Investigation of Learning Trajectory of Mandarin for Tibetan Speakers

Huixia Wang*, Jianwu Dang*,†, Hui Feng*, Hongcui Wang*, Yang Yu*, Kiyoshi Honda*,†
*The Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China
E-mail: I1965843127@163.com, fenghui@tju.edu.cn, hcwang@tju.edu.cn Tel: +86-22-27407757
† Japan Advanced Institute of Science and Technology, Japan
E-mail: jdang@jaist.ac.jp Tel: +81-07-61511235

***Abstract*— When acquiring a spoken language, speakers adapt their speech organs to the articulatory manners and places and thus form a specific articulatory space. For second language (L2) learning, the acquisition process is essentially the one for the learners to approach the standard speech of native speakers in articulatory and acoustical domains by overcoming interferences from their first language. Based on this view, this study investigates learning trajectories in L2 acquisition by measuring the changes in the distance between vowel structures observed from L2 learners and the natives. The vowel structures were constructed for six Mandarin vowels from about 7,200 tokens using the Laplacian eigenmaps. Each vowel was described by a vector with the phase and amplitude of its cluster center and its standard deviation in the vowel structure, and the distance was defined by the difference of the vectors between the learners and the natives. 30 Tibetan Mandarin learners with three Mandarin proficiency levels and 10 Mandarin native speakers participated in this study. The results showed that: i) the higher the Mandarin proficiency of Tibetan speakers is, the closer the structure of their Mandarin vowels is to that of the natives, and ii) the learning trajectory of the six Mandarin vowels, except for vowel /o/, shows a monotonic tendency, that is, as the Tibetan speakers' Mandarin proficiency level increases, the distance reduces for each vowel.**

I. INTRODUCTION

When producing vowels of a certain language properly, speakers adapt their speech organs to vowels' articulatory settings, and thus they form their own vowel structure in articulatory and acoustic domains. In perception of vowels, listeners utilize the limited space within their own oral cavity to adjust the whole acoustic structure of vowels to differentiate the heard vowels. Previous studies have revealed a high correlation between the structural size and the perceptual distinction among vowels [1]. Second language (L2) acquisition process of vowels is essentially the process for the learners' performance to approach the structure of the vowels produced by native speakers.

So far, studies on L2 vowel acquisition concerned more with the characteristics of individual vowels than with the relation among vowels [2], [3], [4]. Considering that a phoneme is only an element in the sound system of a language while maintaining a certain relation to other elements in that system [5], the interrelations among the vowels, not each individual vowel, should be taken into account in the study of developmental changes in L2 vowel production. Due to the difficulties in measuring the interrelations, few studies have been carried out on the developmental changes of Mandarin Chinese (MC) vowels by Tibetan speakers, which may presumably result in the inefficient training of MC vowels on the part of the teachers, and poor performance at learning MC vowels on the part of the students.

The traditional methods to depict the characteristics of vowel acquisition may bring about unnecessary difficulties in the systematic measurement of the developmental changes of L2 learners' performance (L2 learning trajectory) that vary along the learning stages. Of the two frequently used methods, one is to adopt the correctness score for each vowel based on the judgment by native listeners in forced-choice vowel identification, as seen in [6] and [7]. This method may place a heavy burden on native listeners at judgement when the number of speech samples is large. The other is to calculate the Euclidean distance of each vowel between L2 learners and native speakers, but the common vowel representations were only the first two or three formants, as found in [2], [3], and [4]. Such a method fails to take into account the fact that only the characteristics of individual vowels, rather than the interrelations among the vowels, could be obtained by utilizing the method, which in turn may result in an improper standard to judge the accuracy of vowel production by an L2 learner.

Based on the above considerations, we investigate the learning trajectory of Mandarin vowels by Tibetan speakers by measuring the distance of observed vowel structures between the Tibetan speakers and the Mandarin native speakers. We focus on the structure of vowels of the data, rather than each individual vowel. Each vowel is described by a vector with the phase and amplitude of its cluster center, and its standard deviation in the vowel structure, with the distance defined by the difference of the vectors between the learners and the natives.

II. METHOD

This section briefly introduces the principle of the nonlinear analysis method based on the Laplacian eigenmaps that is used for constructing low-dimensional vowel space, and describes the method used in the distance measurement.

