Morphological Normalization: A Study of Vowels for Mandarin and Japanese

Hong Liu[1], Jianguo Wei[1,*], Wenhuan Lu[2], Qiang Fang[3], Liang Ma[4] and Jianwu Dang[1,5]
[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]Department of Chinese Language and Literature, Fudan University,
220 Handan Road, Shanghai 200433, China
[5]School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: wlgc0802@sina.cn; Jianguo.fr@gmail.com

***Abstract*—Reducing the morphological variances of vocal tract across different subjects would benefit articulatory data analysis and modeling. To further test such a hypothesis by a thin-plate spline warping (TPS) method, this study used articulatory data of /a, i, u/ from 3 Chinese subjects and 3 Japanese subjects, which were collected by Electromagnetic Midsagittal Articulographic (EMMA) system. The templates for the normalization of Chinese and Japanese were obtained by averaging the 3 subjects' palates and tongue shapes in each group. The 44 landmarks in each template were then defined by a gridline system of the vocal tract. The results show that the variances among subjects were reduced in both horizontal direction and vertical direction. The similar vowel structures between pre- and post-normalization data indicate that TPS method outperforms the traditional palate-straighten method in that TPS method has reduced mid-sagittal morphological differences among speakers while keeping their individual vowel structures unchanged. The comparison results show that the articulatory differences among the three vowels are consistent with their corresponding acoustic properties.**

I. INTRODUCTION

The articulatory data are not utilized so widely as acoustic data. One of reasons is that the morphological variances among different speakers is a bottleneck for the multi-subject articulatory data study. In order to discover the intrinsic articulatory and kinematic properties of a specific language, the inter-subject normalization of articulatory data is a necessary procedure to reduce the morphological variances across subjects.

Since morphological variances among subjects involve with nonlinear deformations, it is not easy to handle by affine transformation like simple rigid objects. Several vocal tract normalization techniques have been proposed in articulatory space. Bechman *et al.* [1] straightened the vocal tract wall to transform the coordinates of MRI data. Hashi *et al.* [2] normalized the vowel posture for an x-ray microbeam database. Both the two methods straighten the palate wall to normalize the articulatory data, which could not guarantee the intrinsic articulatory space after transformation.

* Corresponding Author;

The vocal tract shape usually reflects local and nonlinear deformations which can be treated as an elastic-deformation. Accordingly, our previous study used a thin-plate spline warping (TPS) [3][4] to normalize inter-subject' Japanese EMMA data. In the present study, Chinese and Japanese EMMA database were used for the normalization. We evaluated the performance of TPS method by not only reducing inter-speaker variances but also keeping speaker-specific characteristics in articulatory space. Finally, a comparison of vowels was carried out between normalized Mandarin and normalized Japanese in articulatory space.

II. ELASTIC DEFORMATION OF VOCAL TRACT

In order to reduce the morphological variances among the speakers, a normalization method [1, 2] was to straighten the vocal tract and then normalize the length of the vocal tract, which maintained the constriction constant. According to the results shown in [5], however, the inter-speaker variances is not only related to the vocal tract length but also the volumes of back and front cavities of the vocal tract. However, this method did not take into account of the non-linear elastic nature of vocal tract deformation. Furthermore, the relative positions of different sensors attached to the articulators were lost after normalization, this possibly lose the kinematic properties of articulators.

The thin-plate spline is a class of non-rigid spline mapping functions with several desirable properties, which meets our purpose. They are globally smooth, separable into affine and non-affine components and transforming source data to target data according to physical feature based landmarks.

Given n points in 2D plane, the TPS is described by 2(n+3) parameters, which include 6 global affine motion parameters and 2n coefficients for correspondences of the control points. These parameters are computed in a linear system [6]. Suppose $(\hat{x}_i, \hat{y}_i) \in \Re^2$, i=1,…,n, are the n control points in a plane, and their corresponding function values are $\hat{v}_i \in \Re$, i=1,2,…,n, then the thin-plate spline interpolation $f(x, y)$ defines a mapping $f : \Re^2 \to \Re$, describing as follows:

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

where $r_i^2 = (x-\hat{x}_i)^2 + (y-\hat{y}_i)^2$.

