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

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

[1]Departments of Head and Neck Surgery and Linguistics, UCLA, Los Angeles, CA, USA
[2]Department of Head and Neck Surgery, UCLA, Los Angeles, CA, USA

Keywords: speaker recognition; acoustic variability; prototypes; personal identity

Introduction

Our recent studies [1] show that acoustic spaces characterizing within- and between-speaker variability in voice quality have similar structures, with a few features (acoustic variability and formant dispersion) that are prominent for all speakers combined with idiosyncratic features characterizing individual talkers. These findings suggest that voice discrimination ("telling voices apart" [2]) should be based on shared acoustic features, while determining which samples come from a single talker ("telling voices together" [2]) should depend on knowledge of each individual's vocal idiosyncrasies. Based on prototype models of voice perception, we hypothesized that errors in telling voices apart would be strongly predictable from distances among those few shared features in the group acoustic space, but that errors in telling voices together would not be associated with these distances. This experiment tested that hypothesis.

Methods

Recordings of one side of unscripted telephone conversations from 49 female speakers were drawn from the UCLA Speaker Variability Database [3]. Based on the acoustic structure of individual and group voice spaces for these speakers, 8 were identified as acoustically prototypical (P1, P2, P3, P4, P5) and aprototypical (A1, A2, A3) for this population of talkers. Fourteen brief samples were excerpted from each conversation for use in this study (phrases or sentences; duration: M = 1.6 s, SD = 0.5 s). Samples were screened to ensure they did not contain identifying content or unusual features (repeated phrases, laughs, speech errors, etc.) and were normalized for intensity.

Stimuli were sorted into 3 sets of 4 trials, such that across sets listeners heard unique pairs of speakers, and across trials each listener heard every speaker, but no listener heard any speaker more than once. Each trial included the 14 speech samples from 2 talkers, represented as randomly colored and shaped icons in a PowerPoint slide (Figure 1). Each set of 4 trials was judged by 10 listeners (30 total).

Listeners were instructed to listen to the voices, and then to drag them into piles on the screen so that each pile corresponded to a single speaker. They were told that there could be different numbers of voices in the piles, and that they could put as many voices as they liked in each pile, depending on how many speakers they thought they heard. (They were not told that there were in fact only 2 speakers.) They were allowed to listen to the samples as often as needed, in any order. The complete experiment lasted about 1 hour.

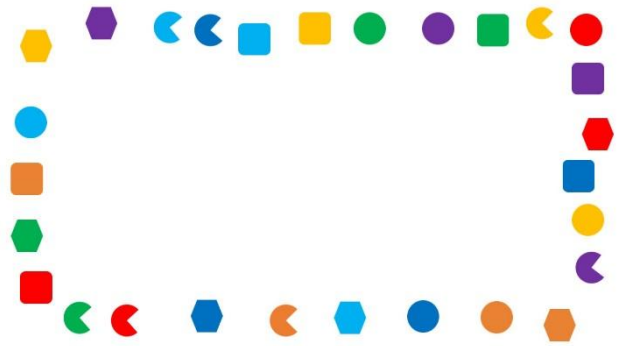

Figure 1: The testing interface. Each icon represents a single voice sample. Listeners used a mouse to drag these icons into piles corresponding to perceived speaker identities.

Preliminary Results and Discussion

Subject testing is currently underway. Most listeners in pilot studies found the task challenging. Results to date are consistent with our hypotheses. Overall, listeners showed fewer errors in the "telling voices apart" task than in the "telling voices together" task. Prototypicality in voices played an important role in voice discrimination. Separate analyses will shed light on the features and strategies involved in these kinds of judgments, with particular attention to features that help listeners tell voices together. Implications for prototype models of voice perception will be discussed.

Acknowledgements

This research was supported by NIH grant DC01797 and NSF grant IIS-1704167.

References

[1]Lee *et al.*, J Acoust Soc Am, 146:1568-1579, 2019.
[2]Lavan *et al.*, Psychon Bull Rev, 36:90-102, 2019.
[3]Keating *et al.*, Proc ICPhS, 2019

*jkreiman@ucla.edu

*jkreiman@ucla.edu