Within- and Between-Talker Variability in Voice Quality in Normal Speaking Situations

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

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

Introduction

Little is known about how, and how much, individual talkers vary their voice quality across normal everyday speaking situations.

In theory, this makes it difficult to specify what we actually mean when we describe a voice as an "auditory pattern."

In practice, lack of knowledge about normal within-talker variability limits ability to predict or explain confusions among voices.

- Forensic applications
- Development and evaluation of automatic recognition systems

Current Objectives

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

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

ACKNOWLEDGMENTS

This work was supported by NSF under grant number IIS 1450992, and by NIH/NIDCD under grant number DC 01797. Thanks to Anya Mancillas and Brenda Garcia for recording the speakers in the database.

Step 1: Develop a database

Database description

- 200 UCLA undergraduate talkers (100 female)
- 3 recording sessions on separate days
- A wide range of speaking tasks chosen to sample normal day-to-day and situation-to-situation variation in voice quality
  - Steady-state vowels (all sessions)
  - Harvard sentences (all sessions)
  - Spontaneous speech
  - Conversational speech
  - Lower- and higher-affect speech
  - Pet-directed speech
- Recorded in a sound-attenuated booth using a ½" Brüel & Kjær microphone suspended from a baseball cap worn by the talker
- Database will be publicly available.

Subset of data used in this study

- 5 females on 3 dates
- 9 tokens of the vowel /a/, 3 from each recording session

Table 1: *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 2: *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

Acoustic Analyses

Selection of measures

- A large set of measures (F0, H1*-H2*, H2*-H4*, H4*-2k*, 2k*-5k*, HNR in 4 frequency bands, H1*-A1*, H1*-A2*, H1*-A3*, RMS energy, cepstral peak prominence (CPP), F1, F2, F3)
- Measured every 100 msec across entire duration of each vowel
- Source measures collected with VoiceSauce and validated with analysis-by-synthesis
- Formant frequencies measured using the Snack option within VoiceSauce

Data reduction

- Correlation and canonical correlation

Final set: F0, H1*-H2*, H2*-H4*, H4*-2K*, 2K*-5K, CPP, F1 - F3

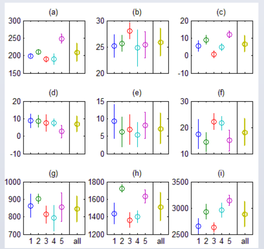

Figure 1: *Mean and standard deviation within-talker and across all talkers for*
(a) F0 (Hz)
(b) CPP
(c) H1-H2* (dB)*
(d) H2-H4* (dB)*
(e) H4-2kHz* (dB)*
(f) 2kHz-5kHz (dB)*
(g) F1 (Hz)
(h) F2 (Hz)
(i) F3 (Hz).

Acoustic distance

- Measures normalized from 0-1 using known ranges
- Average (Euclidean) acoustic distance with equal weighting of measures

Result and discussion

- The average distance between talkers (45.04) exceeded the average distance between tokens from a single talker (34.86) (Tables 1 and 2).
- Ranges for between- and within-speaker acoustic distances overlapped considerably.
- Talkers are not consistently more acoustically similar to themselves than they are to other talkers.

Perceptual Experiment

Method

- 180 "same talker" pairs and 450 "different talker" pairs
- 2 randomized sets, each of which was divided into 3 subsets (total 6 subsets)
- 60 normal-hearing listeners in 6 groups
- Each pair played once in each order (AB/BA)
- Listeners judged whether the voices represented one talker or two different talkers

Results and discussion

- Listener accuracy : Hit rates were quite high overall, but false alarm rates were also high (Table 3), suggesting that listeners had difficulty distinguishing different talkers.
- D' rates in Table 4 confirm that the listening task was difficult.

Table 3: *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 4: *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

The Relationship Between Perceptual and Acoustic Similarity

Acoustic distance significantly predicted confusability between talkers ($R^2 = 0.46$).

Within-talker acoustic variability predicted incorrect "different talker" responses for 4/5 talkers (Table 5).

Parameters implicated in failures of self-similarity varied from talker to talker.

Table 5: *Predicting incorrect "different talker" (Miss) responses.*

Talker	# misses	Predictive variables	R^2
1	135	F0, F1, F2, H2*-H4*, CPP	.62
2	90	F0	.45
3	143	CPP	.21
4	262	F1, F2, H1*-H2*, CPP	.71
5	129	F1, F2, H2*-H4*, 2k*-5k*, CPP	.72

General Discussion

Returning to our original questions:

1) Are talkers consistently more similar to themselves acoustically and perceptually than they are to other talkers?
   - On average, talkers are indeed acoustically and perceptually more similar to themselves than they are to other talkers. Differences are not overwhelming, however, and exceptions abound.

2) What is the relationship between acoustic and perceptual similarity?
   - It depends.
   - Variability in the determinants of self-similarity across talkers is consistent with models that treat voices as complex auditory patterns, so that the importance of any one feature depends on the complete configuration.
   - More detailed analysis of these data will shed light on what it means for a voice to be "a pattern."