HHS Public Access

Author manuscript

J Phon. Author manuscript; available in PMC 2019 May 01.

Published in final edited form as:

J Phon. 2018 May ; 68: 1–14. doi:10.1016/j.wocn.2018.01.003.

Variability of articulator positions and formants across nine English vowels

D. H. Whalen[a,b,c], **Wei-Rong Chen**[a], **Mark K. Tiede**[a], and **Hosung Nam**[a,d]

[a]Haskins Laboratories

[b]City University of New York

[c]Yale University

[d]Korea University

Abstract

Speech, though communicative, is quite variable both in articulation and acoustics, and it has often been claimed that articulation is more variable. Here we compared variability in articulation and acoustics for 32 speakers in the x-ray microbeam database (XRMB; Westbury, 1994). Variability in tongue, lip and jaw positions for nine English vowels (/u, ʊ, æ, ɑ, ʌ, ɔ, ɛ, ɪ, i/) was compared to that of the corresponding formant values. The domains were made comparable by creating three-dimensional spaces for each: the first three principal components from an analysis of a 14-dimensional space for articulation, and an F1xF2xF3 space for acoustics. More variability occurred in the articulation than the acoustics for half of the speakers, while the reverse was true for the other half. Individual tokens were further from the articulatory median than the acoustic median for 40–60% of tokens across speakers. A separate analysis of three non-low front vowels (/ɛ, ɪ, i/, for which the XRMB system provides the most direct articulatory evidence) did not differ from the omnibus analysis. Speakers tended to be either more or less variable consistently across vowels. Across speakers, there was a positive correlation between articulatory and acoustic variability, both for all vowels and for just the three non-low front vowels. Although the XRMB is an incomplete representation of articulation, it nonetheless provides data for direct comparisons between articulatory and acoustic variability that have not been reported previously. The results indicate that articulation is not more variable than acoustics, that speakers had relatively consistent variability across vowels, and that articulatory and acoustic variability were related for the vowels themselves.

Keywords

Variability; articulation; acoustics; vowels; x-ray microbeam; English

Correspondence to: D. H. Whalen.

1. Introduction

Speakers of a language show a great deal of variability in their realization of the distinctive elements of their sound systems. How that variability is related to the underlying goals themselves has been a topic of much debate in the phonetics literature. The goals of speech production have variously been described as articulatory (e.g., Browman & Goldstein, 1992), acoustic (e.g. Guenther, et al., 1999), or some combination of the two (e.g., Ladefoged, DeClerk, Lindau, & Papçun, 1972). The present paper examines vowels that, due to the fact that they can be sustained in a fairly stable way, are often taken as supporting evidence for acoustic accounts (Schwartz, Basirat, Ménard, & Sato, 2012). In particular, the proposition that acoustic variability is less than articulatory variability for vowels (Johnson, Ladefoged, & Lindau, 1993; Ladefoged, et al., 1972) is tested for a large number of speakers in the Wisconsin X-Ray Microbeam Database (XRMB; Westbury, 1994).

Variability must be sufficiently constrained for the effective transmission of linguistic messages. Articulatory accounts such as Articulatory Phonology (e.g., Browman & Goldstein, 1986, 1995; Goldstein & Fowler, 2003) posit that constrictions of the vocal tract, or *gestures*, are the units of phonology and that their acoustic consequences are immediately perceptible by listeners (Goldstein & Fowler, 2003). Gestures do not specify exact tongue shape, and they typically have acoustic consequences that are expected to convey the presence and nature of the gesture, and those gestures that cannot be recovered perceptually from the acoustics are unlikely to become stable phonological units. The synergies among articulators that allow a gesture to achieve its goal even if individual articulators are perturbed are part of the Task Dynamics approach used to model the implementation of the gestures from Articulatory Phonology (Saltzman & Munhall, 1989).

Acoustic accounts claim that some critical features of the sound signal are the targets for phonological units, and that there are many articulatory configurations that can lead to each acoustic target (e.g., Atal, Chang, Mathews, & Tukey, 1978; Diehl & Kluender, 1989; Guenther, Hampson, & Johnson, 1998; Lindblom, 1990; K. N. Stevens, 2002). The existence of articulatory trade-offs that maintain a fairly constant acoustic output, such as lowering the larynx to compensate for retracted lips (Riordan, 1977) or a tube changing the size of the lip constriction (Savariaux, Perrier, & Orliaguet, 1995) have been taken as evidence that only an acoustic target can be implemented. Further evidence for possible acoustic targets is found in online compensations for altered acoustic feedback (e.g., Houde & Jordan, 1998; Munhall, MacDonald, Byrne, & Johnsrude, 2009). While these accounts do not offer an explanation for such results as the perceptibility of silent-center vowels (e.g., Strange, Verbrugge, Shankweiler, & Edman, 1976), the overall dynamic specification of vowels (e.g., Noiray, Iskarous, & Whalen, 2014), or, to a certain extent, the normalization for vocal tract length that is apparent in perception (Traunmüller, 1984), they do provide a strong challenge to articulatory accounts.

Acoustic accounts have also taken reports that there is in some contexts more variability in articulation than in acoustics as important support. This is especially true for English /r/ (e.g., Guenther, et al., 1999), where various tongue shapes result in nearly identical acoustic patterns. (It is less clear that the constrictions differ, however.) Compensation for some

articulations have been proposed for vowels, in service of acoustic targets (Perkell, Matthies, Svirsky, & Jordan, 1995; Savariaux, et al., 1995).

An extension of the argument that the goals of speech are acoustic is that articulatory variability should be greater than acoustic. This was explored by Johnson, Ladefoged and Lindau (1993), who studied tongue positions for five speakers of American English via the x-ray microbeam system (Kiritani, Itoh, & Fujimura, 1975). They found that their speakers had different locations of the tongue for different vowels, even though differences in anatomy did not seem to account for the differences. They then assumed that the acoustic target had to be the goal, even though they did not measure the acoustics to see if, in fact, the same goal was reached. It has since been taken as a general conclusion that articulation is more variable than acoustics (Bouchard, et al., 2016; Flory, 2015, p. 206; Lee, 2014; Magnuson & Nusbaum, 2007; Niebuhr & Michaud, 2015; Yunusova, Weismer, & Lindstrom, 2011) (see also Maeda, 1991).

One of the kinds of variability that was pointed out by Johnson et al. (1993) was the inconsistency across talkers of the height of the tongue for /ɪ/ vs. /e/. Even though the canonical description of /ɪ/ is that it has a higher position than that of /e/, some speakers have an inverse relationship. For example, in Ladefoged et al. (1972), two of six speakers reversed that height, but it was assumed that the acoustics nonetheless were in a typical pattern. In our own work (Noiray, et al., 2014), we found that, indeed, three of seven speakers had the inverted articulatory relationship for /ɪ/ and /e/, with the tongue being higher for the nominally lower vowel, /e/. However, the pattern of their formants was reversed as well. Dynamic changes (even in the nominally monophthongal /ɪ/) in the formants made each of these vowels easily perceived as the intended category. This form of variability points out the risks involved in analyzing only a single point in vowels.

In many cases, increases in variability are taken as decreases in motor control (see, e.g., discussion in Davids, Bennett, & Newell, 2006). Applications of nonlinear dynamic theory make that connection less clear, as will be elaborated on in the Discussion. However, when it comes to making the comparisons objectively, many difficult issues arise. The acoustics at any time point reflect the state of all the articulators and the resulting transfer function, while articulatory measurements are typically sparse and limited. The scales of the two systems are incommensurate, and they account for different amounts of the vocal tract resonances.

We have focused on variability within a speaker, even though variability across speakers is also extensive. Differences between speakers have been attributed to various factors, ranging from different weighting of elements of tongue shape (Harshman, Ladefoged, & Goldstein, 1977) to differences in use of particular timing intervals (Shaw & Gafos, 2015). Variability within a speaker can only be measured if the relevant aspects of articulation are quantified, and these may include compensatory relations between portions of the vocal tract in the same way that the acoustics might contain tradeoffs between, say, F2 and F3. The relative completeness of MRI images might allow us to quantify the entire vocal tract at some point, but the current state of data reduction is focused on finding the most plausible linguistic gestures (Ramanarayanan, Van Segbroeck, & Narayanan, 2016). Despite these difficulties,

we propose a way of comparing variability across domains for three articulators: the tongue, the lips and the jaw.

The present study examined a range of speakers and contexts found in the XRMB database. We took advantage of the availability of simultaneous articulatory and acoustic data. Johnson et al. (1993) examined tongue position for 6–18 tokens of up to 11 English vowels for their 5 speakers across three consonantal contexts; they did not analyze the acoustics. Here, we examined 32 speakers and 24,897 tokens for 9 vowels, /u, ʊ, æ, ɑ, ʌ, ɔ, ɛ, ɪ, i/. We examined both articulatory and acoustic variability. There are many challenges to equating the variability in these two domains. Some will be addressed in the description of our method, while the remainder will be taken up in the General Discussion.

