Interspeech 2022
18-22 September 2022, Incheon, Korea

Linguistic versus biological factors governing acoustic voice variation

Yoonjeong Lee[1,2], *Jody Kreiman*[1]

[1]Department of Head and Neck Surgery, David Geffen School of Medicine at University of California, Los Angeles, Los Angeles, California, USA
[2]Department of Linguistics, University of Michigan, Ann Arbor, Michigan, USA
yoonjeonglee@ucla.edu, jkreiman@ucla.edu

Abstract

This study presents a cross-linguistic investigation of acoustic voice spaces in English, Seoul Korean, and White Hmong, which differ in whether they phonologically contrast phonation type and/or tone. The overarching hypothesis is that acoustic variability in voice will be shaped by biological factors, linguistic factors, and individual idiosyncrasies. By employing principal component analysis on speakers' read speech productions, we identify how individual and population voice spaces are acoustically structured for speakers of these three languages. Results revealed several factors that consistently account for acoustic variability across speakers and languages, but also factors that vary with language-specific phonology.

Index Terms: voice quality, acoustic variability, cross-linguistic comparisons, White Hmong, Seoul Korean, English, read speech

1. Introduction

Acoustic variability within and between speakers is well documented, but poorly understood. It has long been known that speakers vary widely in quality in both the short and long term, due to changes in age, health status, social context, mood, topic, speaking style, intent to deceive, and an array of other factors ([1] for review). Because the input to voice perception is the acoustic voice signal, such variability necessarily has implications for speaker discrimination and recognition. However, it is difficult to establish exactly what these implications are without understanding how, and how much, individuals vary acoustically. Better knowledge of the nature and extent of within- and between-speaker variability is thus critical to formulating models of voice perception and speaker identification.

Our previous work [2] used principal component analysis (PCA) of sentences read by 50 female and 50 male native speakers of English [3] to identify voice quality indices that account for perceptually relevant acoustic variance within individual speakers and for the pooled groups of speakers, using variables drawn from a psychoacoustic model of individual voice quality. Results were remarkably consistent across speakers and groups: Acoustic variability for every speaker in our 100-voice set was best characterized by the balance between high-frequency harmonic and inharmonic energy in the voice (measured using cepstral peak prominence [CPP; 4], a robust harmonics-to-noise ratio) and by formant dispersion. Although these dimensions accounted for the most variance among the extracted principal components, this amounted to only slightly more than half of the variance in the underlying acoustic data. Coherent, shared acoustic patterns did not emerge from the remaining explained variance. Instead, patterns of variability differed idiosyncratically from voice to voice. The measures that best characterized within-speaker acoustic variability also emerged from analyses of groups of female and male talkers, suggesting that individual and population voice spaces have very similar acoustic structures.

These results have replicated for recordings of unscripted telephone conversations from the same speakers [5]. Although spontaneous speech (unsurprisingly) proved more acoustically variable than read speech, the same variables emerged in the first principal component for both speaking styles, with one notable difference: Variability in fundamental frequency (F0) accounted for significant acoustic variation for conversations, but not for read speech. The fact that the same dimensions consistently emerged across speakers and styles suggests that acoustic variability is shaped by factors that are also shared across individuals, possibly through an evolutionary process. Consistent with the existence of such a process, these factors also characterize vocal variability across many species, and appear to provide survival benefits. For example, across species formant dispersion is associated with physical size [6], and the balance of harmonic and inharmonic energy is associated with arousal [e.g., 7, 8]. This information can communicate reproductive fitness or hostile/benign intent, thus providing a survival benefit, and such vocal signaling is common in many animals ([6, 7, 8]; see [1] for review).

