INTERSPEECH 2009 BRIGHTON

Comparison of Vowel Structures of Japanese and English in Articulatory and Auditory Spaces

Jianwu Dang[1] , *Mark Tiede*[2], *and Jiahong Yuan*[3]

[1]Japan Advanced Institute of Science and Technology, Japan;
[2]Haskins Laboratories and MIT R.L.E., USA;
[3]University of Pennsylvania, USA

E-mail: jdang@jaist.ac.jp; tiede@haskins.yale.edu; jiahong@babel.ling.upenn.edu

Abstract

In previous work [1] we investigated the vowel structures of Japanese in both articulatory space and auditory perceptual space using Laplacian eigenmaps, and examined relations between speech production and perception. The results showed that the inherent structures of Japanese vowels were consistent in the two spaces. To verify whether such a property generalizes to other languages, we use the same approach to investigate the more crowded English vowel space. Results show that the vowel structure reflects the articulatory features for both languages. The degree of tongue-palate approximation is the most important feature for vowels, followed by the open ratio of the mouth to oral cavity. The topological relations of the vowel structures are consistent with both the articulatory and auditory perceptual spaces; in particular the lip-protruded vowel /UW/ of English was distinct from the unrounded Japanese /ɯ/. The rhotic vowel /ER/ was located apart from the surface constructed by the other vowels, where the same phenomena appeared in both spaces.

Index Terms: vowels, speech production, speech perception

1. Introduction

Human beings are capable of producing and perceiving speech even under highly adverse conditions, and speech researchers attempt to answer why and how humans can accomplish this [2-5]. Although it is generally believed that the "speech chain" [3] linking speech production and perception in the human brain provides a considerable contribution, there is as yet no consensus about its details. Since vowels constitute the central part of speech, systematic studies on vowels should lead to better understanding of this issue. Accordingly, this study attempts to reveal inherent relations between speech production and perception by investigating inherent vowel structure.

In articulatory phonetic domain, the vowel system is commonly described as a distribution in a coordinate-structured space with a few dimensions such as tongue height and front/backness. In auditory perceptual space, vowels are described by a few formants. The topological compatibility can be seen in articulatory and auditory spaces for isolated vowels. However, if we adopt the same parameters to the vowels extracted from continuous speech, distinctive topology no longer appears in either articulatory or auditory space. In previous studies [1], we proposed a nonlinear analysis method (Laplacian eigenmap) to extract inherent vowel structures from an articulatory database of read speech for Japanese [6]. A consistent topological relation was found in the vowel structures for both the articulatory and auditory perceptual spaces.

Here, a question arises as to whether or not such topological compatibility exists in other languages. In this study, we choose English to test this question, because English has more vowels than Japanese, including the lip-protruded vowel /UW/ and rhotic vowel /ER/ that do not exist in Japanese. In this paper, we apply the same approach to these two languages, and compare the resulting vowel structures between the languages and between the articulatory space and auditory space within each language.

2. Method for exploring vowel structures

In past studies [7-9], a variance based method (PARAFAC) was used to find a few principal components to represent the data variance. However, variance based methods cannot appropriately characterize a dataset with nonlinear characteristics [10]. In contrast, a characterization method based on inherent similarity is a possible approach, since identifying similarity of objects is a basic criterion in human cognition.

Based on the similarity principle, articulations belonging to the same category should have similar properties and be located in a neighboring region of the articulatory space. Therefore, our objective is to find a method that is able to describe the similarity of vowels in articulatory and auditory spaces. For a general description, a vocal tract shape for a vowel is represented as a vector in articulatory space. Thus, all vectors for vowels form a set $\mathbf{X}$, $\mathbf{X}=\{X_i \in \mathrm{R}^n, i=1,2,...N\}$, where N is the data number. The similarity of the vocal tract vectors is described by a non-linear distance between one another, as in (1),

