Acoustical Society of America
2 Dec 2016, Honolulu
Poster 5aSC7

Acoustic similarity among female voices

Patricia Keating[1] and Jody Kreiman[2]

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

UCLA | The Bureau of Glottal Affairs

Introduction

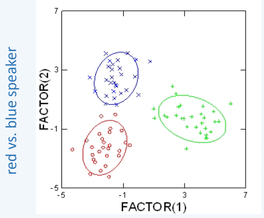
- Voices can be similar or different in many ways
  - many acoustic properties characterize voices
- Typical assumption: a few key parameters are critical in distinguishing individual speakers' voices
  - other parameters will tend to be similar across many speakers
- Some candidates in the literature for key parameters: F0, formant frequencies, nasality, breathiness, jitter, intensity (see references in [1])

Our research question: Are there a few acoustic parameters that do most of the work of characterizing speaker similarity/distinctiveness?

How to look for such parameters across voices?

- a low-dimensional analysis of the voice space for **many voices**?
  - [2] found that **no acoustic parameters** mapped to Multi-Dimensional Scalings of perceived similarity of 80 voices!
- low-dimensional analyses of many small sets of voices?
  - yields **a different parameter set for each set of voices**!
  - because which parameters are important depends on the particular voices compared

Example: preliminary Linear Discriminant Analysis of acoustics of sentence tokens from **3 female speakers** (see next column) – 2 discriminant factors each do some of the work, together giving 100% classification:



Here we focus on distinctiveness rather than similarity:
Which parameters make a voice different from other voices?

- Can a few acoustic parameters distinguish many different voices?
- Or will each additional voice require a new parameter?

Methods

Speakers:

- 50 women from the UCLA Speaker Variability Database
- ages 18-29 (mean=20, SD=1.9 years)
- native English speakers
- fairly homogeneous group with overall similar voices

Speech recordings:

- 5 Harvard sentences, read 6x each over 3 sessions (=30/speaker, total 1475 available tokens)
- recorded in a soundbooth with B&K mic @22k SR
- orthographic transcriptions -> force-aligned (by Penn Forced Aligner) phonemic transcriptions

Speech processing:

- only vowel and approximant intervals
- VoiceSauce [3,4], 42 acoustic parameters
- every 5 ms in 1475 tokens -> ~ 588k data frames
- removed frames with missing or extreme parameter values -> ~193k data frames remain (~.65 sec speech/token)
- for each sentence token, get MEAN and SD of each parameter

Parameter trimming:

- correlated parameters dropped
- -> 26 variables for further analysis (means, SDs):
  - **F0** (from STRAIGHT)
  - **H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k***
    (= the parameters of the source spectrum model of [5])
  - **F1, F2, F3, F4** (from Snack)
  - **Cepstral Peak Prominence (CPP)**
  - Means only: **Energy, Subharmonic-harmonic ratio (SHR)**
  - SDs only: **Harmonic-noise ratios in 4 frequency bands**
- Here, bare parameter name will refer to MEAN
- = a very limited acoustic model, with no dynamics or timing, no info about nasals or obstruents

Analyses:

- Multi-Dimensional Scaling (MDS) for overall map of voices
- Linear Discriminant Analyses (LDA) for voice distinctions

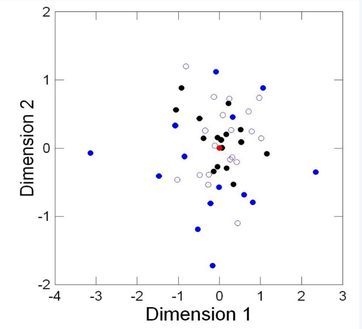
Analyses and Results

1. MDS acoustic space (all speakers)

- normalized variables, unweighted
- all pairwise acoustic distances between speakers
- MDS, 2-D solution **(R^2=.88)**
- correlations of acoustic variables with dimensions

Each dot = 1 voice
Red = acoustic center
Blue = voices most distinguishable below
Black = voices least distinguishable below
White = other voices

Dimension 1 ≈
F0, SHR, F3
Dimension 2 ≈
F4, H1*-H2*, CPP



Voice space: pitch, higher formants, creak/breathiness
But distinctiveness must depend on more than just these

2. LDA of 50 voices (all speakers/tokens)

- 3 eigenfunctions for 49.2% of variance
- **correct classification of tokens by speaker = 68.3%**
- very diverse *misclassifications* of speakers
- (Note: human classification is unknown)
- correlate LDA factors with acoustic variables
- correlated strongly with: **F0; F4, CPP; (then F1, CPP_{SD})**
- just these 5 variables alone classify **22.1%**
- correlated with: **F0, F4**

Top-2 variables on the 2 MDS dimensions above are most important here too
But most of the work is done by **all the other variables**

3. A different approach: 5-speaker subsets

- 198 5-speaker subsets drawn from the 50 voices
- each voice appears in 20 subsets, with all others
- LDA of the 5 speakers in each subset (150 sentences)
- correlate LDA factors with acoustic variables
- **count number of times each acoustic variable correlated above a threshold – how much work is each variable doing across all the 5-voice subsets?**
- most variables are useful only for distinguishing a given voice from just one or two other voices
- a few variables are strongly correlated with LDA factors, making many voice distinctions:

Energy, SHR, F0, F3, (F4), (CPP)

Mostly same variables as before, but re-ordered
And **Energy** now playing major role

Conclusions

- Similar key variables emerge from the different analyses
- a few parameters do more of the work of distinguishing the speakers – though not "most of the work"
- But all parameters contribute to characterizing the full set of voices – many voices require many parameters
- Surprisingly, the voice source spectrum model parameters [5] are not important here
- To pursue: Can results from looking at all voices together (Analyses #1, 2) converge with results from looking separately at many small subsets of voices (#3)?

Acknowledgments

This work is supported by NIH grant DC01797 and NSF grant IIS 1450992; we thank Neda Vesselinova and Naseem Fazeli for help with data analysis.

References

[1] Kreiman, Jody, and Diana Sidtis (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, p. 17
[2] Kreiman, Jody, and Bruce R. Gerratt (1996). "The perceptual structure of pathologic voice quality." *JASA* 100.3: 1787-1795.
[3] Shue, Yen-Liang (2010). The voice source in speech production: Data, analysis and models. PhD Dissertation, Department of Electrical Engineering, UCLA
[4] Shue, Yen-Liang, et al. (2011).VoiceSauce: A program for voice analysis. *Proceedings of ICPhS XVII*: 1846-1949.
[5] Kreiman, Jody, et al. (1996). "Perceptual evaluation of voice source models." *JASA* 138.1: 1-10.