The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality

Jody Kreiman[1], Soo Jin Park[2], Patricia A. Keating[3], and Abeer Alwan[2]

[1] Department of Head and Neck Surgery, School of Medicine,
University of California, Los Angeles, USA
[2] Department of Electrical Engineering, University of California, Los Angeles, USA
[3] Department of Linguistics, University of California, Los Angeles, USA

jkreiman@ucla.edu, sj.park@ucla.edu, keating@humnet.ucla.edu, alwan@ee.ucla.edu

Abstract

Little is known about intraspeaker changes in voice across changing speaking situations in everyday life. In this study, we examined acoustic variations between and within 5 talkers and their effect on the likelihood that voice samples would not be identified as coming from the same talker. Talkers were drawn from a large database recorded to capture everyday variations in vocal characteristics. Nine samples of /a/, recorded on three different days, were examined for each talker. Acoustic characteristics were estimated using VoiceSauce and analysis-by-synthesis, and listeners judged whether pairs of voices came from the same or two different talkers. Results indicate that interspeaker variability in voice quality exceeds intraspeaker variability, but differences are smaller than expected. As predicted by models that treat voice quality as an auditory pattern, the acoustic attributes associated with incorrect "different speaker" responses varied from talker to talker, depending on the particular characteristics of the voice in question.

Index Terms: voice quality, speaker recognition, intra-speaker variability

1. Introduction

Although the study of voice quality has a long history in many disciplines, little is known about the extent to which an individual voice pattern varies across the kinds of speaking situations that arise in normal, every-day life. Although voices are sometimes described as auditory patterns [1], it is difficult to know what this actually means without data describing the acoustic input—and the acoustic variability—that the auditory system transforms into these patterns. This paper describes a preliminary investigation into acoustic variation in individual talkers and its implications for listeners' perceptions of a voice sample as coming from a particular talker. Specifically, we addressed these questions: 1) Are talkers consistently more similar to themselves acoustically and perceptually than they are to other talkers?; and 2) What is the relationship between acoustic and perceptual similarity?

2. Database and Recording Procedures

The first step in addressing these questions is developing a database containing multiple recordings of speakers recorded in a variety of speaking tasks on multiple occasions. To our knowledge none of the existing multi-talker speech databases offers the desired combination of a large number of talkers (both male and female), multiple recording sessions per talker, multiple speech tasks per talker, and very high quality audio (controlled recording conditions, good quality microphone, high sampling rate, etc.). The final database will comprise speech from 200 talkers (100 male, 100 female); the current database includes over 185 complete recordings. All talkers are undergraduate students at UCLA.

Audio recordings are made in a sound-attenuated booth using a ½" Brüel & Kjær microphone suspended from a baseball cap worn by the talker. Each talker participates in three separate recording sessions. All speech is elicited via on-screen displays implemented in Matlab. The first speech task in each session is three utterances of the isolated vowel /a/ (which was chosen because the high F1 reduces errors in estimation of voice source parameters). The second is two repetitions of five Harvard sentences (the same sentences for all talkers and all sessions, randomized for each recording). The sentences task provides samples of read speech, and both of these tasks allow cross-session comparisons.

Each session then includes two further speech tasks, different in each session, for a total of six one-time-only speech tasks. In the first session, talkers are instructed to talk to the research assistant (RA) who is outside the booth, giving her either directions on how to go somewhere, or instructions on how to do something (a set of suggested topics is provided). They are told to speak for at least 30 seconds, and an on-screen display counts out 30 seconds. This task provides a sample of clear but unscripted speech. Next, participants are instructed to repeat to the RA a conversation they had recently that wasn't important – not exciting, not upsetting, just normal. Again, some possible topics are provided, and again, the on-screen display prompts for 30 seconds of speech. This task provides a sample of unscripted low-affect speech.

In the second session, participants are instructed to repeat to the RA a conversation they had recently about something exciting that made them really happy. As before, some possible topics are provided and the on-screen display prompts for 30 seconds of speech. This task may provide a sample of positive-affect speech. Next, participants use their cell phones to call a friend or relative and talk for at least two minutes. Only the participant's side of the conversation is recorded. This task provides a sample of unscripted conversational speech.

In the third session, participants are instructed to repeat to the RA a conversation they had recently about something that really annoyed them. As before, some possible topics are provided and the on-screen display prompts for 30 seconds of speech. This task may provide a sample of negative-affect speech. Finally, participants watch a 1-minute video of cute kittens or puppies, and are asked to talk aloud to the pets as they watch the video. This task provides a sample of pet-directed speech, which typically has exaggerated prosody [2].