Aside from biological factors, the demands of speech production also introduce variation in voice, the particular nature of which varies with the phonology of the language in question. To examine the extent to which phonological structure (versus biological factors) contributes systematically to within- and between-speaker acoustic variability, this study extends our previous work to native speakers of two additional languages: Seoul Korean and White Hmong. These languages differ systematically from English in the linguistic status of phonation contrasts and the linguistic use of F0 (Table 1): Seoul Korean is not a tone language but exhibits regular intonation patterns specific to a phonological phrase [9], and White Hmong contrasts both tone (7 lexical tones) and phonation (a breathy or creaky vowel contrasts with a modal vowel) [10]. We hypothesize that biologically-relevant variables—the interaction of high-frequency harmonic and inharmonic energy plus formant dispersion—will emerge as primary for all speakers, regardless of the language spoken, as they did for speakers of English. We further hypothesize that these shared primary dimensions of variability will be supplemented by additional principal components reflecting the phonology of the specific language being spoken. Specifically, we hypothesize that F0 will emerge from analyses of Korean and Hmong, and that differences in the amplitudes of the first and second harmonics will emerge for Hmong, but not for Korean. Such a result would be consistent with the view that

10.21437/Interspeech.2022-10847

speaker variability is governed by both biological and linguistic factors.

2. Method

2.1. Voice samples

Samples of read speech were drawn from databases for three languages: English [3], Seoul Korean [11], and White Hmong [10] (Table 1). English and Korean samples were collected in a sound-attenuated booth, in which speakers read short sentences. The English recordings were made using a microphone suspended from a baseball cap worn by the speakers, and the Korean recordings were made using a desktop microphone placed about 2 to 4 inches from the speakers. The Hmong corpus was collected in the field (in a quiet room in the speakers' homes). It consists of recordings of story reading. These languages form an organized set with respect to their linguistic use of F0 (whether lexical or phrasal tone) and the phonological status of phonation quality.

Table 1: *Voice samples from read speech.*

Language	Tonal contrast	Phonation contrast	Speaker
English	N	N	50F, 50M
Seoul Korean	Y (phrasal)	N	5F, 5M
White Hmong	Y (lexical)	Y	5F, 3M

2.2. Acoustic measurements

Following Lee et al. [2], we measured the variables listed in Table 2 for all vowels and approximants in each recording in the three language data sets. These variables were selected because as a set they validly quantify the complete sound of a voice [12]. F0 is the fundamental frequency of phonation, usually associated with vocal pitch; F1, F2, F3, and F4 are the frequencies of the first four resonances of the vocal tract (formant frequencies), and FD is formant dispersion, calculated as the average difference in frequency between pairs of adjacent formants [6]. The four measures of the shape of the harmonic voice source spectrum (H1*-H2*, H2*-H4*, H4*-H2kHz*, and H2kHz*-H5kHz) are the relative amplitudes of the specified harmonics, where * indicates that these values were corrected for the influence of formants on harmonic amplitudes [13]. H1* is the first harmonic, H2* is the second harmonic, H4* is the fourth harmonic, H2kHz* is the harmonic nearest in frequency to 2 kHz, and H5kHz is the harmonic nearest in frequency to 5 kHz. CPP is cepstral peak prominence, a robust measure of the relative amounts of harmonic versus inharmonic energy in the voice [4]; energy is the root mean square energy calculated over five phonatory cycles; and SHR is the subharmonics-to-harmonics ratio, which quantifies the extent of period doubling present in the voice [14].

Variables were measured automatically every 5 ms using VoiceSauce software [15]. Data frames with spurious parameter values (e.g., impossible 0s) were removed, after which values for each variable were normalized with respect to the overall minimum and maximum values across the entire set of female or male voice samples, as appropriate. Finally, we calculated moving averages and moving coefficients of variation (CoVs) for all 13 variables (*moving CoV = moving σ/moving μ*), using a smoothing window of 50 ms, for a total of 26 variables. Across speakers, these post-processing steps resulted in 515k data frames for English (F: 266k, M: 249k), 1.05m for Seoul Korean (F: 556k, M: 493k) and 2.03m for White Hmong (F: 1.26m, M: 772k).

Table 2: *Acoustic variables.*

Variable categories	Acoustic variables
Pitch	F0
Formant frequencies	F1, F2, F3, F4, formant dispersion (FD)
Harmonic voice source spectral shape	H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz
Inharmonic source/spectral noise	cepstral peak prominence (CPP), energy, subharmonics-to-harmonics ratio (SHR)
Variability	coefficients of variation (CoVs) for all acoustic measures

2.3. Principal component analysis

To determine which variables characterize acoustic differences within and across speakers, principal component analysis (PCA) was performed as described in [2]. All 26 acoustic variables were simultaneously entered into the analyses. An oblique rotation was applied [16, 17] and the factors with eigenvalues greater than 1 were retained [18]. Each factor was interpreted with respect to variables with loadings at or exceeding 0.32 [19]. PCAs were conducted separately for each speaker (within-speaker analyses) and for complete groups of female and male speakers within that language group (combined speaker analyses). Patterns of acoustic variability found within speakers were largely similar to those for the multi-talker spaces, so only the group analyses will be reported here (see [2, 20, 21] for individual speaker analyses).

