Layers of variance in the speech onion: There's no need to cry over the problem of variability. Evidence from perceptually motivated analyses of fricatives.

Bob McMurray
Dept. of Psychological and Brain Sciences
University of Iowa

Allard Jongman
Dept. of Linguistics
University of Kansas

Speech perception has been classically framed in terms of variability in the acoustic signal (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Perkell & Klatt, 1986). Factors like speaking rate, coarticulation and talker identity affect virtually all phonetic cues (McMurray & Jongman, 2011), and any cue is simultaneously a product of multiple factors (Mermelstein, 1978; Whalen, 1989). However, our understanding of this problem has been built on the basis of small-scale phonetic work, one cue and context at a time. There have been few systematic investigations of the sources of variability across many cues, contexts and talkers.

Recent work offers a powerful approach for understanding variability more systematically. Cole, Linebaugh, Munson, and McMurray (2010) examined vowel-to-vowel coarticulation in a large corpus of productions of /ʌ/ and /ɛ/, recorded in the context of 10 talkers, 6 neighboring consonants, and four subsequent vowels. They showed that a simple regression analysis could account for upwards of 85% of the variance in F1 and F2 as a combination of these factors. These analyses suggested that the massive variability in speech may not be insurmountable, but rather can be described as the simple additive product of multiple known factors. Perception may then operate by simply subtracting the influence of causal factors like talker and coarticulation.

McMurray and Jongman (2011) tested this with a large corpus of the 8 fricatives of English. They measured 24 cues in fricatives spoken by 20 talkers in 6 vowel contexts. A subset of these tokens was also given to listeners to categorize. A logistic regression was trained to predict the category of each sound from the acoustics. Even with all 24 cues, the model performed about 10% poorer than listeners. However, when the influence of talker and neighboring vowel were partialed out of the cues (using similar regressions to Cole et al., 2010), performance matched listeners both in terms of absolute accuracy and the pattern of errors. These regressions implement a model of perception in which listeners use similar processes to explain the factors that underlie a given segment. At any given moment, listeners have expectations about the cues like formant frequencies or fricative spectra based on contextual factors like talker and vowel. For example, they may know that the current talker produces high F1's, or that the preceding rounded vowel predicts low peak frequencies. Perception is based on the difference between the actual cue values and these expectations. This model was recently tested in two experiments that manipulated listeners' expectations to 1) alter the accuracy of fricative identification (Apfelbaum, Bullock-Rest, Rhone, Jongman, & McMurray, 2014), and 2) improve listeners' ability to predict subsequent vowels (McMurray & Jongman, 2015). It seems that listeners are not acting to categorize the input but rather to *explain* the variety of factors that gave rise to the heard form (—much like linear regression). This suggests a contingent categorization model in which listeners categorize one factor (the talker, the vowel) and use it to improve categorization of other factors (the fricative). Once they can explain some portion of current variance as arising from one factor, they can use what is left over to interpret other aspects.

The present talk builds from this work to ask two key questions.

First, we investigated whether a similar strategy may help at the level of phonological features within a phoneme, specifically place of articulation and voicing. That is, if the listener can identify place of articulation, does this improve voicing classification (and vice versa)? We started by analyzing the relative contribution of place and voicing to the variance in cues to fricative identity. While each cue was affected by context (talker and vowel) we found a handful of cues that reflected place, but not voicing, or voicing but not place (Figure 1). We next trained logistic regressions to predict either place or voicing. These were trained either with the raw cues, or with cues from which the effect of place or voicing had been removed. This was meant to simulate a situation in which the listener knows place and may be able to use that to benefit voicing judgements (or vice versa). Surprisingly, once talker and vowel were accounted for there was no additional benefit to knowing place

or voicing. The model trained to predict voicing was 97.53% correct after talker and vowel were accounted for, and 97.58% correct when place was also parsed out. Similarly, the model trained to predict place of articulation was 87.72% correct with talker and vowel parsed from the cues, and 87.56% correct when voicing was also accounted for. Thus, it appears that once the effects of context are accounted for, place and voicing leave largely orthogonal traces on the acoustic form of the fricative. Knowing one does not help the listener figure out the other.

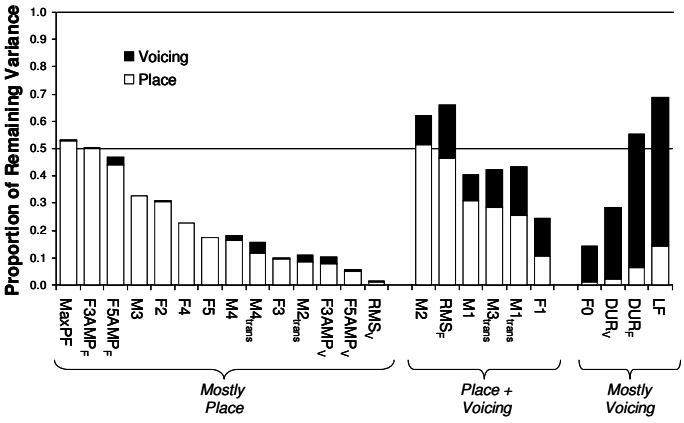

Figure 1. Proportion of remaining variance contributed by Voicing and Place of articulation for each acoustic cue, after effects of context (Talker and Vowel) have been removed.

Second, a critical assumption in this model is that lawful variability in the acoustic signal can be accounted for at the cue level rather than as a function of specific categories. That is, to use the compensation mechanisms proposed here, listeners do not need to track how a talker produces a certain vowel; rather, they track how that talker tends to produce F1 and F2 across all vowels. This may benefit generalization (since the listener doesn't need to know all listener x phoneme combinations), but it may also oversimplifies the problem by ignoring talker-specific articulations of particular phonemes. We thus asked whether the inclusion of talker x vowel interaction terms accounts for new variance over and above the main effects of talker and vowel. Across all 24 cues, the talker and vowel main effects suggest robust effects of context, accounting for an average of 28.1% of the variance. In contrast, the addition of interaction terms accounted for only an average 3.4% additional variance, suggesting modest benefits (if at all) for tracking more specific combinations of factors. Work in progress is comparing the potential benefits of accounting for these interactions during perception using logistic regression. However, these relatively modest effects appear to support the idea that listeners may gain significant leverage by simply tracking the way that cues behave as a series of main effects of contextual factors like talker and coarticulation.

Altogether, our results suggest that variability may not be the daunting problem it is commonly claimed to be. Context factors, as well as phonological features like place and voicing, simply exert additive effects on the cues of any given phoneme. And listeners appear to be equipped with data-explanatory perceptual mechanisms that can subtract the effects of known sources of variance to uncover the target utterance.

References

Apfelbaum, K.S., Bullock-Rest, N., Rhone, A., Jongman, A., & McMurray, B. (2014). Contingent categorization in speech perception. *Language, Cognition and Neuroscience, 29*(9), 1070-1082.

Cole, J.S., 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.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431.

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), 219-246.

McMurray, B., & Jongman, A. (2015). What comes after [f]? Prediction in speech is a product of expectation and signal. *Psychological Science, 27*(1), 43-52.

Mermelstein, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception & Psychophysics, 23*, 331-336.

Perkell, J.S., & Klatt, D. (Eds.). (1986). *Invariance and Variability in Speech Processes.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Whalen, D.H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics, 46*(3), 284-292.