$$w_{ij} = \exp\left(-\left\|X_i - X_j\right\|^2 / \sigma\right) \quad (1)$$

where w_{ij} is the distance between the vocal tracts X_i and X_j. σ is the heat kernel of the data. A vocal tract shape is regarded as a point in the articulatory space. A similarity graph is constructed by connecting the point (vertex) to its neighbors in the given space, where two neighboring vertices are connected by an edge with a weighting coefficient of the distance. Thus, a distance matrix W can be obtained from such a graph as follows,

$$W = [W_1, W_2, W_3, ..., W_i, ..., W_N]$$
$$W_i = \left[w_{i,i(1)}, w_{i,i(2)}, w_{i,i(3)}, w_{i,i(4)}, ..., w_{i,i(k)}\right]^T \quad (2)$$

where $i(k)$ is the k-th nearest neighbor of the vertex i. Based on the vertices and edges, we construct a Laplacian graph to simulate the Laplace-Beltrami operator of the manifold [11, 12]. A "neighborhood keeping" map can be obtained from the discrete graph by minimizing the objective function,

$$L\hat{f}(\mathbf{X}) = \frac{1}{2}\sum_{i,j}\left(\hat{f}(X_i) - \hat{f}(X_j)\right)^2 w_{ij} \quad (3)$$

where L is the Laplacian matrix calculated using

$$L = D - W, \quad d_{ij} = \begin{cases} \sum_{n=1}^{k} w_{i,i(n)} & j = i \\ 0 & else \end{cases} \quad (4)$$

where d_{ij} is the element of matrix D. $\hat{f}$ is a mapping function of the vector vertices, which can be obtained by solving the generalized eigenvalue as

$$(L-\hat{\lambda}D)\hat{f}=0 \qquad (5)$$

The i - th vector can be described in a dimensional reduced space as in (6),

$$X_i \rightarrow \left[\hat{f}_1(X_i),\hat{f}_2(X_i),...,\hat{f}_j(X_i),...\hat{f}_n(X_i)\right]^T \qquad (6)$$

where $\hat{f}_j(X_i)$ is the projection on the space, and n is dimensions of the reduced space. The embedded manifold reflects the most important degrees of freedom of the system derived from the data. In this mapping, the topological relationship of the data can be retained even if using just a few of the principal dimensions.

3. Vowel structures in articulatory and auditory space

To explore vowel structures, the proposed method is applied to articulatory and auditory domains. Measurements of the speech organs during speech are employed in the articulatory case, while acoustic parameters of speech signals are used in the perceptual domain.

3.1. Vowel structure for Japanese

3.1.1.Data set of Japanese

The articulatory data of Japanese used in this study were recorded using the Electromagnetic Midsagittal Articulographic (EMA) system for read speech, and the acoustic signals were recorded simultaneously. The data were collected by a group at NTT communication science laboratories [6]. The sampling frequency was 16 kHz for acoustic signals, and 250Hz for articulatory data. Three Japanese male subjects served as the speakers in the this record.

Measurement points of the articulatory data are: the upper lip, lower lip, lower jaw, and four points on the tongue surface from the tongue tip to tongue rear. Each point is recorded by an x-y coordinate, where x corresponds to the posterior/anterior dimension and y to the inferior/superior dimension. Thus, a vowel articulation is represented by a vector with 14 dimensions. Five Japanese vowels were automatically segmented from stable periods in read speech and extracted from 360 sentences. As a result, the number of extracted vowels is 1,600 for /a/, 1,200 for /i/, 900 for /u/, 800 for /e/ and 1,200 for /o/. Altogether, the articulatory data set has about 5,700 vowels for each subject, which contains most of the phonemic environments of Japanese.

