Vowel Normalization by Articulatory Information

Jingshu Zhang[1], Jianguo Wei[1,*], Wenhuan Lu[2], Qiang Fang[3], Kiyoshi Honda [1,4] and Jianwu Dang[1,4]
[1]Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University,
[2]School of Computer Software, Tianjin University,
92 Weijin Road, Nankai District, Tianjin 300072, China
[3]Chinese Academy of Social Sciences, Beijing,
5, Jianguomennei Dajie, Beijing 100732, China
[4]School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: jingshu@tju.edu.cn; Jianguo@tju.edu.cn

***Abstract*— The differences of vowel sounds among speakers are mainly caused by the morphological differences and different speaking styles of each speaker. The traditional vowel normalization methods in acoustic space are mainly concerned with the variance of acoustic features. There is no information of articulatory space taken into account. This paper proposed an approach to normalize vowel spectra by using articulatory information. The articulatory information of this study is mainly related with morphological variations of speakers. By taking articulatory information into account, the normalization method will have clear physical meaning that which part of acoustic variations has been reduced. This paper proposed a vowel normalization framework by using Thin-plate spline method. Thin-plate spline method was applied to normalize formant frequencies of three Chinese vowels, and was compared with traditional acoustic normalization methods for evaluation. The results show that the variances among different subjects were reduced. Vowel diagrams indicate that this method outperforms other acoustic methods in keeping speakers specific characteristics.**

I. INTRODUCTION

In recent years, the speech database has been increased sharply in number, and the datasets include more speakers. However, the speaker adaptation and the system robustness still remain as the key bottle neck to influence the development. There are morphological variances of the articulatory organs among different speakers that influence variability of speech signals. Furthermore, the speaking styles are different across subjects, and observation methods are limited. Thus, the articulatory databases are not as popular as acoustic databases despite their importance. In order to discover the kinematic properties of articulatory organs among different subjects, appropriate normalization methods for articulatory data are an inevitable procedure to reduce the morphological difference of vocal tract. There are morphological variances of the vocal tract across speakers, and large deformations will happen on the articulatory organs. Hence, it is hard to handle articulatory data by linear transformation of simple rigid objects. Up to date, many normalization method of vocal tract have been proposed. For example, in the articulatory space, Beckman *et al*. [1] adopted straightening of the wall of the vocal tract to transform the coordinate values of MRI data. The normalization of vowel articulation for x-ray microbeam database was also realized by Hashi *et al*. [2]. Studies on normalization in the acoustic space are numerous. For instance in relation to the vocal-tract normalization, Michael Pitz *et al*. processed the length of vocal tract by linear transformation in the frequencies domain [3]. Furthermore, the length normalization of the vocal tract was performed by linear transformation method by Lakshmi Saheer *et al*. [4] in the spectrum space. Among the studies, it can be seen that the common method for normalization in the articulatory and acoustic spaces is the affine transformation of vocal tract length. However, these methods could not reflect the nonlinear relationship among subjects and ensure the articulatory features and characteristics of speaking styles, particularly when the highly local deformations happen on the vocal tract.

The shape of vocal tract often reflects the elastic deformation when speakers pronounce different vowels. Thus, this paper uses a framework to normalize three Chinese vowels by applying a Thin-plate spline (TPS) warping[5][6][7]. The TPS method is a kind of tool which is usually used in processing overstretched data of interpolation surface in the image field. In our study, we propose a vowel normalization framework by using the TPS method, which is applied to normalize formant frequencies of three Chinese vowels. Then, the method is compared with traditional acoustic normalization methods for evaluation. Firstly, we define the landmarks which are based on gridline systems by Electromagnetic Articulographic (EMA) data[6][7]. The framework can maintain the relative position between the palate and tongue during normalization. Then we find the mapping relationship based on landmarks between template and each subject. Finally, formant frequencies are normalized by the mapping function, that is, Thin-plate spline function.

Reducing the morphology difference of vocal tract across speakers would help to analyze the property of the speech organs and the rules of pronunciation. Moreover, it would be expected to enhance the robustness of speech recognition system. In this study, we study single kinematical behaviors of overall group of speakers and analyze the variance of

kinematical behaviors between different groups of speakers. Three subjects' articulatory data and acoustic data are from EMA database. In this study, we evaluate the performance of the TPS method in the acoustic space and compare the results with other methods.

