



# Quantifying Articulatory Distinctiveness of Vowels

Jun Wang<sup>1,2</sup>, Jordan R. Green<sup>2</sup>, Ashok Samal<sup>1</sup>, David B. Marx<sup>3</sup>

<sup>1</sup> Department of Computer Science & Engineering

<sup>2</sup> Department of Special Education & Communication Disorders

<sup>3</sup> Department of Statistics

University of Nebraska - Lincoln, United States

{junwang,samal}@cse.unl.edu, {jgreen4,dmarx}@unl.edu

## Abstract

The articulatory distinctiveness among vowels has been frequently characterized descriptively based on tongue height and front-back position; however, very few empirical methods have been proposed to characterize vowels based on time-varying articulatory characteristics. Such information is not only needed to improve knowledge about the articulation of vowels but also to determine the contribution of articulatory imprecision to poor speech intelligibility. In this paper, a novel statistical shape analysis was used to derive a vowel space that depicted the quantified articulatory distinctiveness among vowels based on tongue and lip movements. The effectiveness of the approach was supported by vowel classification accuracy of up to 91.7%. The theoretical relevance and clinical implication of the derived vowel space were discussed.

**Index Terms:** speech production, articulatory vowel space, Procrustes analysis, multi-dimensional scaling

## 1. Introduction

Clear and intelligible speech is characterized by the ability to produce discernible distinctions between sounds. The acoustic distinctiveness (e.g., based on *F1-F2* formants) of vowels has been studied extensively to investigate a large number of speech phenomena including intelligibility deficits and developmental change in speech [1][2]. Similar constructs are, however, lacking to quantify articulatory distinctiveness of vowels. An articulatory-based measure is needed to improve knowledge about the relation between vowel articulation and acoustics and to quantify the contribution of articulatory imprecision to poor speech intelligibility in persons with speech impairments. Articulatory distinctiveness among vowels has been frequently characterized descriptively by a diagram defined by tongue height and front-back position [3]. Existing empirical work has largely described vowels in terms of static tongue sensor positions [4] or vocal tract geometry [5]. These approaches, however, have rarely accounted for time-varying aspects of vowel production, which may provide an additional source of information for distinguishing vowels.

The goal of this research is to generate a vowel space that is delimited by the articulatory distinctiveness of vowels based on tongue and lip movements. We have developed a method to quantify the spatiotemporal distinctiveness between different vowels. The method was based on Procrustes analysis [6], a statistical shape analysis that has been applied successfully in object recognition [7]. A shape, in Procrustes analysis, is defined by an ordered set of landmarks on the surface of an object. Procrustes distance is the summed Euclidean distances between corresponding landmarks of two shapes after the locational, scaling, and rotational effects are removed. Procrustes analysis is particularly well suited for this application because (1) the sampled motion paths of sensors

attached to the tongue and lips can be viewed as shapes defined in Procrustes analysis; (2) Procrustes analysis provides a direct measure for pair-wise distinctiveness of vowels (Procrustes distance), and (3) in our preliminary study, the classification accuracy of vowels using Procrustes analysis matched that of a machine learning approach using a support vector machine [8]. In this study we have extended the typical use of Procrustes analysis, which was designed to analyze static shapes (i.e., shapes that do not change over time), to the analysis of time-varying shapes (i.e., shapes that change over time).

## 2. Data Collection

### 2.1. Participants

Ten female native American English speakers participated in this study. No speaker had positive history of speech or hearing problems. Each speaker participated in only one data collection session.

### 2.2. Stimuli

Eight major English vowels in consonant-vowel-consonant form, /bɑb/, /bɪb/, /beɪb/, /bæb/, /bʌb/, /boʊb/, /bɒb/, /bub/, were used as stimuli. The eight vowels are representative of the full English vowel set and were chosen because they sufficiently circumscribe the boundaries of articulatory vowel space. Therefore, these vowels provide a good representation of the bounds of tongue and lip movement patterns producing vowels. Each vowel was given a consonant context before and after, which is helpful to preserve the vowel identity. The context /b/, a bilabial, was selected because it is easy to parse and has minimum co-articulation effect on the vowel.

### 2.3. Procedure

The Electromagnetic Articulograph (EMA) (Carstens Inc.) was used to register the 3-D movements of the tongue, lip, and jaw when a subject was talking. The EMA records movements by establishing a calibrated electromagnetic field in a cube that induces electric current into tiny sensor coils that are attached to the surface of the articulators. Dental glue was used to attach the sensors. After the sensors were attached, the participant was seated with his/her head within the electromagnetic cube. When the participant spoke, the 3-D location data of the sensors were recorded and saved. The spatial precision of motion tracking using EMA (AG500) is approximately 0.5 mm [9].

