Inherent Vowel Structures in Speech Production and Perception Spaces

Jianwu Dang[1], Xugang Lu[2], Mark Tiede[3], and Kiyoshi Honda[4]

[1]JAIST, 1-1 Asahidai Nomi Ishikawa 923-1292 Japan; [2]NICT/ATR, SLC; [3]Haskins Laboratories, USA
[4]Laboratoire de Phonetique et de Phonologie, CNRS-Paris Univ.3, France
E-mail: jdang@jaist.ac.jp; xugang.lu@atr.jp

Abstract

To examine relations between speech production and perception in the human brain, we investigated inherent vowel structures in both articulatory space and auditory perceptual space. A nonlinear analysis method (Laplacian eigenmap) was employed to extract inherent vowel structures from an articulatory database of read speech. In articulatory space, the first dimension of the vowel structure represents the degree of tongue-palate approximation, and the second is related to the ratio of lip opening to oral cavity size, while the vowels scatter on a curved surface in the third dimension. When applying the same technique to corresponding acoustic data from the same corpus, a compatible structure was obtained from the vowels in the auditory perceptual space. The results suggest that consistent topological images of the vowel structures exist in both the articulatory and auditory perceptual spaces. This finding supports the hypothesis that an efficient auditory-articulatory linkage exists in the human brain for speech processing.

1. Introduction

Human beings are capable of producing and perceiving speech even under highly adverse conditions. Many studies [1-4] have been conducted to answer why and how. Although it is generally believed that the "speech chain" [2] 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 vowel structure.

In the 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. Thus, a question arises as to whether or not the topological compatibility exists in both spaces of production and perception for continuous speech.

Because the brain develops its capacity for cognizing vowel structure during language acquisition in tandem for both speech production and perception, it seems likely that it would identify and exploit compatible elements of such structure in the two domains. In fact, the Motor Theory of Speech Perception [3] asserts that speech sounds are perceived with reference to articulatory gestures that humans are innately able to produce. As also demonstrated by the McGurk effect [5], articulatory information is an important cue for speech perception. However, to support the hypothesis that a common processing pathway in the brain is applied to both production and perception [3], we must first clarify whether or not a consistent vowel structure exists in both the articulatory and perception spaces, whether the vowels are isolated or drawn from continuous speech. In this study, we reveal such compatible vowel structures using articulatory observations and acoustic measurements.

2. Method for exploring vowel structures

In previous studies [6-8], 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 [9]. In contrast, a characterization method based on inherent

similarity can be expected to be a proper approach, since the 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 vowel articulation in a low dimensional space. 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 $\{X_i \in R^n, i = 1,2,...,N\}$, where N is the data number. The similarity of the vocal tract vectors is described by a non-linear distance between one another as:

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

where w_{ij} is the distance between the vocal tracts X_i and X_j. σ is the variance of the data. A vocal tract shape is regarded as a point in the vocal tract space. A similarity graph is constructed by connecting the point (vertex) of its neighbors in the vocal tract 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. Based on the vertices and edges we construct a Laplacian graph to simulate the Laplace-Beltrami operator of the manifold [9]. A "neighborhood keeping" map can be obtained from the discrete graph by minimizing the objective function:

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

where L is the Laplacian matrix calculated as $L = D - W$. D is a diagonal matrix with entries are column sum of W. $\hat{f}$ is a mapping function of the vector vertices, which can be obtained by solving the generalized eigenvalue as $(L - \hat{\lambda}D)\hat{f} = 0$. The i - th vector can be described in a dimensional reduced space as:

$$X_i \to \left[\hat{f}_1(X_i), \hat{f}_2(X_i), ..., \hat{f}_j(X_i), ...\hat{f}_n(X_i)\right]^T \quad (3)$$

where $\hat{f}_j(X_i)$ is the projection on the space, and n is dimensions. 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 principal dimensions.

3. Vowel structures in articulatory and auditory space

To explore vowel structures, the proposed method is applied to articulatory and perceptual 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 Data set

The articulatory data 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 [10]. The sampling frequency was 16 kHz for acoustic signals, and 250Hz for articulatory data.

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/. All together, the articulatory data set has about 5,800 vowels for each subject, which contains most of the phonemic environments of Japanese.

3.2 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, where six nearest vertices were chosen for each output vertex. A mapping function is obtained generalized eigenvalue decomposition. Finally, a vowel structure with low dimensions is derived from the high dimensional data, while topological relationships are preserved. The 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 the six rotations of the three plotted dimensions, one can see that five Japanese vowels are well clustered into five categories.

In Fig. 1, the top-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/ 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 lips to the oral cavity, while the latter has a smaller ratio. This implies that the second dimension associates with the opening ratio of the lips to the oral cavity.

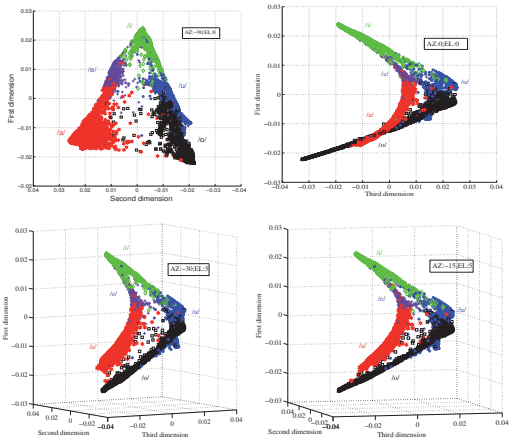

Figure 1. 3D vowel structure in articulatory space.

The top-right panel of Fig. 1 shows the relation between the first and third dimensions. Vowel distribution is not monotonic in the third dimension, which scatter on a curved surface. 3D projections with different orientations show that the vowels distribute on the curved surface regularly. The shape of this surface somehow reflects the vocal tract shape. 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 lip opening ratio. This structure provides a reasonable structure to describe the essential articulatory characteristics of the vowels.