The goal of recording speech from different conditions for the database is to sample normal, daily-life voice variation. We do not try to elicit voice disguises, impersonations, acted emotions, or other dramatic acting. Instead we focus on normal variability in real-life situations, to the extent that these can be elicited in a sound booth. The point of the different conditions is not to study them as such, but simply to enhance the likelihood of sampling realistic amounts of within-talker variability in voice quality.

The voices of five female talkers were selected at random from this database for use in the following experiments. Only the 9 tokens of sustained /a/ were studied, given the preliminary nature of the following experiments.

3. Acoustic Analysis

3.1. Selection of Measures and Data Reduction

Based on a study of patterns of variability across talkers in source spectral shapes and glottal pulse shapes [3], we have developed a spectral model of the voice source that includes six factors: F0, H1-H2 (the amplitude difference between the first and second harmonics), the slope of the harmonic spectrum from H2-H4 (the second to the fourth harmonic), spectral slope from H4-2 kHz, spectral slope from 2 kHz to 5 kHz, and the harmonics-to-noise ratio (HNR), which measures harmonic energy normalized by the spectral noise level [4]. Extensive studies (e.g., [5][6][7][8]) have shown that listeners are perceptually sensitive to all six parameters, and that, as a set, the parameters are sufficient to quantify source contributions to normal personal voice quality, so that the model can be considered perceptually valid.

The literature on voice quality also includes a number of other measures that were included for the sake of completeness. These include measures of HNR in 4 discrete frequency ranges, H1*-A1*, H1*-A2*, H1*-A3* [9], Energy (the Root Mean Square energy, calculated at every frame over a variable window equal to five pitch pulses), and Cepstral Peak Prominence (CPP [10], a measure of signal periodicity). HN* denotes the N-th source spectral harmonic magnitude and AN* denotes the amplitude of the harmonic closest in frequency to the N-th formant. The asterisk (*) indicates a correction for the influence of vocal tract resonances using the formula given in [11]. Finally, measures of F1, F2, and F3 were included because vowel quality differed substantially across (and occasionally within) talkers.

Measures for all parameters were made using 25-msec Hamming window with 1-msec frame interval across the entire duration of each vowel token. The parameters are sampled every 100 msec by averaging over ±50 msec span. . every 100 msec across the entire duration of each vowel token. Source measures were extracted with VoiceSauce [12], and formant frequencies were measured using the Snack option within VoiceSauce [13]. Measures were screened for outliers (which were treated as missing values), and source spectral measures were validated using analysis-by-synthesis [14].

Correlation and canonical correlation were used to examine patterns of association among the many acoustic variables, including F0. The 4 HNR measures were significantly and substantially intercorrelated with each other and with CPP (mean $r = 0.95, p < .001$), so only CPP was retained for subsequent analyses. Similarly, canonical correlation indicated that H1*-H2*, H2*-H4*, H4*-2kHz*, and 2kHz*-5kHz could be predicted as a set from H1*-A1*, H1*-A2*, H1*-A3*, and Energy ($R^2 = 0.88$), so only the first set of variables was retained. Finally, formant frequencies were included in the final set of measures due to prominent differences in vowel quality within and across speakers. Note that the final set of 9 acoustic measures is equivalent to our proposed psychoacoustic model; the observed correlations between model parameters and other variables suggests that adding parameters to the model would not increase its explanatory power.

Figure 1 illustrates within- and across-talker mean and standard deviations for the 9 acoustic measures. Note, for example, that CPP's within-talker standard deviation is larger for talker 4 than the across-talker standard deviation.

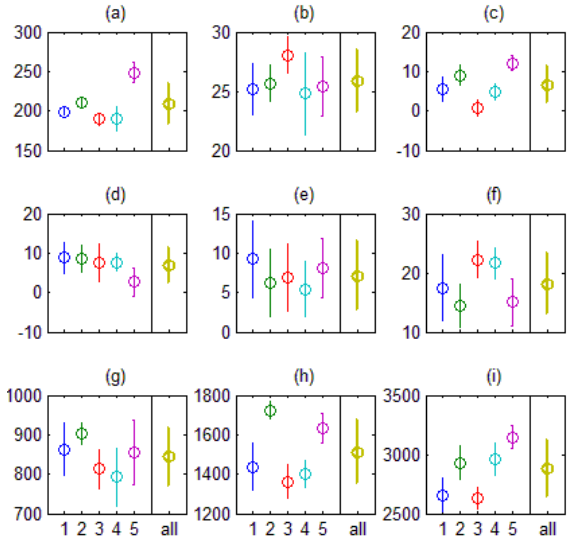

Figure 1: *Mean and standard deviation within-talker and across all talkers for (a) F0 (Hz) (b) CPP (c) H1*-H2* (dB) (d) H2*-H4* (dB) (e) H4*-2kHz* (dB) (f) 2kHz*-5kHz (dB) (g) F1 (Hz) (h) F2 (Hz) (i) F3 (Hz)*