Eq. (1) is the equation of a plate of infinite extent deforming under loads centered at $(\hat{x}_i, \hat{y}_i)$. The plate deflects under the imposition of loads to take values w_i [6]. The interpolation spline function consists of two parts: affine transformation specified by the first three elements, and the last warping part. The function f minimizes the bending energy E_f over the class of such interpolations where E_f is defined as:

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

The following three equations work as constraints in TPS

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

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

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

Constraint (3) shows that the sum of the loads applied to the plate should be zero. This is needed to ensure that the plate would not move under the imposition of the loads but remain stationary. Constraints (4) and (5) require that moments with respect to x and y axes are zero, ensuring that the plate would not rotate under the imposition of the loads.

The TPS parameter vectors a including a_1, a_2 and a_3, and w including w_i, can be computed by solving the following linear equation:

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

where $A_{ij}=r^2_{ij}\ln r^2_{ij}$, i=1,…,n (the number of landmarks), j=1, …,m (the number of raw data to be transformed); the i-th row of P is $(1, \hat{x}_i, \hat{y}_i)$. **O** is a 3×3 matrix of zeros. The 0 is a 3 zero vector in the rightmost part of equation (6). w, a and v are vectors formed from w_i, from a_1, a_2, a_3 and from v_i. The leftmost (n+3)*(n+3) matrix is denoted as K hereafter.

In this research, we focus on mapping points (x, y) of EMMA data to template coordinates (x', y') in light of given landmarks $(\hat{x}_i, \hat{y}_i)$ for one subject's EMMA data vs. $(\hat{x}_i', \hat{y}_i')$ defined for the landmarks of the template. So we are interested in warping 2D points using TPS defined by pairs of control points. Toward that end, we applied TPS functions to x and y coordinates separately. From Equation (6), the TPS warp which maps $(\hat{x}_i, \hat{y}_i)$ to $(\hat{x}_i', \hat{y}_i')$, can be recovered by

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

Where $\hat{x}'$ and $\hat{y}'$ are the vectors formed with $\hat{x}_i'$ and $\hat{y}_i'$ respectively. The w_x and a_x are the parameters for x-dimension, and w_y and a_y are for y-dimension. The transformed coordinates (x_j', y_j') of points (x_j, y_j) are given by

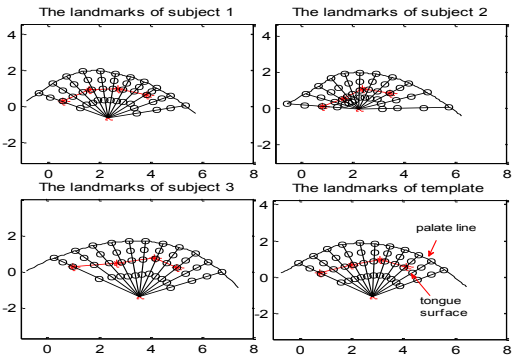

Fig. 1 Landmarks of three Chinese subjects and the template. There shows the palate in the top, the tongue surface and the grid lines. The circles are the landmarks.

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

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. The j-th row of Q is (1, x_j, y_j), and j-th row of the resulting vectors x' and y' are the interpolated x and y coordinates x_j' and y_j', respectively.

III. LANDMARK SELECTION

The articulatory data recorded by EMMA system is difficult to find the corresponding points having clear morphological meaning in the vocal tract. In order to overcome this problem, we define the landmarks in the vocal tract space by a gridline system modified from [7], which has been used to measure the morphology of the vocal tract for describing its acoustical properties.

