TONGUE SHAPE CONVERSION WITH NON-PARALLEL TRAINING DATA

Hao Li, Minghao Yang, Jianhua Tao

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{hli, mhyang, jhtao}@nlpr.ia.ac.cn

ABSTRACT

Articulatory data is an indispensable resource for speech production research. It will facilitate this study if we can convert one speaker's articulatory data to adapt a given target speaker. In this paper, we propose a tongue shape conversion method for non-parallel training data. The method combines thin-plate spline approximation (TPSA) algorithm with codebook mapping. The TPSA is a spatial morph method with landmarks extracted from articulatory data with phonetic segmentations. The landmarks' degree of certainty is evaluated and be considered in the TPSA morph. The proposed method has the advantages of the spatial morph and the codebook mapping by considering both the spatial configuration and the acoustic parameters. The results of our experiments with electromagnetic articulography (EMA) data indicate that the proposed method yields better results than the spatial morph method and the codebook mapping regardless the amount of training data.

Index Terms— tongue shape conversion, non-parallel training, thin-plate spline approximation, codebook mapping

1. INTRODUCTION

In speech production research, many organizations have built their own articulatory database from different speakers, such as MOCHA-TIMIT [1] and mngu0 [2]. A sufficient articulatory database that covers enough motion patterns of articulators is hard to get, sometimes one has to choose a small-scale corpus when establishing articulatory database due to some limitations such as unavailable of speakers. Therefore, integrating different databases and using the articulatory data as if they were recorded from the same speaker will be helpful for the speech production research. However, non-parallel corpus makes it hard to align the data form different databases. Moreover, the shapes of human articulators are different across speakers, which also make the articulatory data conversion a difficult task. Articulatory data conversion is especially needed in the research of audio-visual speech synthesis. To visualize the movement of articulators, graphic models were built based on real speakers' organ, such as the talking head developed by Badin et al.[3], but the articulatory data to drive them may not come from the same speaker, which is the case in our former work [4]. With the conversion method, we can adapt the source articulatory data to the graphic models with prerecorded data of the model speaker, so that other speakers' data will be able to drive the models. In this task, the differences caused by both the effect of speaker-dependent articulation and the recording procedure are need to be normalized. The amount of prerecorded data is usually small, so that the conversion has to be performed with limited training data.

There are many prior works on reducing the variability of inter-speaker articulatory data. One method is to parameterize the articulatory data using speaker-independent articulatory movement representations, for example the tract variable features [5] and the directional relative displacement features [4]. These methods usually need information about the speaker's organs which cannot be obtained if the speaker is unavailable. Codebook mapping method is another solution for conversion problem, and it has been used for non-parallel training of voice conversion [6-8]. Its performance usually depends on the scale and the quality of training data. Speaker adaptive training method has also be used in the prior work by Hiroya and Mochida [9]. Besides the statistical methods the conversion can also be achieved by spatial morph method, such as the vowel posture normalization method proposed by Hashi et al.[10], and Wei et al.[11] normalize inter-speaker articulatory data using thin-plate spline warping [12].

In our work, the conversion is for continuous speech. We combine the codebook mapping and the spatial morph method. These two methods are mutually complementary. The spatial morph method can overcome the data sparsity problem which is the weak point of codebook mapping. For the spatial morph method, it needs spatial configuration of the source and target articulatory space, and the performance is highly dependent on the accuracy of the configuration, this disadvantage can be overcome by combining with the codebook mapping. We use the thin-plate spline approximation (TPSA) [13] method for the spatial morph. MOCHA-TIMIT database is adopted in our experiments.