3.1.2. Vowel structure in articulatory space

To extract the vowel structure, we first construct a discrete graph based on the collected articulatory data of the vowels. The weighting matrix is derived from the graph. In constructing the weighting matrix W using Eq. (2), six nearest vertices were chosen for each output vertex, which is compatible with the number of vowel categories of Japanese. A mapping function is obtained by decomposing the weighting matrix using (5) and (6). Finally, a vowel structure with low dimensions is derived from the high dimensional data, while topological relationships are preserved. The vowel structure is shown in Fig. 1 for one speaker, where each point represents one vowel. The distinctive symbols and colors were used for five different vowels. From this figure, one can see that five Japanese vowels are well clustered into five categories.

In Fig. 1, the left panel shows the relation in the plane of the first and second dimensions. In the first dimension, vowel /i/ is located on the top, vowels /a/ and /o/ are in the bottom, and vowels /e/ and /u/ (/ɯ/) in the middle. With reference to articulation, the first dimension can explain the degree of tongue-palate approximation, i.e., high-front vs. low-back variation. This is consistent with traditional descriptions. In the horizontal direction, vowel /a/ and /e/ have a positive weighting coefficient, and /o/ and /u/ have a negative coefficient. With reference to articulatory configurations, the former has a larger opening ratio of the mouth to the oral cavity, while the latter has a smaller ratio. This implies that the second dimension associates with the opening ratio of the mouth to the oral cavity.

The right panel of Fig. 1 shows a three dimensional (3D) relation for the first three components. Vowel distribution is not monotonic in the third dimension, which scatters on a curved surface. 3D projections with different orientations show that the vowels distribute on the curved surface regularly. The location of the vowels along the surface is similar to that of vowel constrictions along the vocal tract, while the two wings reflect the mouth opening ratio. This structure provides a reasonable structure to describe the essential articulatory characteristics of the vowels.

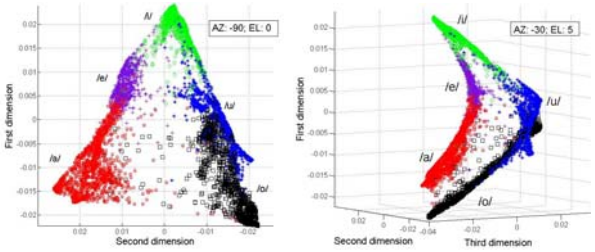

Figure 1: *3D vowel structure in articulatory space for five Japanese vowels from one speaker.*

3.1.3. Vowel structure in auditory space

To clarify the relation between speech production and perception, we investigate the vowel structure in auditory space as well as in articulatory space. Wang *et al.* [13] suggested that the auditory image can be represented by an affine transform of a logarithmic spectrum. Following this suggestion, we adopted the Mel Frequency Cepstral Coefficient (MFCC) as the preliminary parameter in extracting vowel structure in auditory space. Speech signals of the vowels were extracted from the identical period with that used in articulatory data. To keep the uniformity of the algorithm, the number of dimensions for MFCC is also chosen to be 14, which is the same as that of the articulatory data. The same processing approach used for articulatory data is used to explore the inherent vowel structure in auditory space.

Fig. 2 shows the explored vowel structure in the auditory space. The topological relation of the vowel structures is consistent with each other in these two spaces, where the distribution in the auditory space is not distinguished clearly as that in the articulatory space. From 3D projections, we found that the surface consisting of vowel distribution is twisted in the auditory space. As clarified in [1], each vowel is distributed on one plane regularly. However, the whole vowel structure in the auditory space takes on a spiral shape, like a twisted sheet. That is why we cannot find a clear view from any projection in the auditory space.

