INVESTIGATIONS INTO VOWEL AND CONSONANT STRUCTURES IN ARTICULATORY AND AUDITORY SPACES USING LAPLACIAN EIGENMAPS

Jianwu Dang[⋆†] *, Shengbei Wang*[⋆]*, and Masashi Unoki*[⋆]

[⋆] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[†] Tianijn Key Lab. of Cognitive Computing and Application, Tianjin University, China
{jdang, wangshengbei, unoki}@jaist.ac.jp

ABSTRACT

Many studies have investigated the relationship between the articulatory and auditory features for isolated speech sound and vowels. For fully understanding the mechanisms of speech production and perception, it is necessary to investigate the consonants in the same way. For this reason, in this study, we investigate the manifolds of vowels and consonants out of Japanese reading speech using Laplacian eigenmaps. We constructed uniform articulatory and auditory spaces based on the vowels and consonants to investigate their manifolds. It is found that the distribution of consonants in articulatory space could be classified into labial and lingual groups which reflected their articulatory properties, while in auditory space their distribution was clustered according to voiced and unvoiced, plosive and fricative properties. In vowel-consonant acoustic space, the consonants distributed as a hoe-like shape, with voiced consonants located on the blade of the hoe and fused with vowels. We defined average correlation coefficients to measure the similarity of manifold between three speakers. The results indicated that the vowel/consonant structures had high consistency among the three speakers.

Index Terms— Speech production, consonants, vowels, articulatory space, auditory space

1. INTRODUCTION

Speech production and perception are two fundamental functions of human beings that interact with each other in speech communication. For human beings, the process of producing, perceiving, and understanding speech information seems to be quite natural and effortless even under highly adverse conditions. Although our brain is related to and responsible for the complex interpretations of speech [1], many fundamental details about how speech is produced and perceived are still unclear. Since speech perception in auditory space is closely related to speech production in articulatory space [2], one hypothesis is speech communication in the brain may be achieved efficiently by the topological mapping between articulatory and auditory structures [3]. For this reason, the topological structure for vowels has been investigated widely [3, 4, 5, 6]. To understand more details about the interactions between speech production and perception, it is necessary to investigate into the structures of vowels as well as consonants, and the relationship between vowels and consonants in articulatory and auditory spaces.

This work was supported in part by a Grant-in-Aid for Scientific Research by Japan (No. 25240026 and No. 25330190) and by the National Basic Research Program of China (No. 2013CB329301). This study was also supported in part by the National Natural Science Foundation of China (No. 61233009). We would like to thank NTT Communications Science Laboratories for permitting us the use of their articulatory data.

In literature, several existing studies have analyzed the articulations of vowels [7, 8]. However, these methods could only explore the linear characteristics of vowels extracted from isolated speech materials. To investigate vowel structures more accurately, our previous studies [4, 5, 6] have adopted a nonlinear analysis method, viz., Laplacian eigenmaps, to construct the structures for vowels out of continuous Japanese and English speech. We obtained clearer vowel structures than those using linear methods and found that vowels had consistent structures in both articulatory and auditory spaces. This is because all the vowels have the same sound source, and therefore can be described using a few uniform articulatory features: the height and backness of the tongue, and the rounding and protrusion of the lips. Compared with vowels, consonants have more complex situations. They are characterized by both the articulatory location and manner. The articulatory location can be a tight constriction or complete closure within the vocal tract, while the manner can be different types of sound sources: fricative, plosive, and voiced sources, etc. In particular, the manner heavily affects the acoustic characteristics of consonants. It is expectable that the structures of consonants are different in articulatory and auditory spaces. So far, however, no such investigations have been carried out on consonants yet [6].

Since the investigations into consonants would greatly benefit for a full understanding of the mechanisms of speech production and perception. In this study, we investigate the differences of consonant structures between the articulatory and auditory spaces, which are caused by the different sound sources, and clarify the relationship between vowel and consonant structures in the spaces.

2. PRINCIPLES UNDERLYING METHOD OF NONLINEAR ANALYSIS BASED ON LAPLACIAN EIGENMAPS

In previous studies, a linear model of PARAllel FACtor analysis (PARAFAC) [7] has been proposed to analyze the tongue position during articulation. Several approaches developed from this model were applied to analyze the features of tongue position for different languages [8] or different speech databases that were obtained by electromagnetic midsagittal articulography (EMA) [9] or magnetic resonance imaging (MRI) [10]. However, these linear methods are not proper for analyzing the articulatory data out of continuous speech which has heavy non-linear coarticulation [6].

