The role of between- versus within-speaker acoustic variability in vocal identity perception

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

[1]UCLA Bureau of Glottal Affairs; [2]Department of Linguistics, University of California, Los Angeles

jkreiman@ucla.edu; *yoonjeonglee@ucla.edu*

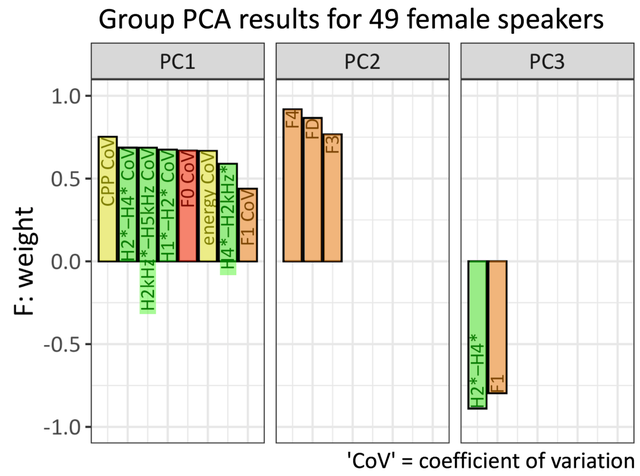
1. Variability in Voice Quality

- The acoustic voice signal
  - input to the perceptual system
  - highly variable
  - critical for formulating models of voice quality and talker recognition
- Current prototype-based models for voice identity perception
  - **Population prototype**: A context-dependent "average" or "ordinary-sounding" voice residing at the center of a multidimensional acoustical voice space
  - **Reference pattern**: Each voice's unique deviations from the group prototype (theoretically underspecified)

What comprises acoustic voice spaces? (Principal Component Analysis by Lee & Kreiman, 2019)

- Acoustic spaces characterizing within- and between-speaker variability in voice quality have similar structures, with a few features that are prominent for all speakers combined with idiosyncratic features characterizing individual talkers.
- This suggests that reference patterns for individual speakers and their population are mainly computed over the balance of **higher-frequency harmonic** versus **inharmonic energy** in the voice, **F0**, and over **formant dispersion**.

Group PCA results for 49 female speakers



How will listeners organize these identified measures of variability into a personal identity?

HYPOTHESIS: Errors in telling voices apart will be strongly predictable from distances among those few shared features in the group acoustic space, but errors in telling voices together will not be associated with these distances.

2. Voice Samples

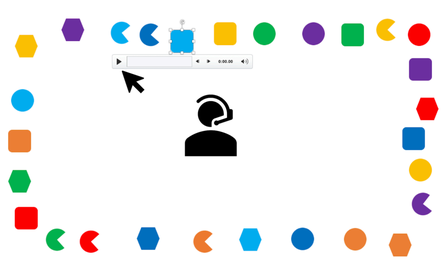
- Recordings of one side of unscripted phone conversation from 49 female speakers of English (F: 49, Age: 18–29) from the **UCLA Speaker Variability Database**
- 8 speakers based on the acoustic structure of voice spaces:
  - **P**rototypical (**P1**, **P2**, **P3**, **P4**, **P5**): similar to the population prototype
  - **A**prototypical (**A1**, **A2**, **A3**): deviated from the population prototype
    - PC1: **formant frequency CoV** (all **A** speakers)
    - PC2: **Formant dispersion** (**A1** & **A3**) or **F2** range (**A2**)
    - PC3: **F1** range (all **A** speakers)

    Note: **A1** & **A3** are closer to the population prototype than **A2**.
- 14 full phrases or sentences from each speaker (*mean duration* = 1.6 s, *SD* = 0.5 s), normalized for intensity

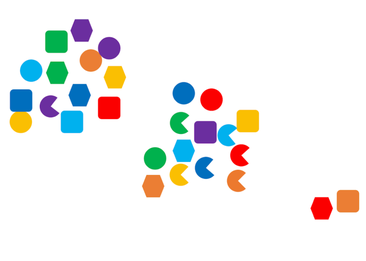
3. Voice Sorting Task

- Voice samples sorted into 3 sets of 4 trials
  - Each trial included samples from 2 of the speakers (28 samples total).
  - Across sets listeners heard unique pairs of speakers.
  - Across trials each listener heard every speaker.
  - No listener heard any speaker more than once.