Figure 1 shows the positions where 12 sensors were attached to a participant's head, face, and tongue. Three of the sensors were attached on a pair of glasses the subject wears. HC (Head Center) was on the bridge of the glasses, and HL (Head Left) and HR (Head Right) were on the left and right

outside edge of each lens, respectively. The movement of HC, HL, and HR were used to calculate the head-independent data of other sensors [10]. Two of the sensors, UL and LL, were attached on the middle positions of upper lip and lower lip. Four of the sensors, T1 (Tongue Tip), T2 (Tongue Body Front), T3 (Tongue Body Back) and T4 (Tongue Root), were attached on the midsagittal line of the tongue. There was a distance of approximately 10 mm between two adjacent tongue sensors [11]. The movements of three jaw sensors, JL (Jaw Left), JR (Jaw Right), and JC (Jaw Center), were recorded for future use.

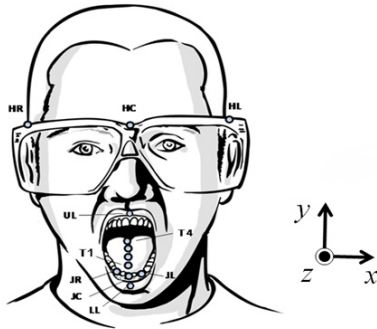


Figure 1: Sensor positions in data collection

Each participant produced the eight vowels sequentially, with small intervals between them. They repeated the sequence at least 20 times. Thus 20 samples for each vowel production for each subject were recorded.

#### 2.4. Data Processing

The time-series data of sensor locations recorded using EMA went through a sequence of preprocessing steps prior to analysis. First, the head movements were subtracted from the tongue and lip locations. The orientation of the derived 3-D Cartesian coordinate system is displayed in Figure 1. Second, a low pass filter of 10 Hz was applied for removing noise [11]. Third, all sequences were parsed to segments that are associated with each vowel. The segmentation was done manually by aligning the movement data with acoustic data recorded synchronously.

Only  $y$  and  $z$  coordinates of the sensors (i.e., UL, LL, T1, T2, T3, T4) were used for analysis because the movement along the  $x$  axis is not significant in normal speech production. Here,  $x$ ,  $y$ , and  $z$  are defined as spatial dimensions width (left-right), height (up-down) and length (front-back) in a 3-D Cartesian coordinate system (Figure 1). The origin (zero point) of the coordinate system is the center of the magnetic cube.

### 3. Method

Procrustes distance between vowel shapes defined by sampled tongue and lip motion paths was proposed as an index of the articulatory distinctiveness between vowels. Pair-wise distinctiveness of vowel shapes formed a distance matrix, which was then used as a dissimilarity matrix to generate a space using multi-dimensional scaling [12]. The effectiveness of distinctiveness measure was validated by vowel classification accuracy.

#### 3.1. Vowel Shape Distance (Vowel Distinctiveness)

In Procrustes analysis, a shape is represented by an ordered set of landmarks on the surface of an object. All shapes have the same number of landmarks. Procrustes distance is the sum of

Euclidean distances between corresponding landmarks of two shapes after location, rotational, and scaling effects are removed from the two shapes. Thus a step-by-step calculation of Procrustes distance between two shapes includes: (1) center the two shapes; (2) scale both shapes to unit size; (3) rotate one shape to match the other and obtain the minimum sum of the Euclidean distances between corresponding landmarks.

A faster method for calculating the Procrustes distance using a complex number representation for the landmark coordinates was used in this experiment. 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 conjugate transpose of  $u$ . Proof of Equation (1) is given in [6].

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

Procrustes analysis was designed for static shape analysis. However, a simple strategy was used to extend Procrustes analysis to time-varying shape analysis. Sampled motion paths of sensors attached on tongue and lips at different time points were spatially integrated as a composite shape representing a vowel before Procrustes distance was calculated. Specifically, motion path (defined by  $y$  and  $z$  coordinates) trajectories of all the six sensors for a vowel 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 [11], thus 10 samples adequately preserve the motion patterns. The composite shape, integration of 10 locations from each of the six sensors, was used to represent a vowel shape. Thus, in Equation (1),  $u$  is a  $1 \times 60$  matrix of complex numbers;  $u^*$  is a  $60 \times 1$  matrix of their complex conjugates; the result  $d_p$  is a real number. A similar strategy of spatially integrating shapes at different time points was used for recognition of human motion represented using images [13].

