Classification of English Stop Consonants: A Comparison of Multiple Models of Speech Perception

Abigail Benecke
abenecke@villanova.edu
Department of Psychology
Villanova University

Joseph Toscano
joseph.toscano@villanova.edu
Department of Psychology
Villanova University

INTRODUCTION

- Speech research has searched for a solution to the lack of invariance.
- Some researchers such as Lisker (1986) have emphasized redundancy in speech as a possible solution: **integrate multiple cues**.
- Some speech cues are highly reliable (e.g. VOT), while others have low reliability (e.g. VL).
- **Purpose:** Measure and test 35 possible cues to test multiple models of word-initial stop consonant categorization (temporal, spectral, and amplitude).
- Corpus analysis to measure cues based on previous literature on stops as well as fricatives (McMurray & Jongman, 2011).
- Create and compare models using different subsets of cues
- Test models that vary in three ways:
  1. Additional cues
  2. Categorizing based on feature or phoneme
  3. Context compensation

METHOD

Measurements & Cue Reliability

- Acoustic measurements of 1,068 speech tokens (35 cues) in Praat (Boersma & Weenik, 2016). 12 talkers in 15 vowel contexts (Schatz et al, 2016).
- Calculated cue reliability using cue-weighting metric from Toscano & McMurray (2011).
- Metric takes into account within-category variance of tokens (σ), distance between category means (μ), and category likelihood (ϕ).

$$r_{cue} = \sum_{i}^{P}\sum_{j}^{P} \frac{(\phi_{ij}\phi_{ij})\,(\mu_{ij}-\mu_{ij})^2}{\sigma_{ij}\sigma_{ij}}$$

Models

- Trained model classifiers using multinomial regression in R (R Core Team, 2016). Each model used a randomly-selected 90% of tokens for training, tested on the remaining 10%, repeated 500 times.

$$phoneme \sim \sum_{i}^{N} cue_i$$

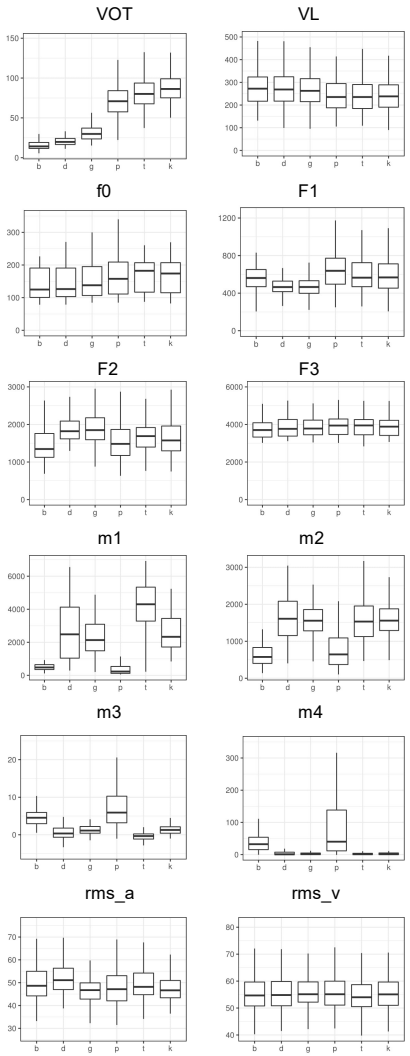
- Feature-based model is comprised of two regression equations: one where the cue(s) are regressed against voicing, and one for place.

$$voicing \sim \sum_{i}^{N} cue_i \qquad place \sim \sum_{i}^{N} cue_i$$

- Context-compensation models used the C-CuRE method for each cue, using the residuals in the model (McMurray & Jongman, 2011).

