

Articulatory Distinctiveness of Vowels and Consonants: A Data-Driven Approach

Jun Wang,^{1, 2} Jordan R. Green,^{1, 2} Ashok Samal,¹ and Yana Yunusova³

1. University of Nebraska–Lincoln
2. Munroe–Meyer Institute, University of Nebraska Medical Center, Omaha
3. University of Toronto, Toronto, Ontario, Canada

Corresponding author — Jun Wang, who is now at Callier Center for Communication Disorders,
University of Texas at Dallas: wangjun@utdallas.edu

Jordan R. Green is now at MGH Institute of Health Professions, Boston, MA

Abstract

Purpose: To quantify the articulatory distinctiveness of 8 major English vowels and 11 English consonants based on tongue and lip movement time series data using a data-driven approach.

Method: Tongue and lip movements of 8 vowels and 11 consonants from 10 healthy talkers were collected. First, classification accuracies were obtained using 2 complementary approaches: (a) Procrustes analysis and (b) a support vector machine. Procrustes distance was then used to measure the articulatory distinctiveness among vowels and consonants. Finally, the distance (distinctiveness) matrices of different vowel pairs and consonant pairs were used to derive articulatory vowel and consonant spaces using multidimensional scaling.

Results: Vowel classification accuracies of 91.67% and 89.05% and consonant classification accuracies of 91.37% and 88.94% were obtained using Procrustes analysis and a support vector machine, respectively. Articulatory vowel and consonant spaces were derived based on the pairwise Procrustes distances.

Conclusions: The articulatory vowel space derived in this study resembled the long-standing descriptive articulatory vowel space defined by tongue height and advancement. The articulatory consonant space was consistent with feature-based classification of English consonants. The derived articulatory vowel and consonant spaces may have clinical implications, including serving as an objective measure of the severity of articulatory impairment.

Keywords: speech production, articulatory vowel space, articulatory consonant space, Procrustes analysis, support vector machine

Intelligible speech is characterized by the ability to produce discernible distinctions between sounds. The acoustic distinctiveness of vowels and consonants has been studied extensively by investigators from a variety of fields, including computer science (i.e., automatic speech recognition), psycholinguistics, neuroscience, and communication sciences and disorders. These studies have been motivated by the need to understand not only the phonetic basis of sounds (Stevens & Klatt, 1974) but also how neuronal processing (e.g., Mitchell et al., 2008), auditory perception (e.g., Johnson, 2000), and speaking rate change as a function of speaking task difficulty (e.g., Tsao & Iqbal, 2005), speaking environment (e.g., noise), and talker characteristics (e.g., age, health; Lindblom, 1990). One commonly used measure of distinctiveness among vowels is the acoustic vowel space area, which is defined by the first and second vowel formants. This measure has been used extensively to investigate declines in speech intelligibility (Kim, Hasegawa-Johnson, & Perlman, 2011; Neel, 2008; Turner, Tjaden, & Weismer, 1995; Weismer, Jeng, Laures,

Kent, & Kent, 2001), articulation rate (Zajac et al., 2006), developmental changes in speech (e.g., Lee, Potamianos, & Narayanan, 1999; Rvachew, Mattock, Polka, & Ménard, 2006), and exaggerated speech directed to infants (Green & Nip, 2010; Green, Nip, Mefferd, Wilson, & Yunusova, 2010; Kuhl et al., 1997; Kuhl & Meltzoff, 1997).

In comparison to acoustic-based measures of phoneme distinctiveness, articulatory-based measures have received little attention because of the logistical difficulty of obtaining articulatory data. Yet articulatory measures have many important clinical and scientific implications, including quantifying the degree of articulatory impairment in persons with speech disorders by articulatory information (rather than by acoustic information), advancing knowledge about articulatory-to-acoustic relations (Mefferd & Green, 2010), and enhancing phoneme recognition accuracy for speech recognition in noisy environments (King et al., 2007; Livescu et al., 2007) and in disordered speech (Rudzicz, 2011), as well as for silent speech recognition from articulatory movements (Denby et al., 2010; Wang,

2011). Moreover, some research has indicated that articulatory control and coordination may not manifest in speech acoustics. For example, the spatiotemporal variations in tongue movement time series are not apparent in associated formant time series (Mefferd & Green, 2010). The development of articulatory-based measures is particularly needed for identifying changes in articulatory control that occur during normal development, treatment, or disease (Wang, Green, Samal, & Marx, 2011).

To date, the articulatory distinctiveness of different phonemes has predominantly been based on the classification of their presumed distinctive articulatory features, such as lip rounding, lip opening, lip height, lip contour, and lip area (Potamianos, Neti, Gravier, Garg, & Senior, 2003; Sadeghi & Yaghmaie, 2006; Shinchi, 1998); tongue tip and tongue body height (Richardson, Bilmes, & Diorio, 2000); lip opening and lip rounding (Richardson et al., 2000; Saenko, Livescu, Glass, & Darrell, 2009); lip width and lip area (Heracleous, Aboutabit, & Beautemps, 2009; Visser, Poel, & Nijholt, 1999); maximum displacement (Yunusova, Weismer, & Lindstrom, 2011); and vocal tract shape geometry (Fuchs, Winkler, & Perrier, 2008; Honda, Maeda, Hashi, Dembowski, & Westbury, 1996). Most of these classification approaches for articulatory data (without using acoustic data) have resulted in only poor to moderate classification accuracy; only a few achieved accuracy of 80% (Yunusova et al., 2011). Two significant limitations of the feature-based approaches are that (a) classification is dependent on accurate feature identification and (b) the approaches assume there are isomorphic, simple mappings between chosen features and phonemes. These approaches are also limited, because they have typically relied on articulatory features, which do not account for time-varying motion pattern information. More direct approaches, such as the one we used in this study, whereby articulatory movement time series are mapped directly to phonemes, may overcome these limitations.

The goal of this project was to provide a better understanding of the articulatory distinctiveness of phonemes, which has been a long-standing empirical challenge—one that required the development of a novel analytic technique for quantifying the subtle across-phoneme differences in articulatory movements. Specifically, we evaluated the accuracy of a direct-mapping approach for classifying and quantifying the articulatory distinctiveness of vowels and consonants based on articulatory movement time series data rather than articulatory features. Classification accuracies using statistical shape analysis (Procrustes analysis) and machine learning (a support vector machine [SVM]) on articulatory movements were obtained as a measure of how well the set of vowels and consonants can be distinguished on the basis of articulatory movements. Procrustes distance was then used to quantify the articulatory distinctiveness of vowel and consonant pairs. Finally, the quantified articulatory distinctiveness of vowels and consonants was used to derive both an articulatory vowel space (an articulatory parallel to acoustic vowel space) and an articulatory consonant space.

Method

Participants

Ten monolingual women, native speakers of English, participated in this study. The average age of the participants was 23.60 years ($SD = 9.48$, range: 19–50). No participant reported hearing and speech problems or a prior history of hearing or speech impairments. They were all from the midwestern region of the United States.

Stimuli

Eight major English vowels in symmetrical consonant-vowel-consonant (CVC) syllables—/bab/, /bib/, /beb/, /bæb/, /bʌb/, /bɔb/, /bob/, /bub/—were used as vowel stimuli. The eight vowels are representative of the English vowel inventory and were chosen because they sufficiently circumscribe the boundaries of the descriptive articulatory vowel space (Ladefoged & Johnson, 2011). Therefore, these vowels provide a good representation of the variety of tongue and lip movement patterns. The consonant context was held constant across stimuli to minimize the influence of consonant coarticulation effects on vowel identity. The context /b/, a bilabial, was selected because it had minimum coarticulation effect on the vowels, compared with other consonants, such as /k/ and /t/ (Lindblom & Sussman, 2012).