3.3 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.* [11] suggested that the auditory image can be represented by an affine transform of a logarithmic spectrum. Following this suggestion, the Mel Frequency Cepstral Coefficient (MFCC) is adopted as the preliminary parameter in extracting vowel structure in auditory space. Speech signals of the vowels were extracted from the identical sampling period with that of the 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 two vowel structures are consistent with each other in these two spaces. That is, the first dimension reflects the degree of tongue-palate approximation, and the second dimension is corresponding to the lip opening ratio. In the top-right panel, the vowels distribute on a curved surface and their location corresponds to the vowel constriction in the vocal tract, where we reversed the third dimension in order to show vowel /a/ clearly. As seen in the top panels, the distribution in the auditory space is not as clearly distinguished as that in the articulatory space. From 3D projections, we found that the surface consisting of vowel distribution is twisted in the auditory space. To clarify the details, we rotated the vowel structure and let the distribution of each vowel be perpendicular to the projection plane, respectively. In the top-right panel, the distribution of vowel /a/ is perpendicular to the projection plane. Comparing with the top-left panel, the distribution of /a/ reduces to a thick line. Similarly, the distribution of /i/ reduces to a line in the panel with the view of (-75, 5), where these are angles of azimuth (AZ) and elevation (EL) respectively. In the view of (-59, 5), the distribution of /u/ reduces to a narrow strip. In this panel, one can clearly see that the vowel structure shows a spiral shape. In the view of (-26, 5), the distribution of /o/ reduces to a narrow strip. In the view of (112, 5), the distribution of /e/ is reduced to an arc, differing somewhat from the others. This means that the distribution of /e/ is along a curved surface. As clarified here, 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.

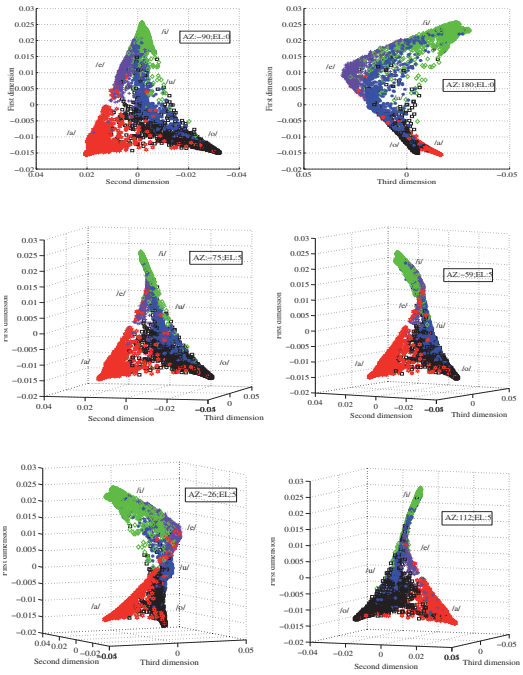

Figure 2. 3D vowel structure in auditory space.

4. Discussion and Conclusions

In this study, we compared vowel structures derived from two modalities based on their similarities, which is commonly exploited in human cognition. Clear vowel structures were obtained in articulatory and auditory spaces. Cross-modal structural consistency has also been confirmed for other subjects. In previous articulation based studies [6-8], vowel structures were explored using the PARAFAC method for different speech materials. The topologies derived from those studies showed large differences. Our preliminary analysis suggests that the different structures can be derived from our vowel structure by rearranging the view point. The clarified auditory vowel structure is consistent with the one in articulatory space. Our structure is consistent with previous work [1, 4] for the first two dimensions. In the third dimension, our result resembles that in [1] which was derived from subjective evaluations. The acoustic parameter used in this study is the entire spectrum. Therefore, our result is more robust than that derived from a few formants alone.

Vowel structures in both articulatory and auditory spaces are consistent with each other in the lower two dimensions, while it shows some difference in the higher dimensions with spiral shapes. The vowel structure in articulatory space can be explained by a clear physical meaning with reference to human articulation. The first three most important factors for vowel production are the degree of tongue-palate approximation, lip opening ratio, and the constriction place in the vocal tract. The consistency of the vowel structures in the two spaces suggests that they are mapped onto a single common image in the human brain, which is derived from our exploratory articulation during language acquisition.

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 share the articulatory data.

References

1. 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.
2. Denes, P., and Pinson, E., *The Speech Chain*. 2nd ed. 1993, New York: W.H. Freeman and Co.
3. Liberman, A., and, Mattingly, G., *The motor theory of speech perception revised.* Cognition, 1985. 21: p. 1-36.
4. Miller, J., *Auditory-perceptual interpretation of the vowel.* J. Acoust. Soc. Am., 1989. 85(5): p. 2114-2134.
5. McGurk, H. and J. MacDonald, *Hearing lips and seeing voices.* Nature, 1976. 264: p. 746-748.
6. Jackson, M., *Analysis of tongue positions: Language-specific and cross linguistic models.* J. Acoust. Soc. Am., 1988. 84(1): p. 124-143.
7. Hoole, P., *On the lingual organization of the German vowel system.* J. Acoust. Soc. Am., 1999. 106(2): p. 1020-1032.
8. 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.
9. Belkin, M. and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural computation, 2003. 15: p. 1373-1396.
10. Okadome, T. and M. Honda, *Generation of articulatory movements by using a kinematic triphone model.* J. Acoust. Soc. Am., 2001: p. 453-463.
11. Wang, K. and S. Shamma, *Spectral Shape Analysis in Central Auditory System.* IEEEE Trans. on Speech and Audio Processing, 1995. 3(5): p. 382-395.