WITHIN AND BETWEEN SPEAKER VARIATION IN VOICES

Yoonjeong Lee, Jody Kreiman

Department of Head and Neck Surgery, University of California, Los Angeles, Los Angeles CA USA 90095

yoonjeonglee@ucla.edu; jkreiman@ucla.edu

ABSTRACT

Little is known about the nature or extent of everyday variability in voice quality within a speaker or how this differs across speakers. Using principal component analysis, we identified measures that account for perceptually relevant acoustic variance within speakers. Based on face-identity studies and cognitive models of speaker recognition, we hypothesized that a few components would be shared across speakers, but that much of what characterizes individual talkers would be idiosyncratic. Fifty female and fifty male speakers of English provided multiple sentence productions recorded over three days. Acoustic parameters were measured for vowels and approximants. For both females and males, the most variance (20%/22%) was accounted for by variability in source spectral shape and by the balance of harmonic versus inharmonic energy in the voice. Formant frequencies accounted for an additional 12%/12% of variance. Remaining variance appeared largely idiosyncratic. Notably, F0 did not emerge. Implications for voice recognition are discussed.

Keywords: voice quality, voice acoustics, variability, principal component analysis

1. INTRODUCTION

What makes your voice yours? The human voice provides significant clues to personal identity. Nevertheless, individual vocal behavior during voice production is highly variable. Although listeners can, to some extent, cope with this variability to establish a stable identity percept, across voices intra-speaker variability makes recognition and discrimination challenging tasks [1–5]. Despite increasing attention to personal voice quality, little is known about the nature or extent of intra-speaker variability, and how it differs across speakers.

Prototype-based approaches are often invoked to account for the computational/neural processes underlying voice identity perception [6–8]. In these accounts, each voice is represented in terms of its deviations from a *prototype* voice, which resides at the center of a multidimensional acoustical 'voice space.' Deviations from prototypicality are stored as unique 'reference patterns' for each identity [9, 10]. While these models account for how listeners tell voices apart, they are underspecified with respect to how within-speaker variation affects formation of reference patterns, and thus the extraction of voice identity [11].

Ample evidence exists for striking similarities between face and voice identity processing [7, 12, 13]. It has become increasingly clear that natural variability within faces (for example, with changes in orientation or emotion) is essential to learning new faces [14–16]. Our study is informed by [17], which used principal component analysis to investigate how images of the same person vary across different photos of that person. The first few components to emerge for analyses of individual faces were consistent across faces of different identities, but the dimensions that appeared in later components did not generalize well from one face to another.

Here, we evaluated voice variation both within and across speakers by employing principal component analysis. The components that emerge from such analyses can be thought of as forming dimensions of an acoustic space specific to a given voice, in which that voice varies. Based on [17] and on cognitive models of voice processing, we hypothesized that a few dimensions would consistently emerge from analyses of individual speakers, but that much more of what characterizes vocal variability within a speaker would be idiosyncratic. We tested this hypothesis against multiple sentence productions from 100 native speakers of English, using a suite of measures that map between acoustics and perception of voice quality [18]. Additionally, we examined the dimensions characterizing acoustic variability across speakers and compared these to within-speaker acoustic variability.

2. METHODS

2.1. Speakers and voice samples

The voices of 50 female and 50 male speakers drawn from the UCLA Speaker Variability Database were used in this experiment [19]. All were native speakers of English, similar in age (F: 18-29, M: 18-26), with no known vocal disorder or speech complaints, and all were undergraduate students at the time of recording. Recordings were made in a sound-attenuated booth at a sampling rate of 22 kHz using a microphone suspended from a baseball cap worn by the speaker.

The database provides significant within- and between-speaker variability. Speakers were recorded on 3 different days and performed multiple speech tasks (e.g., reading, unscripted speech tasks, conversations). The current study used recordings of 5 Harvard sentences [20], read twice each day (a total of 6 repetitions per sentence over 3 recording sessions). As such, variability reported in this paper was calculated over each sentence production and its scope is limited to the reading task.