Eleven consonants in symmetrical vowel-consonant-vowel (VCV) syllables (i.e., /aba/, /aga/, /awa/, /ava/, /ada/, /aza/, /ala/, /ara/, /aʒa/, /adʒa/, /aja/) were used as consonant stimuli. These consonants were selected because they represent the primary places and manners of articulation of English consonants. Consonants were embedded into the /a/ context because this vowel is known to induce larger tongue movements than other vowels (Yunusova, Weismer, Westbury, & Lindstrom, 2008).

Speech Tasks

All stimuli were presented on a large computer screen in front of the participants, and prerecorded sounds were played to help the participants to pronounce the stimuli correctly. Participants were asked to repeat what they heard and put stress on the middle phoneme (rather than the carriers) for each stimulus. Participants were asked to rest (about 0.5 s) between each CVC or VCV production to minimize the coarticulation effect. This rest interval also facilitated segmenting the stimuli prior to analysis. The stimuli were presented in a fixed order (as listed in the Stimuli section) across participants. The stimuli were not presented in a random order, because it draws too much of the participants' attention. Mispronunciations were rare but were identified by the investigator and excluded from the data analysis.

Each phoneme sequence was repeated multiple times by each participant. On average, 20.9 valid vowel samples were collected from each participant, with the number of samples for each vowel varying from 16 to 24 per participant. In total, 1,672 vowel samples with 209 samples for

each vowel were obtained and used for analysis. The average number of valid consonant samples collected from each participant was 19.4, varying from 12 to 24 per participant. In total, 2134 consonant samples (with 194 samples for each consonant) were collected and used for analysis in this experiment.

Data Collection

The Electromagnetic Articulograph (EMA; Model AG500; Carstens Medizintechnik, Inc.) was used to register three-dimensional (3D) movements of the tongue, lip, and jaw during speech. The spatial accuracy of motion tracking using EMA was 0.5 mm (Yunusova, Green, & Mefferd, 2009). EMA registers movements by establishing a calibrated electromagnetic field in a volume that can be used to track the movements of small sensors within the volume. The center of the magnetic field is the origin (zero point) of the EMA coordinate system.

Participants were seated with their head within the calibrated magnetic field. The sensors were attached to the surface of each articulator using dental glue (PeriAcryl Oral Tissue Adhesive). The participants were then asked to produce the vowel and consonant sequences at their habitually comfortable speaking rate and loudness.

Figure 1 shows the placement of the 12 sensors attached to a participant's head, face, and tongue. Three of the sensors were attached to a pair of glasses. The Head Center sensor was on the bridge of the glasses, and the Head Left and Head Right sensors were on the left and right outside edge of each lens, respectively. We used the movements of the Head Center, Head Left, and Head Right sensors to calculate the movements of other articulators independent of the head (Green, Wilson, Wang, & Moore, 2007). Lip movements were captured by attaching two sensors to the vermilion borders of the upper (UL) and lower (LL) lips at midline. Four sensors—T1 (Tongue Tip), T2 (Tongue Blade), T3 (Tongue Body Front), and T4 (Tongue Body Back)—were attached approximately 10 mm from each other at the midline of the tongue (Wang et al., 2011). The movements of three jaw sensors—Jaw Left, Jaw Right, and Jaw Center—were recorded but not analyzed in this study.

Data Preprocessing

Before conducting the analysis, we subtracted the translation and rotation components of head movement from the tongue and lip movements. The resulting head-independent tongue and lower lip sensor positions included movement from the jaw. The orientation of the derived 3D Cartesian coordinate system is displayed in Figure 1. Because the movements for the simple vowels and consonants contain only very low frequency components, a low-pass filter of 10 Hz was applied to the movement traces prior to the analysis (Green & Wang, 2003).

Acoustic signals were recorded simultaneously with kinematic signals directly onto a hard drive of a computer at the sampling rate of 16 kHz with 16-bit resolution. A high-quality lapel microphone (Crown head-worn microphone CM311) was mounted on the forehead approximately 15 cm from the mouth during the recordings. Acoustic recordings

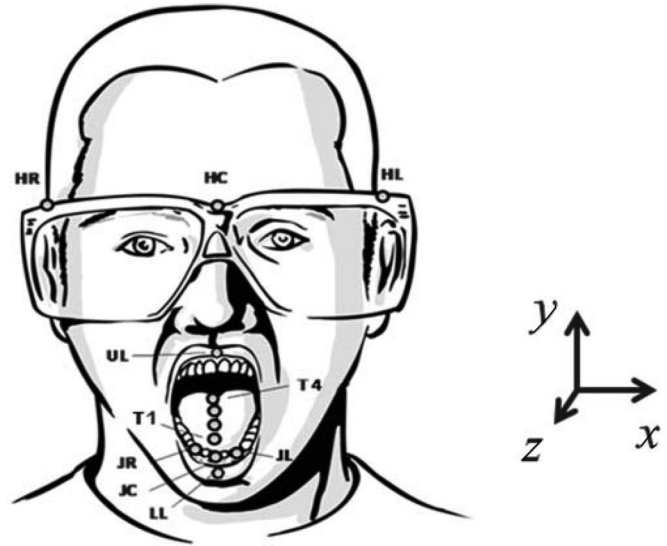


Figure 1. Sensor positions. HR = Head Right; HC = Head Center; HL = Head Left; UL = Upper Lips; T4 = Tongue Body Back; T1 = Tongue Tip; JR = Jaw Right; JL = Jaw Left; JC = Jaw Center; LL = Lower Lips. From "Articulatory-to-Acoustic Relations in Response to Speaking Rate and Loudness Manipulations," by A. Mefferd and J. G. Green, 2010, *Journal of Speech, Language, and Hearing Research*, 53, p. 1209, Rockville, MD: American Speech-Language-Hearing Association. Copyright 2010 by the American Speech-Language-Hearing Association. Adapted with permission.

were used for segmenting articulatory movement data and for extracting F1 and F2 formant values. First, sequences of movements were aligned with acoustic waveforms. Then the onset and offset of the whole CVC and VCV utterances were identified visually on the basis of acoustic waveform data using a customized MATLAB software program. All manual segmentation results were double checked by the investigator. On occasion, erroneous samples were collected because of a sensor falling off during recording or sounds that were not produced correctly. These erroneous samples were excluded in the analysis.

Only y (vertical) and z (anterior-posterior) coordinates of the sensors (i.e., UL, LL, T1, T2, T3, and T4) were used for analysis because the movement along the x (lateral) axis is not significant during speech of healthy talkers (Westbury, 1994).

Analysis

Three analyses were conducted: (a) classification using both Procrustes analysis (Dryden & Mardia, 1998) and SVM (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995), (b) quantifying the articulatory distinctiveness of vowels and consonants using Procrustes distance, and (c) deriving articulatory vowel and consonant space from the distance (distinctiveness) matrices obtained in using multi-dimensional scaling (MDS; Cox & Cox, 1994).

Procrustes analysis. Procrustes analysis is a robust shape analysis technique (Sibson, 1978) that has been successfully applied for object recognition and shape classification (Jin & Mokhtarian, 2005; Meyer, Gustafson, & Arnold, 2002; Sujith & Ramanan, 2005). In Procrustes analysis, a shape is represented by a set of ordered landmarks on the surface of an object. Procrustes distance is calculated as the summed Euclidean distances between the corresponding landmarks of two shapes after the locational, rotational, and scaling effects are removed from the two shapes (called *Procrustes matching*; see Dryden & Mardia, 1998). A step-by-step calculation of Procrustes distance between two shapes includes first aligning the two shapes using their centroids; then scaling both shapes to a unit size; and, last, rotating one shape to match the other and obtaining the minimum sum of the Euclidean distances between their corresponding landmarks (Wang et al., 2011).