2. Method

2.1 Speakers and measurements

The data were taken from the publicly accessible XRMB database (Westbury, 1994), which comprises syllables, words and sentences spoken by 57 speakers of American English. Their productions were recorded with midsagittally placed gold pellets whose three-dimensional movements were converted into a two-dimensional representation (posterior -> anterior in x-axis, inferior -> superior in y-axis). Pellets were tracked with a rasterized focused X-ray sweep that followed these pellets glued to various articulators. These included four points on the tongue (T1 one cm posterior to the tongue apex; T4 at the tongue dorsum, ~five cm posterior from T1; T2 and T3 placed roughly equidistantly between T1 and T4), upper lip (UL), lower lip (LL), and lower incisor (to track the mandible; coded as ‘MANi’ in the database but renamed as ‘JAW’ in this paper). After inspecting the quality of the data, we selected the data from 32 speakers (17 females) in the database for further analysis; we excluded speakers with missing channels and obvious erroneous data (e.g., many tracks going above the palate). Articulatory and acoustic values were extracted from all the speech tasks (regardless of the context) that contain the nine monophthong vowels (/u, ʊ, æ, ɑ, ʌ, ɔ, ɛ, ɪ, i/) with primary stress (e.g., /ɛ/ in ‘special’), identified by the text-to-phone interpreter in the P2FA forced aligner (Yuan & Liberman, 2008). One limitation of XRMB is that it lacks tracking of the back part of vocal tract which contains some information of critical oral constrictions for back vowels (possibly for /æ/ too). Therefore, we carried out a separate analysis for the non-low front vowels /ɛ, ɪ, i/, which are most likely to have the critical constrictions in the front part of the vocal tract, thus providing more direct articulatory evidence.

For the temporal landmark for extracting values for vowels, the center of the vocalic segment, often stable enough to be called a “steady state,” is the best available as it minimizes coarticulation with flanking consonants. The temporal midpoint of the vocalic segment is often used for acoustic analyses, while articulation is often taken as attainment of the articulatory target, a gestural plateau ranging between a set percentage (usually 20%) of peak velocity before and after the maximal constriction (see Gafos, 2002; Shaw & Gafos, 2015). We chose the acoustic midpoint, but we also compared that with the points that would be selected by articulatory criteria. For each sample, articulatory velocity was defined as the gradient of six-dimensional articulatory movements constructed by T2, T3 and T4 by using

MVIEW (Tiede, 2009) and calculated the temporal distance between the acoustic midpoint and the articulatory target. The results were that 80.2% of the samples have identifiable articulatory targets centered around the acoustic midpoint. (Note that in the other 19.8% of the cases, there would be no usable articulatory definition due to continuous movement of the articulators.) 43.1% of the samples have the acoustic midpoint within the articulatory target plateau (whose median duration was 34.3 ms). The median (across all samples) of the absolute differences between the acoustic midpoints and the articulatory targets was 5.4 ms (90% quantile = 51.5 ms), which is less than one acoustic analysis window (25 ms) for formant estimation. Therefore, because the acoustic and articulatory landmarks in this corpus were fairly consistent and because 19.8% of the samples do not have identifiable articulatory targets, we chose to extract values for both articulation and acoustics at the acoustic midpoint of the vocalic segment[1].

We focused on flesh-points of the tongue as well as lip and jaw positions, so each sample consisted of 14 articulatory values (T1x, T1y, T2x, T2y, T3x, T3y, T4x, T4y, ULx, ULy, LLx, LLy, JAWx, JAWy in mm) and three acoustic values (F1, F2 and F3 in mels). The formant frequencies were estimated by Burg LPC method and tracked by Viterbi algorithm in PRAAT (Version 6.0.13; Boersma & Weenink, 2009). To reflect the effect of acoustic variability in human perception, formant frequencies were converted to mel scale (S. S. Stevens, Volkmann, & Newman, 1937). Due to missing values and outlier exclusion, the number of samples per vowel is not consistent across speakers, but the large sample size was assumed to compensate for this. The total number of samples was 30141 and 5244 of them were identified as outliers, leaving us 24897 effective samples. The average number of samples, number and rate of exclusions per speaker for each vowel are summarized in Table 1. Details of outlier identification are provided in Section 2.3.

2.2 Data processing and normalization

In order to compare the variabilities in both articulatory and acoustic domains, the data must be normalized into a space that makes them comparable. The 14 dimensional articulatory space, including lingual, lip and jaw articulators, was converted to three dimensions by using the first three components of a principal component analysis (described in detail below). For each speaker in each domain, the normalization of vowel space takes two steps: 1) Centering (set the origin to the center of the space) and 2) rescaling (rescale the data by the average distance of each data point to the center). The acoustic data of speaker JW24 is shown as an example in Fig. 1 to demonstrate the concept of normalization and the calculation of variability in the acoustic domain. (The actual normalization was performed across multiple dimensions for the articulatory and acoustic spaces, but Fig. 1 shows only two dimensions for illustration.) The first step was to define the center of the vowel space of this speaker. Simply averaging all the data for one speaker as the grand mean can be biased by unbalanced vowel inventory and/or unbalanced coarticulatory contexts. Therefore, in order to minimize these biases, we subset the data of the four corner vowels in English /i ɑ u æ/

[1]A potential issue with measuring articulation at the acoustic midpoint rather than a kinematic inflection is that this may not accurately capture the articulatory target, thus increasing measured variability. We accept this risk to retain the data for which articulatory targets could not be accurately identified (~20%), and minimize it by measuring all data consistently (using vowel acoustic midpoints). We thank an anonymous reviewer for pointing out this issue.

from the full dataset, and restricted the occurrences of those vowels to be only immediately following 1) a placeless segment /h ə/, 2) a labial segment /b p f v/, or 3) a silence, and also excluded those followed by /r/ or a nasal. Then, the center of the vowel space is defined as the grand mean of the four medians for /i ɑ u æ/ in this restricted subset, as indicated by the thick 'X' in Fig 1a. In the second step we calculated the median distance (in the F1xF2xF3 space in this example) from all vowel tokens to the grand mean as the unit length for this space (shown in Fig. 1b). A normalized acoustic space (Fig. 1c) was constructed by subtracting the grand mean from all the acoustic data and dividing the data by the unit length (229 mels in this example) defined in step 2.

Then, we computed the Euclidean distance from each token of each vowel to the median of this vowel, as shown with the dotted vectors in Fig. 1d for the vowel /æ/. The length of one vector in Fig. 1d is the normalized variability of one sample of this vowel, and the average variability of this vowel is the mean length of all vectors for this vowel. Thus vowel category targets in general are one unit from the grand mean, while most individual values depart from their categorical targets by much less than one, since vowel instances are closer to their respective targets than to the center of vowel space. Note that our measure of variability is the absolute distance from the category target in multidimensional space; this measure is similar to the median absolute deviation (MAD) in one-dimensional space, which has been proved to be more robust to the distribution of the data than standard deviation (Hampel, 1974).

For the articulatory data, we need to reduce the degrees of freedom from 14 to three in a normalized space such that the unit length of variability is comparable to the acoustic data. We did this by employing a principal component analysis (PCA) after the normalization procedure.[2] A 14-dimensional articulatory space was first constructed for each speaker by including the eight tongue measurements (T1x, T1y, T2x, T2y, T3x, T3y, T4x, T4y), four lip measurements (ULx, ULy, LLx, LLy) and two jaw measurements (JAWx, JAWy) for all vowels. Then the standard PCA was performed; the first three principal components (PC1, PC2 and PC3) were selected to represent the majority of the structure of those tongue and lip configurations. Fig. 2a shows the implementation of PCA, for speaker JW19, by comparing the original (solid lines) tongue positions with the tongue positions recovered (dashed lines) from the coefficients of the first three PCs, for the vowels /i ɑ u/. The closer the original and recovered tongue positions, the better the first three PCs represent the articulatory data. Across all speakers the mean error (absolute distance) between the original and recovered tongue positions was 0.84 mm (SD = 0.17 mm). This is larger than the static RMS error in the XRMB system (0.15 mm; Westbury 1994: 71), but it is equivalent to the size of the pellets themselves, and the tracking error for moving pellets is unknown. Thus these measures are approximately as accurate as they can be with this approach. The mean variance explained by the first three principal components was 88.9% (SD =2.7%). Fig. 2b visualizes the ranges of ± 2 SDs of PC1 (circle lines), PC2 (triangle lines) and PC3

[2]PCA can be implemented before or after the normalization method; both are theoretically justifiable and should not yield different results since our normalization method only involves re-centering and rescaling and PCA affects none of them. Indeed, we tested both ways with our data and both results are almost the same. We chose to implement PCA after normalization because it yields slightly higer accuracy in a discriminant analysis using deep belief neural network (DBN) models (46.0% vs. 45.6%).

(diamond lines) for speaker JW19. The results were similar to those of Parallel Factor Analysis (PARAFAC) analysis of tongue shape in Harshman, Ladefoged & Goldstein (1977) and Hoole & Mooshammer (2002) in that, roughly speaking, PC1 accounts for the upper-front to lower-back tongue movements and jaw openning, whereas PC2 accounts for the upper-back to lower-front tongue movements and lip/jaw opening as well as lip protrusion, and PC3 tracks complementary raising of the tongue tip with lowering of the root.

