How far can VOT take us? Voicing categorization with and without the use of VOT

Abigail Benecke, Joe Toscano [abenecke, joseph.toscano]@villanova.edu

Department of Psychological and Brain Sciences, Villanova University

http://wraplab.co

INTRODUCTION

- Voice onset time (VOT) and voicing is a model cue-category system for studying how human listeners map sounds onto linguistic categories
- In English, VOT is a primary cue for distinguishing voiced stops (/b,d,g/) from voiceless stops (/p,t,k/) (Lisker & Abramson, 1964)
- VOT is highly reliable, but is it invariant?
- Alternative and secondary voicing cues have been suggested (e.g. F1 onset, f0 onset, vowel length) (Lisker, 1975; Summerfield & Haggard, 1977; Toscano & McMurray, 2012)
- Models must specify which cues listeners use, and match the high level of human accuracy; approx. 99% accurate (Toscano & Allen, 2014)
- To test invariance hypothesis and primary of VOT, evaluate models of speech categorization that:
  1. Include VOT as a cue by itself,
  2. Use VOT in conjunction with other cues, or
  3. Evaluate voicing categorization without VOT

METHOD

Measurements & Cue Reliability

- Acoustic measurements of 1,056 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 (2010)
- Metric takes into account within-category variance (σ) and distance between category means (μ):

$$r = \frac{(\mu_{voiced} - \mu_{voiceless})^2}{\sigma_{voiced}\sigma_{voiceless}}$$

Models

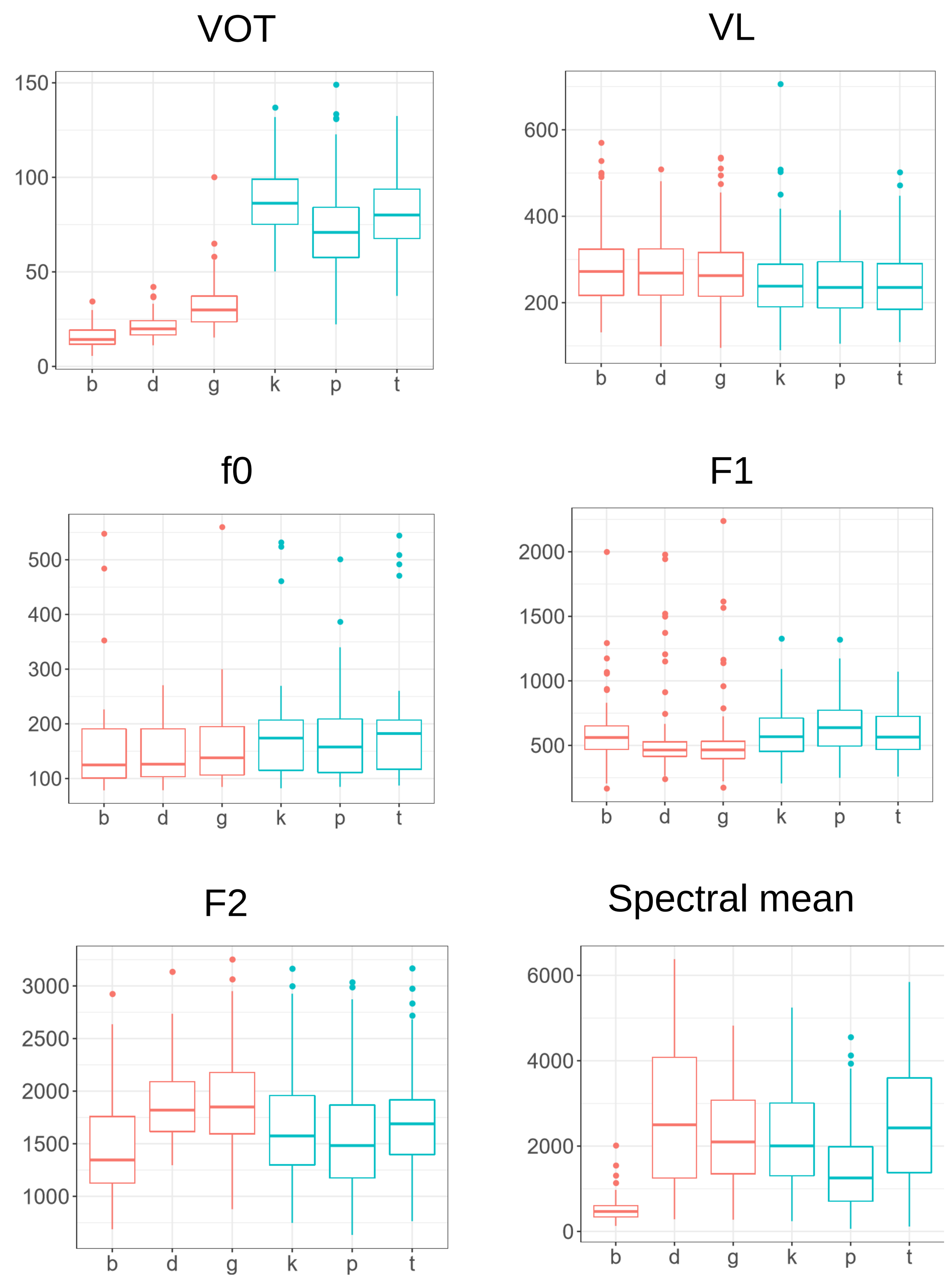
- Multinomial regression classifiers implemented in R (R Core Team, 2016)
- Each model used a randomly-selected 90% of tokens for training, tested on the remaining 10% a total of 500 times.