In this experiment, we used an equivalent but faster method for calculating the Procrustes distance using a complex number representation for the landmark coordinates. Suppose u and v are two centered shapes represented by two sets of complex numbers. Real and imaginary parts of a complex number represent the two coordinates (y and z of sensor locations) of a landmark. The Procrustes distance d_p between u and v is denoted by Equation 1, where u^* denotes the complex conjugate transpose of u . Proof of Equation 1 was given by Dryden and Mardia (1998):

$$d_p(u, v) = \left\{ 1 - \frac{v^* u u^* v}{u^* u v^* v} \right\}^{1/2} \quad (1)$$

Procrustes analysis was designed for analysis of static shapes (i.e., shapes do not deform over time). However, a simple strategy was used to extend Procrustes analysis to time-varying shape analysis. In this study, shapes for phonemes were defined by their sampled motion paths of articulators. First, motion path trajectories (i.e., y and z coordinates) of each articulator were down-sampled to 10 locations spread evenly across time. The predominant frequency of tongue and lip movements is about 2 to 3 Hz for simple CVC utterances (Green & Wang, 2003); thus, 10 samples adequately preserve the motion patterns. Then, the sampled motion paths of all articulators were spatially integrated as a composite shape representing each phoneme. The composite shape, an integration of 10 locations from each of the six sensors, was used to represent a phoneme shape. Thus, in Equation 1, u is a 1×60 matrix of complex numbers; u^* is a 60×1 matrix of the complex conjugates; and the result, dp , is a real number within the range between 0 and 1. Jin and Mokhtarian (2005) used a similar strategy of spatially integrating shapes at different time points for recognition of human motion represented using images. Panel A of Figure 2 gives an example of continuous articulatory movements of /bab/; Panel B of Figure 2 illustrates the corresponding shape in which the 60 circles represent 60 landmarks (10 locations \times 6 sensors) of the movements of six sensors sampled to 10 time points.

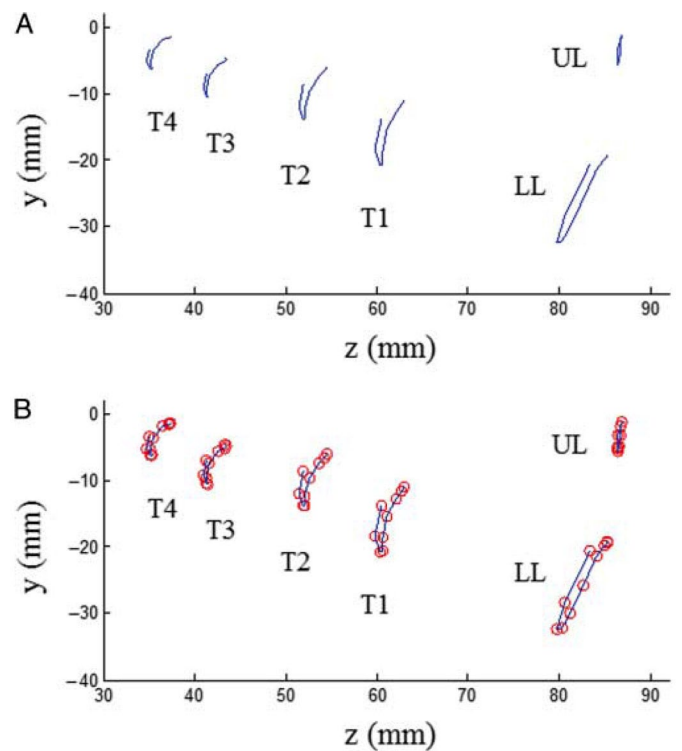


Figure 2. Panel A: continuous articulatory movements of /bab/ produced by a single participant. Panel B: the sampled articulatory movements that form a shape of /bab/ (landmarks are represented by red circles). T2 = Tongue Blade; T3 = Tongue Body Front.

We performed the following three steps, similar to the generalized Procrustes analysis (Gower, 1975), to classify the composite shapes of vowels and consonants for each participant. First, we calculated the average shapes of all samples for each phoneme and used them as references for the phoneme. The average shape of a phoneme is the averaged coordinates of corresponding landmarks of all samples for the phoneme. Second, for each test sample (shape), we calculated the Procrustes distances between it and all the average shapes. Third and last, we considered as the recognized phoneme the one that had the shortest distance between its average shape and the testing sample.

Classification accuracy is defined as number of correctly recognized phoneme samples divided by the total number of samples. We used a classification matrix to show how many of the samples from each vowel or consonant were classified into another vowel or consonant. In a classification matrix, a number at row i and column j in the matrix is the percentage of samples of i th phoneme that were classified as j th phoneme. The classification matrix for a perfect classifier would have 100% along the diagonal and 0% for all the nondiagonal entries.

Then, we calculated Procrustes distances between the average shapes of phoneme pairs and used them as a measure of distinctiveness between the pairs. Two distance (distinctiveness) matrices (for vowels and consonants, respectively) were obtained from a data set from each partic-

ipant. The average distance matrices of all participants defined the quantified articulatory distinctiveness of vowels and consonants (Wang et al., 2011).

SVM. We used a machine learning classifier (i.e., SVM) to provide information on classification accuracy in addition to that gained through Procrustes analysis. We selected SVM rather than other classifiers because our prior work showed that SVM outperformed other approaches, such as neural networks and decision trees, for this application (Wang, Samal, Green, & Carrell, 2009).

In machine learning, a classifier (computational model) predicts classes (or groups, categories) of new data samples on the basis of a training data set, in which the classes are known. In this classification method, a data sample is defined by an array of values (attributes). A classifier makes predictions regarding data classes by analyzing these attributes. The accuracy of the prediction is quantified on the basis of pattern consistency in the data and the classifier's success. SVM is a classifier that tries to maximize the distances between the boundaries of different classes in order to obtain the best generalization of patterns from training data to testing data. SVM classifiers project training data into a higher dimensional space and then separate classes using a linear separator (Boser et al., 1992; Cortes & Vapnik, 1995). The linear separator maximizes the margin between groups of training data through an optimization procedure (Chang & Lin, 2011). A kernel function is used to describe the distance between two samples (i.e., r and s in Equation 2, below). The following radial basis function was used as the kernel function K_{RBF} in this study, where λ is an empirical parameter (Wang, Samal, Green, & Rudzicz, 2012a, 2012b):

$$K_{RBF}(r, s) = \exp(1 - \lambda \|r - s\|) \quad (2)$$

For more details, please refer to Chang and Lin's (2011) article, which describes the implementation of SVM used in this study.

In this study, a sample (e.g., r or s in Equation 2) is a concatenation of time-sampled motion paths of articulators as data attributes. The movement data of each stimulus (a vowel or consonant) initially were time-normalized and sampled to a fixed length (i.e., 10 frames). The length was fixed, because SVM requires the input samples to be in a fixed-width array. The arrays of y or z coordinates for each articulator subsequently were demeaned and concatenated into one sample for each vowel or consonant. Appendix A illustrates how a sample was organized, where $U_{L_y|r}$ one of the attributes, specifies the y coordinate of UL at (normalized) Time Point 1. Overall, each sample contained 120 (6 articulators \times 2 dimensions \times 10 frames) numbers of attributes. An additional integer (e.g., 1 for /a/, and 2 for /i/) was used for labeling the training data (see Appendix A).