After the articulatory space is normalized and reduced to three dimensions, the resulting normalized acoustic and articulatory spaces are on the same scale and comparable.

2.3 Outlier identification: the elbow method

The data in XRMB are subject to noise and measurement errors. Therefore, we carried out an outlier identification method (refered to as the "elbow method"[3] here) to exclude presumably erroneous data. Specifically, for each vowel category produced by each speaker in each of the normalized articulatory and acoustic domains, the Euclidean distance from each data point to the vowel median for that target was constructed and then sorted from the smallest to the largest values, shown as the broken lines (articulatory) and solid lines (acoustic) in the upper panel of Fig. 5. An "elbow" of this array (the triangle marker in the upper panel of Fig. 3), where the variability rate increases, can be identified by detecting the point at which the second derivative of this array passes above a threshold. Fig. 3 demonstrates the elbow method in this study. First we fit a polynomial (the broken circle line in Fig. 3) through the datapoints (the thick curved line) and then differentiated the polynomial twice, as the second panel (first derivative) and the third panel (second derivative) in Fig. 3. A threshold above zero (the cross point of the dotted vertical line and the curve line in the third panel in Fig. 3) for the second derivative was determined heuristically, and the projection of the point onto the original data was defined as the ideal elbow (the triangle marker in the first panel) in this study. The elbow method was applied to both the acoustic and the articulatory distances. Any token that was extreme on either scale was excluded from further analysis.

Recall that our normalization method rescales within-vowel variability by the magnitude of the entire vowel space. Any instance of a vowel with variability of more than one unit on the normalized scale suggests greater within-vowel variability than across-vowel variability, which is less probable than the converse. The exclusion rates are comparable among vowels and speakers. The average exclusion rate was 17.4% across vowels and speakers; of the 17.4%, 9.9% were outliers in the articulatory domain, 9.6% in the acoustic domain, 2% in both. The very small amount of overlap of outliers in both domains indicates that the outliers identified by our elbow method were mostly attributed to measurement errors, not production errors; if most of the exclusions were due to extreme articulations, we might be able to assume that the acoustic compensation was sufficient to allow the correct vowel to be

[3]The notion of "elbow" has been commonly used in determining the number of clusters in K-means clustering algorithm, by selecting the point at which the error decreasing rate drops rapidly (also known as "scree plot"). Chiang et al (2003) proposed a robust outlier detection algorithms based on detecting the "elbow" of the sorted changes in standard deviation. Our outlier identification method in this paper is similar to the one in Chiang et al (2003) but simplified to accommodate a more limited data set.

indicated. The fact that both acoustic and articulatory data were excluded suggests that only measurement errors were involved.

2.4 Statistical analysis

We carried out three statistical analyses: 1) between-speaker correlations of variability between vowels, 2) correlation of variability between the acoustic and articulatory domains, 3) linear mixed modeling of variability predicted by domain and by vowel.

For the correlation between acoustic and articulatory variabilities, multiple Pearson correlation coefficients were calculated by pairing the variabilities of all the five vowels for 32 speakers in the acoustic domain with those in the articulatory domain. We controlled the false discovery rate (FDR) of the multiple tests of null hypotheses by the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) and set the significance level at FDR $q = .05$.

The linear mixed models were computed in R (R Core Team, 2015), using the *lmer* (Bates, Maechler, Bolker, & Walker, 2013) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2016) packages. Candidate models were chosen using log-likelihood comparisons. The selected mixed-effects model predicted *variability* as the dependent variable, from fixed effects of *domain* (two levels: *articulatory* and *acoustic*) and *vowel* (nine levels: /u, ʊ, æ, ɑ, ʌ, ɔ, ɛ, ɪ, i/) and their interaction, with random effects of *speaker* and by-speaker random slopes for the effect of *vowel*. The effect of *Gender* contributed no improvement to the model and was thus excluded during model selection. We also reported the marginal R^2 (variance explained by fixed effects) and conditional R^2 (variance explained by fixed effects and random effects) (Nakagawa & Schielzeth, 2013) as an indication of the sizes of effects and the goodness of fit of the selected model, using the *piecewiseSEM* package (Lefcheck, 2016). For a more intuitive reading in what follows, we renamed the marginal R^2 as 'R^2-fixed' and defined an 'R^2-random' as the difference between conditional R^2 and marginal R^2. We further carried out multiple post-hoc comparisons (the p-values were adjusted by Tukey HSD method) between vowels separately in each of the two domains, by using the *multcomp* package (Hothorn, Bretz, & Westfall, 2008) in R.

3. Results

3.1 Variability compared between articulatory and acoustic domains

Fig. 4 visualizes the normalized variabilities in both articulatory (left panels) and acoustic (right panels) domains; the first two dimensions (PC1 and PC2 for articulatory domain; F1 and F2 for acoustic domain) are plotted in the upper panels, and the first and third dimensions (PC1 and PC3 for articulatory domain; F1 and F3 for acoustic domain) in the lower panels. Each ellipse in Fig. 4 indicates the 95% confidence interval, estimated by PCA, of each vowel target. These ellipses were used only for visualization but not for any of the statistical analyses. As described in Section 2.2, the normalized variability for a given vowel was calculated by averaging the distance of each sample (data point) of the vowel to the vowel median, which is roughly proportional to the area of the ellipse of each vowel in Fig. 4.

We further measured the average difference between articulatory and acoustic variabilities across all vowels for each speaker, for all nine vowels (upper panel in Fig. 5) and for non-low front vowels only (lower panel in Fig. 5). The lines with circles in both panels of Fig. 5 show the mean articulatory variability minus the mean acoustic variability (positive values indicate larger articulatory variability than acoustic variability) for each speaker (scale on the left y-axis); the triangle lines show the percentage of tokens where the acoustic variability is greater than articulatory variability for each speaker (scale on the right y-axis). In the analysis of all nine vowels (upper panel in Fig. 5), half of the speakers show lower articulatory variability (than acoustic variability) and half the reverse. On the other hand, in the analysis of the non-low front vowels only (lower panel in Fig. 5), most speakers show lower articulatory variabilities. The mean difference in variability was −0.01 (units in normalized space) for the nine-vowel analysis and −0.06 for non-low front vowels. The mean percentage of tokens with greater acoustic than articulatory distances to category median was 50% for all nine vowels and 57% for the non-low front vowels. In short, when the comparison of articulatory and acoustic variability is made across all nine vowels, there is no indication of contrast between articulatory and acoustic variabilities. However, when the same analysis is performed for the three non-low front vowels only, there is a trend that acoustic variability is larger than articulatory variability.

Table 2 summarizes the results of the linear mixed-effects model that includes *domain* and *vowel* as well as their interaction as fixed effects, and *speaker* as a random intercept as well as by-speaker random slopes for the effect of *vowel*.[4] The baseline for *domain* effect is 'acoustic domain', and the baseline for the *vowel* effect is the vowel /ɑ/. The coefficient of the main effect *domain* is 0.045 ($p < .01$), indicating that in general the articulatory variability of a vowel is slightly higher than the acoustic variability; however, the significant Vowel interactions show that this effect is not consistent, and the Cohen's D for *domain* is negligible (0.057).

3.2 Correlation of variability between vowels in the articulatory domain

We carried out Pearson correlation analysis to test the correlation among vowel variabilities in the articulatory domain. Table 3 displays the correlation coefficients for each pair of the nine vowels across speakers. Positive values indicate positive correlations between vowel variabilities in articulation. Of the 36 pairs of vowels, two are significant (FDR $q < .01$), another 11 of them have unadjusted p values less than .05, and the rest are not significant.

3.3 Correlation of variability between vowels in acoustic domain

Table 4 displays the correlation coefficients for vowel variabilities in the acoustic domain. Of the 36 pairs of vowels, two are significant (FDR $q < .05$), another six of them have unadjusted p values less than .05. The pairs with significant positive values of correlation coefficients indicate that speakers with larger variability for one vowel are likely to have larger variability for the other vowel in the pair. Compared to the correlations of vowel variabilities in the articulatory domain, the shared pattern is that the variability of /ɔ/ is positively correlated with the variabilities of the other four non-high vowels /ɛ/, /æ/, /ɑ/

[4] *lmer* model syntax: Variability ~ Domain + Vowel +Domain:Vowel+ (Vowel|Speaker)

and /ʌ/ in both articulatory and acoustic domains. Note that the controlled FDR indicates that if there is at least one pair that meets the significant level, then the more general null hypothesis (i.e., no correlation in any one of those vowel pairs) is rejected.