Example testing interface



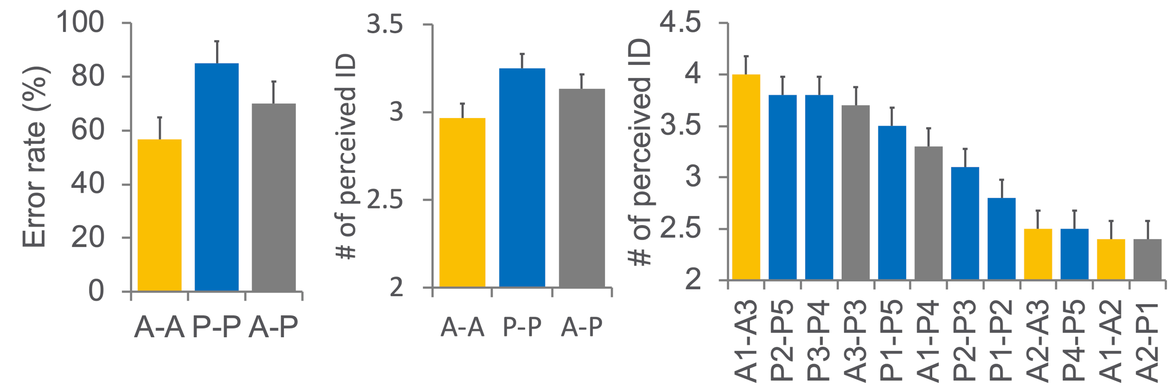
Example listener performance



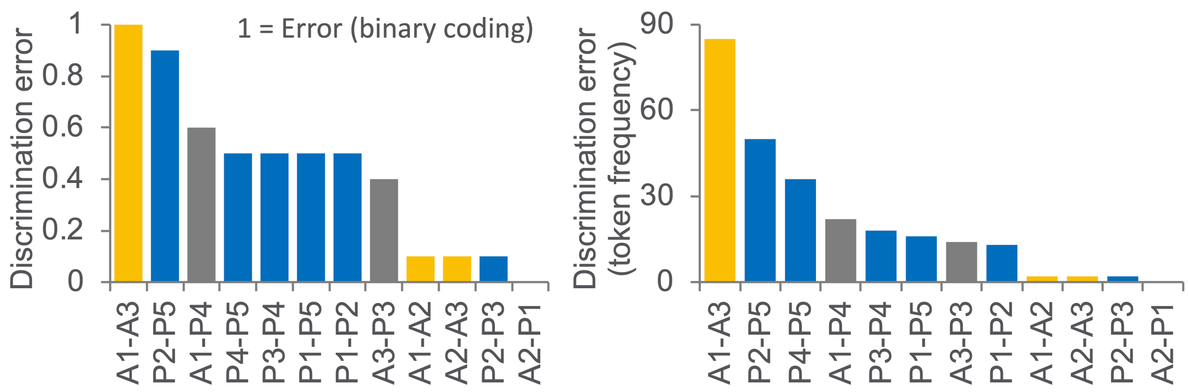
- 10 listeners for each set (30 listeners total)
- Listener task: How many voices did you hear?
  - Listen to the voice samples (icons).
  - Drag icons into piles so that each pile corresponds to a **single** speaker.
  - Listeners were not told that there were in fact only 2 speakers per trial.

4. Results

- Most listeners found the sorting task challenging.
  - Listeners heard 2–7 identities for each pair of speakers.
  - No listener scored 0 error.
  - Listeners showed errors in both "telling voices together" and "telling voices apart."

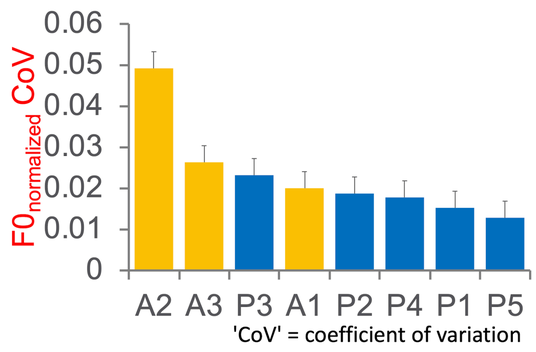


- Overall, **A**prototypical-**A**prototypical pairs were easier to sort than **P**rototypical-**P**rototypical pairs.
  - Exceptions: **A1-A3** (highly confusable) & **P4-P5** (easier to identify)



- Overall, listener discrimination performances were poor for **P**rototypical-**P**rototypical pairs.
  - However, **P2** and **P3** were easily distinguishable from each other.
- Listeners failed to tell **A1-A3** (always), **P2-P5** (90%) and **A1-P4** (60%) apart.
- All speakers paired with Speaker **A2** were easily differentiated from **A2**.

5. Additional Speaker Analysis



- **A2** had a greater mean for **F0 CoV** than the other speakers.
- Consistent with patterns of perceptual confusion

6. Discussion

- Our results confirm that within-speaker acoustic variability in voices introduces challenges in voice identification, particularly in the "telling voices together" task.
- Similarity with respect to the previously identified reference patterns (Lee et al. 2019; Lee & Kreiman, 2019) plays an important role in voice recognition.
  - Voices that are acoustically similarly structured are also perceived similarly to each other.
  - Only a few features are needed to explain many perceptual errors.
- Prototypicality in voices may play a role in voice discrimination, but all individuals in this specific population (and others) seemingly reside in the same acoustic space.
  - Aprototypical voices are just slightly differently structured voices in the same multidimensional acoustic space.
  - It may be possible to explain perceived similarity in terms of position in a shared acoustic space, without reference to a population-wide "average" voice.
  - **Population prototypes may not be a necessary part of the voice perception process.**
- Note: The acoustic space does not align completely with acoustic spaces derived from ratings of perceived similarity (e.g., Baumann & Belin, 2010).
- Going forward, separate analyses will shed light on the acoustic measures and listener strategies involved in these kinds of judgements.
  - Paying attention to F0 might be an important one.

Supported by **NIH DC01797** & **NSF IIS-1704167**