. Soc. Am.* **106**(2), 1064–1077.
- 510 Hill, H., and Bruce, V. (1996). “Effects of lighting on the perception of facial surfaces,” *J.*
511 *Exp. Psychol. Hum. Percept. Perform.* **22**(4), 986–1004.
- 512 Hillenbrand, J., Cleveland, R., and Erickson, R. (1994). “Acoustic correlates of breathy
513 vocal quality: Dysphonic voices and continuous speech,” *J. Speech Lang. Hear. Res.* **37**,
514 769–778.
- 515 Hirson, A., and Duckworth, M. (1993). “Glottal fry and voice disguise: a case study in
516 forensic phonetics,” *J. Biomed. Eng.* **15**(3), 193–200.
- 517 IEEE Subcommittee (1969). “IEEE Subcommittee on Subjective Measurements IEEE Rec-
518 ommended Practices for Speech Quality Measurements,” *IEEE Trans. Signal Process.* **17**,
519 227–246.
- 520 Iseli, M., and Alwan, A. (2004). “An improved correction formula for the estimation of
521 harmonic magnitudes and its application to open quotient estimation,” in *Proceedings of*
522 *IEEE ICASSP*, Vol. 1, pp. 10–13.
- 523 Ives, D. T., Smith, D. R. R., and Patterson, R. D. (2005). “Discrimination of speaker size
524 from syllable phrases,” *J. Acoust. Soc. Am.* **118**(6), 3816–3822.
- 525 Jenkins, R., White, D., Van Montfort, X., and Burton, A. M. (2011). “Variability in photos
526 of the same face,” *Cognition* **121**(3), 313–323.

- 527 Kaiser, H. F. (1960). “The applications of electronic computer to factor analysis,” *Educ.*
528 *Psychol. Meas.* **20**(1), 141–151.
- 529 Keating, P., and Kreiman, J. (2016). “Acoustic similarities among female voices,” *J. Acoust.*
530 *Soc. Am.* **140**, 3393.
- 531 Keating, P. A., Kreiman, J., and Alwan, A. (2019). “A new speech database for within-
532 and between-speaker variability,” in *Proceedings of the ICPhS XIX*.
- 533 Kramer, R. S., Jenkins, R., Young, A. W., and Burton, A. M. (2017). “Natural variability
534 is essential to learning new faces,” *Vis. Cogn.* **25**(4-6), 470–476.
- 535 Kreiman, J., Auszmann, A., and Gerratt, B. (2019). “What does it mean for a voice to sound
536 ‘normal’?,” in *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*,
537 edited by M. Barkat-Defradas, B. Weiss, J. Trouvain, and J. Ohala (Springer, Singapore).
- 538 Kreiman, J., Gerratt, B., Garellek, M., Samlan, R., and Zhang, Z. (2014). “Toward a unified
539 theory of voice production and perception,” *Loquens* **1**(1), 1–9.
- 540 Kreiman, J., Gerratt, B., Precoda, K., and Berke, G. (1992). “Individual differences in voice
541 quality perception,” *J. Speech Lang. Hear. Res.* **35**(3), 512–520.
- 542 Kreiman, J., and Gerratt, B. R. (1996). “The perceptual structure of pathologic voice
543 quality,” *J. Acoust. Soc. Am.* **100**(3), 1787–1795.
- 544 Kreiman, J., and Gerratt, B. R. (2012). “Perceptual interaction of the harmonic source and
545 noise in voice,” *J. Acoust. Soc. Am.* **131**(1), 492–500.
- 546 Kreiman, J., and Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary*
547 *Approach to Voice Production and Perception* (Wiley-Blackwell).