3.4 Correlation between articulatory and acoustic variabilities

To further explore the correlation between articulatory and acoustic variabilities, we carried out Pearson correlation analyses separately for each vowel as well as the correlation across all nine vowels. Table 5 shows the correlation coefficients separately for each vowel. Positive values indicate that greater vowel variability in articulation is accompanied by greater vowel variability in acoustics. The results show that there are positive correlations between articulatory and acoustic variabilities for eight of the nine vowels; five (/i ɪ æ ʊ u/) of them are significant with respect to the FDR level. The vowel /ɑ/ shows no correlation between articulatory and acoustic variability.

Fig. 6 displays the overall correlation between articulatory and acoustic variabilities across vowels and speakers. Each data point in Fig. 6 indicates the mean normalized articulatory and acoustic variabilities of one vowel produced by one speaker. The regression and correlation analyses were carried out separately for all nine vowels (solid line in Fig. 6) and for the non-low front vowels only (dashed line in Fig. 6). The overall correlations are positive and significant ($p < .01$) for both sets of data, and the coefficient of determination (r^2) indicates that the amount of variance in acoustic variability that can be explained by articulatory variability is 35 % in the set of nine vowels and 41% in the set of non-low front vowels.

3.5 Comparing variabilities among vowels

Finally, we compared the variabilities among the nine vowels in both the articulatory and acoustic domains. Fig. 7 displays the distributions of vowel variabilities and summarizes the general comparisons in both articulatory and acoustic domains. The probability density function (curved lines in Fig. 7) for each vowel was fitted across 32 speakers by kernel density estimation (KDE). As we have seen from the high between-domain correlations in the previous section, the general patterns in the distributions are also very similar in both articulatory and acoustic domains: vowel variability is the lowest for /i/, highest for /ɔ/, and low vowels are in general more variable than non-low vowels.

We performed separate analyses on articulatory vowel variabilities and acoustic vowel variabilities, fitting a simple model predicting *variability* with *vowel* as the only fixed effect and random intercepts by *speaker*, separately, and then ran post-hoc comparisons on the *vowel* effect. Table 6 reports the results of the models fitted to the articulatory subset (upper panel) and to the acoustic subset (lower panel). The baseline for the *vowel* effect for both models is /æ/. The coefficients (β) and the associated t values of the two models show that there is at least one vowel that has significantly different variability than the vowel /æ/ in both articulatory and acoustic domains.

Tukey HSD pairwise post-hoc comparisons reveal that the vowel variability decreases in the order: /ɔ/>/ɑ æ/; /ɑ/>/ʌ/; /æ ʌ/ > /ɛ ɪ ʊ/ > /u/ > /i/ in the articulatory domain (‘>’ indicates ‘significantly greater than’ ($p < .05$); implicational law applies); and in the order: /ɔ/>/ɑ/

> /æ ɛ ʊ/; /u ɪ ʌ/ > /ʊ/ > /i/ ($p < .05$) in the acoustic domain. The general pattern of articulatory variability is that low-back vowels have greater variabilities than non-low front vowels. Table 7 further compares the amount of variance explained by fixed effects (R^2-fixed) and random effects (R^2-random) in the two models fitted separately to the articulatory (line 1) and acoustic (line 2) subsets. The results show that the *vowel* effect accounts for 79.6% and 46.8% of the variance in articulatory and acoustic variabilities respectively, whereas the variance explained by the random (speaker) effect is larger in the acoustic domain (R^2-random = 8.9%) than in the articulatory domain (R^2-random = 4.9%).

4. Discussion

Comparison of acoustic and articulatory variabilities for our measures showed near equivalence, with articulation being more variable for half of the speakers. For the non-low front vowels, articulation was less variable than acoustics. These results held despite intrinsic differences in the information provided by our measures for the two domains. The acoustic signal includes the contributions of all the articulators, including side-cavity zeroes, nasal tract coupling, and, most importantly, the posterior and parasagittal tongue beyond the range of the XRMB pellets. Thus aspects of the articulation that might have been important for the acoustic output were not necessarily measured here. However, the predictability of pharyngeal shape from anterior portions of the tongue for English (Whalen, Kang, Magen, Fulbright, & Gore, 1999) appears to have allowed for adequate predictions. On the other hand, it is possible that articulatory variability, such as height of the velum (and, more importantly, amount of nasal coupling) is accurately represented as variable in the acoustic signal but missing from our measurements. Further exploration of this issue, perhaps using real-time MRI (e.g., Narayanan, et al., 2014), is warranted.

Articulation was coded in the experiment via flesh points in the speaker's physiological range. It may be that locating constrictions directly rather than indirectly via tongue and lip pellet positions (as done here) would capture the production more cogently, but the measurement system of the XRMB database was not sufficient to support such a description. Even with such a description, it is possible that our PCA analysis of the articulation might collapse compensatory postures or trading relations (Perkell, et al., 1995; Savariaux, et al., 1995), obscuring some of the articulatory variability that acoustic theories predict. It is a challenge left to future analyses to devise a more global assessment of articulatory variability.

Those speakers who had larger variability in one vowel tended to have larger variability in the other vowels as well. This can be seen in the positive between-vowel correlations in both the articulatory domain (Table 3) and acoustic domain (Table 4). Speakers seem to be either generally variable or generally consistent rather than being variable on individual vowels.

A similar result is that those speakers who had relatively large variability in articulation tended to have relatively large variability in acoustics as well. This can be seen in the positive correlations between articulatory and acoustic variabilities (Table 5 and Fig. 6). It does not seem to be the case that speakers are variable in articulation without also being variable in acoustics. Rather, the two are correlated.

High and non-low front vowels showed less articulatory variability than the other vowels (Fig. 7), while the acoustic variabilities are more similar. This might indicate a difference in articulation, a possible consequence of bracing against the palate for some vowels but not others (Gick, Allen, Roewer-Despres, & Stavness, 2017). The amount of contact with the palate also varies with palate shape (e.g., Brunner, Fuchs, & Perrier, 2009). However, no XRMB pellets were placed in the critical pharyngeal region for the back vowels. We may thus be inflating the measure with the relatively benign variability in tongue position without an accurate measurement of the critical portion. Further, if there were variability in the pharyngeal position that is compensated for by lip rounding, our measurements would require having access to both of those settings in order to see the dependency and thus reduce the overall variability. Such compensation by the lips for changes in tongue position have, of course, been taken as evidence that the vowel's target is acoustic (e.g. Perkell et al., 1995). If such subcomponents are to be included in the measurement of variability, it may be necessary to subdivide the acoustics as well, looking for variability in individual formants or differences between formants, for example. Our procedure for reducing dimensionality takes such synergies into account, as mutual dependencies are projected onto the orthogonal principal components.

There are many remaining issues in normalizing between the articulatory and acoustic domains. As already mentioned, this study only uses the tongue, lips and jaw as measured by the XRMB system. This necessarily excludes any direct measurement of the pharynx, which is crucial for low vowels (e.g., Russell, 1928), although jaw height is relatively well correlated with pharyngeal depth (e.g., Lindblom & Sundberg, 1971). It would be possible to extrapolate from these flesh points to predict the pharyngeal shape with some accuracy (Whalen, et al., 1999), but it is not clear that that would add any information to the results obtained here. The four flesh points that are measured do not necessarily cover every relevant aspect of the tongue, as they may not track the highest point of the tongue (Noiray, et al., 2014). While finding the tongue-to-palate distance would also be useful (Beckman, et al., 1995), the palate traces in XRMB are relatively coarse and do not provide usable data posterior to the hard palate; constrictions in the velar region are likely to include changes in height of the soft palate to some extent. Thus the four tongue pellets are the best estimate of lingual articulation that we have. The placement of the lip pellets on the vermillion border, although standard practice, also made calculations of lip aperture problematic: The inner edges of the lips can, with enough flaring of the outer portion of the lips, make a narrower constriction than indicated by the pellets. Without curling, the relative aperture can be accurate, even though the pellets will be some distance apart even when there is complete closure.

The acoustic signal is not only on a completely different scale, it also includes all the aspects of production, including those that we were unable to quantify from the XRMB data. If it were possible to attribute certain aspects of the acoustic signal exactly to tongue position, then we would be better able to equate our two measures in this regard. Although formants can largely be allocated to front or back cavities (e.g., Apostol, Perrier, & Bailly, 2004; Dunn, 1950), the relationships between particular formants and cavities change with different vowel qualities. We were able to construct a 14 dimensional space for the tongue and lip pellets in part because they were of comparable magnitude and range. Because

additional features of the acoustics (e.g., formant amplitudes) operate over very different scales than formants, they cannot be included in direct reduction via the PCA that we applied to the tongue locations.

If speakers have an acoustic target for vowels, is the target one matched to just their vocal tract, or is it one that fits into the result of speaker normalization, that is, placing their own acoustics into a more general space for the dialect as a whole? On the one hand, the speaker has some reason to tailor the vowel to just the vocal tract that will produce it, and speakers, of course, have vocal tracts of different sizes and acoustic capabilities. On the other hand, the speaker needs to be understood, so the target must be one that a listener can interpret appropriately. If the space is acoustic, it would seem to be acoustic in the sense of a speaker-normalized acoustic space, not in a linear transformation of the formant values.