2. SPATIAL MORPH METHOD

2.1. Thin-plate spline approximation

Among the data acquired by various acquisition methods, electromagnetic articulography (EMA) is an important type, which offers high temporal resolution but sparse spatial information of articulator. Figure 1 shows the distribution of EMA sensors attached on tongue tip (TT), tongue body (TB) and tongue dorsum (TD) in mid-sagittal plane, the EMA sensor's positions cover different areas in different speakers' data because of different settings of recording procedure and variation of human articulator shapes as well as different speaking customs of the speakers. Therefore, the

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027), and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

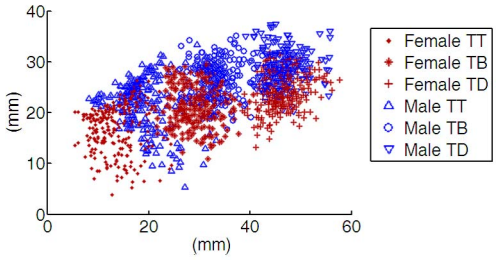

Figure 1. The TT, TB and TD EMA data in MOCHA-TIMIT database, shown in mid-sagittal plane. The positions are randomly selected from all the EMA files.

spatial alignment of EMA data involves rigid and non-rigid transformations, which is exactly one of the properties of the thin-plate spline approximation (TPSA) [13]. The smoothness of TPSA mapping is also an advantage for preserving the movement pattern of the EMA sensors. The use of thin-plate spline for point-based elastic registration was first proposed by Bookstein [12]. We briefly describe TPSA in the context of 2-dimensional (2D) space.

Given two landmark sets each consisting of n landmarks $p_i = (x_i, y_i)^\tau$ and $q_i = (x_i', y_i')^\tau$, $i = 1,...,n$, the thin-plate spline interpolation is to find the transformation u which minimizes a given function $J(u)$ and fulfill the interpolation conditions:

$$u(p_i) = q_i, \quad i = 1,...,n. \tag{1}$$

The function $J(u)$ represents the bending energy or smooth energy, it can be subdivided into two problems for each component z of u, which can be written as:

$$J(z) = \iint_{R^2} \left(\left(\frac{\partial^2 z}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 z}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 z}{\partial y^2} \right)^2 \right) dxdy \tag{2}$$

The solution of the minimization problem can be written as:

$$z(x, y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{n} w_i U\left(\| p_i - (x, y) \| \right) \tag{3}$$

Where U is basic kernel function:

$$U(r) = -r^2 \ln r^2 \tag{4}$$

The coefficients a_0, a_1, a_2 and $W = (w_1, w_2, ...w_n)^\tau$ can be computed through the following linear equations system:

$$\begin{cases} KW + P(a_0, a_1, a_2)^\tau = v \\ P^T W = 0 \end{cases} \tag{5}$$

Where v is corresponding component of q_i, K is a $n \times n$ matrix where $K_{ij} = U(\| p_i - p_j \|)$, $i = 1,...,n \quad j = 1,...,n.$, and P is a $3 \times n$ matrix where $P_{i1} = 1$, $P_{i2} = x_i$, $P_{i3} = y_i$, $i = 1,...,n$.

Equation (3) and (5) are the solution for thin-plate interpolation, which will transform the source landmarks exactly on the position of corresponding target landmarks. To take into account landmark localization errors, we need to weaken the interpolation condition (1). This can be done by combining an approximation criterion, which results in the following objective function

$$J_\lambda(u) = \sum_{i=1}^{n} \| q_i - u(p_i) \|^2 + \lambda J(u) \tag{6}$$

The first term of equation (6) measures the sum of the quadratic Euclidean distances between the transformed source landmarks and the target landmarks. The second term measures the bending energy of the transformation. Hence, the minimization of (6) yields a transformation which approximates the source landmarks to the target landmarks and is sufficiently smooth. The relative weight between the approximation behavior and the smoothness of the transformation is determined by a regularization parameter $\lambda > 0$.

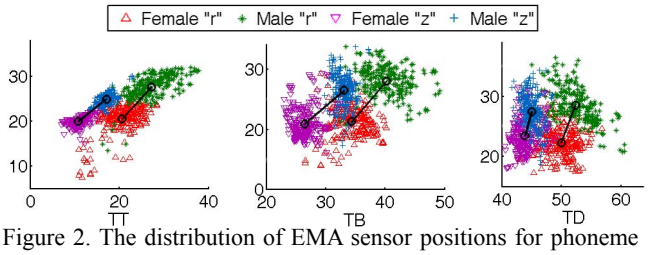

Figure 2. The distribution of EMA sensor positions for phoneme "r" and "z" of the two speakers in MOCHA-TIMIT, the landmarks are denoted by black circles. Corresponding landmarks are connected.