We used cross-validation, a standard procedure to test classification algorithms in machine learning, to evaluate the accuracy of articulatory movement classification using SVM. Training data and testing data are unique in cross-

validation. In this study, Leave- N -out cross-validation was conducted, in which N (= 8 or 11) is the number of vowels or consonants, respectively. In each execution, one sample for each stimulus (totally N samples) in the data set was selected for testing and the rest were used for training. There was a total of m executions, in which m is the number of samples per phoneme. The average classification accuracy of all m executions was considered the overall classification accuracy (Wang, 2011).

MDS. We used MDS (Cox & Cox, 1994) to derive articulatory vowel and consonant spaces based on the distinctiveness matrices of vowels and consonants. MDS is widely used to visualize high-dimensional data in a lower dimensional space. Given a set of items and their pairwise distances (in a symmetric distance matrix), MDS can generate the locations of the points in a coordinate system in which the distance relationships between the items are preserved. The orientation of the space is random and hence does not hold any physical significance. Green and Wang (2003) also used MDS to generate a consonant space based on pairwise covariance of movements of pellets attached on the midsagittal line of tongue (also named T1, T2, T3, and T4) tracked using x-ray microbeam.

In our use of MDS, the number of dimensions was specified with the input data (i.e., dissimilarity matrix), and then MDS output optimized results in the given number of dimensions. Given an input dissimilarity matrix of phonemes (diagonal numbers are zeros), MDS assigns a location to each phoneme in an N -dimensional space, where N is prespecified by the user; that is, if $N = 2$, MDS will visualize the data in a two-dimensional (2D) space; if $N = 3$, MDS will visualize the data in a 3D space. In this study, the distance matrices between the phonemes were used as dissimilarity matrices. The implementation of MDS in MATLAB was used in this analysis. The effectiveness of an MDS outcome can be evaluated by an R^2 value resulting from a linear regression between the distance matrix obtained from the MDS outcome and the original distance matrix. R^2 (between 0 and 1) indicates the similarity between the two distance matrices. A larger R^2 value indicates a better fit between the MDS outcome and the original distance matrix.

Results

Classification Accuracy of Vowels

The average classification accuracies of vowels computed across individual speakers were 91.67% ($SD = 5.34$) and 89.05% ($SD = 11.11$) using Procrustes analysis and SVM, respectively. We applied a two-tailed t test on the classification accuracies using the two approaches for each participant. The t test result showed that there was no significant difference ($p < .26$) between the accuracies obtained using Procrustes analysis and SVM, which means Procrustes analysis has power similar to a widely used classifier (i.e., SVM) in vowel classification.

Table 1. Average vowel classification matrix (in percentage) of all participants using Procrustes analysis.

Actual	Classified							
	/a/	/i/	/e/	/æ/	/ʌ/	/ɔ/	/o/	/u/
/a/	90.53			0.43	3.52	5.08	0.43	
/i/		98.24	0.91	0.43				0.42
/e/		4.19	94.29		0.63			0.89
/æ/	3.01		1.25	92.48	2.20	1.06		
/ʌ/	1.97			1.48	89.47	5.57	1.51	
/ɔ/	4.76			0.43	7.89	81.09	5.39	0.43
/o/	1.06				5.07	2.93	88.34	2.59
/u/					0.63		0.43	98.94

Zeros are not displayed. Diagonal numbers are in bold.

The average classification matrices (in percentage) of all participants, using Procrustes analysis and SVM, are shown in Tables 1 and 2.

Articulatory Distinctiveness of Vowels

The average distance matrix (articulatory distinctiveness), computed across all participants, is shown in Table 3. A larger distance between a vowel pair indicates that they are more articulatory distinct. For example, the distances between /a/ and /i/ and that between /a/ and /u/ (.2506 and .2024, respectively) are the largest, suggesting that these vowels are the most articulatory distinct; the distances among /ʌ/, /ɔ/, and /u/ are the shortest, suggesting that these vowels are least articulatory distinct.

Quantitative Articulatory Vowel Space

The symmetric distance matrix shown in Table 3 was used as a dissimilarity matrix for generating a vowel space using MDS. Panel A of Figure 3 shows the derived 2D quantitative articulatory vowel space. As explained previously, in this derived space the two coordinates are the two optimized dimensions of an MDS solution. Pairwise distances obtained from the derived space accounts for a large amount of the variance in the original distances, as indicated by a regression that yielded a very high R^2 value, .98. MDS can also generate a 3D space (not shown in this article). However, the third dimension did not contribute significantly to the vowel distinctiveness (R^2 also = .98).

Acoustic Vowel Space

The first and second formants (F1 and F2) of the same eight major English vowels obtained from the synchronously collected acoustic data were used to derive an acoustic vowel space (see Panel C, Figure 3). The vowel formant values obtained in this study were consistent with those in literature (e.g., Bunton & Story, 2010; Neel, 2008; Rosner & Pickering, 1994; Tsao & Iqbal, 2005; Turner, Tjaden, & Weismer, 1995). Possible slight variation between the formants in this study and those in literature may be due to the dialect or accent effects. As

Table 2. Average vowel classification matrix (in percentage) of all participants using a support vector machine.

Actual	Classified							
	/a/	/i/	/e/	/æ/	/ʌ/	/ɔ/	/o/	/u/
/a/	89.03			1.74	1.25	6.73	0.63	0.63
/i/		95.57	4.01					0.42
/e/		2.35	97.03					0.63
/æ/	2.80		0.63	92.75	1.08	2.75	0.00	
/ʌ/	4.32		0.43	0.43	80.36	12.31	1.51	0.63
/ɔ/	6.79			1.04	9.29	75.42	6.83	0.63
/o/	0.63			0.43	4.18	5.93	85.71	3.13
/u/	0.63			1.29		0.45	1.06	96.57

Zeros are not displayed. Diagonal numbers are in bold.

mentioned previously, all our participants are from the midwestern United States. The formant values in Panel C of Figure 3 are provided in Appendix B.

Classification Accuracy of Consonants

The across-talker average accuracies of consonant classification were 91.37% ($SD = 4.04$) and 88.94% ($SD = 6.07$) using Procrustes analysis and SVM, respectively. A one-tailed t test showed that the accuracy obtained using Procrustes analysis was significantly higher than that obtained using SVM ($p < .01$). The average classification matrices using Procrustes analysis and SVM, respectively, are shown in Tables 4 and 5.

Articulatory Distinctiveness of Consonants

The average distance (articulatory distinctiveness) matrix for consonant pairs computed across all participants is shown in Table 6. A larger distance between a consonant pair indicates they are more articulatory distinct. The distance between /b/ and /j/ (.2586) was the largest, representing the greatest articulatory contrast between any two consonants. The distance between /z/ and /dʒ/ was the shortest distance (.0641), representing the least amount of articulatory distinctiveness among any two consonants.

Articulatory Consonant Space

We used the distance matrix shown in Table 6 as a dissimilarity matrix for generating an articulatory consonant space using MDS. Panel A of Figure 4 gives the derived 2D articulatory consonant space. Similar to the derived vowels space, the two coordinates in the consonant space are the two optimized dimensions in an MDS solution, which contributed most to the distinctiveness of consonants. An R^2 value of .94 was obtained in a regression between the pairwise distances obtained from the derived space (see Panel A, Figure 4) and the original distance matrix (see Table 6). A 3D articulatory consonant space was also generated using MDS (see Panel B, Figure 4). Pairwise distances between consonants obtained from the 3D space yielded an R^2 value of .98.

Table 3. Average articulatory distinctiveness between vowel pairs across participants.