It is believed that human beings categorize and perceive speech based on a number of articulatory and auditory cues so that one phoneme could be differentiated from the others based on their similarities/differences. Therefore, phonemes with closer similarities should distribute in the same category or a neighboring region in the articulatory and auditory spaces. In a general view, either acoustic space or articulatory space possibly has a manifold structure, in

which a local area can be treated as a linear region. Based on this concept, our previous works [4, 5, 6] adopted a nonlinear method, i.e., Laplacian eigenmaps, to explore the vowel structures in both articulatory and auditory spaces for vowels out of continuous speech. In this study, we use the same approach to investigate the consonant structures in articulatory and auditory spaces, and re-examine the vowel structures in a uniform space with both vowels and consonants included since the structures are data-dependent to some extent.

The principles underlying Laplacian eigenmaps are briefly described below. In speech production, a phoneme is represented using a number of feature points. Mathematically, the feature points of a phoneme can be treated as a vector in an articulatory space, where phonemes with similar properties are located in a neighboring region. All the vectors of the phonemes form a set, $\mathbf{X}$, in Eq. (1):

$$\mathbf{X} = \{X_i \in R^n, i = 1, 2, 3, ..., N\}, \tag{1}$$

where N is the number of phonemes.

The similarity between two vectors X_i and X_j can be described using a nonlinear distance with Eq. (2):

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

where w_{ij} is the distance between X_i and X_j, and σ is the heat kernel of the data.

In the articulatory space, a configuration of the vocal tract can be regarded as a point. A similarity graph is constructed by connecting a point (vertex) to its neighbours in the given space, where two neighbouring vertices are connected by an edge with a weighting coefficient of the distance. Thus, a distance matrix $\mathbf{W}$ can be obtained from a graph as:

$$\mathbf{W} = [W_1, W_2, W_3, ..., W_N], \tag{3}$$

and W_i is calculated with Eq. (4):

$$W_i = [w_{i,i(1)}, w_{i,i(2)}, w_{i,i(3)}, ..., w_{i,i(k)}]^T, \tag{4}$$

where $i(k)$ is the k-th nearest neighbour of vertex i.

Based on the vertices and edges, a Laplacian graph is constructed to simulate the Laplace-Beltrami operator of the manifold. A "neighborhood keeping" map can be obtained from the discrete graph by minimizing the objective function:

$$L\hat{f}(\mathbf{X}) = \frac{1}{2}\sum_{i,j}(\hat{f}(X_i) - \hat{f}(X_j))^2 w_{ij}, \tag{5}$$

where L is the Laplacian matrix calculated using:

$$L = D - \mathbf{W}, \quad d_{ij} = \begin{cases} \sum_{n=1}^{k} w_{i,i(n)}, & j = i \\ 0, \quad \text{otherwise}, \end{cases} \tag{6}$$

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

$$(L - \hat{\lambda}D)\hat{f} = 0. \tag{7}$$

The i-th vector can be described in a dimension reduced space as:

$$X_i \rightarrow [\hat{f}_1(X_i), \hat{f}_2(X_i), ..., \hat{f}_j(X_i), ..., \hat{f}_n(X_i)]^T, \tag{8}$$

where $\hat{f}_j(X_i)$ is the projection on the space, and n represents the dimensions of the reduced space. The embedded manifold reflects the most important degrees of freedom derived from the data set. In this mapping, the topological relationships of the data can be preserved, even if only a few principal dimensions are used. For more details about this method, please refer to our previous study [4].

3. DATA SET AND METHODS

3.1. Data set

The data set used in this study is NTT articulatory database including articulatory data and acoustic data [11]. Three Japanese male subjects, i.e., MH, TM, and TO, uttered 360 Japanese sentences at normal speech rates for the data collection [12]. The articulatory data were recorded using EMA system with 11 channels including three reference points. The acoustic data were simultaneously recorded. The sampling frequencies were 250 Hz for articulation and 16 kHz for acoustic data. Five Japanese vowels /a/, /i/, /u/, /e/, /o/ and 20 consonants /b/, /c/, /d/, /f/, /g/, /h/, /j/, /k/, /m/, /n/, /p/, /r/, /s/, /t/, /w/, /x/, /y/, /z/, /T/, and /N/ were manually segmented and labelled in the reading speech, where /w/ and /y/ are semi-vowels, whose counterparts are /u/ and /i/, respectively. Note that in this study, /T/ is the assimilated sound of /t/ and /N/ denotes the velar nasal. All the phonemes are finally refined by acceleration of the jaw during articulatory movement (see [11] for more details).

