Overcoming talker variability when learning speech sound categories: A computational approach

Taylor Curley (taylor.curley@gatech.edu)
School of Psychology, Georgia Institute of Technology

Joe Toscano (joseph.toscano@villanova.edu)
Department of Psychology, Villanova University

WRAP Lab http://wraplab.co/

INTRODUCTION

- Listeners must map context-dependent acoustic cues onto phonetic categories
- However, contextual variability (e.g., effects of talker identity) is considerable
  (Peterson & Barney, 1952; Hillenbrand et al., 1995)
- Current models address this by specifying context *a priori*, a type of supervised learning
  (Cole, Linebaugh, Munson, & McMurray, 2009; McMurray & Jongman, 2011)
- But this is not developmentally plausible: infants use *unsupervised* statistical learning to acquire speech sound categories
  (Saffran, Aslin, & Newport, 1996; Maye, Werker, & Gerken, 2003)
- **Goal:** Develop a model that compensates for contextual differences using developmentally-realistic learning processes

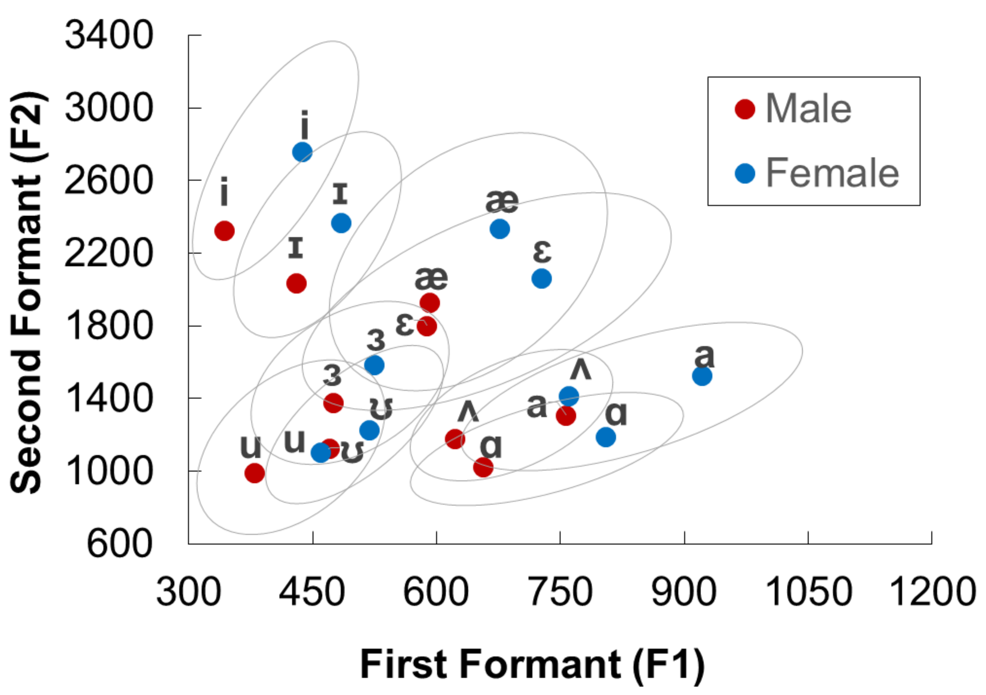
METHOD

Model Architecture:

- Acoustic-phonetic categories represented as a mixture of Gaussians
  (McMurray et al., 2009; Toscano & McMurray, 2010; Vallabha et al., 2007)
- Vowel categories represented by bivariate (2D) Gaussians (F_1xF_2 space); gender categories as univariate (1D) Gaussians (F_0)
- Each 1D Gaussian has 3 parameters: mean (μ), standard deviation (σ), and likelihood (ϕ)
- 2D Gaussians have μ- and σ-values for each dimension as well as a parameter for the correlation between F_1 & F_2 (ρ)