3. Results

Across languages, analyses produced from 7 to 9 principal components (PCs) and accounted for 64-71% of the variance in the underlying data. The first three PCs were largely shared across speakers, together accounting for about 50% of the explained variance in the underlying acoustic data. The remaining PCs, cumulatively explaining about 20% of the variance, differed widely across speaker groups (i.e., genders and languages).

Results are summarized in Table 3. The first 4 rows of this table list variables that explain significant variability in all 3 languages. Variables in the first three rows define the first PC and explain the most variance for all groups of speakers. This PC weighs on the coefficients of variation in H1*-H2*, H2*-H4*, H4*-H2kHz*, and H2kHz*-H5kHz. The second row (CoVs for F1 and F2 values) and third row (energy CoV) also emerged for female and male speakers of Korean and White Hmong (but not English), and represented variability in vowel quality and noise across utterances.

The variables listed in the next two rows consistently appeared in PC2 or PC3 across speaker groups. The fourth row represents the shape of the higher-frequency part of harmonic voice source spectrum and F2 (the mid-frequency range). The fifth row includes formant dispersion and high formant frequency measures, which emerged in PC2 or PC3 for most speaker groups but accounted for less variance for the male Korean group (PC4), and even less variance for the male Hmong group. For the male Hmong group, the formant

dispersion measures appeared after the component that consisted of energy, H2*-H4*, and F0.

Table 3: *Cross-linguistic comparison of acoustic variables that weighed on the first three PCs.*

Variables	English		Korean		Hmong	
	F	M	F	M	F	M
Harmonic spectral shape CoV + CPP CoV	✓	✓	✓	✓	✓	✓
F1 CoV + F2 CoV	✗	✗	✓	✓	✓	✓
Energy CoV	✗	✗	✓	✓	✓	✓
H4*-H2kHz* + H2kHz*-H5kHz + F2	✓	✓	✓	✓	PC4	✓
F3 + F4 + FD	✓	✓	✓	PC4	✓	✗
F0 CoV	✗	✗	✓	✓	✓	✓
F0	✗	✗	✓	✓	✗	✗
H1*-H2* (+ SHR)	✗	✗	✗	✗	✓	✓

The bottom three table lines apply to subsets of the languages under study. English does not weigh heavily on any other variables after the first 5 lines of Table 3. Korean and Hmong both weighed heavily on variability in F0 (which emerged in the first principal component), consistent with their phonologies. Korean additionally weighed on mean F0 values that emerged in PC2. Analyses for both female and male speakers of Hmong weighed on H1*-H2*, and the male Hmong group also weighed on SHR, which emerged together with H1*-H2*. Additionally, for the Korean male group H2*-H4* emerged together with F1 and H1*-H2* in PC3.

4. Discussion

Variability conveys a wide array of information. Our previous studies of large groups of female and male speakers of English [2, 5] showed remarkably consistent patterns of acoustic variability, both within and between speakers, combined with idiosyncratic personal details. The present results replicate this finding and extend it by demonstrating that acoustic voice variability is partially universal in that certain dimensions of variability are shared by all speakers examined to date, and partially determined by language-specific factors.

The finding that certain aspects of variability recur so consistently across speakers, regardless of sex or language spoken, suggests that these aspects convey biologically important information. In our analyses, the first principal component to emerge always reflected variations in the balance of harmonic and inharmonic energy in the voice source. This combination of parameters is often associated with a quality continuum from "strained" or "pressed" to "breathy" [6, 22], which signals arousal across many species ([6, 7]; see e.g., [1, Table 9.1], for review of this property in human phonation). This information is employed across species for assessing the hostile or friendly intent of another animal and for communicating one's own intent, potentially altering the behavior of another animal as a result [23, 24]. Similarly, formant dispersion emerged consistently from earlier principal components. This sexually dimorphic parameter varies with the size of the vocal tract and is an important signal of both dominance and reproductive fitness across many species [5].