Vowel	/a/	/i/	/e/	/æ/	/ʌ/	/ɔ/	/o/	/u/
/a/	—	0.2506	0.1960	0.1265	0.1087	0.0891	0.1358	0.2024
/i/	0.2506	—	0.1042	0.1940	0.1911	0.2339	0.2089	0.1461
/e/	0.1960	0.1042	—	0.1411	0.1504	0.1858	0.1681	0.1406
/æ/	0.1265	0.1940	0.1411	—	0.1227	0.1248	0.1521	0.1817
/ʌ/	0.1087	0.1911	0.1504	0.1227	—	0.0739	0.0814	0.1255
/ɔ/	0.0891	0.2339	0.1858	0.1248	0.0739	—	0.0999	0.1636
/o/	0.1358	0.2089	0.1681	0.1521	0.0814	0.0999	—	0.1028
/u/	0.2024	0.1461	0.1406	0.1817	0.1255	0.1636	0.1028	—

Discussion

High classification accuracies obtained using Procrustes analysis for both vowels and consonants (similarly high as those obtained using SVM, a widely used classifier) indicate that Procrustes analysis is well suited for this articulation analysis. The articulatory distinctiveness of eight English vowels and 11 consonants were then quantified using Procrustes analysis on sparsely sampled lip and tongue movements represented as time series. The dissimilarity matrices for vowels and consonants, when visualized using MDS, were consistent with descriptive schemes that are commonly used to distinguish phonemes based on their unique features (Ladefoged & Johnson, 2011). The scientific and clinical implication of the derived articulatory vowel and consonant spaces are also discussed below, as are limitations of our approaches.

Classification of Vowels and Consonants

Articulatory position time-series data from multiple articulators were directly mapped to vowels and consonants. This approach differs from prior efforts to classify phonemes from articulatory information, which have primarily been based on extracted articulatory features. The use of statistical shape analysis (i.e., Procrustes analysis) to quantify the differences among phonemes in their articulatory movements also is novel.

The results of this study indicate that both methods (i.e., Procrustes analysis and SVM) were able to classify vowels and consonants accurately and consistently across talkers. The data presented in the classification matrices (see Tables 1 and 2) and the distance matrix (see Table 3) for vowels indicated that /i/, /e/, /æ/, and /u/ were easier to distinguish than were /ʌ/, /ɔ/, /o/, and /a/. This result supports the previous findings that low tongue vowels (e.g., /a/) have more articulatory variation than high tongue vowels (e.g., /i/ and /u/; see Perkell & Cohen, 1989; Wang, Green, Samal, & Carrell, 2010). More specifically, our results suggest that high and front vowels (i.e., /i/, /e/, /æ/, and /u/) are more articulatory distinct than low and back vowels (i.e., /ʌ/, /ɔ/, /o/, and /a/). Neel (2008) found that high vowels tend to be more acoustically distinct than low vowels based on the first and second formants of 10 representative vowels. Our findings then suggest that

more acoustically distinct vowels are also articulated more distinctly, which also agreed with a previous finding in a study on formants and tongue tip locations of two vowels /a/ and /i/ (Mefferd & Green, 2010).

The classification matrices (see Tables 4 and 5) and distance matrix (see Table 6) for consonants using both approaches indicated that errors occurred most frequently between /r/, /z/, /dʒ/, and /j/; this result might be because these sounds are produced with a similar, but not identical, place of lingual articulation.

The high classification accuracies obtained in this study motivates further inquiry into the usefulness of classification for a variety of applications. For example, additional research is required to determine whether classification accuracy is a sensitive metric for quantifying the severity of speech impairment or the articulatory changes that occur under different speaking conditions (Mefferd & Green, 2010). In addition, further work is planned to determine whether the classification approaches are suitable as the recognition engine for silent speech interfaces (Denby et al., 2010; Fagan, Ell, Gilbert, Sarrazin, & Chapman, 2008; Hueber et al., 2010; Wang et al., 2010; Wang et al., 2012a, 2012b) to facilitate oral communication in persons with moderate to severe speech or voice impairments. Finally, although only female talkers were investigated in this study, we anticipate that the classification of male talkers' vowels and consonants would produce similar results.

Quantified Articulatory Vowel and Consonant Spaces

Although the quantitative articulatory vowel space (see Panel A, Figure 3) was remarkably consistent with existing qualitative depictions of articulatory vowel space (Panel B, Figure 3), the /u/ appeared to be closer to the /i/ in the quantitatively derived articulatory vowel space than in the descriptive articulatory vowel space (Panel B, Figure 3). This finding might be interpreted to suggest that, compared to the /u/, the other back vowels are produced with a more posterior tongue posture. Another explanation, however, may be that the backing feature of /u/ was not adequately captured because our most posterior sensor was only on the back of the tongue body and not on the root.

The articulatory vowel space (see Panel A, Figure 3) was also strikingly similar to the acoustic vowel space

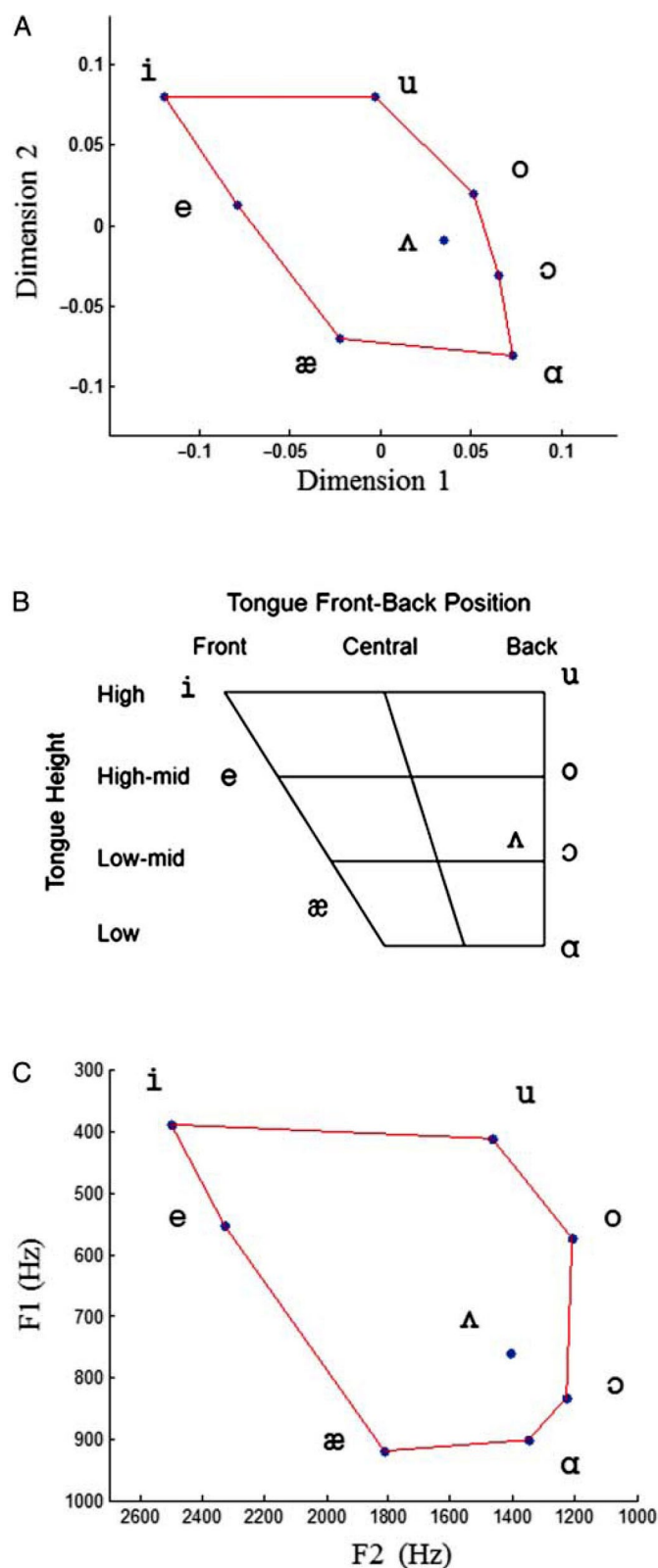


Figure 3. Quantified (Panel A) and descriptive (Panel B) articulatory vowel spaces, and (Panel C) acoustic vowel space including eight major English vowels. Dimensions in Panel A are the results of the multidimensional scaling solution. See Appendix B for the formant values in Panel C.

obtained from the same participants (Panel C, Figure 3). These similarities suggest that, despite the extensive processing of the articulatory movement data, the distinguishing aspects of vowel articulation were preserved in vowel acoustic output.