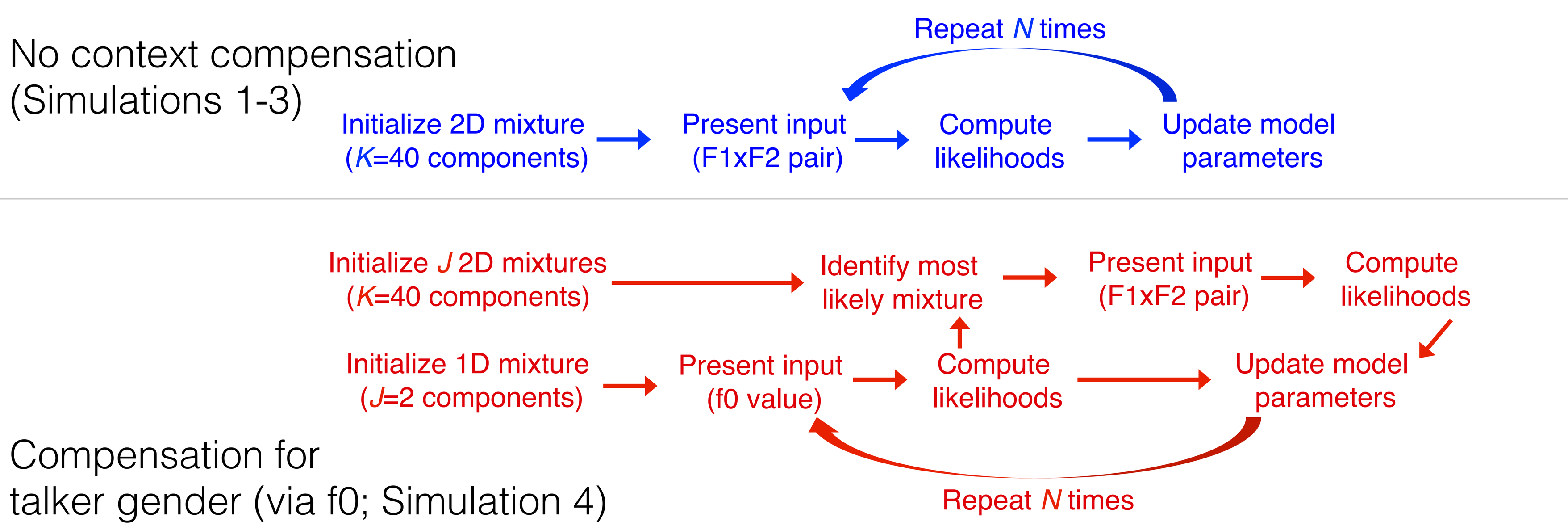
Training data:

- Model trained on distributions of English vowel sounds based on measurements from from Hillenbrand et al. (1995)

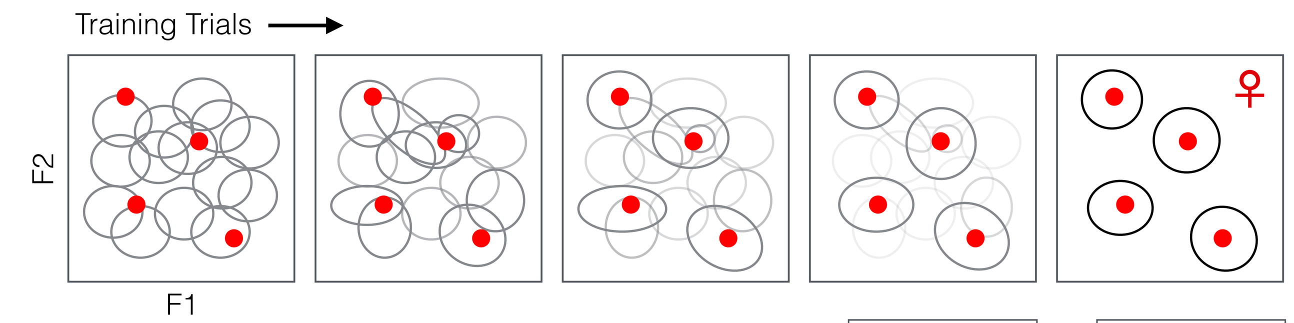


RESULTS

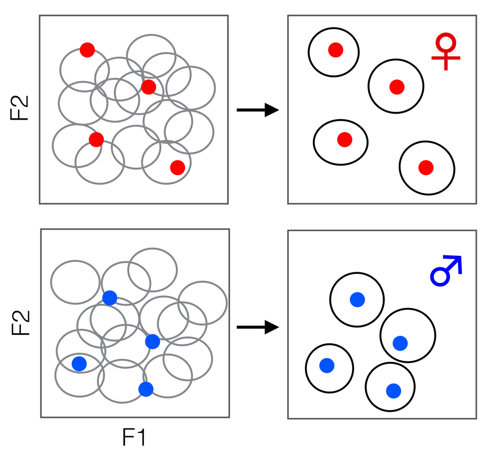
Training procedure



Predicted developmental trajectory

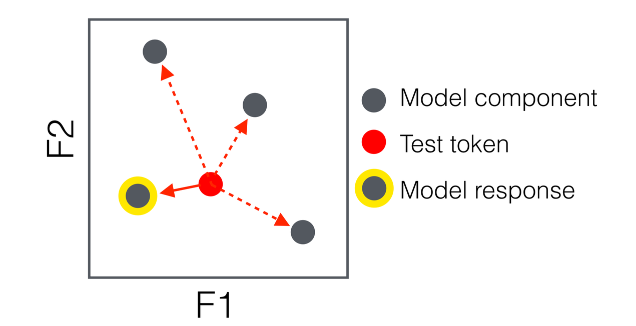


- Model determines number of vowel categories by pruning away unnecessary components in mixture
- Context compensation model should learn talker-specific vowel spaces