Two other factors—variability in the frequencies of the first and second formants and in energy—also emerged in the early principal components for many speaker groups. Variability in F1 and F2 is associated with changes in vowel quality across utterances, and variability in energy reflects changes in the loudness of the utterances. Both of these dimensions emerged for Hmong and Korean, but not for English. These differences likely reflect differences in the construction of the corpora and in recording conditions. In particular, mouth-to-microphone distance was controlled for the English recordings, but not for Hmong or Korean.

The appearance of additional acoustic variables in the first three components varied with the phonological structure of the language being spoken. As predicted, variability in fundamental frequency (F0) accounted for significant acoustic voice variability for White Hmong and Seoul Korean, whose phonological systems employ the tonal variable, but not for English, which is non-tonal. In addition to F0 variability, Seoul Korean, but not Hmong, also weighed heavily on mean F0. Korean and Hmong differ in whether the pitch variable is manifested for lexical contrast or prosodic phrasing. Our analyses seem to capture the language-specific use of a variable that is fine-tuned by the language's phonological structure.

Lack of F0 differences among English speakers may also reflect the use of read sentences in these analyses, which tend to be highly stylized across speakers. F0 variability did emerge from parallel PCAs using recordings of spontaneous speech from the same English speakers [4], consistent with this explanation.

Such comparisons also reveal that differences in the amplitudes of lower harmonics explain significant variance for Hmong voices, but not for English or Korean. White Hmong exhibits a linguistically defined set of contrasts in phonation quality: H1*-H2* is associated with the language's breathiness contrast, and SHR measures the creakiness associated with low-falling tones [9]. These variables explained significant amounts of acoustic variability for Hmong, although differences between male and female speaker groups in the importance of SHR remain to be explained.

These results suggest that as phonological complexity increases (i.e., from no tone/phonation to tone/no phonation to tone/phonation), the acoustic voice space for a language also increases in complexity. We note that, although acoustic variables associated with arousal always emerged as the first PC in our analyses, formant dispersion sometimes emerged before language-specific features, and sometimes after these features. Understanding the factors that govern order of emergence in these analyses requires further data from a broader range of languages. The manner in which acoustic voice spaces systematically vary with phonological structure is also consistent with the existence of an "own language" advantage in speaker discrimination and recognition [e.g., 25], in which listeners are more accurate in identifying or telling apart speakers whose language the listeners know. These results suggest a potential mechanism explaining this effect: Listeners have detailed knowledge of the manner and extent to which speakers of their native language differ from one another, based on life-long experience listening to voices. Hypothetically, these expectations are generalized when the listener is confronted with speakers of other languages, with the success of that generalization depending on how close the languages are phonologically [e.g., 26]. This hypothesis remains to be explored in detail.

Finally, we note that this shared acoustic structure accounts for only about a half of the acoustic variability in the individual and group data, with remaining variability being idiosyncratic. These idiosyncratic details appear to be critical for "telling people together," that is, for determining that different voice samples were spoken by the same person. In contrast, shared features like those that emerged in the first 3 PCs appear more important for "telling voices apart," which requires assessing the position of two samples in a shared voice space [27].

Several limitations to this study must be noted. Our sample of languages is incomplete, in that it does not include a language with lexical tones, but no phonation contrast, or one with a phonation contrast, but no tone. We have begun analyses of additional languages in response to this issue. There are also differences between data sets in the number of speakers recorded, the recording conditions, and the texts speakers produced, which add additional variability to the data. It is notable that even lacking perfectly matched stimuli and rigorous experimental controls, results clearly show the effects of both biology and the language spoken.

In summary, there are certain variables (listed in the first 3 lines of Table 3) that seem to characterize acoustic variability in voice no matter who is speaking or what language they speak. Other variables are needed to account for the variability introduced by features of the relevant linguistic system; but this process is additive, not subtractive: linguistically determined variability exists in voice spaces alongside biologically derived variability. Acoustic voice spaces are structured first by biologically driven factors, with linguistic factors adding additional complexity to the acoustic spaces for individual languages.