Table 1: *Normalization ranges for the variables*

F0	CPP	H1*-H2*
93 ~ 275Hz	15 ~ 32	-2.64 ~ 21.6 dB
H2*-H4*	H4*-2kHz	2kHz*-4kHz
-0.39 ~ 29.2 dB	0 ~ 37.7 dB	0 ~ 46.87 dB
F1	F2	F3
522 ~ 1163 Hz	963 ~ 2701 Hz	2138 ~ 3490 Hz

All nine measures were normalized to a 0-1 scale using the range for that variable observed in previous studies [6][15][16]. The normalization ranges are shown in Table 1. Averages were then calculated for each talker and assembled into vectors. The average (Euclidean) acoustic distance between talkers was calculated from these vectors under the

assumption of equal weighting of measures. Note that each talker was represented by an average on each measure, so that acoustic variability within tokens was not captured in our final analyses.

3.2. Results and Discussion

Average acoustic distances between talkers are given in Table 1Table 2, and average acoustic distances between the nine tokens from each individual talker are given in Table 3. The average distance between talkers (45.04) did exceed the average distance between tokens from a single talker (34.86), but ranges overlapped considerably, with variability within an individual often exceeding differences between talkers. Thus it is not the case that talkers are consistently more acoustically similar to themselves than they are to other talkers. For example, talker 1's voice samples differed more from one another, on average, than talker 1 differed from talkers 2, 3, or 4. Talkers 3 vs 5 were by far the most distinctive pair acoustically, and in general talker 5 differed the most from the other talkers.

Table 2: *Average acoustic distances between the 5 talkers.*

	Talker B			
Talker A	2	3	4	5
1	36.9	31.1	32.4	59.5
2		55.6	34.7	35.4
3			38.6	75.4
4				50.8

Table 3: *Average acoustic distances between the 9 tokens for each individual talker.*

	Talker				
Distance	1	2	3	4	5
Average	41.8	29.3	31.9	41.1	30.2
SD	1.00	0.90	0.80	1.20	1.00

4. Perceptual Experiment

4.1. Method

The nine /a/ vowels recorded from the five randomly-selected female talkers described above were used as stimuli in this experiment. To ensure that idiosyncratic vocal features like final creak or pitch declination were represented, the full unedited vowel samples were used. Given nine recordings from five talkers, the stimulus set included a total of 180 "same talker" pairs and 450 "different talker" pairs, for a total of 630 possible comparisons among stimuli. Two different randomizations of this set were created, each of which was divided into thirds to create 6 subsets of 210 listening trials.

Listeners were normal-hearing UCLA students and staff members. They received payment or class credit for their participation. Ten listeners were assigned at random to each of the stimulus subsets, for a total of 60 listeners in 6 groups; but across groups, each pair of stimuli was judged by 20 listeners. Listeners heard the pairs of stimuli over Etymotic insert earphones (model ER-1) at a comfortable constant listening level (interstimulus interval = 250 msec). Each pair could be played only once in each presentation order (AB/BA). Listeners were not told how many speakers were represented in the trials. For each pair of stimuli, they judged whether the voices represented one talker or two different talkers, and reported their confidence in their response on a scale from 1 (positive) to 5 (wild guess). The experiment was self-paced and listeners were encouraged to take as many breaks as needed. Testing lasted about 45 minutes on average.

4.2. Result and Discussion

4.2.1. Listener Accuracy

The probability of a correct "same talker" response (the hit rate) and the probability of an incorrect "same talker" response (the false alarm rate) are given in Table 4 for the 5 target stimuli. Each listener's same/different responses were combined with their confidence ratings to create a scale ranging from 1 (positive that voices are the same) to 10 (positive that voices are different), and d' was calculated from these recoded responses (Table 5).

Although hit rates are quite high overall, false alarm rates are also rather high, suggesting that listeners had difficulty distinguishing different talkers. This difficulty is reflected in the d′ rates in Table 5, particularly for speaker 1 vs. 3 and 1 vs. 4, which were not discriminable at above chance levels.

