Integrating acoustic cues to phonetic features: A computational approach to cue weighting

Joseph C. Toscano and Bob McMurray
University of Iowa
Dept. of Psychology

A central issue in phonology and speech perception is the relationship between acoustic cues in the speech signal and the features they correspond to. Voice onset time (VOT), for example, serves as a primary cue to the phonetic feature voicing. However, these relationships are more complicated when we consider that for any given phonetic feature, there are multiple acoustic cues that contribute to its perception (Lisker, 1978). Listeners' use of multiple cues has been demonstrated for a large number of phonological distinctions by examining changes in the location of category boundaries along one dimension as a function of a second or third dimension (see Repp, 1982, for a review).

Previous perceptual experiments have demonstrated that listeners weight individual acoustic cues differently as they are combined to form a phonological dimension or feature. For example, in determining voicing, VOT is a primary cue, while F0, F1 and vowel length make minor contributions. To date, there have been no theoretical accounts that make specific predictions about how each cue might be weighted during perception. Gestural approaches, for example, might posit that the particular cue weights used for a phonetic feature are determined by the relationship between the articulatory gestures that a speaker uses to produce a particular speech sound and the acoustic effects of that gesture. However, this account requires listeners to have knowledge about the relationships between particular articulatory gestures, their acoustic counterparts, and the relevant phonetic distinctions associated with them. Moreover, given that many relationships between cues differ cross-linguistically (e.g. the relationship between VOT and vowel length is different for languages in which vowel length is phonemic), an account in which cue weights could be learned may be preferable.

We propose that these questions can be answered without regard to their gestural origins by weighting acoustic cues as a function of their statistical reliability. That is, a cue that is more reliably correlated with phonetic categories along a given dimension should be weighted higher than less reliable ones. For example, VOT provides an excellent cue to word-initial voicing because VOT values in speech production tend to cluster into distinct categories (Lisker & Abramson, 1964). Vowel length, while also distinguishing word-initial voicing categories (Allen & Miller, 1999), is more variable, and is therefore less reliable than VOT. Thus, a system using the reliability of acoustic cues to weight them would weight VOT higher. This reflects the relationship observed in perceptual experiments between these cues (Summerfield, 1981).

We present a method for computing the reliability of acoustic cues on the basis of their statistical distributions and apply this method in a computational model. The model represents individual cues as a mixture of Gaussians (MOG) along an acoustic dimension. Each Gaussian distribution represents one acoustic-phonetic category. The mean and variance of these Gaussians can be extracted from phonetic data on the particular cues being modeled. This yields categories structured as graded prototypes along an acoustic dimension. Each individual cues is then weighted using a metric that takes into account the distance between categories, the variability of individual categories, and the relative frequency of each category along each acoustic dimension. Figure 1 shows examples of distributions that are reliable (1A) and unreliable (1B) using this metric. After computing each cue's reliability, cues-values are linearly combined to form a graded phonological dimension that is based on the weighted combination of acoustic inputs and corresponds to a particular phonetic feature. From this more abstract

dimension, new Gaussian categories are learned and used to categorize speech sounds for that phonetic distinction. Thus, phonetic features are represented as graded distributions along a single dimension that are computed from weighted inputs of different acoustic cues.

Simulations of several trading relations based on different acoustic cues and phonetic features are presented. The model is able to determine cue weights that produce behavior similar to that of human listeners for these different trading relations. Figure 2 shows the results of an identification task for voicing category with stimuli varying in VOT and vowel length. The results indicate a similar trading relation for human listeners and for models trained on these acoustic cues. This suggests that these cue weights can be determined from the statistical properties of the acoustic input without needing to posit innate knowledge about how particular cues should be weighted or what their distributions should be. Further, it provides a computationally explicit account of how the speech system can weight and integrate these cues during speech perception. Finally it suggests that feature dimensions can be constructed solely on the basis of the bottom-up statistical reliability of the acoustic cues they are based on.

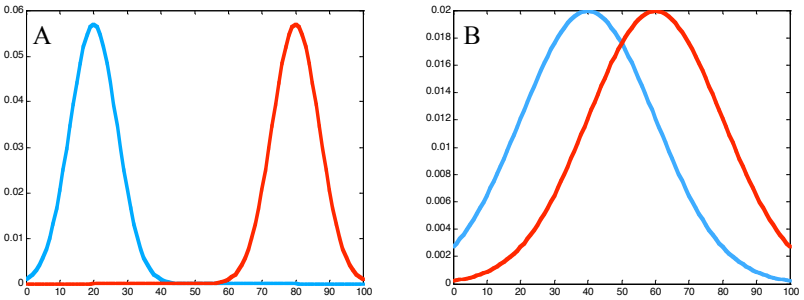

Figure 1. (A) A reliable cue dimension in which the categories are distinct. (B) A cue dimension with highly overlapping categories that would be less reliable using the cue weighting metric.

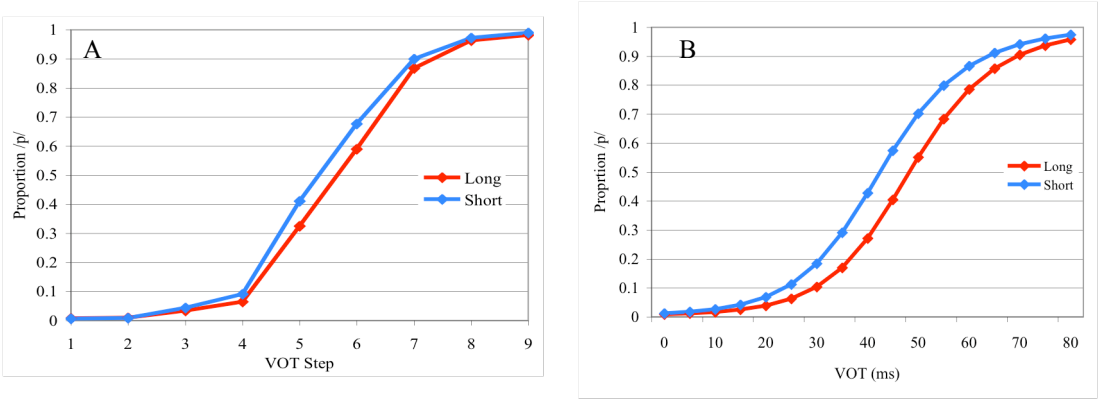

Figure 2. Identification functions along a VOT continuum for two different vowel lengths (long and short). (A) Behavioral data from human listeners. (B) Simulation results from the model.

References

Allen, J. S. & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *J. Acoust. Soc. Am., 106*, 2031-2039.

Lisker, L. (1978). *Rapid* vs. *rabid*: a catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, SR-54*, 127-132.

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

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92*, 81-110.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform., 7*, 1074-1095.