The articulatory data were collected at seven points on the speech organs: upper lip, lower lip, lower jaw, and four points on the tongue's surface from the tip to the rear located on the midsagittal plane. Each point was recorded in x-y coordinate, where x is the posterior/anterior dimension and y is the inferior/superior dimension. As a result, one phoneme was represented by a vector with 14 dimensions. The five targeted Japanese vowels and 20 targeted consonants were extracted from the reading speech according to the labels. About 2,800 occurrences of the targeted vowels and 2,500 occurrences of the targeted consonants were obtained for each subject. Features extracted from the 20-ms period around the middle part of a given phoneme was used to represent the phoneme, since this part is considered to be stabe to represent the phoneme. The extracted vowels and consonants have a large variety since the speech materials reflect most coarticulation environments in Japanese.

3.2. Method of constructing the articulatory structure

To construct vowel and consonant structures in articulatory space, each phoneme is treated as one point in the 14-dimension articulatory space, which is the average over six samples (basically enough) within the 20-ms period. Thus, a discrete graph is constructed with all the points of the phonemes. To discover the inherent manifold, a given vertex of the graph is allowed to connect eight nearest vertices in its neighboring region by weighted edges. The other vertices have no connection. Based on the Laplacian eigenmaps, the weighting matrix $\mathbf{W}$ is calculated with Eqs. (2) and (3), and the eigenvectors are obtained with Eqs. (5)-(7) [13]. Accordingly, we can choose a few important eigenvectors to construct a low dimensional space, in which the topological relationships in the manifold properties can be preserved. By mapping the vowel and consonant vectors into the low dimensional space, we can visualize their structures in two- and/or three-dimensional representations.

3.3. Method of constructing the auditory structure

Similarly, vowel and consonant structures are also constructed in auditory space to investigate the relationship between articulation and auditory aspects. Since the auditory image can be represented with an affine transform of a logarithmic spectrum, we adopted the Mel Frequency Cepstral Coefficient (MFCC) as the preliminary parameter for describing vowels and consonants in auditory space [14]. The speech signals of vowels and consonants are extracted from the same period as that used in articulatory data, and the dimensions for

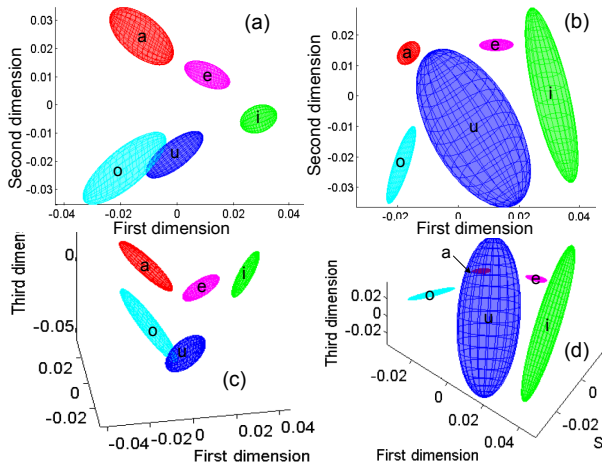

Fig. 1. Vowel structures in articulatory and auditory spaces: (a) and (c) are articulatory structures, (b) and (d) are auditory structures for speaker MH.