Figure 2 gives an example shape of /bab/ in which each circle is a landmark at a time point. A shape has totally 60 landmarks (10 locations  $\times$  6 sensors).

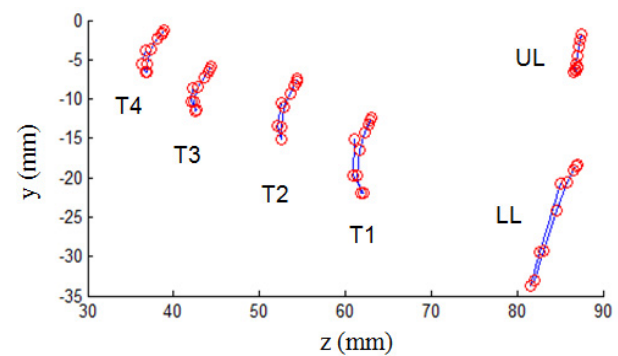


Figure 2: A shape of /bab/ produced by a subject

Finally, Procrustes distances between average shapes of vowel pairs were calculated and used as a measure of articulatory distinctiveness between the vowel pairs. Average shape of a vowel is the averaged coordinates of corresponding landmarks of all samples for each vowel.

#### 3.2. Vowel Space Generation

The articulatory vowel space was generated by organizing the vowels and preserving their distance relationships with each other. The distance between each vowel pair was stored in a symmetric distance matrix. Treating this as a dissimilarity matrix, multi-dimensional scaling (MDS) was used to generate the vowel space. Given a set of data points and distances between them, MDS can generate a space in which all distances between points are preserved. The orientation of the space is random and hence does not hold any physical significance. In this derived quantitative space, the distance between a vowel pair represents the articulatory distinctiveness between the vowel pair.

### 3.3. Vowel Classification

The effectiveness of the Procrustes distance as a measure of vowel distinctiveness was evaluated using a vowel classification/identification paradigm. The effectiveness of the approach would be supported by a high classification accuracy (i.e., greater than 90%). Vowels were classified using the following procedure: (1) the average shape for each vowel was calculated based on the average positions of corresponding landmarks of all samples for the vowel. The average shape serves as the reference for a vowel; (2) for each test sample (shape), the Procrustes distances between it and all the average shapes were calculated; (3) the vowel with the shortest distance to the test sample was considered as the recognized vowel.

The performance of the classification approach is shown in a classification matrix (or confusion matrix) and is measured by classification accuracy.

## 4. Results & Discussion

Vowel datasets collected from ten subjects were used in this experiment. The average number of samples was 20.9 for each vowel from each subject, although not every dataset contained 20 samples for each vowel. The number of samples for each vowel ranged from 16 to 24 from each subject. In all, 1672 vowel samples with 209 samples for each vowel were obtained and used for analysis.

### 4.1. Vowel Distinctiveness

Table 1 gives the average distance matrix across the ten subjects. The distance in this table was used as a measure of articulatory distinctiveness between vowels. For example, the distances between / $\alpha$ / and /i/ and that between / $\alpha$ / and /u/ (0.251 and 0.202 respectively) are the largest; the distances

among / $\Delta$ /, / $\circ$ /, and / $\text{o}$ / are shortest. This is consistent with classic phonetic knowledge that / $\alpha$ /, /i/ and /u/ are more distinct; / $\Delta$ /, / $\circ$ /, and / $\text{o}$ / are less distinct than others.

Table 1. Distance matrix of quantified articulatory distinctiveness of vowels.

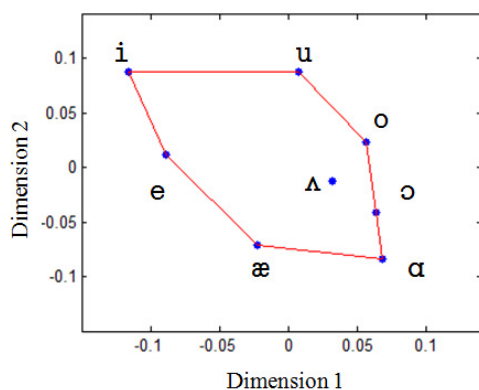
	$\alpha$	i	e	$\text{æ}$	$\Delta$	$\circ$	$\text{o}$	u
$\alpha$	0	0.251	0.196	0.126	0.109	0.089	0.136	0.202
i	0.251	0	0.104	0.194	0.191	0.234	0.209	0.146
e	0.196	0.104	0	0.141	0.150	0.186	0.168	0.141
$\text{æ}$	0.126	0.194	0.141	0	0.123	0.125	0.152	0.182
$\Delta$	0.109	0.191	0.150	0.123	0	0.074	0.081	0.126
$\circ$	0.089	0.234	0.186	0.125	0.074	0	0.100	0.164
$\text{o}$	0.136	0.209	0.168	0.152	0.081	0.100	0	0.103
u	0.202	0.146	0.141	0.182	0.126	0.164	0.103	0