2.2. Measurements and data processing

Acoustic measurements were made automatically every 5 ms on vowels and approximants excerpted from each sentence, using VoiceSauce [21]. The acoustic parameters included: fundamental frequency (**F0**); the first four formant frequencies (**F1**, **F2**, **F3**, **F4**) and formant dispersion (**FD**, often correlated with vocal tract length [22]), calculated as the average difference in frequency between each adjacent pair of formants; the relative amplitude of the cepstral peak prominence in relation to the expected amplitude as derived via linear regression (**CPP**) [23]; root mean square energy calculated over five pitch pulses (**energy**); the amplitude ratio between subharmonics and harmonics (**SHR**) [24]; the relative amplitudes of the first and second harmonics (**H1*-H2***), the second and fourth harmonics (**H2*-H4***), the spectral slopes from the fourth harmonic to the harmonic nearest 2 kHz in frequency (**H4*-H2kHz***), and from the harmonic nearest 2 kHz to the harmonic nearest 5 kHz in frequency (**H2kHz*-H5kHz**). Values of harmonics marked with * were corrected for the influence of formants on harmonic amplitudes [24, 25]. As a set, these measures constitute a psychoacoustic model of voice quality [18].

Frames with missing or obviously erroneous parameter values (for example, impossible 0 values) were removed. Next, for each speaker, the obtained values of each acoustic variable were normalized with respect to the overall minimum and maximum values from that speaker's entire set of samples, so that all variables ranged from 0 to 1. Then, for each sentence production, a smoothing window of 50 ms (10 observations) was used to calculate moving averages of the 13 variables during that sentence. The corresponding moving coefficients of variation were also calculated as estimates of signal variability, so that the input to the principal component analysis included both steady-state and time-varying aspects of voice quality. Across speakers, the above winnowing and post-processing steps resulted in about 515k data frames (F: 266k, M: 249k).

2.3. Principal component analysis

In principal component analysis (PCA), variables that are correlated with one another but relatively independent of other subsets of variables are combined into components, with the goal of reducing a large number of variables into a smaller set which are thought to reflect internal structures that have created the correlations among variables. As moderate correlations were expected between variables, we employed an oblique rotation to create the simplest possible factor structure for our data [26, 27]. For within-speaker analyses, PCA was performed separately on each individual talker's measurement data to reveal the dimensions of the voice variability space for that particular voice. For combined speaker analyses, PCA was performed separately on data from females and males, pooling the 50 speakers' data in each analysis. PCs were restricted to the resulting factorial solutions with eigenvalues greater than 1 [29], which was also visually confirmed with Scree plots [30]. In our study, the combination of variables with loadings at or exceeding 0.32 on a given component were considered to form a principal component [31].

3. RESULTS

Although all acoustic variables were entered simultaneously into the analyses, for brevity they are grouped into 5 categories, following [32]: i) **F0**; ii) **formant frequencies** (F1, F2, F3, F4, FD); iii) **spectral noise** (CPP, energy, SHR); iv) **source spectral shape** (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz); and v) **variability** (coefficients of variation for all measures).

3.1. Within-speaker PCA: Common dimensions for individual speakers

Across individual speakers the total number of retained components (PCs) having eigenvalues greater than 1 ranged from 6 to 9. These components accounted for 65%-74% (*M*=69%) of the cumulative variance for individual female speakers and 62%-73% (*M*=68%) for individual male speakers.

We counted the number of times each acoustic category appeared in a within-speaker solution for each of the 100 speakers. Fig. 1 shows the distribution of variables and weights for the variables that emerged in the first two components (PC1, PC2). The first component accounted for 17%-23% (*M*=20%) and 20%-25% (*M*=22%) of the variance for females and males, respectively. For both females and males, the most frequently emerging variable in PC1 for individual speakers is **variability** (dark grey bars in PC1). (Detailed sub-analyses appear below.)

For female speakers, PC2 accounted for 10%-16% (*M*=12%) of variance. The variable most frequently associated with this component for each of the speakers was **formant frequencies** (black bars in PC2). For male speakers, the second component accounted for 10%-14% (*M*=12%) of the variance.