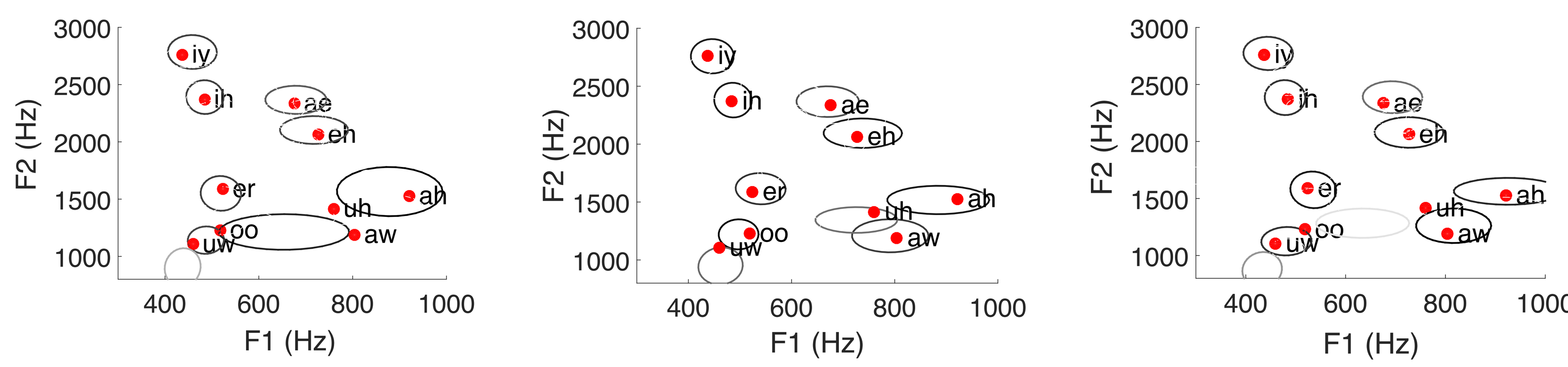
Testing procedure

- Success evaluated by measuring (1) number of above-threshold categories after training and (2) classification accuracy for a set of 500 test tokens
- Euclidean distance between model component means and test token calculated; shortest distance corresponds to model classification



