Acoustical Society of America
4 Dec 2017, New Orleans
Poster 1pSC9

Acoustic similarities among voices. Part 2: Male speakers

Jody Kreiman[1,2], Patricia Keating[2] and Neda Vesselinova[2]

[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

UCLA

The Bureau of Glottal Affairs

Introduction

Voices can be similar or different in many ways [1]; many acoustic features can be measured and used to model voice similarity [e.g. 2-9].

Research question: Do some acoustic features do most of the work of characterizing voice similarity/ distinctiveness?

- Baumann & Belin [2]: voice quality features not important
- but in our previous work (Keating & Kreiman [10]): for 50 women's voices, *all* features mattered, with each feature key in distinguishing at least some voices

Research question: *Which* features are most important in characterizing voice similarity/ distinctiveness?

- Baumann & Belin [2] and Nolan et al. [11]: listeners' similarity ratings of men's voices were most related to F0, then higher formant(s)
- our previous work [10]: the same parameters were most important for distinguishing women's voices acoustically

Research question: Do men's and women's voices differ in this respect?

- [12], an early study on listeners' similarity ratings: F0 was more important for female voice similarity, but formants were more important for male voice similarity
- [2]: While F0 is most important for both men's and women's voices, the next most important feature for women's voices is F1 (whereas it is F4 and F5 for men – see above)
- but in our previous work [10]: best features for women's voices already were same as previously shown for men

Here we pursue these questions by analyzing 50 men's voices in the same way as we previously analyzed 50 women's voices

- Our analyses of women's and men's voices use more speakers, and more acoustic voice features, than most previous studies.

Methods

Speakers:

- 50 men from the UCLA Speaker Variability Database [15]
- UCLA undergraduate students
- native English speakers
- fairly homogeneous group with overall similar voices

Speech recordings:

- 5 Harvard sentences [16], read 6x each (over 3 sessions) (=28-30/speaker, total 1461 available tokens)
- recorded in a soundbooth with B&K mic @22k SR
- orthographic transcriptions -> force-aligned phonemic transcriptions (by Penn Forced Aligner, [17])

Speech processing:

- only the vowel and approximant intervals in the sentences
- VoiceSauce [18,19], 12 acoustic parameters every 5 ms (below)
- removed frames with missing or extreme parameter values -> ~262k data frames remain
- for each sentence token, get MEAN and Coefficient of Variation (COV) of each parameter **(→24 variables for analyses below)**

Acoustic parameters:

- F0 (from STRAIGHT)
- H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k*
  (= the parameters of the source spectrum model [14])
- F1, F2, F3, F4 (from Snack)
- Cepstral Peak Prominence (CPP)
- Energy
- Subharmonic-harmonic ratio (SHR) (~ creaky voice)

= a very limited acoustic model, with no dynamics or timing, no information about nasals or obstruents

Analyses:

- Linear Discriminant Analyses (LDA) (Discriminant in SPSS) – determine % correct classification of tokens by speakers
- correlate acoustic variables with each dimension of LDA solution – which variables relate most strongly to the dimensions doing the most work in classification?

Analyses and Results

1. LDAs of 50 voices (entire dataset)

Variables used	# eigen-vectors (R^2)	% tokens correctly classified by speaker
All 24	3 (57.2)	78.3
F0 only	1 (100)	7.9
All minus F0	3 (51.5)	73.0
5 highest-correlated only = F0 on #1; F4 on #2; H1*-H2*, SHR, SHR_{COV} on #3	3 (91.7)	28.3

- all **24 variables** together give respectable but not perfect classification of the 1461 sentence tokens (78.3%)
- **F0** is the variable that classifies the best on its own (8%), and contributes the most in addition to other variables (5%); H1*-H2* is next best; but even these variables do relatively little work
- even all 5 variables with high correlations on the 3 eigenvectors do not, by themselves, classify the tokens well (28.3%)

2. A different approach: 5-speaker subsets

- 198 5-speaker subsets drawn from the 50 voices
- each voice appears in 20 quintuples, with all other voices
- LDA of the 5 speakers in each quintuple (~150 sentences)
- correlate LDA eigenvectors with acoustic variables, as before
- **count # times each acoustic variable was the most important in distinguishing a speaker from the other 4 in a quintuple – how much work is each variable doing across all pairs in quintuples?**
- clear winner is **F0** (269 times); then **Energy** (86 times), **H1*-H2*** (85 times), **F4** (79 times), **$F4_{COV}$** (60 times); but every variable does some work

Variables used (from quintuples analyses)	# eigen-vectors (R^2)	% tokens correctly classified by speaker
5 best (F0, Energy, H1*-H2*, F4, $F4_{COV}$)	3 (83.9)	36.9
5 best minus F0 (Energy, H1*-H2*, F4, $F4_{COV}$)	4 (100)	24.8