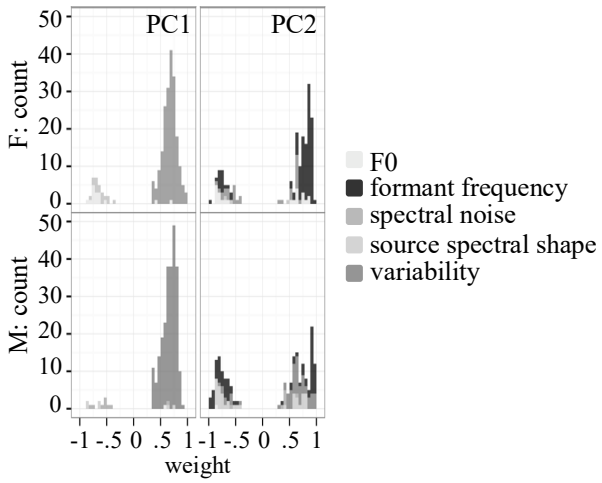

Figure 1: Distribution of acoustic parameters plotted (stacked histogram) against the rotated component loadings (weight).

While the measures associated with this component varied somewhat across the individual speakers, formant frequencies and variability appeared more frequently than other variables. After these two PCs, the remaining variance was largely idiosyncratic, with the dimensionality of the voice space differing for different speakers. For example, the component in which F0 emerged ranged from PC3 to PC8 across individuals. Similar patterns were observed for the rest of the measures.

Recall that the **variability** measure included coefficients of variation derived from all 13 acoustic variables. Table 1 summarizes the observed patterns of weights for individual variables on the rotated component loading for the first two PCs. The column labelled 'primary variable' shows the variable with the largest weight within each component. The next most weighted variable in that component is listed as the secondary variable. The shaded cells indicate the variability measures.

This closer analysis of the variability measures revealed that for many speakers (both female and male) the most variance is accounted for by the combination of **variability in source spectral shape and spectral noise**. An additional analysis further revealed that while across speakers all 4 measures of source spectral variability (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz) emerged in the first component, **spectral slope variability in the higher frequencies — H2kHz*-H5kHz — was the dominant measure**. The other important emerging variable, spectral noise variability, was mostly related to **variability in CPP**. Additionally, for many female speakers variability in formant frequencies was the most representative variable in the first component.

For both female and male individual speakers, **formant frequencies and their variability** emerged as the second PC. In this component, the **F4 and FD** measures appeared most important across speakers. Notably, except for a single male speaker, F0 did not emerge in the first two components.

Table 1: Patterns of weights on the rotated component loading for acoustic variables in the first two PCs across speakers.

PC	Primary variable	Secondary variable	# of speakers
1	spectral shape variability	noise variability	F: 40/50 M: 44/50
	formant frequency variability	noise variability	F: 3/50 M: 3/50
	formant frequencies	*N/A*	F: 4/50
	formant frequency variability	noise	F: 3/50
	noise	spectral shape variability	M:2/50
	spectral shape	*N/A*	M: 1/50
2	formant frequencies	formant frequency variability	F: 50/50 M: 15/50
	formant frequency variability	formant frequencies	M: 34/50
	F0	spectral shape	M: 1/50

3.2. Between-speaker PCA: "General" voice spaces

For both female and male speakers, the total number of extracted PCs was 8, accounting for 67% of the cumulative variance for female speakers and 66% for male speakers. Not surprisingly given how consistent results were across individual speakers, patterns of acoustic variability in these multi-talker spaces largely mirrored the patterns found within speakers. The first PC was composed of **variability in source spectral shape and spectral noise**, accounting for 18% and 20% of variance across females and males, respectively. As was the case with within-speaker analyses, **variability in H2kHz*-H5kHz and CPP** were the most important variables for this PC.

The second component accounted for 11% of variance in female voices, and corresponded to **formant frequencies**. For males, **spectral slope in the higher frequencies and F2** accounted for 10% of variance in the combined acoustic data. The opposite was observed for the third component: an additional 10% of the variance was accounted for by the higher frequencies and F2 for females; formant frequencies accounted for 9% of the variance in male voices. F0 only emerged in the later components (PC5 for females, PC4 for males) and accounted for very little variance in the data (6% for females, 7% for males).

4. DISCUSSION AND CONCLUSIONS

Variability is a key factor in models of voice perception and speaker identification. Using principal

component analysis, this study identified voice quality measures that accounted for perceptually relevant acoustic variance both within individual talkers and for the pooled groups of speakers. Unlike previous studies of within-talker variations in voice (e.g., [33]), this study included multiple sustained utterances from large numbers of male and female speakers, and included variation within and across utterances and over time.