A. Method for exploring vowel space

Wang et al. [8] suggested that the auditory image (acoustic space) can be represented by an affine transform of a logarithmic spectrum. Following this suggestion, the Mel Frequency Cepstral Coefficient (MFCC) was adopted as a set of acoustic features in constructing the vowel space. Since Chinese and Tibetan are tonal languages, another set of features used in this study is the fundamental frequency (F0). Speech segments were extracted from stable periods in monosyllables for the six vowels, /a, o, e, i, u, y/, and were used to calculate the MFCC parameters. The 14-dimensional MFCC was used, while the zeroth-order MFCC was discarded because it only represents the energy information of the vowels. The F0 for each vowel is represented by three values that extracted from the initial, middle, and final parts of each vowel segment. Thus, the acoustic feature of each vowel includes 16 dimensions.

In this study, a vowel space is constructed using the Laplacian eigenmaps to reduce the 16 dimensions of the feature vector to three dimensions. This characterization method based on inherent similarity is appropriate because the topological relationship of the vowel data could be retained even if only a few of the principal dimensions are used [10]. For a general description, a vowel is represented as a vector in the acoustic space. Thus, all the vectors for the vowels form a set $X = \{X_i \in R^n, i = 1,2,...,N\}$, where N is the data number. The similarity of the vowel vectors is described by a non-linear distance between one and another as,

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

where w_{ij} is the distance between the vowel's acoustic data X_i and X_j. σ is the heat kernel of the data. A vowel's acoustic feature vector is regarded as a point in the acoustic space. A similarity graph is constructed by connecting the point (vertex) to its neighbors in the given space, where two neighboring vertices are joined by a connector (edge) with a weighting coefficient of the distance [8]. Thus, a distance matrix W can be obtained from the graph as follows,

$$\begin{aligned} W &= [W_1, W_2, ..., W_i, ..., W_N] \\ W_i &= [W_{i,i(1)}, W_{i,i(2)}, ..., W_{i,i(k)}] \end{aligned} \tag{2}$$

where $i(k)$ is the k-th nearest neighbor of the vertex i. Based on the vertices and edges, a Laplacian graph is constructed to simulate the Laplace-Beltrami operator of manifold [9]. Then, a "neighborhood keeping" map can be obtained from the discrete graph by minimizing the objective function,

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

where L is the Laplacian matrix calculated using

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

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

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

The i-th vector can be described in a dimensionally-reduced space as,

$$X_i \to [\hat{f}_1(X_i), \hat{f}_2(X_i), ..., \hat{f}_j(X_i), ..., \hat{f}_{n_0}(X_i)] \tag{6}$$

where $\hat{f}_j(X_i)$ is the projection on the space, and n_0 is the dimensions of the reduced space. n_0 value is chosen to be three in this study. From the above, the acoustic space of vowels is visualized in three dimensions.

B. Speech materials and subjects

Speech sounds produced by Tibetan speakers and native Mandarin speakers are collected in a sound-proof studio. The recordings were stored with a sampling frequency of 16 kHz. Speech materials are Mandarin monosyllables, and the six foundational Mandarin vowels /a, o, e, i, u, y/ were segmented for this study. Subjects in our experiment are 30 male Tibetan speakers and 10 male native Mandarin speakers. The Tibetan speakers are college students with different Mandarin proficiency levels. The native Mandarin speakers are from Beijing with the average age of 23.

Mandarin proficiency scores of Tibetan subjects followed the normal distribution, whose mean value was 53, and the standard deviation was 14. Then, boundaries to classify the subjects' levels were set at halves of the standard deviation below and above the mean. As a result, the 30 Tibetans were divided into three groups according to their Mandarin proficiency levels. The subjects in Group 2 were chosen with their score between 47 and 60. The scores of the subjects in Group 1 were lower than 47, and those in Group 3 were higher than 60. Details of the three groups are shown in Table 1.

TABLE I
DETAILS OF SUBJECTS IN THE THREE GROUPS

Information / Groups	Subjects' Number	Mean score	Mandarin proficiency level
Group 1	10	36	Beginner level (25<score<=46points)
Group 2	10	54	Intermediate level (46<score<=60points)
Group 3	10	68	Advanced level (60<score<80points)
Control Group	10	—	Native level

Finally, the extracted vowel segments were about 450 for /a/, 410 for /e/, 320 for /i/, 210 for /o/, 250 for /u/, and 100 for /y/ in each learner group (Group 1, Group 2 and Group 3). The segments in Control Group were about 510 for /a/, 440 for /e/, 360 for /i/, 260 for /o/, 290 for /u/ and 190 for /y/. All the vowel segments were together used for constructing the vowel space based on the method described in Section IIA. The four structures of vowel were obtained by projecting the data of the four groups in the space separately.