### 4.2. Quantitative Articulatory Vowel Space

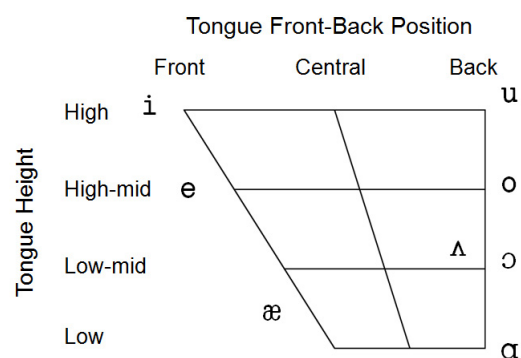
The symmetric distance matrix shown in Table 1 was used as a dissimilarity matrix for generating a vowel space using multi-dimensional scaling. Figure 3a gives the derived 2D articulatory vowel space. In this derived space, the two coordinates are the optimized two dimensions in non-classical MDS solution. Pair-wise 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 (i.e., 0.98). The distinctiveness between two vowels is indicated by the distance between them. The orientation of the space is random and does not hold any physical significance. Surprisingly but reasonably, the space resembles the descriptive articulatory vowel space [3] (Figure 3b). More interestingly, our space showed some differences, for example, /i/ and /u/ are closer than that indicated in the descriptive space.

MDS can also generate a 3D space (not shown in this paper). However, the third dimension does not help much in distinguishing the vowels ( $R^2$  is also 0.98). This is not quite surprising, because it is widely known that there are two major factors (tongue height and front-back positions) that distinguish vowels.

The articulatory vowel space may be affected by context, environment, speaker gender, speaking rate, and speech intelligibility, etc. Therefore, derived measures like the area of



a. Quantitative articulatory vowel space



b. Descriptive articulatory vowel space

Figure 3. Quantitative and descriptive articulatory vowel space (for eight major English vowels).

the convex hull (Figure 3a) that circumscribes the vowel space may serve as an indicator of speech intelligibility or overall speech severity. It has been reported that acoustic vowel space area explains some of the variance in intelligibility scores for speakers with dysarthria related to amyotrophic lateral sclerosis (ALS) [1]. It was also suggested that the distinctiveness among neighboring vowels (in acoustic vowel space) is more important in determining vowel intelligibility than (acoustic) vowel space area [2]. Future work is required to determine the association between acoustic and articulatory vowel space, and potential associations between articulatory vowel space and speech intelligibility. The average quantitative vowel space area of the ten healthy subjects in this experiment is 0.025 with a standard deviation 0.009.

### 4.3. Vowel Classification

Vowel classification was conducted to validate the effectiveness of our proposed measure for articulatory distinctiveness of vowels. To reduce the variation across speakers, vowel classification was conducted on each speaker individually.

Table 3 gives the average classification matrix in percentage across all subjects. A number at row  $i$  and column  $j$  in the matrix is the percentage of samples of  $i$ 'th vowel that was classified as  $j$ 'th vowel. Zeros are not displayed in Table 3. The classification matrix indicated / $\alpha$ /, / $i$ /, / $e$ /, / $\text{æ}$ /, and / $u$ / are easier to distinguish than / $\Delta$ /, / $\text{ɔ}$ /, and / $o$ /, which is consistent with the finding of vowel identity based on acoustic vowel space from female speakers [2].

The average classification accuracy of individual speakers was up to 91.7%. The standard deviation of classification accuracies across subjects was as low as 5.3%, which means our classification method works consistently across speakers. The high vowel classification accuracy indicated the effectiveness of our proposed vowel distinctiveness measure.