As hypothesized, results paralleled [17]'s for within-face variability. The first two PCs, which accounted for the most acoustic variability (but not the majority of variability) within a speaker, were shared by nearly all speakers. For both females and males, the combination of higher-frequency harmonic and inharmonic energy (associated with the degree of perceived breathiness or brightness [34]) accounted for the most variance within talkers. Formant frequencies and their variability also appeared to explain considerable within-voice acoustic variance; formant dispersion, associated with vocal tract length [22], appeared to be the core variable in this component. However, the majority of within-person acoustic variability was in fact idiosyncratic—the talker-specific dimensionality of the derived voice spaces differed for different talkers.

Similar dimensions also emerged when data from all male and female speakers were analyzed in two group PCAs. Although this finding may appear trivial given the homogeneity of the individual results, in fact there is no a priori reason why individual solutions should have coincided as they did, and hence no a priori reason why individual and "general" acoustic spaces should be so similar.

The fact that F0 did not emerge in the early retained components for either the within-speaker or group analyses is a bit puzzling given that listeners are often sensitive to even small differences in F0 during voice processing [32, 34–37]. This finding might be due to the use of read speech in the present study, during which F0 variability is often limited. While F0 is useful for telling people apart, it might have ignorable perceptual weighting in tasks such as "telling people together" [5], which depend on within-, rather than between-speaker differences in voice. Alternatively, the lack of a major F0 component in our results may be an artefact of our normalization technique, which was based on acoustic ranges, but did not take into account differences in perceptual sensitivity to different variables. However, we note that previous studies reporting an F0 factor have used similar normalization procedures and steady-state vowels (e.g., [33]). This apparent discrepancy between acoustic structure and perceptual data requires further consideration.

These results have implications for current prototype models of voice processing [6, 7, 38], which are underspecified with respect to within-person variability in voice. Our results suggest that the within-person reference patterns are mainly computed over the balance of harmonic versus inharmonic energy and formant frequencies in the voice. Likewise, group analyses suggest that the "general" voice spaces are formulated with reference to the same attributes. These and the remaining idiosyncratic vocal behaviors form a person-specific voice space. These individual voice spaces could then enter into a "general" voice space, in which between-speaker differences in voice are evaluated.

Perceptual processes can only make use of the acoustic input they receive, so understanding the structure of acoustic voice spaces can provide insight into why voice perception functions as it does. Our results suggest that perception of unfamiliar voices is a two-part process. The fact that individual and group voice spaces have a similar acoustic structure suggests that in one part, listeners find the position of the voice in a "general" voice space relative to the overall prototype for that population of speakers. The large amount of idiosyncratic acoustic variance suggests that a second stage of processing is needed, in which these rather under-specified individual voice patterns are separated from their nearest neighbors in the general voice space, using ad hoc featural analysis. Such a scenario is broadly consistent with the finding that voice discrimination requires both right (pattern recognition) and left (featural analysis) hemisphere participation [40].

In conclusion, our study applied principal component analysis to identify measures that characterize variability in voice quality within *and* between speakers. A few components were shared across speakers, but most patterns of within-speaker acoustic variability in voice were idiosyncratic. Our results further showed that the measures that were frequently shared by individual speakers also characterized voice variation across speakers, suggesting that individual and "general" voice spaces are indeed composed of a similar acoustic structure. Our results have implications for unfamiliar voice perception and processing, in particular, providing evidence for what comprises a reference pattern in individual and universal voice spaces. Going forward, it will be essential to consider how these identified measures of within-person variability would be used in listeners' identity processing behaviors and how they interact with between-speaker differences.

5. ACKNOWLEDGMENTS

This work was supported by NIH grant DC01797 and NSF grant IIS-1704167. We thank Meng Yang for her help with VoiceSauce analyses. We are grateful to Pat Keating and Abeer Alwan for useful discussion.

6. REFERENCES

[1] Read, D., Craik, F.I.M. 1995. Earwitness identification: Some influences on voice recognition. *J Exp Psychol Appl*. 1, 6–18.

[2] Saslove, H., Yarmey, A.D. 1980. Long-term auditory memory: Speaker identification. *J Appl Psychol*. 65, 111–116.

[3] Wester, M. 2012. Talker discrimination across languages. *Speech Commun*. 54, 781–790.

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

[5] Lavan, N., Burston, L.F.K., Garrido, L. 2018. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *Br J Psychol*.