C. *Distance measure of a vowel between two different vowel structures*

After constructing the vowel structures, the distance of vowels between the vowel structures of Tibetan speakers and native speakers was used to evaluate the acquisition trajectory of Mandarin vowels by the Tibetan speakers. It is expected that the smaller distance means that the vowels produced by L2 learners (Tibetan speakers) are acoustically closer to the standard vowels produced by native speakers.

Here, we denote the Mandarin vowel structure obtained from Tibetan speakers X space and call the native speakers' vowel structure Y space. Then, a vowel p in X space is formed as a set X_{Ap}, $X_{Ap} = \{X_{ip} \in R^3, i = 1,2,..., M_1\}$ where M_1 is the data number of vowel p in X space. In the same way, the vowel p in Y space is formed as a set Y_{Ap} $Y_{Ap} = \{Y_{ip} \in R^3, i = 1,2,..., M_2\}$.

To calculate the distance of each vowel between X space and Y space, it is necessary to extract the vowel structure from the vowel space. The vowel structures with the six vowels (a, i, u, o, e, y) are shown in Fig. 1. In the vowel structures, each vowel distributes as a cluster. To describe the spatial relation among the vowels, the first to consider are the location and dispersity of the cluster in the vowel structure. To do so, we first define the center point of the whole structure as the origin of the space, denoted by C, and then find out the central point for each vowel cluster, represented by p. The vowel location can be described using a vector from the origin to the central point of its cluster using the amplitude and phase values. The standard deviation (SD) of the cluster is used as the dispersity of the vowel. Now, we define $Xp = \overrightarrow{Cp}$ (p is one of six vowel central points), and the Xp_i represents the value in the i-th dimension of Xp. The definition in Y space is the same as in X space. According to the above, the amplitude discrepancy of each vowel between X and Y spaces can be obtained by,

$$A_{p_{XY}} = \frac{\| Xp \| - \| Yp \|}{\| Xp \| + \| Yp \|} \tag{7}$$

The phase (angle) discrepancy of each vowel between X and Y spaces can be obtained by,

$$\beta_{p_{XY}} = s_p * \sin(\arccos(\frac{X_p Y_p}{\| X_p \| . \| Y_p \|})) \tag{8}$$

where s_p is the direction of angular deflection. Compared to the Y space, when the direction is clockwise rotation, the value of s_p is -1, otherwise it is 1. The SD discrepancy of each vowel between X and Y spaces can be obtained by,

$$S_{p_{XY}} = \frac{SD(X_{Ap}) - SD(Y_{Ap})}{SD(X_{Ap}) + SD(Y_{Ap})} \tag{9}$$

Finally, the distance of each vowel p between X and Y spaces can be obtained by,

$$D_{p_{XY}} = \sqrt{A_{p_{XY}}^2 + \beta_{p_{XY}}^2 + S_{p_{XY}}^2} \tag{10}$$

Fig. 1 also illustrates an example for the three distances of the vowel /u/ between X and Y spaces. Therefore, the distance between Tibetan-Mandarin (T-M) and Standard-Mandarin (S-M) vowels can be a measure for the vowel structure for different speakers.

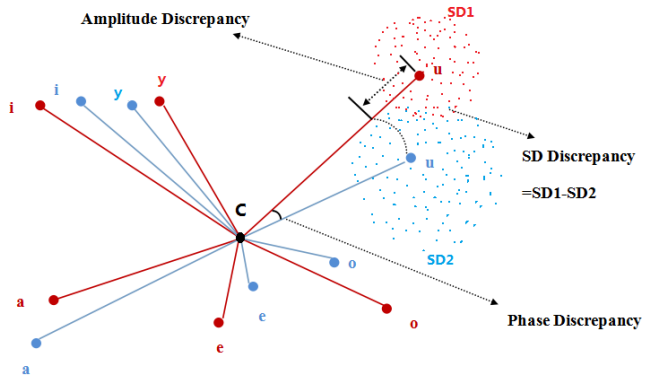

Fig. 1 2D structure of six Mandarin vowels for native speakers (red) and Tibetan speakers (blue).