The difficulties in making articulation and acoustics comparable have always existed, even though some authors have been willing to assert that variability is greater in the articulatory domain than the acoustic (Johnson, et al., 1993, and citations in the Introduction). Although the decisions we have made about ways of equating the scales or variability have their benefits and drawbacks, they have the virtue of addressing the challenge directly. It is not enough to compute standard deviations for each formant and each pellet and compare the results. The overall range of possibilities is not commensurate between the two domains, and thus any such comparisons are suspect. There are, no doubt, improvements that can be made on our measures, but any future comparisons should address the complexities explicitly.

Quantal theory (e.g., K. N. Stevens, 1989) predicts that certain vowels, especially /i ɑ u/, will be more stable acoustically than articulatorily. Evidence for this possibility has been somewhat mixed (Beckman, et al., 1995; Gendrot & Adda-Decker, 2007; Pisoni, 1980; Syrdal, 1985), though generally suggesting that caution is needed. For the three vowels /i ɑ u/ in the current data, only /ɑ/ is less variable in the acoustics than in the articulation (Fig. 7).

The assumption that greater variability indicates lesser control is deeply embedded in current theoretical models, but it is not always the case. There are times when increased control *increases* variability (Riley & Turvey, 2002). For example, Riley et al. (1999) found, using Recurrence Quantification Analysis (RQA), that two aspects of postural sway (essentially, the anterior-posterior sway and the lateral sway) responded to task difficulty in two rather paradoxical ways: lateral sway became less deterministic but also less variable, while anterior-posterior sway increased in determinism but also in variability. Somewhat similar results have also been found in speech, where adults who stutter increased their determinism (as measured by RQA) but did not reduce their variability in certain conditions (Jackson, Tiede, Beal, & Whalen, 2016). As RQA cannot be applied to datasets that are not time series, the challenge remains in knowing when to attribute variability in vowel targets to lack of control and when, instead, to an excess of control. However, RQA analysis holds promise for analyzing the trajectories that vowels take rather than measuring single points along that trajectory, as done here and in many other studies. It remains to be determined whether the examples of increased control leading to increased variability are unusual or indicate that variability should always be analyzed more fully. For the present, we will continue to

assume that increased variability correlates with lesser control. The current results indicate that variability in the production domain is not counteracted, in general, in the acoustic domain as assumed by, e.g., Johnson et al. (1993). Such an outcome does not contradict the ability of speakers to compensate for perturbations, but it demonstrates, for a fairly large number of speakers in sentential context, that the control parameters we might expect for the two domains are similar if not, indeed, exactly the same.

5. Conclusion

Vowels in running speech of 32 American English speakers were found to be approximately equal in production variability (as measured on principal components derived from positions of the jaw, lips and flesh points on the anterior portion of the tongue) and in acoustics (measured by the first three formants). The principal components capture basic synergies of linguistic gestures (see Fig. 2). Because the articulators measured here are not the only ones affecting acoustic output, the contribution of unmeasured aspects (pharyngeal shape, nasal coupling, etc.) are inferred from redundancy in the articulators that were measured. A speaker's degree of variability in production for one vowel correlated significantly with their degree of variability in other vowels; that is, more variable speakers tended to be more variable for all vowels measured. The same correlation was also observed for acoustic variability. The results are consistent with theories that take articulation and acoustics as intimately linked and equally important in conveying information via speech.

Acknowledgments

This work was funded by NIH grant DC-002717 to Haskins Laboratories. We thank Eric S. Jackson, Jason Shaw, Phil Hoole, Marija Tabain and two anonymous reviewers for helpful comments. Audiences at the University of Western Sydney, LaTrobe University, Macquarie University, the DYMO workshop at the University of Cologne also provided feedback. Dani Byrd provided the impetus for major improvements in the experiment, as well as editorial comments. None of them bears responsibility for the final product.

References

Apostol L, Perrier P, Bailly G. A model of acoustic interspeaker variability based on the concept of formant–cavity affiliation. Journal of the Acoustical Society of America. 2004; 115:337–351. [PubMed: 14759026]

Atal BS, Chang JJ, Mathews MV, Tukey JW. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. Journal of the Acoustical Society of America. 1978; 65:1535–1555.

Bates D, , Maechler M, , Bolker B, , Walker S. lme4: Linear mixed-effects models using Eigen and S4 R Core Team; 2013 Retrieved from

Beckman ME, Jung TP, Lee S, De Jong K, Krishnamurthy AK, Ahalt SC, Cohen KB, Collins MJ. Variability in the production of quantal vowels revisited. Journal of the Acoustical Society of America. 1995; 97:471–490.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57:289–300.

Boersma P, , Weenink D. Praat: doing phonetics by computer [Computer program] 2009 Retrieved from http://www.praat.org/

Bouchard KE, Conant DF, Anumanchipalli GK, Dichter B, Chaisanguanthum KS, Johnson K, Chang EF. High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings. PLoS ONE. 2016; 11:e0151327. [PubMed: 27019106]

Browman CP, Goldstein LM. Towards an articulatory phonology. Phonology Yearbook. 1986; 3:219–252.

Browman CP, Goldstein LM. Articulatory phonology: An overview. Phonetica. 1992; 49:155–180. [PubMed: 1488456]

Browman CP, , Goldstein LM. Dynamics and Articulatory Phonology. In: Port RF, , Gelder Tv, editorsMind as motion Cambridge, MA: MIT Press; 1995 175193

Brunner J, Fuchs S, Perrier P. On the relationship between palate shape and articulatory behavior. Journal of the Acoustical Society of America. 2009; 125:3936–3949. [PubMed: 19507976]

Chiang LH, Pell RJ, Seasholtz MB. Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control. 2003; 13:437–449.

Davids K, , Bennett S, , Newell KM. Movement system variability Human kinetics; 2006

Diehl RL, Kluender KR. On the objects of speech perception. Ecological Psychology. 1989; 1:121–144.

Dunn HK. The calculation of vowel resonances, and an electrical vocal tract. Journal of the Acoustical Society of America. 1950; 22:740–753.

Flory Y. Unpublished PhD dissertation University of Cambridge; 2015 The impact of head and body postures on the acoustic speech signal.

Gafos A. A grammar of gestural coordination. Natural Language and Linguistic Theory. 2002; 20:269–337.

Gendrot C, , Adda-Decker M. Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. Proceedings of the 16th International Congress of Phonetic Sciences; Germany: Institute of Phonetics, Saarland University Saarbrüken; 2007 14171420

Gick B, Allen B, Roewer-Despres F, Stavness I. Speaking tongues are actively braced. Journal of Speech, Language and Hearing Research. 2017; 60:494–506.

Goldstein LM, , Fowler CA. Articulatory phonology: A phonology for public language use. In: Schiller N, , Meyer A, editorsPhonetics and phonology in language comprehension and production: Differences and similarities Berlin: Mouton de Gruyter; 2003 159207

Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, Perkell JS. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. Journal of the Acoustical Society of America. 1999; 105:2854–2865. [PubMed: 10335635]

Guenther FH, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech movements. Psychological Review. 1998; 105:611–633. [PubMed: 9830375]

Hampel FR. The influence curve and its role in robust estimation. Journal of the American Statistical Association. 1974; 69:383–393.

Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. Journal of the Acoustical Society of America. 1977; 62:693–707. [PubMed: 903511]

Hoole P, , Mooshammer CM. Articulatory analysis of the German vowel system. In: Auer HP, Gilles P, , Spiekermann H, editorsSilbenschnitt und Tonakzente Tübingen: Niemeyer; 2002 129152

Hothorn T, Bretz F, Westfall PH. Simultaneous inference in general parametric models. Biometrical Journal. 2008; 50:346–363. [PubMed: 18481363]

Houde JF, Jordan MI. Sensorimotor adaptation in speech production. Science. 1998; 279:1213–1216. [PubMed: 9469813]

Jackson ES, Tiede MK, Beal DS, Whalen DH. The impact of social-cognitive stress on speech variability, determinism, and stability in adults who do and do not stutter. Journal of Speech, Language, and Hearing Research. 2016; 59:1295–1314.

Johnson K, Ladefoged P, Lindau M. Individual differences in vowel production. Journal of the Acoustical Society of America. 1993; 94:701–714. [PubMed: 8370875]

Kiritani S, Itoh K, Fujimura O. Tongue-pellet tracking by a computer-controlled x-ray microbeam system. Journal of the Acoustical Society of America. 1975; 57:1516–1520. [PubMed: 1141500]

Kuznetsova A, , Brockhoff PB, , Christensen RHB. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-32 2016 Retrieved from https://cran.r-project.org/package=lmerTest

Ladefoged P, DeClerk J, Lindau M, Papçun GJ. An auditory-motor theory of speech production. UCLA Working Papers in Phonetics. 1972; 22:48–75.

Lee HN. Unpublished PhD dissertation University of Hawaii at Manoa; Honolulu: 2014 A grammar of Baba Malay with sociophonetic considerations.

Lefcheck JS. piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. Methods in Ecology and Evolution. 2016; 7:573–579.