$$P(\text{VOICING}) = \frac{1}{1 + exp\left(\beta_0 + \sum_{i=1}^{N} \beta_i C_i\right)}$$

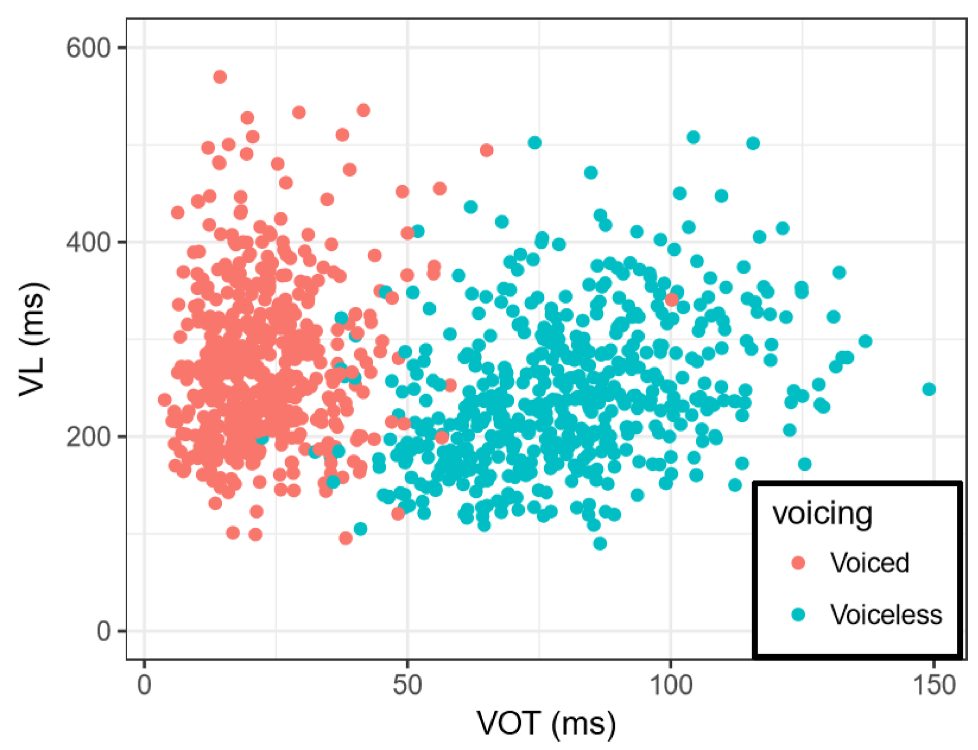
- Evaluated the following models with VOT:
  1. VOT alone
  2. VOT and vowel length (VL)
  3. Set of phonetically-relevant cues (VOT, VL, F1 onset, F2 onset, and f0 onset)
  4. All 35 cues
- Evaluated the following models without VOT:
  1. Second-most reliable cue- initial spectral mean
  2. Spectral mean (SM) and VL
  3. SM, VL, F1 onset, F2 onset, and f0 onset
  4. All 35 cues without VOT