Table 4: *Overall rates of correct (hit) and incorrect (false alarm; FA) "same talker" responses for the 5 talkers.*

Talker	Hit Rate	FA Rate
1	.89	.43
2	.91	.34
3	.89	.37
4	.67	.43
5	.83	.22

Table 5: *d' values for pairs of talkers, with a hit defined as a correct "same talker" response.*

	Talker B			
Talker A	2	3	4	5
1	1.50	0.83	0.75	1.22
2		1.66	1.01	1.64
3			1.08	2.00
4				1.16

4.2.2. The Relationship between Perceptual and Acoustic Similarity

The correlation between the acoustic distances between talkers and d' equaled 0.68 ($p < .001$), suggesting that a similar set of cues mediates confusions between different talkers. To examine whether the parameters associated with incorrect "different talker" responses are also consistent across talkers, we used stepwise multiple regression to predict the likelihood of such responses from acoustic distance between the 9 tokens from each talker. This method was chosen over standard multiple regression because of the exploratory nature of these analyses, and because we wished to observe the order of entry

of the different variables into the equation. Results are given in Table 6. While this relationship is quite strong for three talkers (1, 4, 5), it is notably unexplanatory for talker 3. Further analysis is needed to understand why (by what acoustic measures) this talker sometimes sounded more like someone else than like herself.

Table 6: *Predicting incorrect "different talker" (Miss) responses.*

Talker	# misses	Pred. vars.	R^2
1	135	F0, F1, F2, H2H4, CPP	.62
2	90	F0	.45
3	143	CPP	.21
4	262	F1, F2, H1H2, CPP	.71
5	129	F1, F2, H2H4, 2k5k, CPP	.72

5. General Discussion

Returning to the two questions listed in the Introduction:

1. Are talkers consistently more similar acoustically and perceptually to themselves than they are to other talkers?

On average, talkers were acoustically more self-similar than they were similar to other talkers. Perceptually, Table 4 shows that listeners produced many more correct than incorrect "same talker" responses, indicating that (at least for the small sample of talkers studied here), talkers are indeed perceptually more similar to themselves than they are to other talkers. The differences are not overwhelming in either case, however.

2. What is the relationship between acoustic and perceptual similarity?

This question is perhaps best answered by, "It depends." Confusions between different talkers were reasonably well predicted (R^2 = .46) from average values of a set of static acoustic measures. However, results were much more variable when we attempted to predict cases where two voice samples were not identified as coming from the same talker. Although prediction was relatively good for 4/5 talkers, the parameters associated with rates in incorrect "different talker" responses differed from talker to talker. This result is consistent with models of voice quality that treat voices as patterns, so that the importance of any one acoustic parameter in determining the quality of a voice depends on the values of the other parameters that make up the pattern.

Further studies will include more detailed analysis of the acoustic vectors that characterize each talker, and how they are related to confusion between talkers. The study will also be expanded to analyzing a larger data set with more talkers (including males) and various kinds of speech.

6. Acknowledgements

This work was supported by NSF under Grant number IIS 1450992 and by NIH/NIDCD under Grant number DC 01797.

7. References

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

[2] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? On talking to babies and animals," *Science*, vol. 296, no. 5572, p. 1435, 2002.

[3] J. Kreiman, B.R. Gerratt, and N. Antoñanzas-Barroso, "Measures of the glottal source spectrum," *J. Speech Lang. Hear. Res.,* vol. 50, no. 3, pp. 595-610, 2007.

[4] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech Hear. Res*., vol. 36, no. 2, pp. 254-266, 1993.

[5] J. Kreiman and B.R. Gerratt, "Perceptual sensitivity to first harmonic amplitude in the voice source," *J. Acoust. Soc.Am.,* vol. 128, no. 4, pp. 2085-2089, 2010.

[6] J. Kreiman and B.R. Gerratt, "Perceptual interactions of the harmonic source and noise in voice," *J. Acoust. Soc.Am.*, vol. 131, no. 1, pp. 492-500, 2012.

[7] M. Garellek, P. Keating, C. M. Esposito, and J. Kreiman, "Voice quality and tone identification in White Hmong," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1078-1089, 2013a.

[8] M. Garellek, R. Samlan, J. Kreiman, and B. R. Gerratt, "Perceptual sensitivity to a model of the source spectrum," *Proc. Meet. Acoust.,* vol. 9, no. 1, 060157, 2013b.

[9] H. M. Hanson and E.S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.,* vol. 106, no. 2, pp. 1064-1077, 1999.

[10] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech Hear. Res.*, vol. 37, no. 4, pp. 769-778, 1994.

[11] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *J. Acoust. Soc. Am.,* vol. 121, no. 4, pp. 2283-2295, 2007.

[12] Y.-L. Shue, *The voice source in speech production: Data, analysis and models*, UCLA dissertation, 2010.

[13] 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.

[14] J. Kreiman, N. Antoñanzas-Barroso, and B. R. Gerratt, "Integrated software for analysis and synthesis of voice quality," *Behav. Res. Meth.,* vol. 42, no. 4, 1030-1041, 2010.

[15] R. J. Baken, *Clinical Measurement of Voice and Speech*, Boston: College Hill, 1996.

[16] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099-3111, 1995.