II. MATERIALS AND METHOD

A. EMA database and subjects

The dataset used in this study is a Chinese EMA database includes articulatory and acoustic data of /a, i, u/ from 3 Chinese subjects. In the first experiment, we extracted 192 configurations of the vocal tract for each vowel in different sentences from Chinese EMA database. The articulatory data is mainly referred to the two-dimension data (horizontal direction and vertical direction) collected from 4 tongue sensors in EMA system. The subjects include two male and one female and Chinese EMA database includes articulatory and acoustic data.

B. Preprocessing

In order to transfer acoustic data by using the transfer function obtained from articulatory space, we normalized the articulatory data and formant frequencies by Min-Max Normalization method, so as to make both articulatory and acoustic data in the same range. Min-Max Normalization is a kind of linear transformation method. We can see that the standard deviations of raw data reduced the same times after Min-Max-Normalization from Table I and Table II. We applied Thin-plate spline method to normalize articulatory data then acoustic data were normalized. The principle of Thin-plate spline method will be shown as followed.

C. Thin-plate spline method

TABLE I
STD OF RAW ARTICULATORY DATA

Coordinate value	Vowels		
	a	i	u
x	268.44	519.07	494.46
y	177.06	127.74	99.68

TABLE II
STD OF MIN-MAX NORMALIZATION ARTICULATORY DATA

Coordinate value	Vowels		
	a	i	u
x	0.27	0.51	0.49
y	0.15	0.11	0.08

In this paper, the formant frequencies were normalized by Thin-plate spline method based on articulatory information, that is, we apply the same idea which is used to normalize the vocal tract data [6] to normalize formant frequencies. We define the landmarks in the vocal tract space by a gridline system, then we find the mapping relationship among the template and subjects. Thus, the parameters of normalization need to be calculated according to the marked point of vocal tract gridline system. Suppose a set of n corresponding point were defined in 2D plane, the Thin-plate spline warp is defined by (2n+3) parameters which include 2n corresponding coefficients of reference point and 6 global affine motion transformation parameters. These parameters are calculated by solving a linear system. We suppose that $(\hat{x}_i^{'}, \hat{y}_i^{'}) \in \Re^2$, ($i$=1,.. n) are the n control points in a 2D plane and their corresponding function values are $\hat{v}_i \in \Re$, i=1,…n, The definition of Thin-plate spline interpolation $f(x, y)$ is a mapping of $f : \Re^2 \to \Re$, the formula of Thin-plate spline function is shown as followed :

$$f(x, y) = a_1 + a_2 x + a_3 y + \sum_{i=1}^{n} w_i r_i^2 \ln r_i^2 \qquad (1)$$

where $r_i^2 = (x - \hat{x}_i)^2 + (y - \hat{y}_i)^2$. (1) express that the equation is based on $(\hat{x}_i, \hat{y}_i)$ as the center and is a kind of infinite extent deformation under loads [8]. The coefficient w_i denotes the nonlinear transformation, that is, the weight value and a_1, a_2 and a_3 mean the linear conversion coefficient. In the end, r_i is the distance between the target point to the original point. The plate deflects to take values under the imposition. The Thin-plate spline interpolate on function includes two parts: three elements composed by a_1, a_2 and a_3 of a particular linear conversion coefficient. The goal is to ensure linear parallelism parallel and collinear points and do not cause changes in the shape of an object. And the other part is nonlinear part in the formula. In our study, according to the theory of TPS method, firstly, we need to find the mapping relation between the template and the three subjects. So we need to solve the (2n+3) parameters in the (1). The Thin-plate spline method is a interpolation spline function, which tries to seek a smooth surface which can pass through all the control points with minimum bending degree and ensures that the data points can be correctly matched. Hence, we add an energy constraint and three weight coefficient constraints. The minimum bending defined by an energy function is shown below:

$$E_f = \iint_{\Re} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dxdy \qquad (2)$$