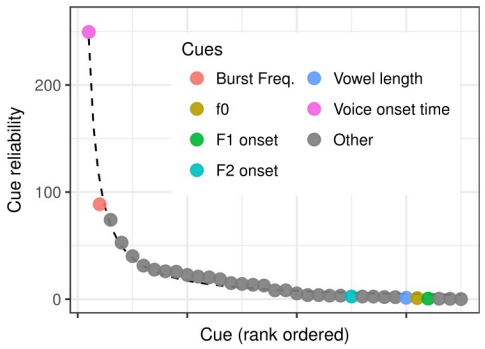
RESULTS

Acoustic Measurements



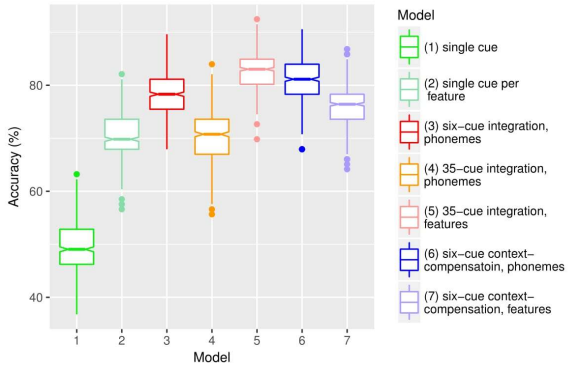
- Some cues (e.g. VOT, VL) distinguish voicing while others (e.g. m1, m3) distinguish place
- Some cues highly reliable (VOT); others much less reliable (VL)
- No single cue is invariant

Cue Reliability

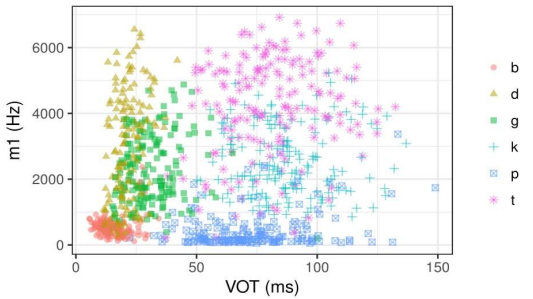


- All 35 cues sorted by their reliability follow a power curve (dashed line). Note that VOT is the most reliable cue by far, followed by BF; other cues previously reported in the literature are much less reliable.
- BF refers to mean frequency of the burst (m1), not spectral shape as in Stevens and Blumstein (1978).

Model results



- Percent correct for each model over the 500 simulation runs. Cue integration models perform better than single cue (1,2) models.
- Feature-based models (2,5) tend to perform better than phoneme-predictor counterparts (1,4), with the exception of the context compensation models (6,7).
- Single cue model (VOT alone) achieved 49% accuracy, and the single cue per feature model (voicing = VOT, place = BF) showed 78% accuracy.
- The best model (5) performed at 83%. This is the 35-cue integration model using the feature-based classifier.



- All tokens plotted in a VOT x BF space
- There's still significant overlap, showing the need for large-scale cue integration

DISCUSSION

- Two best cues are VOT and BF. Other cues previously cited for stop consonants (F0, F1, F2, and VL) lie on the tail end of the distribution. In general, additional cues increased accuracy, suggesting support for cue-integration models
- Feature-based models out-performed phoneme models
- Context compensation (talker and vowel) has slight impact on categorization, but not for feature-based models, in contrast to fricatives (McMurray & Jongman, 2011) or vowels (Cole, Linebaugh, Munson, McMurray, 2010)
- Overall, results demonstrate that large-scale cue-integration may be sufficient for overcoming contextual variability. However, human performance on similar categorization tasks is still much higher (~95% accuracy; Toscano, 2014).

ACKNOWLEDGEMENTS & REFERENCES

Boersma, P. & Weenik, D. (2016). Praat: Doing phonetics by computer. http://www.praat.org.

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(2), 167-184.

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech, 29*(1), 3-11.

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*(2), 219.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schatz, T. et al. (2015). *Articulation Index LSCP LDC2015S12*. Linguistic Data Consortium.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America, 64*(5), 1358-1368.

Toscano, J. C., & Allen, J. B. (2014). Across-and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research, 57*(6), 2293-2307.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science, 34*(3), 434-464.