III. Results and Discussions

A. *Mandarin vowel spaces for Tibetan speakers and native speakers*

Using all the collected data produced by the subjects in four groups as introduced in section II, we first construct a discrete graph with each acoustic sample as one of the vertices of the graph for the four groups. For a given vertex on the graph, we choose a neighboring region patch and establish edges between the given vertex and the vertices in this patch [9], [11]. In this study, we choose eleven nearest neighboring vertices as the neighboring region patch for the given vertex. Then, using the eigenvalue decomposition based on (2), (3), (4), (5) and (6), the mapping function is obtained, which is used for low-dimensional embedding. Then, the four vowel structures were obtained by projecting the acoustic data of the four groups in the space separately. Finally, the four vowel spaces with the low dimensions are derived from the original high dimensional data, while their topological relationships are preserved. The vowel space for all the acoustic data in each group is shown in Fig. 2 (in 2D) and Fig. 3 (in 3D).

From Fig. 2 and Fig. 3, the vowel distribution in Standard-Mandarin (S-M) vowel space of the native speakers is distinguished much clearly and shows smaller overlapping. Vowel distribution in Group 1 with the beginner level is located in a narrow band comparing with Group 2 (intermediate) and Group 3 (advanced). Group 3 demonstrated a more clearly distinguished structure than Group 2. Compared with the graphic pattern of the vowel space in Group 1, the pattern in Group 2 is closer to that in Control Group (native speakers). Compared with Group 2, Group 3 is further closer to Control Group.

Thus, as the Tibetan speakers' Mandarin proficiency level increases, the vowel structure of learners approaches the Standard-Mandarin vowel structure.

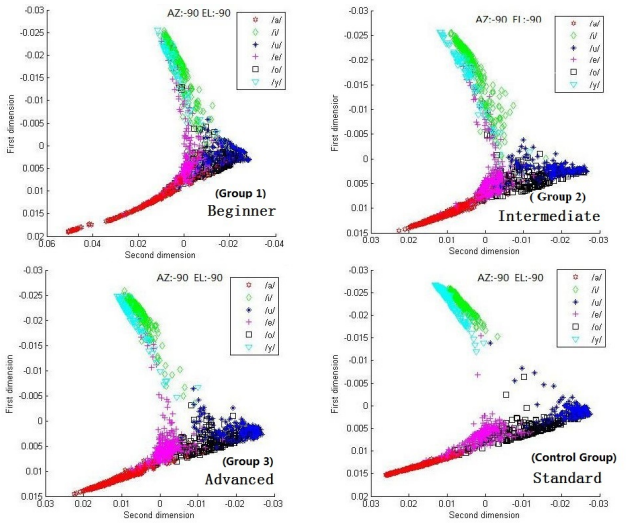

Fig. 2: 2D vowel spaces for six Mandarin vowels in four groups with different Mandarin proficiency levels (beginner level, intermediate level, advanced level, and native level).

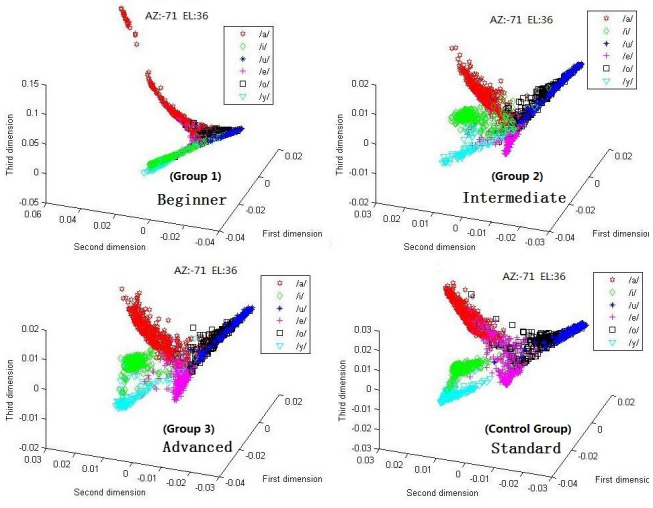

Fig. 3: 3D vowel spaces for six Mandarin vowels in four groups with different Mandarin proficiency levels (beginner level, intermediate level, advanced level, and native level).