A mean shape of the vocal tracts was used as a template, which was obtained from all EMMA data of Chinese database and Japanese database, respectively. A set of landmarks were first defined in the template, and then the corresponding landmarks were also determined on EMMA data for each subject. In the processing, we first calculated the average tongue posture (from tongue tip to tongue rear) over three vowels.

The gridline system was constructed based on the tongue surface and its centroid point, which had equal fan sections to cover the tongue movement regions. Consequently, ten sub-fans' edges intersected the palate line, the middle line, the tongue surface and the line below the tongue surface. There were 44 intersection points in total, which served as the landmarks in the normalization. The landmarks of each subject were defined under the same procedure. The results are shown in Fig. 1.

IV. EXPERIMENTS

Chinese EMMA database includes articulatory and acoustic data of /a, i, u/ in different sentences speaking by 3 Chinese subjects. Japanese EMMA database contains the same vowels from 3 Japanese subjects. 100 configurations were extracted of the vocal tracts for each vowel from Chinese EMMA

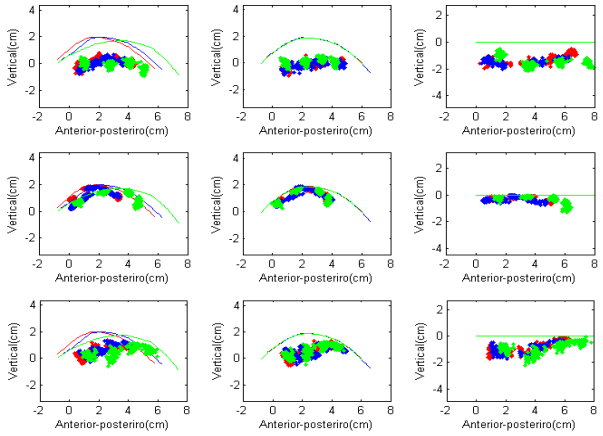

Fig. 2 Chinese EMMA data. The left column corresponds to the raw data, middle column corresponds to the normalized data by TPS method and right column corresponds to the normalized data by straightening palate-based method. Three rows correspond to vowels /a, i, u/. Three subjects were denoted by different colors.

database, and a total of 192 configurations from Japanese EMMA database.

The TPS method is used to define landmarks and the template for each language. Fig. 2 shows the raw data, normalized data by TPS, and normalized data by straightening palate-based method of the three Mandarin vowels /a, i, u/, respectively. Each row in Fig. 2 refers to the raw data and normalized data of one Mandarin vowel. Fig. 3 is the corresponding data for Japanese vowels.

Comparing the raw data with normalized data by TPS method in Fig. 2 and Fig. 3, one can see that the variances among different subjects are reduced. The palate curves of each subject almost overlapped the palate of the templates.

V. EVALUATIONS

A. Evaluation in articulatory space

The straightening palate wall method [1] is used as a baseline method for comparison to evaluate the TPS method. Fig. 2 reveals that TPS method yields smaller variances in the distribution of Mandarin vowels. Fig. 4 shows that the cross-subject standard deviations have been reduced by about 1.5 mm of horizontal direction and by 0.6 mm of vertical direction over three Mandarin vowels.

We also apply TPS method to Japanese database in the second experiment, and similar results are found with Japanese vowels. Specifically, the standard deviations have been reduced by 0.2 mm of horizontal direction and by 1.5 mm of vertical direction.