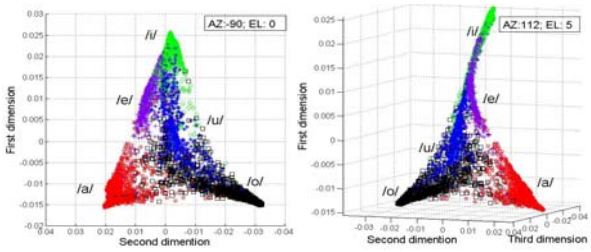

Figure 2: *3D vowel structure in auditory space for five Japanese vowels for the same speaker as shown in Figure 1.*

3.2. Vowel structure for American English

3.2.1.Data set of English

The articulatory and acoustic data of English used in this study were selected from the X-ray Microbeam Speech Production Database of Waisman Center at University of Wisconsin, USA [14]. The articulatory and acoustic data were recorded simultaneously. The sampling frequency was 21739Hz for acoustic signals, and 146Hz for articulatory data. Speech materials are read speech. The alignment of the phonemes in the read sentence is carried out using a forced alignment algorithm [15]. 12 American English vowels were automatically segmented based on the forced alignment in read speech for 21 speakers. As a result, the extracted samples were from 150 to 180 for the vowels.

Measurement points of the articulatory data were: the upper lip, lower lip, lower jaw, and four points on the midsagittal tongue surface from the tongue tip to tongue rear. Each pellet is recorded by an x-y coordinate, where x corresponds to the posterior/anterior dimension and y to the inferior/superior dimension. Thus, a vowel articulation is represented by a vector with 14 dimensions.

3.2.2.Vowel structures in articulatory and auditory spaces

To extract the vowel structures of English vowels, we applied the same approach to the data selected from the U.W. X-ray Microbeam Database. In constructing the weighting matrix *W* using Eq. (2), 12 nearest vertices were chosen for each output vertex, which is compatible with the number of vowel categories of English. Figure 3 shows the vowel structure obtained from one speaker. In the left panel of Fig. 3, the topological relation shows that the high vowels are located in the upper region of the structure and the lower vowels in the bottom region. The vowels with a larger open ratio of the mouth to the oral cavity are located on the right side of the distribution region, while the vowels with a smaller ratio are located on the left side. The right panel shows that most vowels are located on a curved surface, while the rhotic vowel /ER/ is located apart from the other vowels. In the articulatory space, one can see that except for the rhotic vowel, the structure for the other 11 English vowels has the same topological relation as the Japanese vowels. That is, the first dimension can explain the degree of tongue-palate approximation, i.e., high-front vs. low-back variation. In the second dimension, the vowels with a larger open ratio of the mouth to the oral cavity are located on one side, and the smaller ratio on the opposite side.

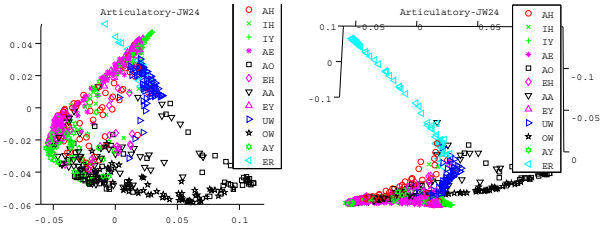

Figure 3: *3D vowel structure in articulatory space for 12 English vowels from one speaker.*

We extract the vowel structure of English in auditory space using the same methods as in 3.1.3, where 12 nearest vertices were chosen for each output vertex. Figure 4 shows the vowel structure in auditory space. The topological relation in auditory space is consistent with that in articulatory space for the other vowels. The surface of the vowels is somewhat spiral but is not clear as that seen in Japanese vowels. The topological relation of the vowels is consistent with the result of Miller [5], in which the vowel /ER/ is located out of the surface constructed by the other vowels. This implies that the rhotic vowel may have some special characteristics differing from the other vowels.

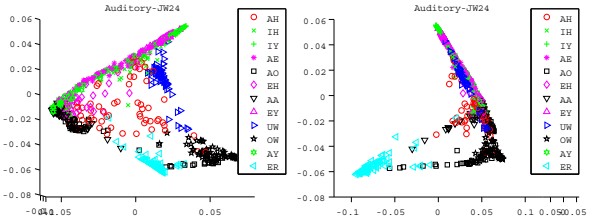

Figure 4: *3D vowel structure in auditory space for 12 English vowels from the same speaker as that shown in Figure 3.*

4. Comparison and Discussion

In the above results, one can see that the vowel structures are basically consistent with articulatory and auditory spaces for both Japanese and English. The topological relation of the vowel structures is similar between two languages. Since the vowel systems are different between the two languages, some results need to be investigated further.