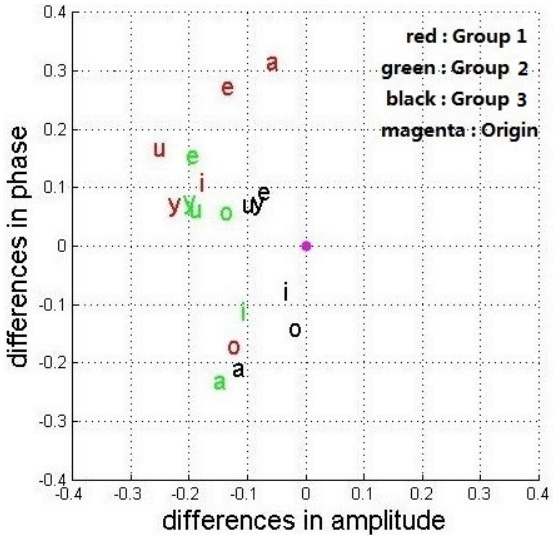

Fig.4: Amplitude difference vs. Phase difference of each Mandarin vowel between subjects in the learners' groups and native subjects in Control Group.

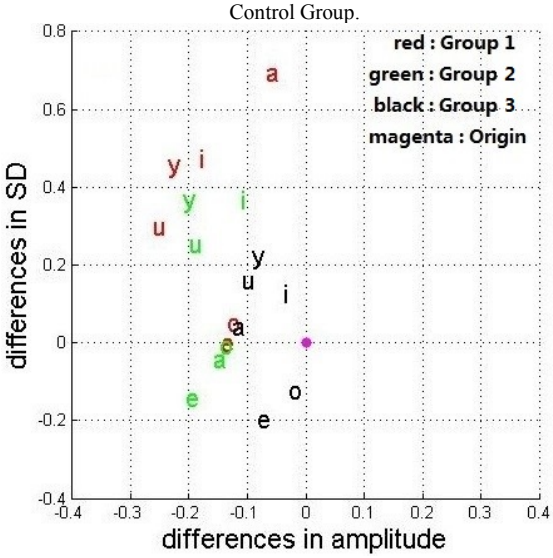

Fig.5: Amplitude difference vs. SD difference of each Mandarin vowel between subjects in learners' groups and native subjects in Control Group.

B. Learning trajectory of each Mandarin vowel produced by learners

Using the proposed method described in section IIC, we measured the distance of Mandarin vowel between the three Tibetan groups and Control Group. Fig. 4 and Fig. 5 show the relations of the amplitude, phase and standard deviation (SD). In these figures, the vowels close to the origin mean that the vowels have similar distribution to the native vowels in the vowel space.

In Fig. 4, the position of the vowels closely approaches the origin in the order of Group 1, Group 2, and Group 3. This means that Group 3 is closest to Control Group in the plane of the phase and amplitude among the three Tibetan groups.

Fig. 5 shows the amplitude-SD plane. The vowels, except for vowels /o/ and /e/, have the same tendency as those in Fig. 4. That is, the position of the vowels approaches the origin closely in the order of Group 1, Group 2, and Group 3. However, vowel /o/ shows similar distance for the three groups. Vowel /e/ in Group 2 has larger distance than in Group 1 and Group 3, while it shows similar distance for Group 1 and Group 3.

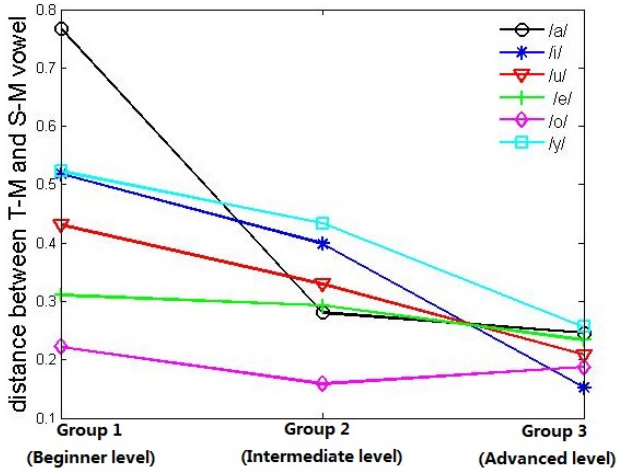

Fig.6: Distance between each Tibetan-Mandarin (T-M) and Standard-Mandarin (S-M) vowel for three groups of learners with different Mandarin proficiency level.