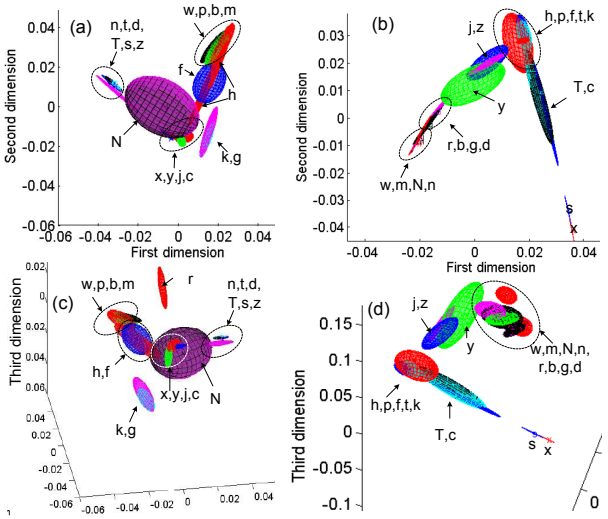

Fig. 2. Consonant structures in articulatory and auditory spaces: (a) and (c) are articulatory structures, (b) and (d) are auditory structures for speaker MH.

MFCC are chosen to be 14, the same dimensions as that of the articulatory data. The same process used for articulatory data is applied to explore the inherent structure in auditory space.

4. VOWEL AND CONSONANT STRUCTURES IN ARTICULATORY AND AUDITORY SPACES

To obtain the intrinsic relationship, in this study, we place both consonants and vowels together to find the common eigenvectors, and thus construct a uniform coordinate system for vowels and consonants. In this section, we separately extract the vowel and consonant structures from the uniform space, and then investigate their properties, as well as examine the relationship between the vowel structure and consonant structure within the same coordinate system.

4.1. Vowel structures in articulatory and auditory spaces

Since the spaces constructed by Laplacian eigenmaps are data dependent, in this section, we re-examine the vowel structures in the uniform articulatory and auditory spaces with vowels and consonants put together. Figure 1 plots the vowel structures for one speaker (MH), where the left panels (a) and (c) show the articulatory structures at different viewpoints, and the right panels (b) and (d) show the auditory structures at different viewpoints. We used ellipsoid to represent the distribution of the vowels, in which about 68% of vowels within the standard deviation can be included. Ellipsoids in different colors correspond to different vowels. From Fig. 1, one can see that the vowels have a relatively consistent structure in articulatory and auditory spaces. That is, the vowels shaped a triangle in the first and second dimensions, where /i/ is located in the vertex while /e/, /a/ and /u/, /o/ are located in the two wings, respectively. With reference to vowel articulation, the first dimension is related to the degree of tongue-palate approximation, i.e., low-back vs. high-front, and the second dimension corresponds to the opening ratio of the mouth to the oral cavity. This result is almost the same as that in Dang et al. [4, 5], which indicates the basic structure of the vowels was not affected when consonants were added.

4.2. Consonant structures in articulatory and auditory spaces

We extract consonant structures from the uniform articulatory and auditory spaces, and show them in Fig. 2 for speaker MH. The left panels (a) and (c) show articulatory structures in different viewpoints, and the right panels (b) and (d) show auditory structures in different viewpoints.

In Figs. 2(a) and 2(c), as indicated by the labels, the consonants are well clustered into several regions in the articulatory space. The palatal consonants /k/ and /g/, which have a closure at tongue dorsum with the soft palate, are distributed with heavy overlap. With reference to the palatal consonants, roughly speaking, the labial-related consonants /b/, /p/, /m/, /w/, and /f/, /h/ are located on the right side in the 2D view of (a), while the lingual-related consonants are on the left side. For the lingual-related consonants, the apical consonants /d/, /t/, /n/, /s/, /z/, and /T/, which obstruct the tongue apex, are located on the tip of the wing, while the laminal-related consonants /x/, /y/, /j/, /c/, and /r/, which obstruct the air passage with the blade of the tongue, are located between the apical and palatal consonants. It seems that the distance of the consonants to the reference consonants /k/ and /g/ is related to their constriction location to the reference to some extent. The relation indicates that consonant structure in the articulatory space preserved the inherent properties of their articulation. In contrast, /N/ and /h/ are distributed much wider than other consonants since their articulatory constrictions are not sensitive so that they are largely affected by coarticulations.

In Figs. 2(b) and 2(d), the consonants are also clustered into several regions in the auditory space. However, the distribution is largely different from the one in the articulatory space. For example, in articulatory space, /p/ and /b/, /t/ and /d/, /k/ and /g/ respectively belong to the same groups, while in acoustic space they are recombined into a voiceless consonant group (/p/, /t/, and /k/) and a voiced group (/d/, /b/, and /g/), and located distantly. By treating /j/ and /z/ as a break-point, voiced consonants are located on the left and voiceless consonants are on the right in the 2D view of Fig. 2(b).