E_f is the energy constraint for $f(x, y)$ to ensure the minimum second derivative of accumulation of each data point on the surface of the item. Moreover, three weight constraints are defined by:

$$\sum_{i=1}^{n} w_i = 0 \qquad (3)$$

$$\sum_{i=1}^{n} \hat{x}_i w_i = 0 \qquad (4)$$

$$\sum_{i=1}^{n} \hat{y}_i w_i = 0 \qquad (5)$$

Constraint (3) , (4) and (5) are the constraint of the weight coefficient, equation to the right value is to control the degree of smooth surface. Commonly, the value is zero. (3) shows the sum of weights should be zero for palate's smooth. Constraints (4) and (5) request that moment of the palate should not be rotate after the increase in weight respectively in the x axis and y axis direction. The parameter vector $\boldsymbol{a}$ of TPS contains a_1, a_2 and a_3, and the vector $\boldsymbol{w}$ consists of w_i , these coefficients are calculated by solving linear equation:

$$\begin{bmatrix} A & P \\ P^T & O \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix} \qquad (6)$$

where $A_{ij} = r_{ij}^2 \ln r_{ij}^2$, i=1,…n. (n is the number of reference points) j=1,…m (m is the number of normalized transformation for the original data points); the $(1, \hat{x}_i, \hat{y}_i)$ is the *i*-th row of P, and zero matrix O is 3×3. There is three zero vectors in the rightmost part of (6). The vectors of *w*, *a* and *v* are composed by w_i , by a_1, a_2 , a_3 and by v_i. The leftmost (n+3) × (n+3) matrix is corresponded as K hereafter. In our study, we concentrate on mapping points (x, y) of EMA data to template coordinates (x', y') in light of given landmarks $(\hat{x}_i, \hat{y}_i)$ for one subject's EMA data vs. So we solved the coefficient based on the mapping relation and applied the TPS interpolation function to normalize formant frequencies again. The landmarks of template is defined by $(\hat{x}_i', \hat{y}_i')$. Hence, we are interested in using Thin-plate spline method which was defined by pairs of control points in warping 2D plane points. Therefore, we used Thin-plate spline method to horizontal axis and vertical axis respectively. According to the (6), the mapping of $(\hat{x}_i, \hat{y}_i)$ to $(\hat{x}_i', \hat{y}_i')$ by Thin-plate spline warp can be regained by

$$\begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} = K^{-1} \begin{bmatrix} \hat{x}' & \hat{y}' \\ 0 & 0 \end{bmatrix} \qquad (7)$$

where $\hat{x}'$ consists of vectors $\hat{x}_i'$ and $\hat{y}'$ consists of $\hat{y}_i'$ separately. The parameters of the x axis direction vector are w_x and a_x and the parameters of the y axis direction are w_y and a_y. The coordinates (x_j', y_j') which was transformed from points (x_j, y_j) are shown by

$$\begin{bmatrix} x' & y' \end{bmatrix} = \begin{bmatrix} B & Q \end{bmatrix} \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} \qquad (8)$$

where $B_{ji} = ((x_j - \hat{x}_i)^2 + (y_j - \hat{y}_i)^2 \ln((x_j - \hat{x}_i)^2 + (y_j - \hat{y}_i)^2)$, i=1,…, n, j=1,…, m. $(1, x_j, y_j)$ is the j-th row of Q and the coordinate x_j' and y_j' which are the j-th row of the interpolated x and y coordinate x' and y' respectively[9]. As we all know, the formant frequencies are related to tongue position. The first formant frequency is correlate to the high-low of tongue position and the second formant frequency is relevant to the back and forth of tongue position. Thus, in our paper, x' and y' are the second and first formant frequencies.

III. RESULTS

In our experiments, 192 configurations were extracted of the vocal tract and formant frequencies for three vowels (a / i / u) from Chinese EMA database. The selection of landmark was referenced by [7][10]; the vocal tract gridline system for three subjects are shown in Fig.1. According to the previous study [6][7], we conclude that the variances of different subject are reduced in articulatory space.

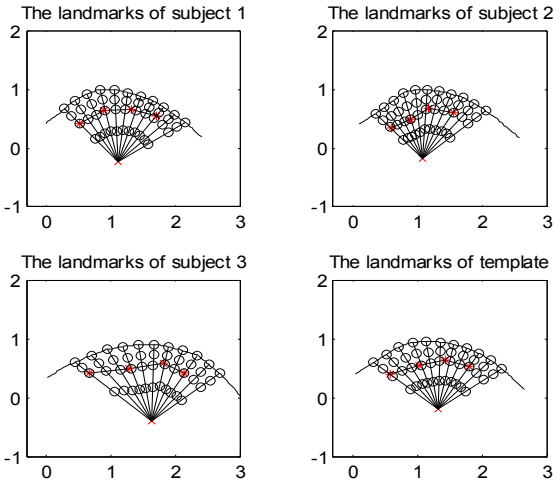

Fig. 1 Landmarks of template and three Chinese subjects. Four sensors on tongue surface, the palate on the top and the gridlines are shown. The position of the mouth is on the left side.

