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

The relationship between voice onset time (VOT) as a phonetic cue and voicing as a phonological feature serves as a model *cue-category system* for studying how human listeners map sounds onto meaningful linguistic categories. In English, VOT is a primary cue for distinguishing unaspirated stops (/b,d,g/; referred to hereafter as *voiced*) from aspirated stops (/p,t,k/; referred to as *voiceless*). VOT is an extremely reliable cue, such that there is little overlap word-initially between voiced and voiceless tokens. Thus, perhaps VOT is sufficient by itself. On the other hand, other cues have been proposed as alternatives, and there are a number of secondary cues for voicing judgments, including f0, formant onsets, and vowel length [1, 2], suggesting that a cue-integration approach, where listeners combine multiple cues [3], might be more effective.

Here, we investigate the extent to which listeners can accurately determine voicing categories on the basis of VOT and other cues by evaluating models of speech categorization that include VOT as a cue by itself, use VOT in conjunction with other cues, or evaluate voicing categorization without VOT. We trained a series of logistic regression classifiers to make voicing judgments using phonetic data from a corpus of stop consonants produced by N=12 speakers of American English [4]. We addressed the following questions: (1) is VOT alone sufficient for voicing categorization, (2) does the addition of other cues substantially increase categorization accuracy, and (3) in the absence of VOT, can other cues suffice? Results from the models are compared against listener data, which sets a high bar for the model; listeners make few errors in consonant recognition and are particularly accurate when making voicing judgments [as high as 98.9% in quiet; 5].

Method. Thirty-five potential cues were derived from those identified in previous phonetic studies of stop consonants, as well as cues used in studies of other types of consonants [e.g., spectral moments in different time windows for fricatives; 6]. Sounds were produced by 12 talkers in 15 vowel contexts. Recordings were coded in Praat [7] to identify two acoustic landmarks: (1) burst onset and (2) vocoid onset; a Praat script then automatically extracted the cues. Tokens with undefined f0 measurements were removed, leaving 1056 tokens for analysis. Based on the distributional statistics of the cues, we computed their reliability ($r_{voicing}$), using the metric,

$$r_{voicing} = \left(\mu_{voiced} - \mu_{voiceless}\right)^2 / \sigma_{voiced}\sigma_{voiceless}, \tag{1}$$

where $r_{voicing}$ is the statistical reliability of the cue (i.e., how well it distinguishes the voicing categories), μ is the mean of a category, and σ is its standard deviation [8]. This yields a unitless measure, where higher values correspond to less overlapping distributions, akin to d'.

We then evaluated several models of voicing categorization based on these measurements using logistic regression classifiers, implemented in R [9]. Each model was trained to predict VOICING (voiced vs. voiceless) based on different subsets of the cues. The general form of the classifier,

$$P(\text{VOICING}) = 1/1 + exp\left(\beta_0 + \Sigma_{i=1}^{N}\beta_i C_i\right), \tag{2}$$

evaluates the additive effect of different sets of acoustic cues (C), similar to other cue-integration models [3, 8]. VOICING is the predicted category (VOICED=0; VOICELESS=1), N is the number of cues, β_i is the regression coefficient for cue C_i, and β_0 is the overall intercept of the model (i.e., bias towards voiced vs. voiceless responses). Classifiers were trained on a random 90% of the tokens and tested on the remaining 10%; this process was repeated 500 times. We evaluated the following models *with* VOT: (1) VOT alone, (2) VOT and VL, a purported secondary voicing cue uncorrelated with VOT [10]; (3) VOT and several other cues cited in the literature (VOT, VL, F1 onset, F2 onset, and f0 onset); (4) all 35 cues. Similarly, models were evaluated *without* VOT: (1)

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 205-206.
Salt Lake City, Utah, January 4-7, 2018

the second-most reliable cue after VOT (spectral mean [SM] during the first 40ms); (2) SM and VL; (3) SM, VL, F1 onset, F2 onset, and f0 onset; and (4) all 35 cues except VOT.