- these 5 variables (including Energy rather than SHR) do slightly better at classification than the 5 above (36.9 vs 28.3%)
- they have already been selected on the basis of their ability to classify within quintuples of voices – here they scale up

Discussion

1. Men's voices

- No one feature or small set of features does most of the work of classification; instead, many variables are needed for good classification
- Most important acoustic variables for classifying men's voices are **F0, F4, and H1*-H2***, seen in both kinds of analysis above
- Better performance by features derived from pairwise speaker comparisons is in accord with what we know about voice perception: both pattern-matching based on a small set of core features, and ad-hoc analysis of many features, play important roles [1]

2. Comparison with women's voices

- **F0 and F4** are most important parameters for classifying voices of both sexes, and **Energy, H1*-H2*, and SHR** are also important for both; however, men's classification is better (78.3 vs 68%)
- No major difference between the feature set needed for men's vs. women's voices, unlike results in earlier literature.
- Quintuples of the women's voices had required **all** acoustic parameters for reasonable voice discrimination, but some parameters do **no** work in the men's quintuples.
- F0 does relatively more work in distinguishing pairs of men's voices than pairs of women's voices.

Acknowledgments

This work is supported by NIH grant DC01797, and NSF grants IIS 1704167 and IIS 1450992; we thank Anita Auszmann and Jordan Shavalian for help with data analysis.

References

[1] Jody Kreiman & Diana Sidtis (2011) *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons. **[2]** Oliver Baumann & Pascal Belin (2010) "Perceptual scaling of voice identity: common dimensions for different vowels and speakers", *Psychological Research 74*: 110-120. **[3]** J. Vaňková & R. Skarnitzl (2014) "Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality", *Proc. Speech Prosody 2014*: 1081–1085. **[4]** Finnian Kelly, Anil Alexander, Oscar Forth, Samuel Kent, Jonas Lindh & Joel Akesson (2016) "Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features", *Proc. Interspeech 2016:* 1567-68. **[5]** Hanyong Park & Noah H. Silbert (2016) "Speaker similarity, acoustic properties, and perceptual learning", poster presented at the Fall ASA meeting in Honolulu. **[6]** Eugenia San Segundo, Athanasios Tsanas & Pedro Gomez-Vilda (2017) "Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics", *Forensic Science International 270*: 25-38. **[7]** Abraham Woubie, Jordi Luque & Javier Hernando (2016) "Improving i-vector and PLDA based speaker clustering with long-term features" *Proc. Interspeech 2016*: 372-376. **[8]** Tomi Kinnunen & Haizhou Li (2009) "An overview of text-independent speaker recognition: From features to vectors", *Speech Communication 52*: 12-40. **[9]** Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia A. Keating & Abeer Alwan (2017) "Using voice quality features to improve short-utterance, text-independent speaker verification systems", *Proc. Interspeech 2017*: 1522-1526. **[10]** Patricia Keating & Jody Kreiman (2016) "Acoustic similarity among female voices", poster presented at the Fall ASA meeting in Honolulu. **[11]** F. Nolan, K. Mcdougall & T. Hudson (2011) "Some acoustic correlates of perceived (dis) similarity between same-accent voices", *Proc. ICPhS XVII*: 1506-1509. **[12]** T. Murray & S. Singh (1980) "Multidimensional analysis of male and female voices" J. *Acoust. Soc. Am. 68*: 1294–1300. **[13]** M. Garellek, R. Samlan, B. R. Gerratt & J. Kreiman (2016) "Modeling the voice source in terms of spectral slopes", *J. Acoust. Soc. Am. 139*(3): 1404-1410. **[14]** J. Kreiman, S. J. Park, P. A. Keating & A. Alwan (2015) "The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality" *Proc. Interspeech 2015*: 2357–2360. **[15]** IEEE Subcommittee on Subjective Measurements (1969) "IEEE Recommended Practices for Speech Quality Measurements" *IEEE Trans. Audio Electroacoust. 17* (297): 227–246. **[16]** I. Rosenfelder, J. Fruehwald, K. Evanini & J. Yuan (2011) FAVE (forced alignment and vowel extraction) program suite. *URL http://fave. ling. upenn. edu*. **[17]** Yen-Liang Shue (2010) *The voice source in speech production: Data, analysis and models*. Unpublished PhD dissertation, Department of Electrical Engineering, UCLA. **[18]** Y.-L. Shue, P. Keating, C. Vicenik & K. Yu (2011) "VoiceSauce: A program for Voice Analysis" *Proc. ICPhS XVII*: 1846-1849.