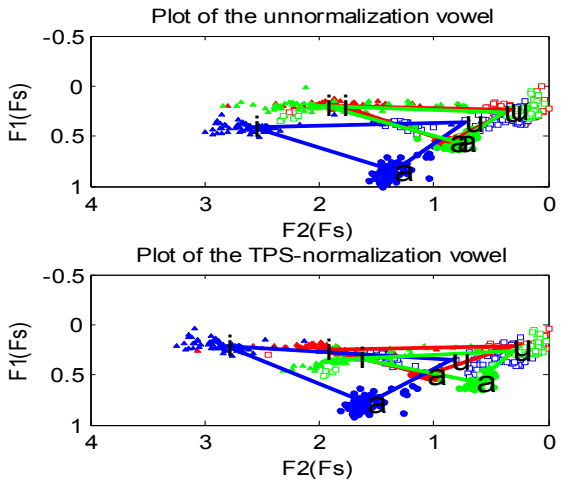

Fig. 2 Vowel space diagram of three Chinese speakers before and after the TPS normalization. Three colors correspond to the three subjects, the filled triangle denotes the distribution of 'i', the dot denotes the distribution of 'a' and the quadrate denotes the distribution of 'u'. Each closed triangle depicts different subjects' vowel space.

In our paper, the Thin-plate spline method is applied to define the relationship of landmark between the template and three subjects. Then, according to the mapping relationship, that is, the transformation coefficient, we can get the normalization results of formant frequencies. The results are shown in Fig.2.

IV. Evaluations and Discussions

In our study, we evaluate our result with raw vowel diagrams as shown in Fig.2, indicating that three vowel spaces are normalized to a horizontal dimension so that the vowel space resemble with each other Table III and IV show the formant frequencies convergence of pre-normalization and post-normalization. In this study, the convergence of data was examined by the percentage of standard deviation and average value. On the one hand, we see that the convergence of formant frequencies in Table IV are mostly lower than that in Table III. Hence, we can ensure the effect of normalization. On the other hand, comparing the pre-normalization and post-normalization of the articulatory space in Fig.3, we can clearly observe that the shapes of vowel triangles are in a close resemblance between original data and normalized data. In addition, the vowel triangle is almost overlapping after the TPS normalization, which agrees with the acoustic space. Furthermore, in order to evaluate our method on the normalization of formant frequencies, we applied standard methods in the field as a comparison. Fig.4 shows the normalization results of two vowel normalization methods that include Nearey's and Bark Difference Metric methods [11]. As we can see from the Fig.4, the coordinates denote the formants with different meaning, because there are different theories between two methods. Comparing with raw formant frequencies, we can see that the vowel triangles are highly similar to each other. However, the speaker specific characteristics do not have a good performance.

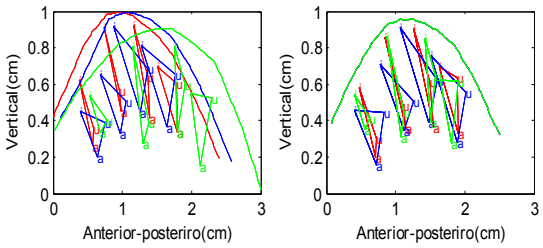

Fig.3 Vowel structure before normalization (left panel)and after TPS normalization(right panel). The curves are the subject's palate. The red and green indicate male data, and the blue shows female data. The triangles are for Chinese vowels /a, i , u/. The x-axes denotes the horizontal direction of vocal tract and the y-axes denotes the vertical direction of vocal tract.