RESULTS

Acoustic Measurements



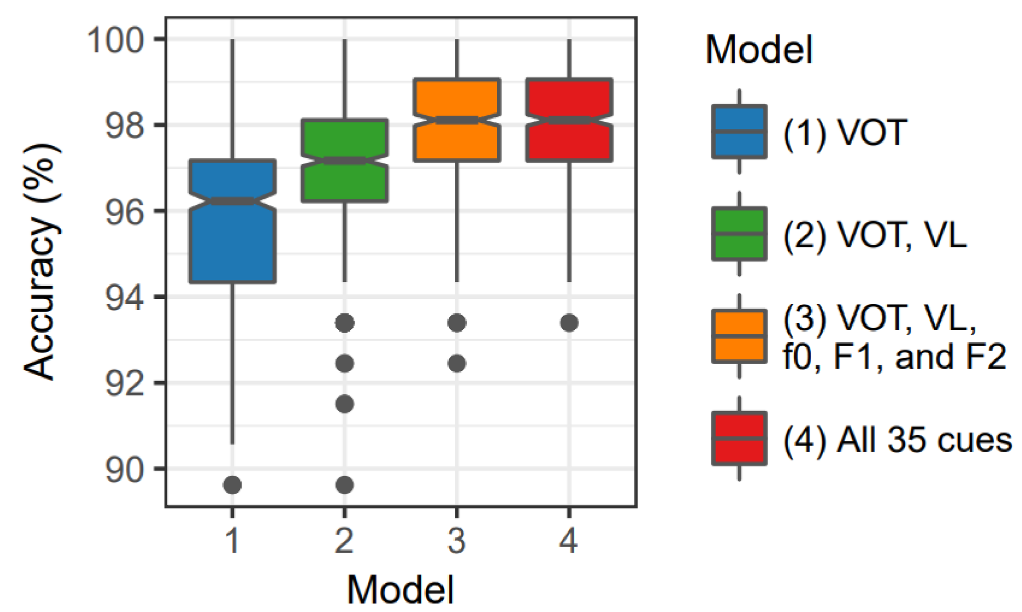
- Distributional statistics for six voicing cues from previous phonetic and perceptual studies



- All tokens plotted in a VOT x VL space (i.e., two most reliable uncorrelated cues)
- Very little overlap between categories, suggesting categorization accuracy should be high based on just these two cues