- 548 Kreiman, J. E., Keating, P., and Vesselinova, N. (2017). “Acoustic similarities among voices.
549 Part 2: Male speakers,” **142**, 2519.
- 550 LaRiviere, C. (1975). “Contributions of fundamental frequency and formant frequencies to
551 speaker identification,” *Phonetica* **31**(3-4), 185–197.
- 552 Latinus, M., and Belin, P. (2011a). “Anti-voice adaptation suggests prototype-based coding
553 of voice identity,” *Front. Psychol.* **2**(175), 1–12.
- 554 Latinus, M., and Belin, P. (2011b). “Primer: Human voice perception,” *Curr. Biol.* **21**(4),
555 R143–R145.
- 556 Lattner, S., Meyer, M. E., and Friederici, A. D. (2005). “Voice perception: Sex, pitch, and
557 the right hemisphere,” *Hum. Brain Mapp.* **24**(1), 11–20.
- 558 Lavan, N., Burston, L. F., and Garrido, L. (2018). “How many voices did you hear? Natural
559 variability disrupts identity perception from unfamiliar voices,” *Br. J. Psychol.* **110**, S76–
560 S93.
- 561 Lavan, N., Burston, L. F. K., Ladwa, P., Merriman, S. E., and Knight, S. (2019). “Breaking
562 voice identity perception : Expressive voices are more confusable for listeners,” *Q. J. Exp.*
563 *Psychol.* 1–9.
- 564 Lavner, Y., Rosenhouse, J., and Gath, I. (2001). “The prototype model in speaker identifi-
565 cation by human listeners,” *Int. J. Speech Technol.* **4**(1), 63–74.
- 566 Legge, G. E., Grosman, C., and Pieper, C. M. (1984). “Learning unfamiliar voices,” *J.*
567 *Exp. Psychol. Learn. Mem. Cogn.* **10**(2), 298–303.
- 568 Lively, S. E., Logan, J. S., and Pisoni, D. B. (1993). “Training japanese listeners to identify
569 english /r/ and /l/. ii: The role of phonetic environment and talker variability in learning

- 570 new perceptual categories,” *J. Acoust. Soc. Am.* **94**, 1242–1255.
- 571 Maguinness, C., Roswadowitz, C., and von Kriegstein, K. (2018). “Understanding the
572 mechanisms of familiar voice-identity recognition in the human brain,” *Neuropsychologia*
573 **116**, 179–193.
- 574 Murphy, J., Ipser, A., Gaigg, S. B., and Cook, R. (2015). “Exemplar variance supports
575 robust learning of facial identity,” *J. Exp. Psychol. Hum. Percept. Perform.* **41**(3), 577–
576 581.
- 577 Murry, T., and Singh, S. (1980). “Multidimensional analysis of male and female voices,” *J.*
578 *Acoust. Soc. Am.* **68**, 1294–1300.
- 579 Murry, T., Singh, S., and Sargent, M. (1978). “Multidimensional classification of normal
580 voice qualities,” *J. Acoust. Soc. Am.* **64**, 81–87.
- 581 O’Toole, A. J., Edelman, S., and Bühlhoff, H. H. (1998). “Stimulus-specific effects in face
582 recognition over changes in viewpoint,” *Vis. Res.* **38**(15-16), 2351–2363.
- 583 Papçun, G., Kreiman, J., and Davis, A. (1989). “Long-term memory for unfamiliar voices,”
584 *J. Acoust. Soc. Am.* **85**(2), 913–925.
- 585 Patel, A. D. (2008). “Music and the brain: Three links to language,” in *The Oxford Hand-*
586 *book of Music Psychology*, edited by S. Hallam, I. Cross, and M. Thaut, 1st ed. (Oxford
587 Univ. Press, Oxford), pp. 208–216.
- 588 Patterson, K. E., and Baddeley, A. D. (1977). “When face recognition fails,” *J. Exp. Psy-*
589 *chol. Learn. Mem. Cogn.* **3**(4), 406–417.
- 590 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O’Connor, J. J. M., Röder, S., Andrews, P. W.,
591 Fink, B., DeBruine, L. M., Jones, B. C., and Feinberg, D. R. (2014). “Vocal indicators of