B. Evaluation in vowel space

In order to evaluate the TPS method, the vowel diagrams of raw data and normalized data of each subject were plotted for three Japanese speakers as shown in Fig. 5. One can see that for each subject, the size and shape of the four triangles for the tip, blade, dorsum and rear of the tongue in the normalized

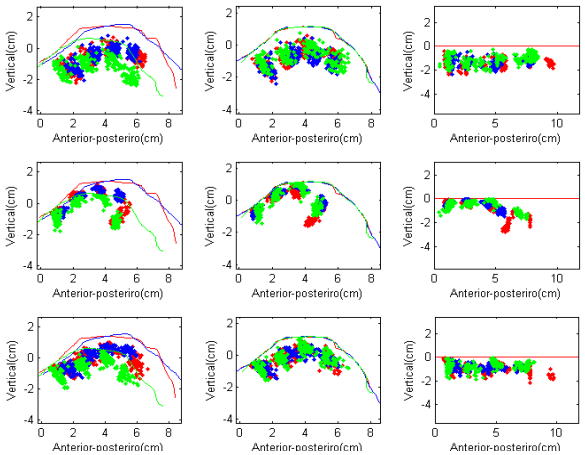

Fig. 3 EMMA data was from Japanese EMMA database. The column and row had the same meaning with Fig. 2.

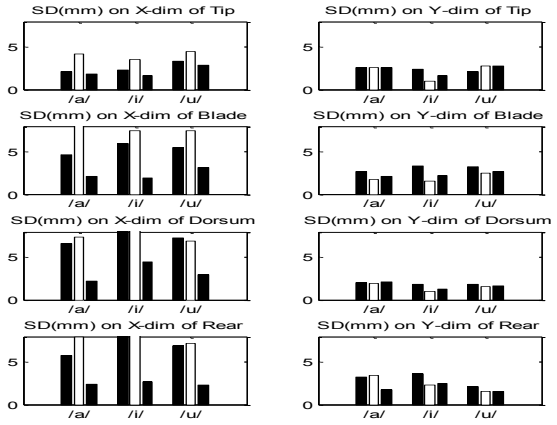

Fig. 4 Comparisons of cross-subject standard deviations of raw data and normalized data from Chinese EMMA data. The left bars denote the raw. The middle bars denoted the straightening palate method, the right bars for the TPS-based method.

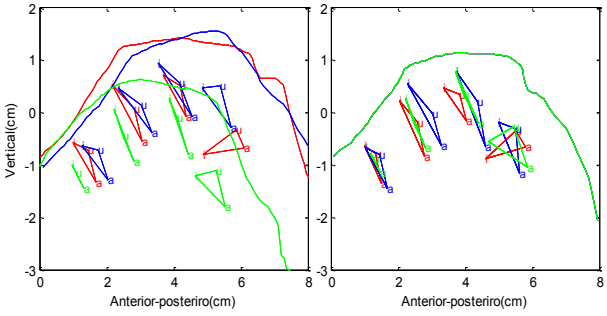

Fig. 5 The vowel diagrams of raw data (left panel) and TPS-normalized data (right panel) from Japanese EMMA database. The triangles are tongue tip, blade, dorsum and rear for Japanese vowels /a, i, u/.

data are highly similar to the original ones, which indicates that TPS method maintains more speaker-specific characteristics. Furthermore, for three subjects, the triangles for the tongue tip, tongue blade, tongue dorsum and tongue

rear almost overlapped, respectively, which reveals that TPS method can effectively reduce inter-speaker variances range. Since the TPS method is effective in different languages, the vowel diagrams for Mandarin are omitted here.

C. *Comparisons between Mandarin and Japanese vowels*

Reducing the morphological variances of the vocal tracts would facilitate the comparison of articulatory properties in different languages. In order to study the differences between Mandarin and Japanese vowels, we normalize the vocal tracts of six subjects in this study. The template was built based on the average palate and average tongue shape of all six subjects.