Fig. 6 shows the learning trajectory of Mandarin vowels for the three groups, where the horizontal axis is the Mandarin proficiency level, and the vertical axis is the distance between L2 learners and natives. The distance between T-M and S-M decreases as the proficiency level increases. This implies that the learning trajectory of the six Mandarin vowels monotonically changes with the proficiency levels, where vowel /o/ was not improved with the level changes. For the regression problem in the acquisition process of vowel /o/ by learners, Ellis (1994) argues that making error is common in the compulsory period in the acquisition process [12]. Thus, we believe that the regressing in the acquisition process of vowel /o/ is reliable.

IV. Conclusion

In our study, we investigated the learning trajectories of Mandarin vowels in second language acquisition by Tibetan speakers. 30 Tibetan speakers were divided into three groups according to their Mandarin proficiency levels. We constructed acoustically-based vowel structures for the three groups, and the native Mandarin speakers are used as a control group. The distances were measured between the L2 learners and the natives according to their vowel structures. It is found that i) the higher the Mandarin proficiency of Tibetan speakers is, the closer the structure of their Mandarin vowels is to the structure of standard Mandarin vowels, and ii) the developmental trajectories of the six Mandarin vowels follow a linear pattern, that is, except for vowel /o/, as the Tibetan speakers' Mandarin proficiency level increases, the distance for each vowel between their Mandarin and standard Mandarin decreases. The trajectory of Mandarin vowel /o/ by Tibetan speakers shows subtle regression in the acquisition process, that is, Tibetan speakers with an intermediate level of Mandarin proficiency perform slightly better than those with the advanced level in the production of Mandarin /o/.

To conclude, this study gives the first insight of manifold structure of Mandarin vowels produced by Tibetan speakers. We believe that the new research method employed in this study will be a reference to the studies on the developmental changes of phonetic acquisition by L2 learners.

Acknowledgment

The research is supported by the National Basic Research Program of China (No. 2013CB329301), and the National Natural Science Foundation of China (No. 61303109).

References

[1] N. Minematsu, S. Asakawa, and K. Hirose, "Paralinguistic information represented as distortion of the acoustic universal structure in speech," *Proc. Int. Conf. Acous. Speech and Signal Processing. Toulouse,* pp. 261-264, May 2006.

[2] L. Zhao, H. Feng, H. Wang, "Acoustic features of Mandarin monophthongs by Tibetan speaker," *Proc. Asian Language Processing (IALP). Malaysia,* pp.147-150, 2014.

[3] C. Chen, K. Lin, C. and C. Wu, et al. "Acoustic study in mandarin-speaking children: developmental changes in vowel production," *Chang Gung Med J*. vol. 31, pp. 503-509, 2008.

[4] V. Dommelen and W A, "The production of Norwegian vowels by French and Russian speakers," *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS). Germany*, pp.1633-1635, 2007.

[5] H. A. Gleason, *An introduction of descriptive linguistics*, New York: Holt, Rinehart & Winston, 1961.

[6] S. Y. Cheon, "Effects of phonetic similarity and L2 experience: production of English /s/-/□/ by adult Korean ESL learners," *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS). Germany*, pp.1697-1700, 2007.

[7] V. Hazan, "Second language acquisition and exemplar theory," *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS). Germany*, pp.43-48, 2007.

[8] K Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Processing*. vol. 3, pp. 382-395, 1995.

[9] J. Dang, M. Tiede and J. Yuan, "Comparison of vowel structures of Japanese and English in articulatory and auditory spaces," *Proc. INTERSPEECH. United Kingdom*, pp. 2815-2818, 2009.

[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*. vol. 15, pp. 1373-1396, 2003.

[11] X. Lu and J. Dang, "Vowel production manifold: Intrinsic factor analysis of vowel articulation," *IEEE Transactions on Audio, Speech, and Language Processing*. vol. 18, pp.1053-1062, 2010.

[12] E. Rod, *The study of second language acquisition*. Oxford University Press, Oxford, 1994.