The 2D articulatory consonant space (see Panel A, Figure 4) clustered consonants on the basis of place of articulation along Dimension 1. For example, bilabial sounds (i.e., /b/ and /w/), alveolar sounds (i.e., /l/, /z/, and /d/), and postalveolar sounds (i.e., /ʒ/ and /j/) were grouped from left to right along Dimension 1. The 3D articulatory consonant space (see Panel B, Figure 4) clustered the consonants on the basis of the place of articulation as well. For example, alveolar sounds (i.e., /l/, /z/, and /d/), postalveolar sounds (i.e., /ʒ/), and bilabial sounds (i.e., /b/, and /w/), were grouped by place of articulation. On the basis of the data clusters, the manner of articulation did not appear to be represented in either the 2D or 3D space. Future efforts that encode differences among consonants in their duration may provide a basis for improving the detection of manner differences; duration information was not preserved in our kinematic signals because the articulatory movements were time normalized to the same length prior to classification. In addition, we could not determine whether the approaches could distinguish among voiced and voiceless consonants, because our speech samples did not include voice cognates.

The observation that consonants tend to cluster based on place of articulation is not surprising and is consistent with findings reported by Green and Wang (2003), who compared differences among consonants based on tongue and lip movement coupling patterns. Green and Wang also derived a 3D articulatory consonant space using MDS, but obtained an R^2 value of only .70, which was much lower than the R^2 (.98) obtained for the 3D fit in our study. One possible reason why our approach has achieved a better fit than theirs is that our approach relied on two dimensions of articulatory movements, rather than only the vertical dimension that Green and Wang used.

Another interesting finding was that two principal components were sufficient to capture the variance in articulatory vowel space ($R^2 = .98$), but three components were required to capture the variance in articulatory consonants space ($R^2 = .98$ for 3D space as compared to .94 for 2D space). This finding is also consistent with feature-level descriptions of phonemes, which emphasize that two major factors (i.e., tongue height and tongue front-back position) determine the distinctiveness of vowel production, but more factors (e.g., manner of articulation, place of articulation, voiced and voiceless, nasality) contribute to the distinctiveness of consonants.

Limitations

The analysis used in the current study provided only a coarse-level analysis of the patterns of classification. Additional work is needed to investigate the patterns of misclassification, which may provide more details about the articulatory distinctiveness between those phonemes.

Table 4. Average consonant classification matrix (in percentage) of all participants, using Procrustes analysis.

Actual	Classified											
	/b/	/g/	/w/	/v/	/d/	/z/	/l/	/r/	/ʒ/	/dʒ/	/j/	
/b/	94.96		2.88	2.17								
/g/		94.70	0.50		1.36	0.50		0.53		0.50	1.92	
/w/	4.03	0.50	93.75	1.31				0.42				
/v/	1.81		1.03	96.14	0.48			0.56				
/d/		1.05		0.50	91.17	2.48	1.06	0.53	2.19	1.03		
/z/				0.56		96.92	1.05	1.06	0.42			
/l/		0.50		0.50	1.43	2.55	94.61		0.42			
/r/		0.48	0.89	1.47		2.11		88.41	3.40	2.17	1.08	
/ʒ/		2.98	0.56		0.83	0.50		1.03	83.25	10.32	0.53	
/dʒ/					1.58	2.79	0.53	1.06	12.49	81.56		
/j/	1.08	3.41	1.05		0.53	1.92			2.36	1.56	88.09	

Zeros are not displayed. Diagonal numbers are in bold.

Table 5. Average consonant classification matrix (in percentage) of all participants, using a support vector machine.

Actual	Classified											
	/b/	/g/	/w/	/v/	/d/	/z/	/l/	/r/	/ʒ/	/dʒ/	/j/	
/b/	92.32	0.43	3.33	1.89				0.53			1.50	
/g/	0.42	89.69	0.00	0.00	2.19	0.50	0.50	2.69	0.95	0.50	2.55	
/w/	2.52	0.43	90.89	1.74	0.50			2.41	0.50		1.00	
/v/	1.64	0.43	1.48	90.21	0.98	0.56		3.83		0.45	0.42	
/d/		0.53		0.50	86.67	1.43	2.48	4.05	1.89	1.54	0.92	
/z/	0.83	0.87		0.56	1.93	91.58	2.81	0.56	0.87			
/l/	0.42	0.50		0.45	2.45	3.00	90.90	0.83	0.45		1.00	
/r/	0.42	1.01	0.56		1.33	1.94	0.50	91.29	2.42		0.53	
/ʒ/		1.90			0.53	1.03	0.45	4.24	81.45	9.87	0.53	
/dʒ/	0.42				1.48			4.01	9.79	82.78	1.53	
/j/	0.42	1.82			0.00	0.42		4.33	0.45	2.06	90.51	

Diagonal numbers are in bold.

Duration and temporal information play an important role in distinguishing a number of vowels and consonants. However, Procrustes analysis, which is a spatial analysis, may not encode important temporal features based on, for example, manner of articulation. In Procrustes analysis, shapes are required to have the same numbers of data points. Thus, we sampled the articulatory movements for all phonemes to a fixed length (i.e., 10 data points) and consequently lost the duration and temporal information when the phonemes were compared in this study. Future efforts should consider extending standard Procrustes analysis to compare time-varying shapes with different lengths.

Consonant classification may be enhanced by including distinguishing features such as voicing and nasality. These additions, however, would require the integration of data from sensors that record information about voice and resonance.

In addition, because all of our speech stimuli were embedded in either a /b/ context (e.g., /bab/) or an /a/ context (e.g., /aba/), the extent to which the current findings generalize to other consonant and vowel contexts is unknown.