Results. As expected, VOT was the most reliable cue ($r_{voicing}$=81.73; voiced: 22±10 ms; voiceless: 80±20 ms; Fig. 1). The second most reliable cue was the SM ($r_{voicing}$=5.85; voiced: 1095±904 Hz; voiceless: 2405±1698 Hz). This cue was correlated with VOT (r=0.50); thus, it may provide a good alternative. VL was most reliable cue uncorrelated with VOT (r=-0.04).

To test classifiers, we first investigated VOT as the only voicing cue. The classifier did extremely well, with a mean accuracy of 95± 2%, well above chance. However, performance still falls short of human listeners' (≈99%). The second model included VOT and VL as two independent cues and performed significantly better than the VOT-only model at categorization ($t = 12.76$, $p < .0001$), with a mean accuracy of 97±2%. However, the model still falls short of human listeners by approx. 2%. Next, the classifier with 5 phonetically-motivated cues had a mean accuracy of 98±1%, which is almost at listener level. Lastly, inputting all 35 cues yielded the same results as the previous model (98±1%), suggesting those cues do not provide more information (Fig. 2).

In order to see how well listeners can identify voicing *without* VOT, we substituted the second best cue (SM) for VOT. This cue is not nearly as reliable ($r_{voicing}$=5.85); as such, the model performed more poorly, with 71±4% correct (Fig. 3). Subsequent simulations evaluated models with SM and VL (72±4%), SM and other phonetically-motivated cues (78±4%), and a model with all 34 cues except VOT (87±3%). These results provide evidence that 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. Thus, VOT appears to be a necessary, but not sufficient, cue for voicing judgments.

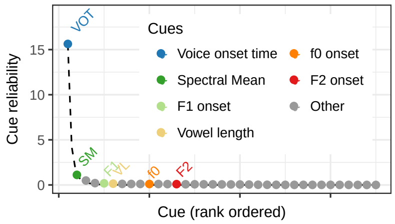

Figure 1: Cue reliabilities.

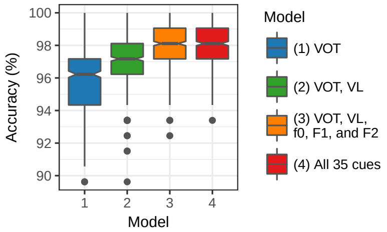

Figure 2: Models with VOT.

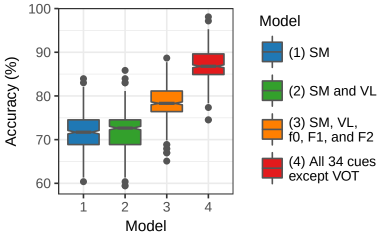

Figure 3: Models without VOT.

References. **1.** JS Allen and JL Miller. Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *J Acoust Soc Am*, 106:2031–2039, 1999. **2.** RN Ohde. Fundamental frequency as an acoustic correlate of stop consonant voicing. *J Acoust Soc Am*, 75(1):224–230, 1984. **3.** GC Oden and DW Massaro. Integration of featural information in speech perception. *Psychol Rev*, 85:172–191, 1978. **4.** T Schatz et al. *Articulation Index LSCP LDC2015S12*. Linguistic Data Consortium, 2015. **5.** JC Toscano and JB Allen. Across- and within-consonant errors for isolated syllables in noise. *J Speech Lang Hear Res*, 57:2293–2307, 2014. **6.** A Jongman, R Wayland, and S Wong. Acoustic characteristics of English fricatives. *J Acoust Soc Am*, 108:1252–63, 2000. **7.** P Boersma and D Weenik. Praat: doing phonetics by computer. http://www.praat.org/, 2016. **8.** JC Toscano and B McMurray. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Sci*, 34:434–464, 2010. **9.** R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. **10.** JC Toscano and B McMurray. Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attn Percep Psychophys*, 74(6):1284–1301, 2012.