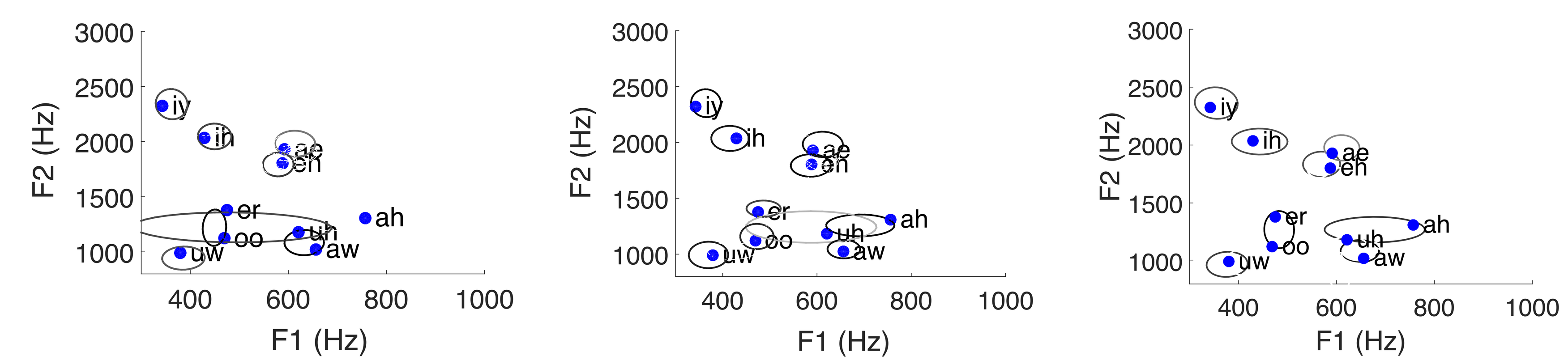
Simulation 1: Female talkers only

- Baseline performance for female talkers
- Learned correct number of categories



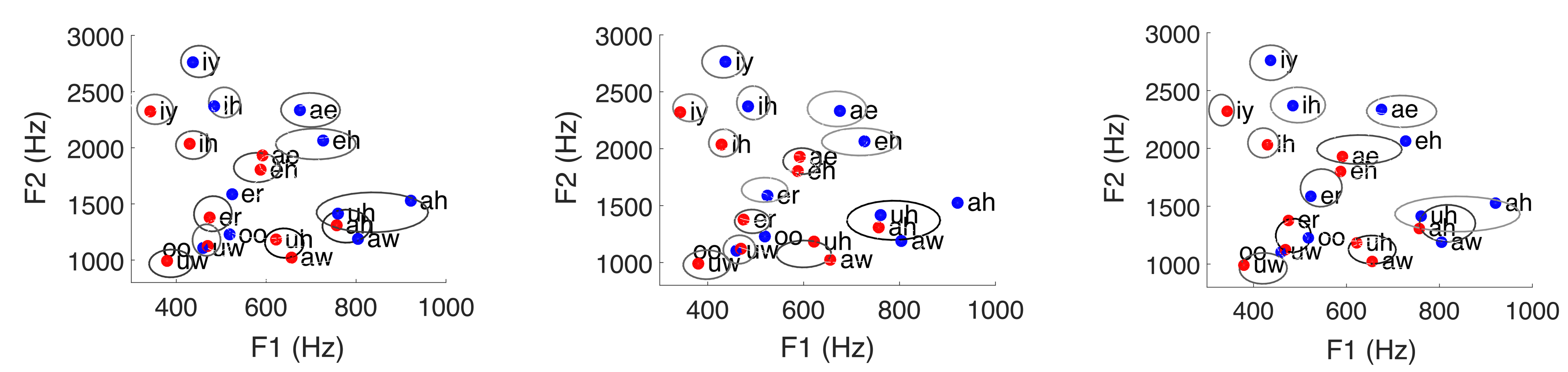
Simulation 2: Male talkers only

- Baseline performance for male talkers
- Some vowel mergers (e.g., /ɑ/ and /ʌ/)



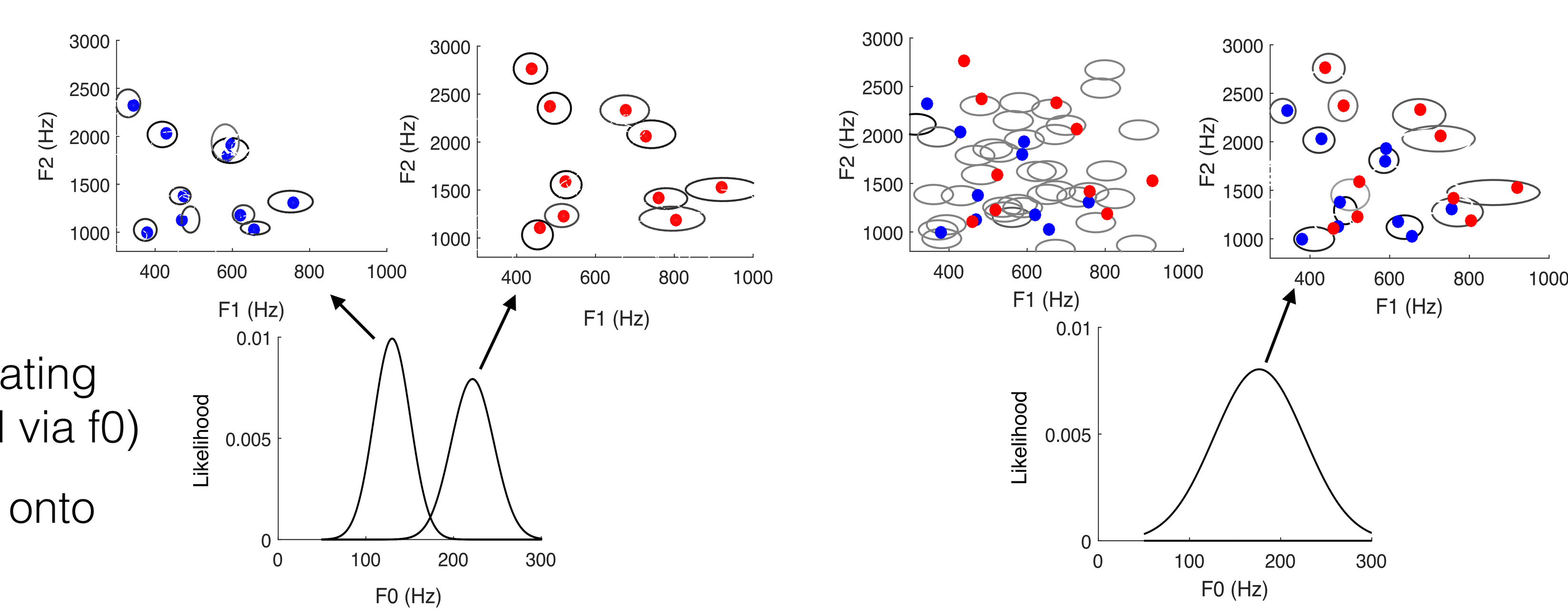
Simulation 3: Both groups of talkers; no context compensation