Table 3. Average classification matrix across subjects (in percentage)

		Classified							
		$\alpha$	$i$	$e$	$\text{æ}$	$\Delta$	$\text{ɔ}$	$o$	$u$
Actual	$\alpha$	<b>90.5</b>			0.4	3.5	5.1	0.4	
	$i$		<b>98.2</b>	0.9	0.4				0.4
	$e$		4.2	<b>94.2</b>		0.6			0.9
	$\text{æ}$	3.0		1.3	<b>92.5</b>	2.2	1.1		
	$\Delta$	2.0			1.5	<b>89.5</b>	5.6	1.5	
	$\text{ɔ}$	4.8			0.4	7.9	<b>81.1</b>	5.4	0.4
	$o$	1.1				5.1	2.9	<b>88.3</b>	2.6
	$u$					0.6		0.4	<b>98.9</b>

## 5. Conclusion & Future Work

This paper proposed a method based on Procrustes analysis for quantifying the articulatory distinctiveness of vowels and derived a quantitative articulatory vowel space based on tongue and lip movements using multi-dimensional scaling. In this derived space, the distance between vowel pairs indicated the articulatory distinctiveness of the pairs. Experimental results using datasets collected from ten speakers showed the effectiveness of our proposed measure for articulatory distinctiveness of vowels. The average speaker-dependent vowel classification accuracy was up to 91.7%. Surprisingly but reasonably, the quantified articulatory vowel space

resembles the widely used descriptive articulatory vowel space (both shown in Figure 3).

The current findings demonstrate (1) the possibility of quantifying the articulatory vowel distinctiveness using time-varying tongue and lip movements, rather than static tongue positions; and (2) the quantified articulatory vowel space strongly parallels both acoustic vowel space and the long-standing descriptions of vowel space defined by tongue height and front-back position [3].

The following research topics will be investigated as future directions: (1) apply this method to derive a quantitative articulatory space for consonants; (2) extend Procrustes analysis by addressing the dependence of articulatory movements at different time points; (3) investigate the relation between the quantified articulatory vowel space and acoustic vowel space; (4) investigate the scientific and clinical implications of the quantified articulatory vowel space.

## 6. Acknowledgements

This work was in part funded by the Barkley Trust, Barkley Memorial Center, University of Nebraska-Lincoln and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States). We would like to thank Dr. Tom D. Carrell, Dr. Mili Kuruvilla, Dr. Lori Synhorst, Cynthia Didion, Rebecca Hoelsing, and Katie Lippincott for their contribution to subject recruitment, data collection, and data processing.

## 7. References

- [1] Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R. D., and Kent, J. F., "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatrica et Logop.*, 53(1): 1-18, 2001.
- [2] Neel, A. T., "Vowel space characteristics and vowel identification accuracy", *Journal of Speech, Language, and Hearing Research*, 51(3):574-585, 2008.
- [3] Ladefoged, P., *A course in phonetics*, (2nd Ed.), Fort Worth: Harcourt Brace Jovanovich Publishers, 1982.
- [4] Fuchs, S., Winkler, R., and Perrier, P., "Do speakers' vocal tract geometries shape their articulatory vowel space?", *Proc. of the 8th International Seminar on Speech Production*, 333-336, 2008.
- [5] Honda, K., Maeda, S., Hashi, M., Dembowski, J.S., and Westbury, J.R., "Human palate and related structures: their articulatory consequences", *Proc. of International Conference on Spoken Language Processing*, 2:784-787, 1996.
- [6] Dryden, I. L., and Mardia, K. V., *Statistical shape analysis*, John Wiley, Chichester, 1998.
- [7] Meyer, G. J., Gustafson, S. C., and Arnold, G. D., "Using procrustes distance and shape space for automatic target recognition", *Proc. of SPIE*, 4667:66-73, 2002.
- [8] Wang, J., Green, J. R., Samal, A., and Carrell, T. D., "Vowel recognition from continuous articulatory movements for speaker-dependent applications", *IEEE Intl. Conf. on Signal Processing and Communication Systems*, 1-7, 2010.
- [9] Yunusova, Y., Green, J. R., and Mefferd, A., "Accuracy assessment for AG500 electromagnetic articulograph," *Journal of Speech, Language, and Hearing Research*, 52(2):547-555, 2009.
- [10] Green, J. R., Wilson, E. M., Wang, Y. and Moore, C. A., Estimating mandibular motion based on chin surface targets during speech, *Journal of Speech, Language, and Hearing Research*, 50(4):928-939, 2007.
- [11] Green, J. R. and Wang, Y. T., "Tongue-surface movement patterns during speech and swallowing," *Journal of Acoustical Society of America*, 113(5):2820-2833, 2003.
- [12] Cox, R.F., and Cox, M.A.A., *Multidimensional Scaling*, Chapman & Hall, 1994.
- [13] Jin, N., and Mokhtarian, F., "Human motion recognition based on statistical shape analysis", *IEEE Conference on Advanced Video and Signal Based Surveillance*, 4-9, 2005.