If all data were plotted on the figure, we would see that the distribution of consonants forms a hoe-like shape, where the voiced consonants /b/, /g/, /d/, the nasal consonants /m/,/n/, /N/, and /w/, /r/ are fully distributed on the blade of the hoe; the voiceless fricatives

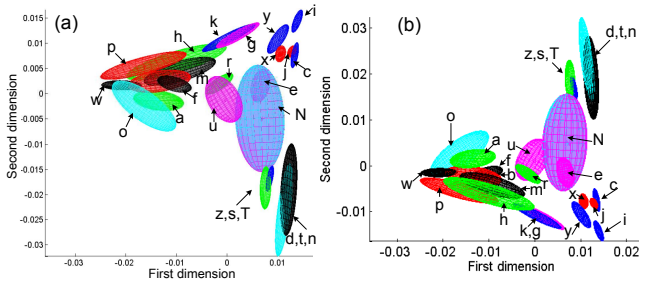

Fig. 3. Vowel and consonant structures in articulatory space.

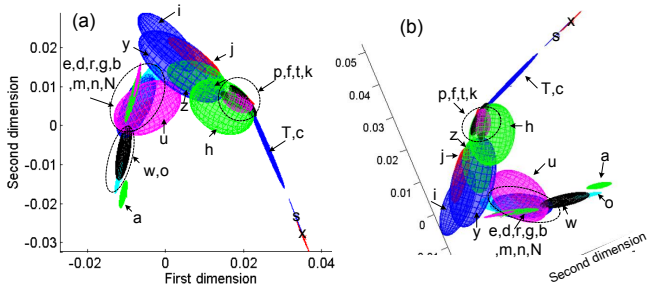

Fig. 4. Vowel and consonant structures in auditory space.

/s/, /x/, /c/ are located at the far end of the hoe handle; the voiceless consonants /h/, /p/, /f/, /t/, /k/ are next to the fricatives, and /y/, /j/, /z/ are located in the connection between the blade and the handle. In general, the properties of the sound source dominate the acoustic space, which is unique to consonants.

4.3. Relationships between vowel and consonant structures

To investigate the relationship between vowel structure and consonant structure, we plotted both vowels and consonants together in the uniform articulatory and auditory spaces, respectively.

Figures 3(a) and 3(b) show the articulatory structures of vowels and consonants at different viewpoints. Since consonant structure is introduced to the same space, the vowel structure is relatively reduced. The vowels are located on the periphery of the vowel-consonant space. The curvature presented in the vowel-alone space cannot be found in this space. However, both vowels and consonants maintain their original structures in the uniform coordinate system. In articulatory space, the consonants /x/, /y/, /j/ and /c/ are located near /i/ because their articulatory places are closer although their manners are quite different. For the same reason, /w/ is close to /o/.

Figures 4(a) and 4(b) show the auditory structures of vowels and consonants in different views. As mentioned earlier, the distribution with all consonants still forms a hoe-like shape in auditory space, where the vowels are fused with the blade of the hoe in their original triangle shape. In a close-up view, the semi-vowel /w/ is located in the neighboring regions of /o/, since they have similar acoustic features. For the same reason, semi-vowel /y/ is closed to vowel /i/.

4.4. Similarity between the speakers

The previous sections demonstrated the vowel and consonant structures, and analyzed the relationship between vowels and consonants using the data from speaker MH. In this section, we use a similarity

Table 1. The similarities of the structures between three speakers.

Spaces		MH vs. TM	MH vs. TO	TM vs. TO
Vowels	Arti.	0.9996	0.9999	0.9998
	Acous.	0.9996	0.9999	0.9992
Consonants	Arti.	0.9991	0.9983	0.9995
	Acous.	0.9253	0.9263	0.9993
Vow. and Cons.	Arti.	0.9998	0.9149	0.9146
	Acous.	0.9136	0.9979	0.9137

to verify whether or not the vowel and consonant structures are universal for other speakers. To do so, we construct the structures for the other two speakers TM and TO, and define an average correlation coefficient (ACC) to measure the similarity between the structures.