- Representative model runs shown
- Numerous mergers; low classification accuracy
- Same vowel often mapped onto separate categories (one male, one female)



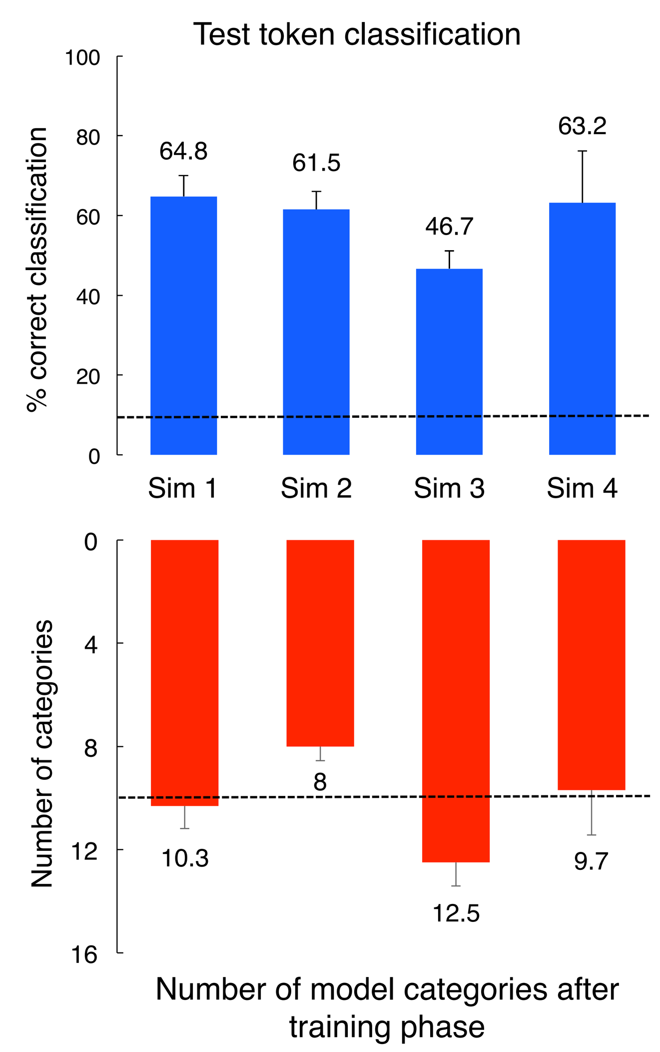
Simulation 4: Both groups of talkers; context compensation via f0 mixture

- Representative model runs shown
- Generally, model was successful in separating vowels based on talker gender (estimated via f0)
- Categories in each F1xF2 space mapped onto correct vowels for each group of talkers



Summary of results

- All simulations demonstrated above chance classification accuracy, but poorer performance than would be predicted from human listeners
- Context compensation mechanism (f0 mixture) yielded performance equivalent to training on each group of talkers separately, in terms of accuracy (63% correct) and acquisition of the correct number of categories (mean: 9.7 out of 10)



DISCUSSION

- Accounting for talker variability allowed the model to successfully learn vowel categories
- When categories are mapped to separate vowel spaces according to gender, classification accuracy improves 16.5% percent (**Sim 4**) relative to performance without compensation (**Sim 3**)
- Provides a developmentally-plausible learning mechanism that makes minimal assumptions about pre-existing knowledge to account for contextual variability in speech

ACKNOWLEDGEMENTS & REFERENCES

This work was supported by a Psychonomic Society Graduate Student Travel Award to TC.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167-184.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099- 3111.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*, 369-378.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434-464.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S., (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273-13278.