Lindblom BE. Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle WJ, , Marchal A, editorsSpeech production and speech modeling Dordrecht: Kluwer Academic Publishers; 1990 403439

Lindblom BE, Sundberg J. Acoustical consequences of lip, tongue, jaw, and larynx movement. Journal of the Acoustical Society of America. 1971; 50:1166–1179. [PubMed: 5117649]

Maeda S. On articulatory and acoustic variabilities. Journal of Phonetics. 1991; 19:321–331.

Magnuson JS, Nusbaum HC. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of Experimental Psychology: Human Perception and Performance. 2007; 33:391–409. [PubMed: 17469975]

Munhall KG, MacDonald EN, Byrne SK, Johnsrude IS. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. Journal of the Acoustical Society of America. 2009; 125:384–390. [PubMed: 19173425]

Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution. 2013; 4:133–142.

Narayanan S, Toutios A, Ramanarayanan V, Lammert A, Kim J, Lee S, Nayak K, Kim YC, Zhu Y, Goldstein LM, Byrd D, Bresch E, Ghosh P, Katsamanis A, Proctor M. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). Journal of the Acoustical Society of America. 2014; 136:1307–1311. [PubMed: 25190403]

Niebuhr O, Michaud A. Speech data acquisition: the underestimated challenge. KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik. 2015; 3:1–42.

Noiray A, Iskarous K, Whalen DH. Variability in English vowels is comparable in articulation and acoustics. Laboratory Phonology. 2014; 5:271–288. [PubMed: 25101144]

Perkell JS, Matthies ML, Svirsky MA, Jordan MI. Goal-based speech motor control: A theoretical framework and some preliminary data. Journal of Phonetics. 1995; 23:23–35.

Pisoni DB. Variability of vowel formant frequencies and the Quantal Theory of Speech: A first report. Phonetica. 1980; 37:285–305.

R Core Team. R: A language and environment for statistical computing 2015 Retrieved from https://www.r-project.org/

Ramanarayanan V, Van Segbroeck M, Narayanan SS. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Computer Speech and Language. 2016; 36:330–346. [PubMed: 26688612]

Riley MA, Balasubramaniam R, Turvey MT. Recurrence quantification analysis of postural fluctuations. Gait and Posture. 1999; 9:65–78. [PubMed: 10575072]

Riley MA, Turvey MT. Variability and determinism in motor behavior. Journal of Motor Behavior. 2002; 34:99–125. [PubMed: 12057885]

Riordan CJ. Control of vocal-tract length in speech. Journal of the Acoustical Society of America. 1977; 62:998–1002. [PubMed: 908793]

Russell GO. The vowel: Its physiological mechanism as shown by x-ray Columbus, OH: Ohio State University Press; 1928

Saltzman E, Munhall KG. A dynamical approach to gestural patterning in speech production. Ecological Psychology. 1989; 1:333–382.

Savariaux C, Perrier P, Orliaguet JP. Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. Journal of the Acoustical Society of America. 1995; 98:2466–2474.

Schwartz JL, Basirat A, Ménard L, Sato M. The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics. 2012; 25:336–354.

Shaw JA, Gafos AI. Stochastic time models of syllable structure. PLoS ONE. 2015; 10:e0124714–e0124714. [PubMed: 25996153]

Stevens KN. On the quantal nature of speech. Journal of Phonetics. 1989; 17:3–45.

Stevens KN. Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of the Acoustical Society of America. 2002; 111:1872–1891. [PubMed: 12002871]

Stevens SS, Volkmann J, Newman EB. A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America. 1937; 8:185–190.

Strange W, Verbrugge RR, Shankweiler DP, Edman TR. Consonant environment specifies vowel identity. Journal of the Acoustical Society of America. 1976; 60:213–224. [PubMed: 956528]

Syrdal AK. Aspects of a model of the auditory representation of American English vowels. Speech Communication. 1985; 4:121–135.

Tiede MK. MVIEW: software for visualization and analysis of concurrently recorded movement data. 2009 Retrieved from.

Traunmüller H. Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. Speech Communication. 1984; 3:49–61.

Westbury JR. X-ray microbeam speech production database user's handbook Madison, WI: Waisman Center, University of Wisconsin; 1994

Whalen DH, Kang AM, Magen HS, Fulbright RK, Gore JC. Predicting pharynx shape from tongue position during vowel production. Journal of Speech, Language and Hearing Research. 1999; 42:592–603.

Yuan J, Liberman MY. Speaker identification on the SCOTUS corpus. The Journal of the Acoustical Society of America. 2008; 123:3878.

Yunusova Y, Weismer GG, Lindstrom MJ. Classifications of vocalic segments from articulatory kinematics: Healthy controls and speakers with dysarthria. Journal of Speech, Language, and Hearing Research. 2011; 54:1302–1311.

Highlights

- Variability in two domains, articulatory and acoustic, were compared for 9 English vowels produced by 32 speakers in the x-ray microbeam database (XRMB; Westbury, 1994).
- Individual tokens were closer to the acoustic median than to the articulatory only about half the time, indicating balance of the two factors for each speaker (range: 40–60%).
- Speakers who were relatively variable on one vowel were relatively variable on the other vowels as well.
- Acoustic and articulatory variability were positively correlated.

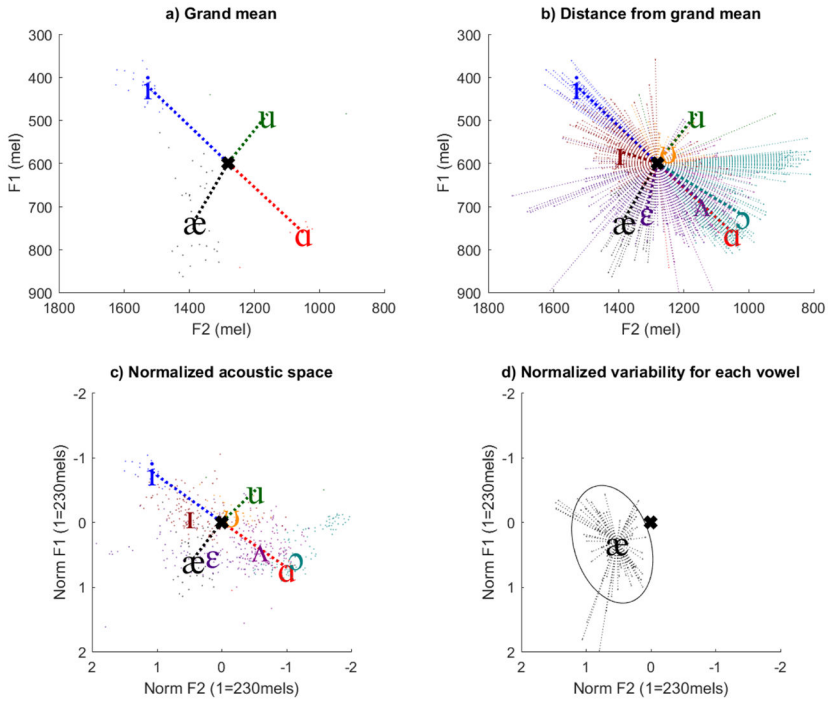

Fig 1.

Concept of the normalization method and the calculation of variability illustrated in two dimensions. The third dimension of F3 is not visualized here, but was included in the calculation. The black cross in each subplot indicates the location of the grand mean of the entire vowel space. The ellipse indicates the 95% confidence interval of each vowel estimated by PCA. a): Define the grand mean of the acoustic (F1xF2xF3) space as the average of the medians of the four corner vowel distributions. b): Calculate the median of the distance from each data point to the grand mean as the unit length of the acoustic space. c): Remove the grand mean from the data and divide each data point by the unit length. d): vowel variability is defined as the mean Euclidean distance of all data points to its vowel target in the normalized acoustic space.

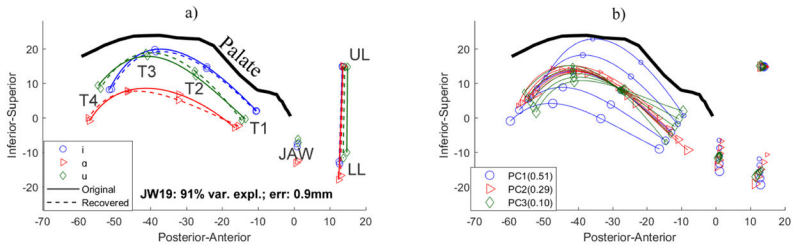

Fig. 2.

A conceptual figure for visualizing PCA. Left: Original tongue positions (solid lines) vs. Recovered tongue positions (dashed lines) from three PCs for the vowels /i ɑ u/. Right: Ranges from −2 (smaller markers) to +2 (larger markers) SDs of the first component (PC1, blue circles), second component (PC2, red triangles) and the third component (PC3, green diamonds). The numbers in the parentheses following PC1, PC2 and PC3 in the legend show the variances explained by each PC. [Color online.]