Fig. 6 shows the comparison of vocal tracts between Mandarin /a, i, u/ and Japanese /a, i, u/ after normalization. Each tongue shape is a mean shape over 64 samples of a single vowel from one subject. Although speaker specific characteristics still exist, differences can be observed for each vowel. For vowel /a/, tongue dorsum for Japanese /a/ is higher than that Mandarin /a/, and the tongue shapes for Japanese /a/ across all three subjects are a little higher than those of three Mandarin speakers. The oral cavity of Mandarin /a/ is larger than that for Japanese /a/ can be speculated. Due to the larger oral cavity, the F1 (746Hz) [8] of Mandarin /a/ is higher than that the F1 (625Hz) of Japanese /a/. For vowel /i/, the tongue shapes match well between two groups, which means the palates constrictions F1 (311Hz) and F2 (2138Hz) for Japanese vowel /i/ and F1 (337Hz) and F2 (2161Hz) of Mandarin /i/ were very similar. For vowel /u/, the tongue shapes of Japanese speakers are more forward than Mandarin /u/. In other word that the tongue retraction for Mandarin vowel /u/ is more retracted than Japanese vowel /u/. The larger the oral cavity becomes as the tongue is retracted, the lower the F2 that will be resonated. Owing to the larger oral cavity, the F2 (703Hz) of Mandarin /u/ is lower than that the F2 (1375Hz) of Japanese /u/.

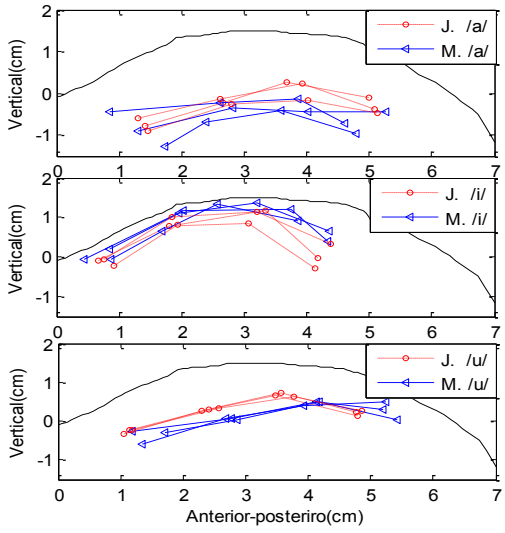

Fig. 6 Comparison of tongue configurations (tongue tip, blade, dorsum and rear) between Mandarin (M.) vowels /a, i, u/ and Japanese (J.) vowels /a, i, u/ after normalization.

VI. CONCLUSIONS

In this study, we have applied the thin-plate spline (TPS) method to normalize EMMA data across subjects. The evaluation shows that the inter-subject variances are reduced in articulatory space. The average standard deviations have been reduced for both Mandarin vowels (by around 1.5 mm of horizontal direction and by 0.6 mm of vertical direction) and Japanese vowels (by around 0.2 mm of horizontal direction and by 1.5 mm of vertical direction). Vowel triangles do not change much in both pre- and post- normalizations, which indicates that TPS method, compared with traditional straightening palate-based method, can effectively maintain more speaker-specific characteristics during normalization. The comparison results show that Mandarin vowels and Japanese vowels can be easily analyzed in the same space. The articulatory differences among the three vowels of Mandarin and Japanese are consistent with their corresponding acoustic properties.

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., 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 revisited," *J. Acoust. Soc. Am.*, vol. 97, pp. 471-490, 1995.

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

[3] B. FL, "Principal warps: Thin plate splines and the decomposition of deformations," IEEE Trans Pattern Anal. Mach. Intell, vol. 11, pp. 567-85, 1989.

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

[5] Yang, C.-S. and Kasuya, H., "Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child," *IEICE Trans. On Inf. & Syst.*, Vol.E78-D, No.6, pp.732-737, 1995.

[6] L. Zagorchev and A. Goshtasby, " A comparative study of transformation functions for nonrigid image registration," IEEE Trans. Image Processing, vol. 15, pp. 529-538, 2006.

[7] 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, 16, 27-47, 1995.

[8] Yuguang Wang, Jianwu Dang, Xi Chen, Jianguo Wei, Kiyoshi Honda and Hongcui Wang, " An MRI-based Acoustic Study of Mandarin Vowels." Interspeech, August 2013.