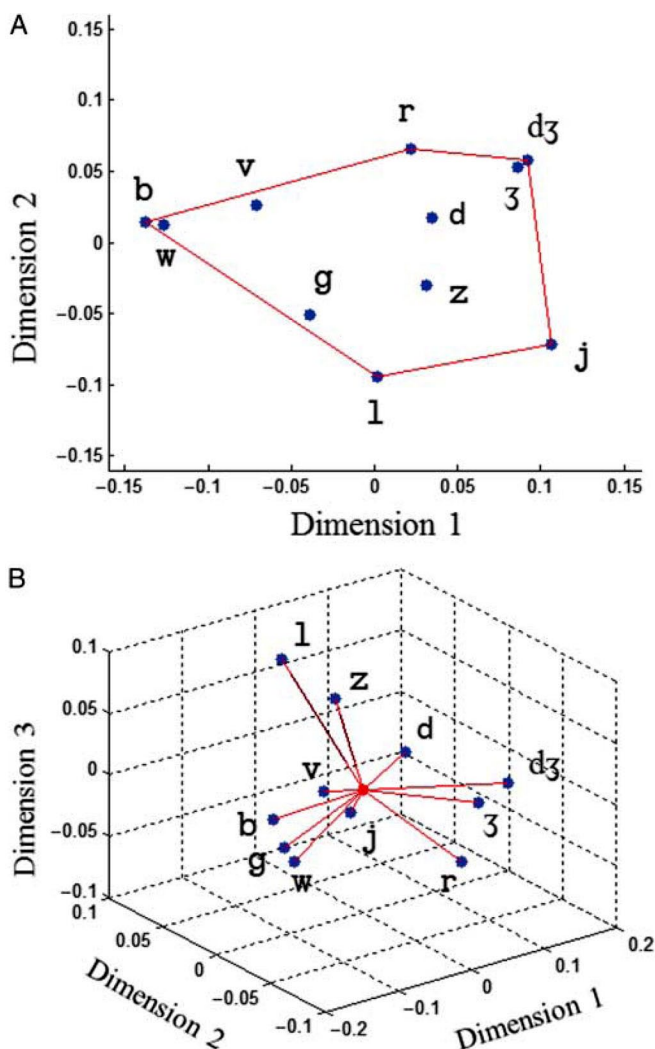
Additional research is required to determine potential context effects.

Clinical and Scientific Implications of the Derived Articulatory Vowel and Consonant Spaces

The current investigation was conducted not only to improve knowledge about the articulatory distinctiveness of vowels and consonants but also to develop articulation-based methods that could be used in future studies to quantify the severity of speech-motor impairment (Ball, Willis, Beukelman, & Pattee, 2001; Wang et al., 2011). Just as acoustic vowel space has been extensively used to explain the variance in intelligibility scores for speakers with dysarthria (e.g., Higgins & Hodge, 2002; McRae, Tjaden, & Schoonings, 2002; Tjaden & Wilding, 2004; Weismer et al., 2001), the derived articulatory spaces may also contribute to understanding intelligibility deficits in clinical populations. In contrast to acoustic analyses, the articulatory level of analysis can be used to directly determine the contribution of specific, compromised articulators to the speech impairment (Yunusova, Green, Wang, Pattee, & Zinman, 2011).

Table 6. Average articulatory distinctiveness of consonant pairs across participants.

Consonant	/b/	/g/	/w/	/v/	/d/	/z/	/l/	/r/	/ʒ/	/dʒ/	/j/
/b/	—	0.1761	0.1013	0.1035	0.1967	0.1972	0.1957	0.2098	0.2295	0.2546	0.2586
/g/	0.1761	—	0.1604	0.1608	0.1554	0.1637	0.1682	0.1785	0.1914	0.2121	0.1593
/w/	0.1013	0.1604	—	0.1068	0.2035	0.2110	0.2157	0.1819	0.2215	0.2485	0.2434
/v/	0.1035	0.1608	0.1068	—	0.1666	0.1607	0.1740	0.1643	0.1910	0.2156	0.2233
/d/	0.1967	0.1554	0.2035	0.1666	—	0.1207	0.1416	0.1386	0.1112	0.1257	0.1787
/z/	0.1972	0.1637	0.2110	0.1607	0.1207	—	0.1029	0.1712	0.1483	0.1639	0.1647
/l/	0.1957	0.1682	0.2157	0.1740	0.1416	0.1029	—	0.1978	0.1870	0.2003	0.1846
/r/	0.2098	0.1785	0.1819	0.1643	0.1386	0.1712	0.1978	—	0.1166	0.1371	0.1791
/ʒ/	0.2295	0.1914	0.2215	0.1910	0.1112	0.1483	0.1870	0.1166	—	0.0641	0.1730
/dʒ/	0.2546	0.2121	0.2485	0.2156	0.1257	0.1639	0.2003	0.1371	0.0641	—	0.1828
/j/	0.2586	0.1593	0.2434	0.2233	0.1787	0.1647	0.1846	0.1791	0.1730	0.1828	—

**Figure 4.** Quantitative articulatory consonant space. Dimensions are the results of the multidimensional scaling solution, which maintains the distance relationships between the data points. 2D = two dimensional; 3D = three dimensional.

Summary

Classification of eight vowels and 11 consonants based on articulatory movement time-series data were tested using two novel approaches, Procrustes analysis and SVM. Experimental results using a data set obtained from 10 healthy native English speakers demonstrated the effectiveness of the proposed approaches. The articulatory distinctiveness of the vowels and consonants were then quantified using Procrustes analysis. The quantified articulatory distinctiveness was then used to derive articulatory vowel and consonant spaces, which provided a visual representation of the distinctiveness of vowels and consonants. The clustering of those vowels and consonants in the derived spaces was consistent with feature-level descriptions of differences among the vowels and consonants. The approaches used in this study to quantify articulatory distinctiveness may be relevant to the continued efforts to improve differential diagnosis of speech disorders and to augment computer-based interventions of speech.

Acknowledgments – This work was funded in part by the Barkley Trust, Barkley Memorial Center, University of Nebraska-Lincoln, and National Institutes of Health Grant R01 DC009890/DC/NIDCD NIH HHS. We thank Tom D. Carrell, Mili Kuruvilla, Lori Synhorst, Cynthia Didion, Rebecca Hoising, Kayanne Hamling, Katie Lippincott, and Kelly Veys for their contribution to participant recruitment, data collection, and data processing.

References

- Ball, L. J., Willis, A., Beukelman, D. R., & Pattee, G. L. (2001). A protocol for identification of early bulbar signs in ALS. *Journal of Neurological Sciences*, 191, 43–53.
- Boser, B. E., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In D. Haussler, ed., *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York, NY: Association for Computing Machinery.