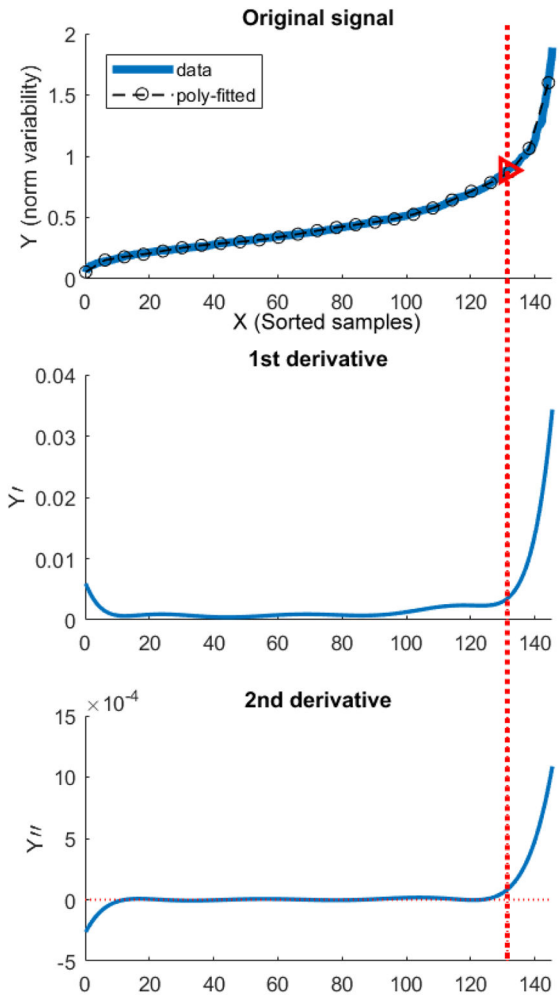

Fig. 3.

Concept of the elbow method. The thick curve line in the first panel represents the original data points, and the broken circle line is the result of a polynomial fitted to the data. The second panel presents the first derivative of the polynomial, and the third panel shows the second derivative. The dotted vertical line through the three panels indicates the location of the "elbow."

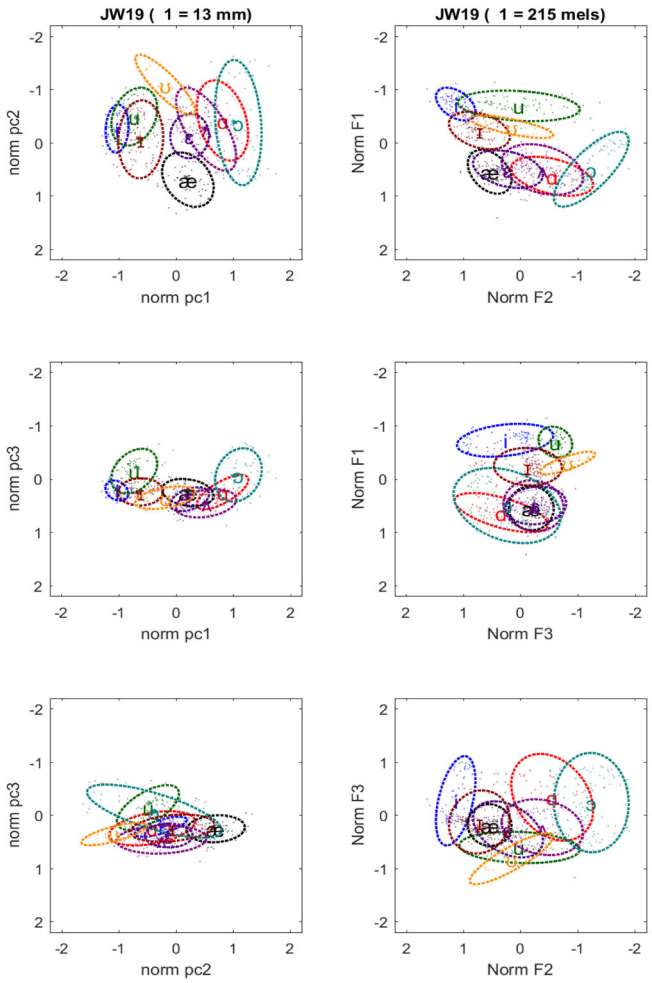

Fig. 4.

Visualization of the normalized variabilities of vowels for one speaker (JW19) in both articulatory (left) and acoustic (right) domains. Larger area of ellipse indicates greater variability. Each ellipse indicates the 95% confidence interval of each vowel estimated by PCA.

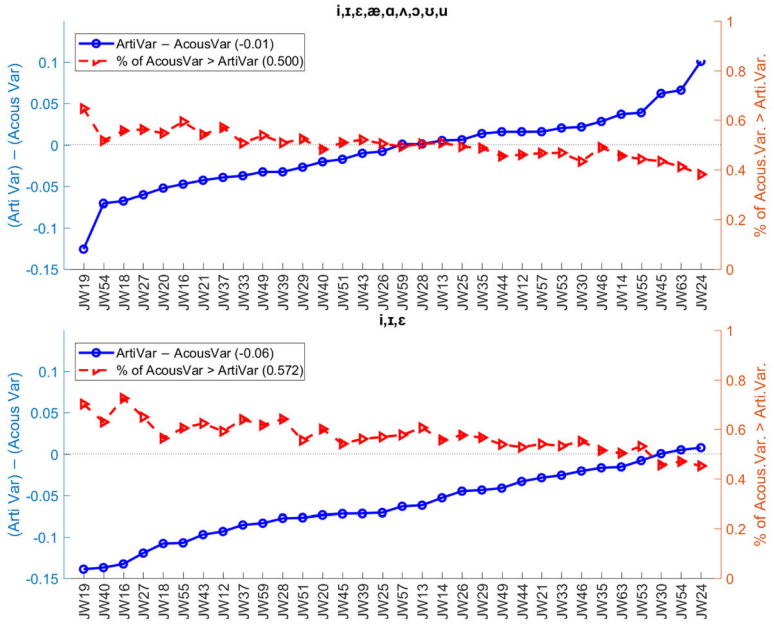

Fig. 5.

Comparisons of variabilities in acoustic and articulatory domains calculated separately for all nine vowels (upper panel) and for the non-low front vowels only. The blue circle line indicates the mean difference between articulatory and acoustic variabilities across vowels for each speaker (scale on the left y-axis; positive values indicate higher articulatory variability than acoustic variability); the red triangle line shows the proportion of tokens where the acoustic distance to the category median was greater than that distance to the articulatory median (scale on the right y-axis). [Color online.]

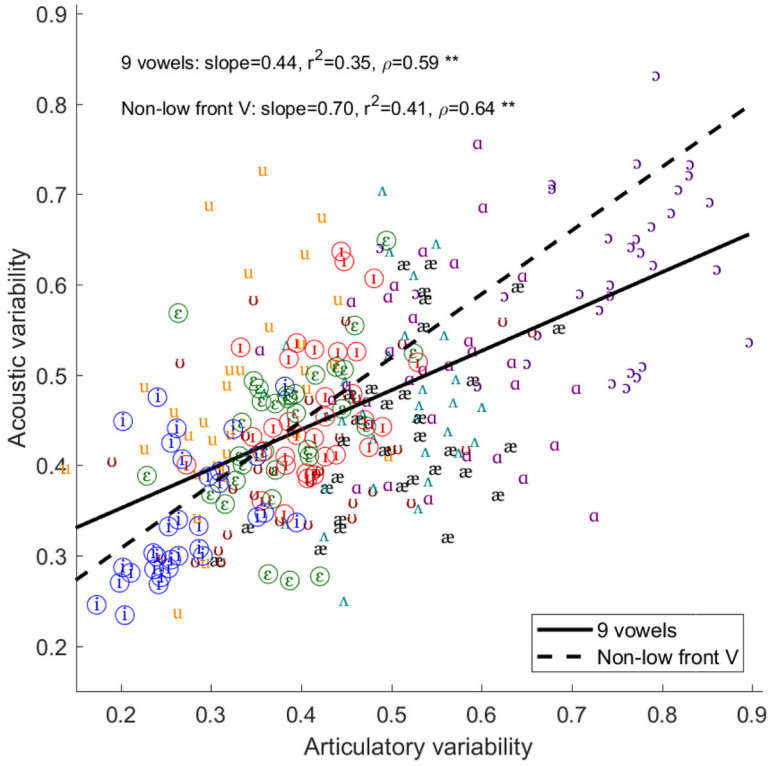

Fig. 6.

Scatter plot of acoustic variability against articulatory variability across speakers and vowels. Each dot represents the mean articulatory and acoustic variabilities of one vowel produced by one speaker. Non-low front vowels are circled. The solid line is the regression line drawn through all nine vowels, and the dashed line through the non-low front vowels only.