Table III
Convergence of raw formant frequencies

Vowels	Formant frequency	convergence
a	F2	0.27
	F1	0.24
i	F2	1.01
	F1	0.23
u	F2	0.40
	F1	0.30

Table IV
Convergence of TPS-normalized formant frequencies

Vowels	Formant frequency	Convergence
a	F2	0.37
	F1	0.29
i	F2	0.87
	F1	0.23
u	F2	0.36
	F1	0.29

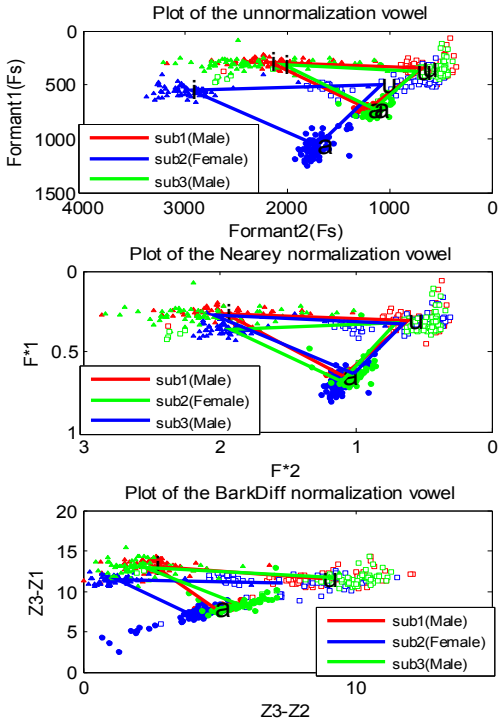

Fig.4 Vowel space diagram of three Chinese before and after normalization, three colors correspond to three subjects, the filled triangle denotes the distribution of 'i', the dot denotes the distribution of 'a' and the quadrate denotes the distribution of 'u'. Each closed triangles denote different subjects' vowel space. The coordinates denote the formants with different meanings.

V. Conclusions

In our paper, we obtained the transfer function of Thin-plate spline (TPS) method by the articulatory data and applied the TPS function to normalize the formant frequencies across three subjects. In the acoustic space, the result of normalized data showed that the variance was reduced across subjects that indicated that TPS method can apply to the acoustic space. There is not much change of vowel triangle in both pre-normalization and post-normalization. The evaluation demonstrated the Thin-plate spline based method could reduce variations across subjects and has clear physical meaning. Felicitously reducing the morphological variations of speech organs would be beneficial to data fusion across speakers and building articulatory model.

ACKNOWLEDGMENT

This work was supported in part by the National basic Research Program of China(No. 2013CB329305), and in part by grants from the National Natural Science Foundation of China(General Program No. 61175016, and Key Program No. 61233009).

REFERENCES

[1] M.E. J. Beckman, T.-P. Jung, S.-h. Lee, K. d. Jong, A. K. Krishnamurthy, S. C.Ahalt,K.B.Cohen, and M. J. Collins, Variability in the production of quantal vowels revisite-d, J. Acoust. Soc. Am,1995, 471-490.

[2] M. Hashi, J. R. Westbury, and K. Honda, Vowel posture normalization,JASA,1998, 104:. 2426–2437.

[3] Michael Pitz, Frank Wessel; Hermann Ney, Improved mllr speaker adaptation using confidence measures, The Proceedings of the 6th International Conference on Spoken Language Processing (Volume Ⅳ),2000.

[4] Lakshmi Saheer, J. Yamagishi, P.N. Garner, J. Dines, Combining vocal tract length normalization with hierarchical linear transformations, IEEE Journal of Selected Topics in Signal Processing (Impact Factor: 3.3). 2013.

[5] B.FL, Principal warps: Thin plate splines and the decomposition of deformations, IEEE Trans Pattern Anal. Mach. Int63ell, 1989,11: 567-585.

[6] Jianguo Wei and Jianwu Dang, Morphological normalization of vocal tract shape, IEEE Acoustics Speech and Signal Processing, 2010, 4186-4189.

[7] Hong Liu, Jianguo Wei, Morphological Normalization: A Study of Vowels for Mandarin and Japanese,apsipa,2013.

[8] L. Zagnonrigid orchev and A. Goshtasby, A comparative study of transformation functions for image registration, IEEE Trans. Image Processing, 2006,15: 529-538.

[9] J. Lim and M. H. Yang, A Direct Method for modeling Non-rigid Motion with Thin Plate Spline, IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2005

[10] Beautemps, D., Badin, P., and Laboissière, R..Deriving vocal-tract area function from midsagittal profiles and formants frequencies: A new model for vowels and fricative consosnants based on experimental data. Speech Communication,1995 ,16: 27-47.

[11] The Vowel Normalization and Plotting Suite. http://ncslaap.lib.ncsu.edu/tools/norm/index.php.