- Bunton, K., & Story, B. (2010). Identification of synthetic vowels based on a time-varying model of the vocal tract area function. *The Journal of Acoustical Society of America*, 127, EL146-EL152.
- Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27.
- Cortes, C., & Vapnik, V. (1995). Support-vector network. *Machine Learning*, 20, 273-297.
- Cox, R. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London, UK: Chapman & Hall.
- Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, 52, 270-287.
- Dryden, I. L., & Mardia, K. V. (1998). *Statistical shape analysis*. Hoboken, NJ: Wiley.
- Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., & Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics*, 30, 419-425.
- Fuchs, S., Winkler, R., & Perrier, P. (2008). Do speakers' vocal tract geometries shape their articulatory vowel space? In *Proceedings of the 8th International Seminar on Speech Production* (pp. 333-336); online at <http://issp2008.loria.fr/Proceedings/PDF/issp2008-77.pdf>
- Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40, 33-51.
- Green, J. R., & Nip, I. S. B. (2010). Organization principles in the development of early speech. In B. Maeseen & P. H. H. M. van Lieshout, eds., *Speech motor control: New developments in basic and applied research* (pp. 171-188). Oxford, UK: Oxford University Press.
- Green, J. R., Nip, I. S. B., Mefferd, A. S., Wilson, E. M., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, 53, 1-14.
- Green, J. R., & Wang, Y. (2003). Tongue-surface movement patterns during speech and swallowing. *The Journal of Acoustical Society of America*, 113, 2820-2833.
- Green, J. R., Wilson, E. M., Wang, Y., & Moore, C. A. (2007). Estimating mandibular motion based on chin surface targets during speech. *Journal of Speech, Language, and Hearing Research*, 50, 928-939.
- Heracleous, P., Aboutabit, N., & Beautemps, D. (2009). Lip shape and hand position fusion for automatic vowel recognition in cued speech for French. *IEEE Signal Processing Letters*, 16, 339-342.
- Higgins, C. M., & Hodge, M. M. (2002). Vowel area and intelligibility in children with and without dysarthria. *Journal of Medical Speech-Language Pathology*, 10, 271-277.
- Honda, K., Maeda, S., Hashi, M., Dembowski, J. S., & Westbury, J. R. (1996). Human palate and related structures: Their articulatory consequences. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 784-787). Baixas, France: International Speech Communication Association.
- Hueber, T., Benaroya, E-L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52, 288-300.
- Jin, N., & Mokhtarian, F. (2005). Human motion recognition based on statistical shape analysis. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance* (pp. 4-9). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Johnson, K. (2000). Adaptive dispersion in vowel perception. *Phonetica*, 57, 181-188.
- Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, 63, 187-194.
- King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121, 723-742.
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., . . . Lacerda, F. (1997, August 1). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277, 684-686.
- Kuhl, P. K., & Meltzoff, A. N. (1997). Evolution, nativism, and learning in the development of language and speech. In M. Gopnik, ed., *The inheritance and innateness of grammars*, (pp. 7-44). New York, NY: Oxford University Press.
- Ladefoged, P., & Johnson, K. (2011). *A course in phonetics* (6th ed.). Independence, KY: Cengage Learning.
- Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105, 1455-1468.
- Lindblom, B. (1990). Explaining variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal, eds., *Speech production and speech modeling* (pp. 403-439). Dordrecht, the Netherlands: Kluwer Academic.
- Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, 40, 1-19.
- Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., . . . Saenko, K. (2007). Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 621-624). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- McRae, P. A., Tjaden, K., & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in Parkinson disease. *Journal of Speech, Language, and Hearing Research*, 45, 35-50.
- Mefferd, A. S., & Green, J. R. (2010). Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research*, 53, 1206-1219.
- Meyer, G. J., Gustafson, S. C., & Arnold, G. D. (2002). Using Procrustes distance and shape space for automatic target recognition. *SPIE 4667, Image Processing: Algorithms and Systems*, 66-73. doi: 10.1117/12.468018
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008, May). Predicting human brain activity associated with the meanings of nouns. *Science*, 320, 1191-1195.
- Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, 51, 574-585.

- Perkell, J. S., & Cohen, M. H. (1989). An indirect test of the quantal nature of speech in the production of the vowels /i/, /a/ and /u/. *Journal of Phonetics*, 17, 123–133.
- Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of IEEE*, 91, 1306–1326.
- Richardson, M., Bilmes, J., & Diorio, C. (2000). Hidden-articulator Markov models for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 713–716). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Rosner, B. S., & Pickering, J. B. (1994). *Vowel perception and production*. Oxford, UK: Oxford University Press.
- Rudzicz, F. (2011). Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 947–960.
- Rvachew, S., Mattock, K., Polka, L., & Ménard, L. (2006). Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French. *The Journal of the Acoustical Society of America*, 120, 2250–2259.
- Sadeghi, V. S., & Yaghmaie, K. (2006). Vowel recognition using neural networks. *International Journal of Computer Science and Network Security*, 6, 154–158.
- Saenko, K., Livescu, K., Glass, J., & Darrell, T. (2009). Multi-stream articulatory feature-based models for visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 1700–1707.
- Shinchi, T. (1998). Vowel recognition according to lip shapes by using neural networks. In *Proceedings of the IEEE International Joint Conference on Computational Intelligence* (Vol. 3, pp. 1772–1777). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of Royal Statistical Society: Series B*, 40, 234–238.
- Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of Acoustical Society of America*, 55, 653–659.
- Sujith, K. R., & Ramanan, G. V. (2005, October). *Procrustes analysis and Moore–Penrose inverse based classifiers for face recognition*. Paper presented at the International Workshop on Biometric Recognition Systems, Beijing, China.
- Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 47, 766–783.
- Tsao, Y. C., & Iqbal, K. (2005). Can acoustic vowel space predict the habitual speech rate of the speaker? In *Proceedings of the IEEE International Conference on Engineering in Medicine and Biology* (pp. 1220–1223). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 38, 1001–1013.
- Visser, M., Poel, M., & Nijholt, A. (1999). Classifying visemes for automatic lip-reading. In *Lecture notes of computer science* (Vol. 1692, pp. 349–352). Berlin, Germany: Springer.
- Wang, J. (2011). Silent speech recognition from articulatory motion. Ph.D. dissertation, University of Nebraska–Lincoln.
- Wang, J., Green, J. R., Samal, A., & Carrell, T. D. (2010). Vowel recognition from continuous articulatory movements for speaker-dependent applications. In *Proceedings of the IEEE International Conference on Signal Processing and Communication Systems* (pp. 1–7). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Wang, J., Green, J. R., Samal, A., & Marx, D. B. (2011, August). *Quantifying articulatory distinctiveness of vowels*. Paper presented at InterSpeech, Florence, Italy; online at <http://digitalcommons.unl.edu/specedfacpub/74/>
- Wang, J., Samal, A., Green, J. R., & Carrell, T. D. (2009). Vowel recognition from articulatory position time-series data. In *Proceedings of the IEEE International Conference on Signal Processing and Communication Systems* (pp. 1–6). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Wang, J., Samal, A., Green, J. R., & Rudzicz, F. (2012a). Sentence recognition from articulatory movements for silent speech interfaces. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 4985–4988). Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Wang, J., Samal, A., Green, J. R., & Rudzicz, F. (2012b, September). *Whole-word recognition from articulatory movements for silent speech interfaces*. Paper presented at InterSpeech, Portland, OR; online at <http://digitalcommons.unl.edu/specedfacpub/76/>
- Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*, 53, 1–18.
- Westbury, J. (1994). *X-ray microbeam speech production database user's handbook*. Unpublished manuscript, University of Wisconsin–Madison.
- Yunusova, Y., Green, J. R., & Mefferd, A. (2009). Accuracy assessment for AG500 electromagnetic articulograph. *Journal of Speech, Language, and Hearing Research*, 52, 547–555.
- Yunusova, Y., Green, J. R., Wang, J., Pattee, G., & Zinman, L. (2011). A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (ALS). *Journal of Visualized Experiments*, 48, e2422; doi: 10.3791/2422
- Yunusova, Y., Weismer, G., & Lindstrom, J. (2011). Classification of vocalic segments from articulatory kinematics: healthy controls and speakers with dysarthria. *Journal of Speech, Language, and Hearing Research*, 54, 1302–1311.
- Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. (2008). Articulatory movements during vowels in speakers with dysarthria and in normal controls. *Journal of Speech, Language, and Hearing Research*, 51, 596–611.
- Zajac, D. J., Roberts, J. E., Hennon, E. A., Harris, A. A., Barnes, E. F., & Misenheimer, J. (2006). Articulation rate and vowel space characteristics of young males with fragile X syndrome: Preliminary acoustic findings. *Journal of Speech, Language, and Hearing Research*, 49, 1147–1155.

Appendix A. Sample data format in machine learning approach (n = 10).

Attribute												Label					
UL _{y1} , UL _{y2} , ... UL _{yn}				UL _{z1} , UL _{z2} , ... UL _{zn}				...		T1 _{y1} , ... T1 _{yn}		...		T4 _{z1} , ... T4 _{zn}		Phoneme	

Appendix B. Means and standard deviations of F1 and F2 values (Hz) across participants in Figure 3, Panel C.

Formant	/a/		/i/		/e/		/æ/		/ʌ/		/ɔ/		/o/		/u/	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
F1	901	67	391	101	553	71	919	87	759	60	834	96	575	63	412	43
F2	1,349	107	2,450	362	2,329	220	1,812	139	1,408	146	1,227	80	1,210	101	1,469	179