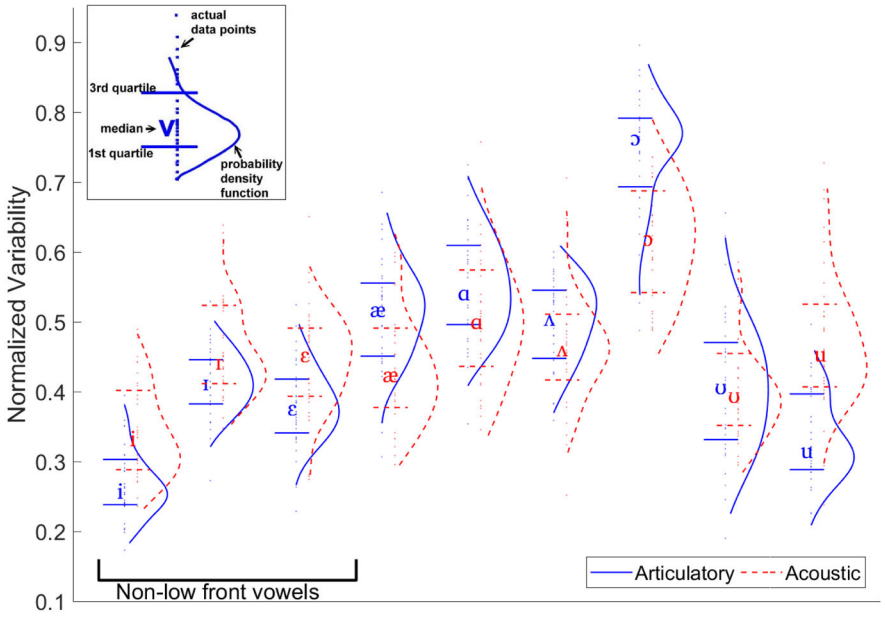

Fig. 7.

The distributions (across 32 speakers) of articulatory and acoustic variabilities for nine vowels. Blue solid lines indicate articulatory variabilities and red broken lines acoustic variabilities. Curved lines are the probability density functions for each distribution fitted by kernel density estimation. The vowel letter indicates its median value, the upper bar above the vowel letter the third quartile, and lower bar below the vowel letter the first quartile.

Table 1

Mean number of samples, number of excluded samples (outliers) and exclusion rate per speaker for each vowel.

Vowel	# samples	# of exclusion	proportion exclusion
/ɑ/	75.7	11.9	0.16
/æ/	119.2	19.3	0.16
/ʌ/	121.9	21.3	0.17
/ɔ/	121.4	24.9	0.21
/ɛ/	92.9	14.3	0.15
/ɪ/	142.8	23.1	0.17
/i/	125.1	23.3	0.19
/ʊ/	32.3	4.8	0.15
/u/	110.6	21.0	0.19
	Total: 30141 Avg: 104.7	Total: 5244 Avg: 18.2	 Avg: 0.17

Table 2

The fixed effect in the linear mixed-effects model. The baseline of the *domain* effect is 'Acoustic', and the baseline of *vowel* effect is /ɑ/.

	Model coefficients (β)	Estimated DOF	*t* value	*p* value	
(Intercept)	0.507	77.9	36.4	0.000	**
Domain Articulatory	0.045	372.0	2.7	0.007	**
Vowel /æ/	−0.064	74.3	−3.2	0.002	**
Vowel /ʌ/	−0.039	84.9	−2.1	0.043	**
Vowel /ɔ/	0.111	96.3	6.1	0.000	**
Vowel /ɛ/	−0.065	98.2	−3.5	0.001	**
Vowel /ɪ/	−0.044	182.3	−2.6	0.011	**
Vowel /i/	−0.163	149.5	−9.4	0.000	**
Vowel /ʊ/	−0.097	74.2	−4.8	0.000	**
Vowel /u/	−0.030	72.6	−1.5	0.135	
DomainArticu:Vowel /æ/	0.023	372.0	1.0	0.332	
DomainArticu:Vowel /ʌ/	−0.015	372.0	−0.6	0.524	
DomainArticu:Vowel /ɔ/	0.079	372.0	3.4	0.001	**
DomainArticu:Vowel /ɛ/	−0.107	372.0	−4.6	0.000	**
DomainArticu:Vowel /ɪ/	−0.096	372.0	−4.1	0.000	**
DomainArticu:Vowel /i/	−0.119	372.0	−5.1	0.000	**
DomainArticu:Vowel /ʊ/	−0.047	372.0	−2.0	0.046	**
DomainArticu:Vowel /u/	−0.193	372.0	−8.3	0.000	**

** : $p < .01$;

* : $p < .05$;

† : $p < .1$.

Table 3

Pearson correlation coefficients (ρ) for vowel variabilities in the articulatory domain. The indications of significance were based on the controlled FDR levels

	/ɪ/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɔ/	/ʊ/	/u/
/i/	0.29	0.24	−0.12	0.28	0.49†	0.17	0.31	0.62**
/ɪ/		0.58**	0.28	0.29	0.38†	0.13	−0.02	0.43†
/ɛ/			0.25	0.47†	0.37†	0.43†	0.05	0.36†
/æ/				−0.07	0.29	0.39†	−0.21	−0.29
/ɑ/					0.13	0.43†	0.38†	0.19
/ʌ/						0.39†	0.04	0.37†
//ɔ/							−0.07	−0.11
/ʊ/								0.30

**: $p \leq .0005$ (FDR $q < .01$);

†: p (unadjusted) < .05.

Table 4

Pearson correlation coefficients (ρ) for vowel variabilities in the acoustic domain The indications of significance were based on the controlled FDR levels

	/ɪ/	/ɐ/	/æ/	/ɑ/	/ʌ/	/ɔ/	/ʊ/	/u/
/i/	0.11	0.17	0.02	0.39†	−0.07	0.26	−0.08	0.16
/ɪ/		−0.01	0.07	0.02	−0.20	0.01	−0.27	−0.01
/ɐ/			0.40†	0.20	0.59*	0.41†	0.06	−0.09
/æ/				0.28	0.42†	0.37†	−0.18	0.26
/ɑ/					0.17	0.42†	0.08	0.03
/ʌ/						0.54*	0.02	0.02
/ɔ/							0.08	0.12
/ʊ/								0.20

*: $p \leq .0015$ (FDR $q < .05$);

†: p (unadjusted) $< .05$.

Table 5

Pearson correlation between articulatory and acoustic variabilities separately for each vowel. The indications of significance were based on the controlled FDR levels

	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɔ/	/ʊ/	/u/
ρ	0.47*	0.39*	0.36†	0.49*	−0.1	0.19	0.24	0.45*	0.43*
Unadjusted p	*.01*	*.03*	*.04*	*.01*	*.65*	*.29*	*.19*	*.01*	*.02*
FDR Adjusted q	*.03*	*.05*	*.06*	*.03*	*.65*	*.33*	*.25*	*.03*	*.03*

*: $p \leq .027$ (FDR $q < .05$);

†: p (unadjusted) < .05.

Table 6

Tables of the fixed effects in two linear mixed-effects models fitted separately to articulatory and acoustic subsets. The baseline for the *vowel* effect for both model is /æ/.

Model:	**Variability in articulatory domain**				
ArticuVar ~ Vowel+(1|Speaker)	**Model coefficients (*β*)**	**Estimated DOF**	***t* value**	***p* value**	
(Intercept)	0.51	214.0	37.7	0.000	**
Vowel /i/	−0.24	248.0	−14.0	0.000	**
Vowel /ɪ/	−0.10	248.0	−5.7	0.000	**
Vowel /ɛ/	−0.13	248.0	−7.6	0.000	**
Vowel /ɑ/	0.04	248.0	2.4	0.018	*
Vowel /ʌ/	−0.01	248.0	−0.8	0.439	
Vowel /ɔ/	0.23	248.0	13.4	0.000	**
Vowel /ʊ/	−0.10	248.0	−5.9	0.000	**
Vowel /u/	−0.18	248.0	−10.6	0.000	**
Model:	Variability in acoustic domain				
AcoustVar ~ Vowel+(1|Speaker)	Model coefficients (*β*)	Estimated DOF	*t* value	*p* value	
(Intercept)	0.44	239.5	28.0	0.000	**
Vowel /i/	−0.10	248.0	−4.8	0.000	**
Vowel /ɪ/	0.02	248.0	1.0	0.341	
Vowel /ɛ/	0.00	248.0	−0.1	0.960	
Vowel /ɑ/	0.06	248.0	3.1	0.002	**
Vowel /ʌ/	0.02	248.0	1.2	0.245	
Vowel /ɔ/	0.17	248.0	8.4	0.000	**
Vowel /ʊ/	−0.03	248.0	−1.6	0.112	
Vowel /u/	0.03	248.0	1.6	0.106	

** : $p < .01$;

* : $p < .05$.

Table 7

Comparisons of explained variance (R^2) between fixed and random effects for the models fitted separately to articulatory (first row) and acoustic (second row) variabilities. R^2-fixed indicates the amount of variance explained by fixed effects, and R^2-random that explained by random effects.

Model	# obs.	R^2-fixed	R^2-random	AIC
ArticuVar ~ Vowel+(1|Speaker)	288	**79.6%**	4.9%	−674
AcoustVar ~ Vowel+(1|Speaker)	288	**46.8%**	8.9%	−575