5. Conclusions

Our findings suggest that both biologically and phonologically relevant factors shape acoustic voice spaces.

6. Acknowledgements

This research was funded by NIH DC01797 and NSF IIS-1704167. NSF SPRF-2105410 and NIH DC011300 supported YL for manuscript preparation. We are grateful to Marc Garellek and Christina Esposito for kindly providing their data for White Hmong.

7. References

[1] J. Kreiman and D. Sidtis, *Foundations of Voice Studies*. Malden, MA: Wiley, 2011.

[2] Y. Lee, P. Keating, and J. Kreiman, "Acoustic voice variation within and between speakers," *Journal of the Acoustical Society of America*, vol. 146, pp. 1568–1579, 2019.

[3] P. Keating, J. Kreiman, A. Alwan, A. Chong, and Y. Lee, *The UCLA Speaker Variability Database*. LDC 2021S09. Web Download. Philadelphia: Linguistic Data Consortium. doi: 10.35111/c5gk-6j49, 2021.

[4] J. Hillenbrand, and R.A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, pp. 311–321, 1996.

[5] Y. Lee and J. Kreiman, "Acoustic voice variation in spontaneous speech," *Journal of the Acoustical Society of America*, vol. 151, pp. 3462–3472, 2022.

[6] W.T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *Journal of the Acoustical Society of America*, vol. 102, pp. 1213–1222, 1997.

[7] A. Anikin, "A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations," *Phonetica,* vol. 77, pp. 327–349, 2020.

[8] J.V. Congdon, A.H. Hahn, et al., "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," *Journal of Comparative Psychology*, vol. 133, pp. 520–541, 2019.

[9] S.A. Jun, *The Phonetics and Phonology of Korean Prosody*. Columbus, OH: Ohio State University, 1993.

[10] M. Garellek, P. Keating, C. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *Journal of the Acoustical Society of America*, vol. 133, pp. 1078–1089, 2013.

[11] M. Oh and D. Byrd, "Data for: syllable-internal corrective focus in Korean," *Mendeley Data*, vol. 2, doi: 10.17632/mt6fph6v2c.2, 2019.

[12] J. Kreiman, Y. Lee, M. Garellek, R. Samlan, and B.R. Gerratt, "Validating a psychoacoustic model of voice quality," *Journal of the Acoustical Society of America*, vol. 149, pp. 457–465, 2021.

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

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

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

[16] R.B. Cattell, The scree test for the number of factors. *Multivariate Behaviroal Research*, vol. 1, pp. 245–276, 1966.

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

[18] H.F. Kaiser, "The applications of electronic computer to factor analysis," *Educational and Psychological Measurement*, vol. 20, pp. 141–151, 1960.

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

[20] Y. Lee and J. Kreiman, "Language effects on acoustic voice variation within and between talkers," *Journal of the Acoustical Society of America*, vol. 148, p. 2473, 2020.

[21] Y. Lee, M. Garellek, C. Esposito, and J. Kreiman, "A cross-linguistic investigation of acoustic voice spaces," *Journal of the Acoustical Society of America*, vol. 150, p. A191, 2021.

[22] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B.R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation." *Journal of the Acoustical Society of America*, vol. 132, pp. 2625–2632, 2012.

[23] C.T. Snowdon, "Expression of emotion in nonhuman animals," in *Handbook of Affective Sciences*, edited by R.J. Davidson, K.R. Scherer, and H.H. Goldsmith. Oxford: Oxford University Press, pp. 457–480, 2003.

[24] J.-A. Bachorowski and M.J. Owren, "Sounds of emotion: Production and perception of affect-related vocal acoustics," *Annals of the New York Academy of Science*, vol. 1000, pp. 244–265, 2003.

[25] T.K. Perrachione, S. N. Del Tufo and J. D. Gabrieli, "Human voice recognition depends on language ability," *Science*, vol. 333, p. 595, 2011.

[26] C.P. Thompson, "A language effect in voice identification," *Applied Cognitive Psychology* vol. 1, pp. 121–131, 1987.

[27] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech," *Journal of the Acoustical Society of America*, vol. 151, pp. 1393–1403, 2022.