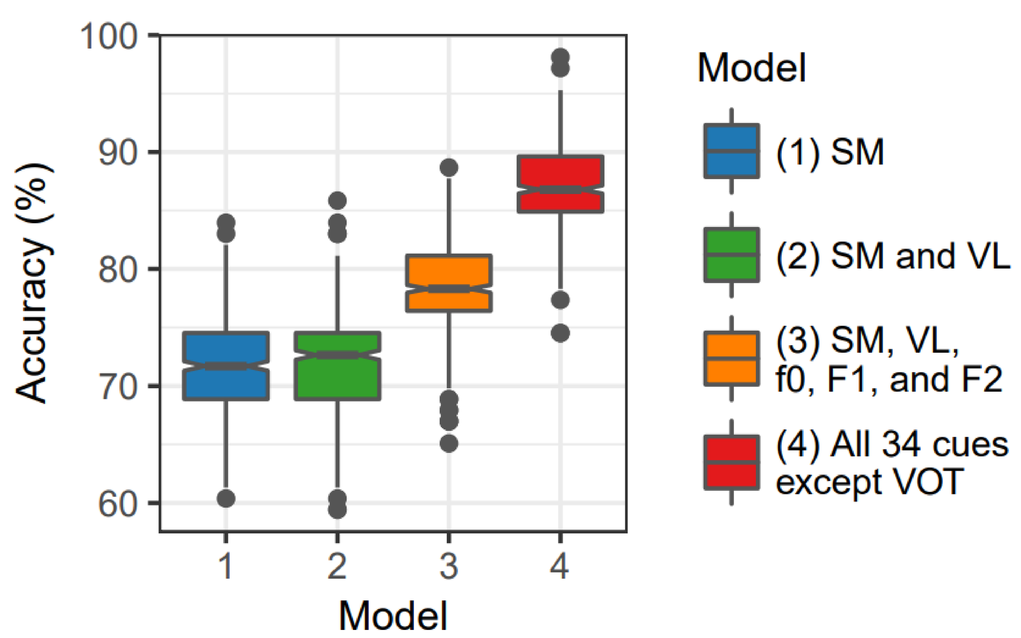
Model results

Models with VOT



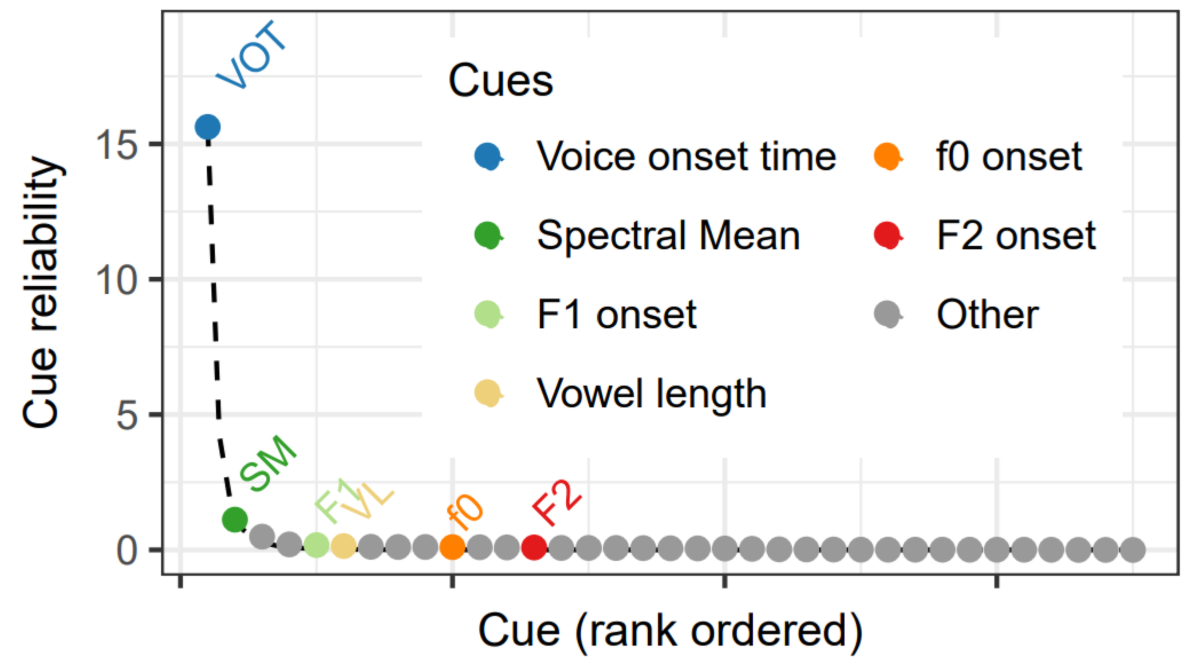
- With VOT as the only voicing cue, classifier did extremely well: mean accuracy of 95%
- However, performance still short of human listeners
- VOT and VL model performed significantly better than the VOT-only model at categorization (t=12.76, p<.0001), with a mean accuracy of 97%
- 5-cue model (VOT, VL, f0, F1, F2) performed at approximately listener level (98% accuracy)
- 35=cue model (all measured cues) performed similar to the 5-cue model

Models without VOT



- To test classification accuracy without VOT, we first tested a model with the second-best cue alone (spectral mean; SM).
- This classifier fell far short of the VOT-alone model, reaching only 71% accuracy
- Addition of VL brought accuracy to 72%
- 5-cue model (SM, VL, f0, F1, F2) performed at 78%, which still falls far short of human listeners
- 34-cue model (all cues except VOT) reached accuracy of 87%, much lower than classifiers with VOT included

Cue Reliability



- Cue reliability metric reveals VOT as substantially more reliable than other cues
- Sorted reliability follows a power law function

DISCUSSION

- VOT is the most reliable cue by far. Even the second-best cue (SM) is much lower in reliability
- Models with VOT and multiple secondary cues reach human-level performance (98%)
- Secondary cues can give the listener some information to assist in accurate categorization, but the classifier is never able to achieve human-like performance without VOT
- However, a VOT-only model still falls short—a cue-integration approach including VOT and secondary cues, offers the best model of categorization
- VOT appears to be a necessary, but not sufficient, cue for voicing judgments

REFERENCES

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

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384-422.

Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America, 57*(6), 1547-1551.

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.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America, 62*(2), 435-448

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.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics, 74*(6), 1284-1301.