4.1.1.Effects of the articulatory features

For about half of the 21 AE speakers, vowel /UW/ is located anterior to the /IY/ in the apex of the structure in the articulatory space, as shown in the left panel of Fig. 5. The causes may be related to the articulatory features of the lips. For testing, we constructed a new vowel structure by removing the components of the lips from the articulatory data, where the other components are exactly same. Figure 5 shows the vowel structures in articulatory space using full data (left) and without lip features (right). One can see that after the lip components were removed vowel /UW/ moved downward and backward, while almost no effect was seen on the other parts of the structure. The topological relation for /UW/ is close to the structure as shown in Figure 3. For the data used in Figure 3, we also investigated the effects of the lips by removing the lip components. There is no change in the topological relation while the distribution of /UW/ and /AO/ widened somewhat. These results show that the locations of /UW/ in the vowel structure in articulatory space depends on the lip features.

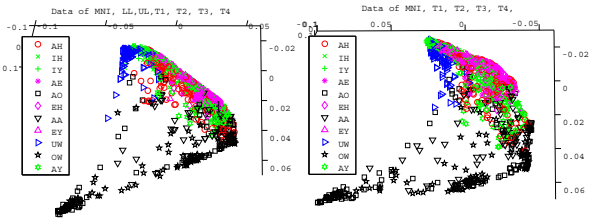

Figure 5: *Vowel structures in articulatory space using the full data (left) and without the lip features (right) for one speaker.*

4.1.2.Effects of the rhotic vowel

As shown in the articulatory and auditory spaces, vowel /ER/ is distinct from the other vowels in the vowel structures. To investigate the effects of the rhotic vowel on the structure, a comparison is made between the vowel structures with and without the rhotic vowel. Figure 6 shows the vowel structures without /ER/ in articulatory space (upper) and auditory space (lower); left for front view and right for lateral view. For the front view in the articulatory space, there is no difference between the structures with /ER/ in Fig. 3 and without /ER/ in upper panels of Fig. 6. For the lateral view, one can see that /ER/ is clearly apart from the other vowels, while it did not affect the topology so much by its inclusion. For the auditory space, one can see that /ER/ did not affect the distribution of the other vowels by comparing the vowel structure in Fig. 6 and the one in Fig 4. The results suggest that the explored vowel structures are intrinsic ones that are less dependent on the particular data set. In a small number of structures, however, the /ER/ was not so outstanding from the other vowels. It may be concerned with the articulation approach of particular speakers.

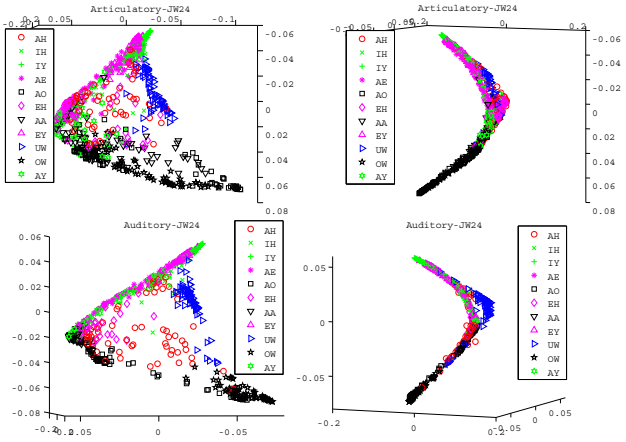

Figure 6: *Vowel structures without /ER/ in articulatory space (upper) and auditory space (lower); left for front view and right for lateral view from the same speaker as that shown in Figure 3.*

5. Summary

In this study, the vowel structures of Japanese and English were extracted and compared with respect to both articulatory and auditory structure. Although Japanese and English have different vowel systems, their vowel structures essentially have common characteristics. Vowel location in the structures basically reflects the articulatory features for both languages. The first dimension represents the degree of tongue-palate approximation. The second dimension reflects the open ratio of mouth to oral cavity by manipulating the articulatory features. The topological relations of the vowel structures are consistent with both the articulatory and auditory perceptual spaces.