The ACC is defined as follows. For a given vowel and/or consonant structure, we calculate the centroid point of the whole structure and the center point of the distribution for each phoneme. Thus, the vectors are obtained for all phonemes from the centroid point of the structure to the phonemes. Then, we choose the vector of a phoneme with a larger number of data to be a reference vector. For other phonemes, we calculate the included angle of their phoneme vectors to the reference vector and normalize the amplitude of the phoneme vectors. As the result, each phoneme is described by a new vector with a normalized amplitude and a cosine value of included angle in the vowel and consonant structures. To measure the similarity of the two structures, we calculate the correlation coefficient for the corresponding phonemes in different structures and take the average of all the phonemes, named an average correlation coefficient (ACC). The ACC is ranged between -1 to 1, where -1 represents completely the opposite, and 1 represents exactly the same.

The similarities were calculated between every two speakers, i.e., MH and TM, MH and TO, TM and TO. Table 1 lists the ACC for speakers in articulatory and auditory spaces under several conditions. For vowel condition, the ACCs are approximately equal to 1, which means the three speakers have the same vowel structure in both articulatory and acoustic spaces. For consonant space and vowel-consonant space, the ACCs ranged between 0.9136 and 0.9998, which indicates that the three speakers have almost the same structures under all conditions.

5. CONCLUSIONS

In this study, we constructed the uniform articulatory space and auditory space for vowels and consonants using Laplacian eigenmaps. Based on the results, we found that (1) In the uniform space, the vowel structure was not affected when consonants were introduced to the same space; (2) The distribution of consonants in articulatory space was dominated by the articulatory properties, which could be classified into labial and lingual groups, while the auditory space was governed by the property of sound source, whose distribution was clustered according to voiced and unvoiced, plosive, and fricative properties; (3) In vowel-consonant acoustic space, the consonants distributed as a hoe-like shape, with voiced consonants located on the blade of the hoe and fused with vowels; and (4) Average correlation coefficients, i.e., ACC, indicated the vowel and consonant structures have high consistency for the three speakers.

This study gave the first insights into the manifold structures of consonants in articulatory and auditory spaces, and clarified the relation between the vowels and consonants in articulatory and auditory spaces. How to apply such findings in speech signal processing is one of future works.

6. REFERENCES

[1] A. Liberman and G. Mattingly, "The motor theory of speech perception revised," in *Cognition*, vol. 21, pp. 1-36, 1985.

[2] E. Casserly and D. Pisoni, "Speech perception and production," in *Wiley Interdisciplinary Reviews: Cognitive Science*, pp. 629-647, 2010.

[3] K. Honda, "Organization of tongue articulation for vowels," in *J. Phonetic*, vol. 24, pp. 39-52, 1996.

[4] J. Dang, X. Lu, M. Tiede, and K. Honda, "Inherent vowel structures in speech production and perception spaces," in *Proceedings of ISSP*, Strasburg, France, 2008.

[5] J. Dang, M. Tiede, and J. Yuan, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," in *Proceedings of Interspeech*, pp. 2815-2818, 2009.

[6] X. Lu and J. Dang, "Vowel production manifold: intrinsic factor analysis of vowel articulation," in *IEEE Trans. on Audio, Speech, Language Processing*, vol. 18, no, 5, pp. 1053-1062, 2010.

[7] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," in *J. Acoust. Soc. Am.*, vol. 62, no. 3, pp. 693-707, 1977.

[8] M. T. Jackson, "Analysis of tongue positions: Language-specific and cross linguistic models," in *J. Acoust. Soc. Am.*, vol. 84, no. 1, pp. 124-143, 1988.

[9] P. Hoole, "On the lingual organization of the German vowel system," in *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 1020-1032, 1999.

[10] Y. Zheng and H. J. Mark, "Analysis of the three dimensional tongue shape using a three-index factor analysis model," in *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 478-486, 2002.

[11] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," in *J. Acoust. Soc. Am.*, vol. 110, no. 1, pp. 453-463, 2001.

[12] J. Dang, M. Honda, and K. Honda, "Investigation of coarticulation in continuous speech of Japanese," in *Acoustical Science and Technology*, vol. 25, no. 5, pp. 318-329, 2004.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," in *Neural Comput.*, vol. 15, no. 6, pp. 1373-1396, 2003.

[14] K. Wang and S. Shamma, "Spectral shape analysis in central auditory system," in *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 382-395,1995.