Prediction of articulator positions from subsets of natural variability

D. H. Whalen[1,2,3], Mark K. Tiede[1], and Wei-Rong Chen[1]
[1] Haskins Laboratories; [2] City University of New York; [3] Yale University

The shape of the tongue and its location within the vocal tract largely determines which vowel is being articulated. Data from three or four points on the midsagittal tongue are typically enough to distinguish the shapes, allowing, for example, successful automatic classification of x-ray microbeam data (XRMB; Westbury, 1994). In a training paradigm for second language learners of English, it was found that reasonable shape targets for American English /æ/ could be predicted from three overlapping vowels in Japanese (/i/, /u/, /a/). When speakers matched those targets, this resulted in improved pronunciation (Suemitsu, Dang, Ito, & Tiede, 2015). Here, we further explore how natural variability contributes to accurate predictions. Specifically, how close to the true center of a vowel's articulatory distribution does the estimate of its centroid have to be in order to generate reasonably accurate estimates of individual vowel tokens?

Speech, as we know, is highly variable, even within a speaker. So, do we need to get our estimates just right? Or is there a great deal of latitude in these estimates? We explored this issue using the XRMB database. We examined articulation for English vowels produced by 35 speakers. The tongue pellet positions for 8 vowels (/i ɪ ɛ æ ʌ ɑ ɔ u/) were sampled at the acoustic midpoint of the vocalic segment. The position of two pellets (T2 and T3) were used to predict a tongue blade (TB) position; the two original positions are often on either side of the critical constriction. The tongue tip (TT (=T1)) and tongue dorsum (TD (=T4)) pellets were also used.

There were several steps. First, we plotted the x/y position of each pellet for each vowel to assess the centroid of the distribution, its 1.96 sd. [95% coverage] confidence ellipse, and the orientation of that centroid relative to the centroid of the entire vowel space (Fig. 1). We then identified 21 measurement points ranging from -2 to +2 the length of the confidence ellipse major axis (PC1), projected onto the line connecting the vowel centroid with the overall vowel space centroid (negative sign indicates the direction toward the overall vowel space centroid). Each of these points represented an increasingly centralized or peripheral variant from the true centroid of the distribution. We went beyond the 95% confidence interval deliberately: While those points would necessarily result in worse predictions than the ones within the interval, this nonetheless allowed us to see the shape of the resulting averaged distance functions. Would the accuracy fall off linearly as measurement points deviated from the true centroid? Or is there flexibility in the location that would show tolerance for deviation, perhaps as far as the 95% level itself?

An example of a model measurement function (averaged distance from all the individual tokens to one of those centralized/peripheral measurement points) is shown in Figure 2, for the vowel /æ/, speaker JW12. It is typical of the patterns we found. The TT usually has the largest difference and the steepest functions. But in general, the functions do indeed show a fair amount of flexibility, resulting in essentially the same degree of accuracy through most of the 95% range. To put the distances into a better perspective: The average distance between centroids for neighboring vowels is about 4 mm. Thus accuracy within 2 mm is still fairly discriminative.

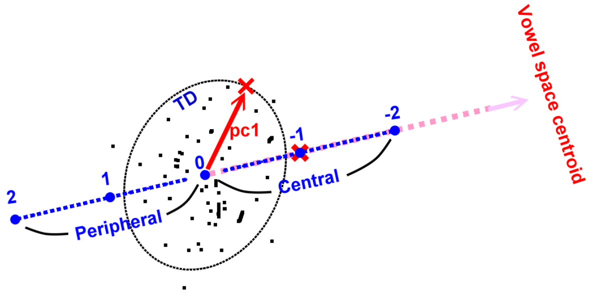

Fig. 1. Example of one vowel's calculations. Dots are individual tokens. Ellipse is 95% confidence (PCA). "pc1" is length of major axis. This is projected onto line connecting centroid ("0") with overall vowel space centroid (not shown). Blue dots represent measurement points for calculating Euclidean distances to data.

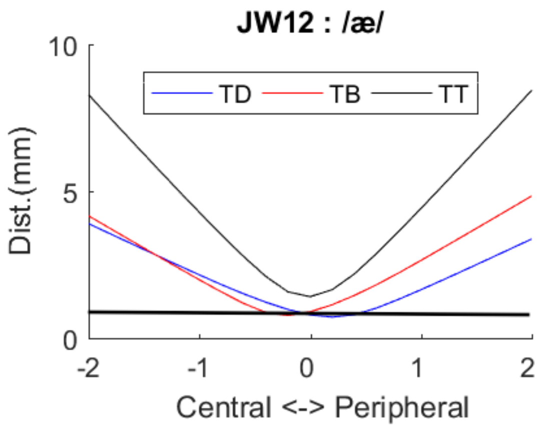

Fig. 2. Example of functions resulting from the 21 measurement points (Fig. 1) for a single vowel for a single speaker. Black line at 1 mm of distance is included for reference (values cannot reach 0 given the distribution of data points).

The results will be further analyzed along these lines, but our preliminary examination of 5 speakers shows patterns very much like those in Figure 2. This indicates that the natural variability seen in productions by any one speaker, even if sparsely sampled, should give rise to reasonably accurate predictions about the shape of the vowel space.

Listeners, of course, do not have direct access to articulator positions, and must recover them from acoustics. Further, exact tongue positions are not likely to be recovered, but, instead, constriction locations and degrees. These have been shown to be recoverable if formant amplitudes are incorporated with formant frequencies (Iskarous, 2010).

The results indicate that variability in itself does not curtail the ability to make sufficiently accurate assessments of the articulatory vowel space.

References:

Iskarous, K. (2010). Vowel constrictions are recoverable from formants. *Journal of Phonetics, 38*, 375-387.

Suemitsu, A., Dang, J., Ito, T., & Tiede, M. K. (2015). A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *Journal of the Acoustical Society of America, 138*, EL382-EL387. doi: doi:http://dx.doi.org/10.1121/1.4931827

Westbury, J. R. (1994). X-ray microbeam speech production database user's handbook: Madison, WI: Waisman Center, University of Wisconsin.