The major difference between two languages is that English has more number of vowels than Japanese; in particular lip-protruded /UW/ and rhotic /ER/. When the lip features are removed from articulatory data of English, the vowel structure became similar to that of Japanese. Miller [5] indicated that rhotic vowel /ER/ was distinct from the surface constructed by the other vowels in auditory space. This study shows that the same phenomenon also exists in articulatory space. The rhotic vowel is located apart from the other vowels but its inclusion did not affect the structure of the other vowels. This implies that the structure reflects the intrinsic properties of vowels but less depends on the data set. However, the lip-protruded vowel /UW/ showed different topologies in the articulatory and auditory spaces. This inconsistency remains an issue for future study.

6. Acknowledgements

This study is supported in part by SCOPE (No. 071705001) of Ministry of Internal Affairs and Communications (MIC) and in part by Grant-in-Aid for Scientific Research of Japan (No. 20300064). We would like to thank NTT communication science laboratories for permitting us to use their articulatory data. We also appreciate the opportunity to work with the X-ray microbeam database collected at the University of Wisconsin by John Westbury.

7. References

1. Dang, J., et al. *Inherent Vowel Structures in Speech Production and Perception Spaces*. in *International Seminars on Speech Production*. 2008. Strasburg, France.
2. Pols, L., L. van der Kamp, and R. Plomp, *Perceptual and Physical Space of Vowel Sounds.* J Acoust. Soc. Ame., 1969. **46**: p. 458-467.
3. Denes, P., and Pinson, E., *The Speech Chain*. 2nd ed. 1993, New York: W.H. Freeman and Co.
4. Liberman, A., and, Mattingly, G., *The motor theory of speech perception revised.* Cognition, 1985. **21**: p. 1-36.
5. Miller, J., *Auditory-perceptual interpretation of the vowel.* J. Acoust. Soc. Am., 1989. **85**(5): p. 2114-2134.
6. Okadome, T. and M. Honda, *Generation of articulatory movements by using a kinematic triphone model.* J. Acoust. Soc. Am., 2001: p. 453-463.
7. Jackson, M., *Analysis of tongue positions: Language-specific and cross linguistic models.* J. Acoust. Soc. Am., 1988. **84**(1): p. 124-143.
8. Hoole, P., *On the lingual organization of the German vowel system.* J. Acoust. Soc. Am., 1999. **106**(2): p. 1020-1032.
9. Zheng, Y., M. Hasegawa-Johnson, and S. Pizza, *Analysis of the three dimensional tongue shape using a three-index factor analysis model.* J. Acoust. Soc. Am., 2003. **113**(1): p. 478-486.
10. Venna, J. and S. Kaski, *Local multidimensional scaling.* Neural Networks, 2006. **19**: p. 889-899.
11. Belkin, M. and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural computation, 2003. **15**: p. 1373-1396.
12. Rosenberg, S., *The Laplacian on a Riemannian manifold.* 1997: Cambridge University Press.
13. Wang, K. and S. Shamma, *Spectral Shape Analysis in Central Auditory System.* IEEE Trans. on Speech and Audio Processing, 1995. **3**(5): p. 382-395.
14. Westbury, J., *X-RAY MICROBEAM SPEECH PRODUCTION DATABASE USER'S HANDBOOK*. 1994, Waisman Center, University of Wisconsin: Madison, USA. p. 1-100.
15. J. Yuan and M. Liberman. *Speaker Identification on the SCOTUS corpus*. in *Prceedings of Acoustics 2008*.