[6] Lavner, Y., Rosenhouse, J., Gath, I. 2001. The prototype model in speaker identification by human listeners. *Int J Speech Technol*. 4, 63–74.

[7] Yovel, G., Belin, P. 2013. A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*. pp. 263–271.

[8] Latinus, M., Belin, P. 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. *Front Psychol*. 2, 175.

[9] Papcun, G., Kreiman, J., Davis, A. 1989. Long-term memory for unfamiliar voices. *J Acoust Soc Am*. 85, 913–925.

[10] Latinus, M., Belin, P. 2011. Human voice perception. *Curr Biol*. 21.

[11] Lavan, N., Burton, A.M., Scott, S.K., McGettigan, C. 2018. Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin and Review*. 1–13.

[12] Maguinness, C., Roswandowitz, C., von Kriegstein, K. 2018. Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*. 116, 179–193.

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

[14] Murphy, J., Ipser, A., Gaigg, S.B., Cook, R. 2015. Exemplar variance supports robust learning of facial identity. *J Exp Psychol Hum Percept Perform*. 41, 577–581.

[15] Kramer, R.S.S., Jenkins, R., Young, A.W., Burton, A.M. 2017. Natural variability is essential to learning new faces. *Vis Cogn*. 25, 470–476.

[16] Ritchie, K.L., Burton, A.M. 2017. Learning faces from variability. *Q J Exp Psychol*. 70, 897–905.

[17] Burton, A.M., Kramer, R.S.S., Ritchie, K.L., Jenkins, R. 2016. Identity from variation: Representations of faces derived from multiple instances. *Cogn Sci*. 40, 202–223.

[18] Kreiman, J., Gerratt, B.R., Garellek, M., Samlan, R., Zhang, Z. 2014. Toward a unified theory of voice production and perception. *Loquens*. 1, 1–9.

[19] Keating, P., Kreiman, J., Alwan, A. 2019. A new speech database for within- and between-speaker variability. *Proc of the 19th ICPhS*.

[20] IEEE Subcommittee. 1969. IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. *IEEE Trans Audio Electroacoust*. 17, 227–246.

[21] Shue, Y.-L., Keating, P., Vicenik, C. 2009, VOICESAUCE: A program for voice analysis. *J Acoust Soc Am*. 126, 2221.

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

[23] Hillenbrand, J., Houde, R.A. 1996. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *J Speech Lang Hear Res*. 39, 311.

[24] Sun, X. 2002. Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. *Proc IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 333–336.

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

[26] Iseli, M., Alwan, A. 2004. An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. *Proc IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 10–13.

[27] Thurstone, L.L. 1947. *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind*. Chicago, IL, US: University of Chicago Press.

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

[29] Kaiser, H.F. 1960. The applications of electronic computer to factor analysis. *Educ Psychol Meas*. 20, 141–151.

[30] Cattell, R.B. 1966. The scree test for the number of factors. *Multivariate Behav Res*. 1, 245–276.

[31] Tabachnick, B.G., Fidell, L.S. 2013. *Using Multivariate Statistics*. Pearson.

[32] Kreiman, J., Auszmann, A., Gerratt, B.R. in press. What does it mean for a voice to sound "normal?" In: Ohala, J., et al. (eds), *Vocal Attractiveness*. Springer

[33] Baumann, O., Belin, P. 2010. Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychol Res*. 74, 110–120.

[34] Samlan, R.A., Story, B.H., Bunton, K. 2013. Relation of perceived breathiness to laryngeal Kinematics and acoustic measures based on computational modeling. *J Speech Lang Hear Res*. 56, 1209–1223.

[35] Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A., Schwartz, D.M. 1978. Correlates of psychological dimensions in talker similarity. *J Speech Hear Res*. 21, 265–275.

[36] Murry, T., Singh, S., Sargent, M. 1978. Multidimensional classification of normal voice qualities. *J Acoust Soc Am*. 64, 81–87.

[37] Murry, T., Singh, S. 1980. Multidimensional analysis of male and female voices. *J Acoust Soc Am*. 68, 1294–1300.

[38] Kreiman, J., Gerratt, B.R., Precoda, K., Berke, G.S. 1992. Individual differences in voice quality perception. *J Speech Lang Hear Res*. 35, 512–520.

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

[40] Van Lancker, D., Kreiman, J. 1987. Voice discrimination and recognition are separate abilities. *Neuropsychologia*. 25, 829–834.