- 592 body size in men and women: a meta-analysis,” *Animal Behav.* **95**, 89–99.
- 593 Read, D., and Craik, F. I. (1995). “Earwitness identification: Some influences on voice
594 recognition,” *J. Exp. Psychol. Appl.* **1**(1), 6–18.
- 595 Reich, A. R., and Duke, J. E. (1979). “Effects of selected vocal disguises upon speaker
596 identification by listening,” *J. Acoust. Soc. Am.* **66**(4), 1023–1028.
- 597 Reich, A. R., Moll, K. L., and Curtis, J. F. (2005). “Effects of selected vocal disguises upon
598 spectrographic speaker identification,” *J. Acoust. Soc. Am.* **60**(4), 919–925.
- 599 Ritchie, K. L., and Burton, A. M. (2017). “Learning faces from variability,” *Q. J. Exp.*
600 *Psychol.* **70**(5), 897–905.
- 601 Samlan, R. A., Story, B. H., and Bunton, K. (2013). “Relation of perceived breathiness to
602 laryngeal kinematics and acoustic measures based on computational modeling,” *J. Speech*
603 *Lang. Hear. Res.* **56**(4), 1209–1223.
- 604 Saslove, H., and Yarmey, A. D. (1980). “Long-term auditory memory: Speaker identifica-
605 tion,” *J. Appl. Psychol.* **65**(1), 111–116.
- 606 Schweinberger, S. R. (2001). “Human brain potential correlates of voice priming and voice
607 recognition,” *Neuropsychologia* **39**(9), 921–936.
- 608 Schweinberger, S. R. (2013). “Audiovisual integration in speaker identification,” in *Integrat-*
609 *ing Face and Voice in Person Perception*, edited by P. Belin, S. Campanella, and T. Ethofer
610 (Springer Science + Business Media, New York), pp. 119–134.
- 611 Schweinberger, S. R., Herholz, A., and Sommer, W. (1997). “Recognizing famous voices:
612 Influence of stimulus duration and different types of retrieval cues,” *J. Speech Lang. Hear.*
613 *Res.* **40**(2), 453–463.

- 614 Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). “VoiceSauce: A program for voice
615 analysis,” in *Proceedings of the ICPHS XVII*, pp. 1846–1849.
- 616 Smith, D., Patterson, R., Turner, R., Kawahara, H., and Irino, T. (2005). “The processing
617 and perception of size information in speech sounds,” *J. Acoust. Soc. Am.* **117**(1), 305–318.
- 618 Stevenage, S. V., Neil, G. J., Parsons, B., and Humphreys, A. (2018). “A sound effect:
619 Exploration of the distinctiveness advantage in voice recognition,” *Appl. Cogn. Psychol.*
620 **32**(5), 526–536.
- 621 Sun, X. (2002). “Pitch determination and voice quality analysis using Subharmonic-to-
622 Harmonic Ratio,” in *Proceedings of IEEE ICASSP*, Vol. 1, pp. 333–336.
- 623 Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th ed. (Pearson,
624 Boston).
- 625 Thurstone, L. L. (1947). *Multiple-Factor Analysis: A Development and Expansion of The*
626 *Vectors of Mind* (University of Chicago Press, Chicago).
- 627 Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). “Familiar voice recognition: Pat-
628 terns and parameters. Part I: Recognition of backward voices,” *J. Phon.* **13**, 19–38.
- 629 Wagner, I., and Köster, O. (1999). “Perceptual recognition of familiar voices using falsetto
630 as a type of voice disguise,” in *Proceedings of the ICPHS XI*, pp. 1381–1384.
- 631 Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M.
632 (1978). “Correlates of psychological dimensions in talker similarity,” *J. Speech Lang. Hear.*
633 *Res.* **21**(2), 265–275.
- 634 Yovel, G., and Belin, P. (2013). “A unified coding strategy for processing faces and voices,”
635 *Trends Cogn. Sci.* **17**(6), 263–271.