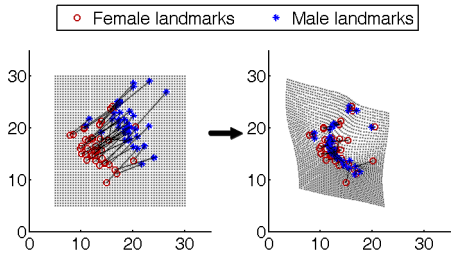

Figure 3. 2D spaces and landmarks before and after TPSA morph for TT of the male speaker, the female's data is target.

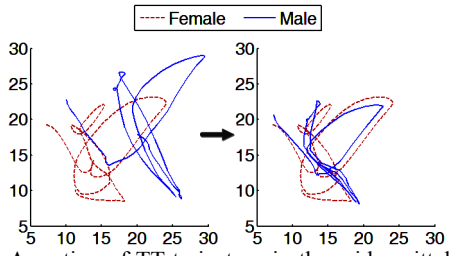

Figure 4. A section of TT trajectory in the mid-sagittal plane before and after TPSA morph for the data of the male speaker.

The solution to the approximation problem can be written as:

$$\begin{cases} (K + \lambda I)W + P(a_0, a_1, a_2)^\tau = v \\ P^T W = 0 \end{cases} \tag{7}$$

Equation (3) and (7) are the TPSA method in 2D space.

2.2. Landmarks and degree of certainty

The conversion is for EMA data, it is straightforward to use the EMA sensors' corresponding positions as landmarks if possible. For non-parallel training, our solution is to segment all the data into phoneme sections, and use the mean positions of the sensors for each phoneme as landmarks. These phonemes must appear in both the target and the source training data. Suppose the number of appeared phonemes is N, Let A and B stand for a pair of corresponding EMA sensor of source speaker and target speaker; m_i^A and m_i^B denote the total number of frames belong to phoneme i; a_{ij} and b_{ik}, $i = 1,...,N \quad j = 1,...,m_i^A$, $k = 1,...,m_i^B$ denote the sensor positions in frame j or k of phoneme i of source or target speaker, respectively, then the landmarks can be represented by

$$p_i = \frac{1}{m_i^A} \sum_{j=1}^{m_i^A} a_{ij}, \quad q_i = \frac{1}{m_i^B} \sum_{k=1}^{m_i^B} b_{ik}, \quad i = 1,...,N \tag{8}$$

Figure 2 shows the corresponding landmarks of the TT, TB and TD EMA data of the male and female speakers in MOCHA-TIMIT database. Two phonemes "r" and "z" are selected to

demonstrate the sensor's distributions of different phonemes; their mean positions are landmarks. As we can see, the EMA data of phonemes are with large variances. Therefore, we need to consider the degree of certainty of those landmarks. The generalize approximation scheme to take account the degree of certainty in the TPSA as described in [13] is to weight data term $\| q_i - u(p_i) \|^2$ in equation (6) by the inverse variances of landmark *i*, but the variances are reliable only when there are enough samples. If the database is not sufficient, there may be not enough frames for some phonemes. Hence, for each EMA sensor, we give all the landmarks a same weight ω to represents their degree of certainty:

$$\omega = \min\left\{\frac{1}{\det(\Sigma^A)}, \frac{1}{\det(\Sigma^B)}\right\} \tag{9}$$

where

$$\Sigma^A = \frac{1}{\sum_{i=1}^{N} m_i^A} \sum_{i=1}^{N} \left(\frac{m_i^A}{m_i^A - 1} \sum_{j=1}^{m_i^A} (a_{ij} - p_i)(a_{ij} - p_i)^{\mathsf{T}} \right) \tag{10}$$

$$\Sigma^B = \frac{1}{\sum_{i=1}^{N} m_i^B} \sum_{i=1}^{N} \left(\frac{m_i^B}{m_i^B - 1} \sum_{k=1}^{m_i^B} (b_{ik} - q_i)(b_{ik} - q_i)^{\mathsf{T}} \right) \tag{11}$$

Σ^A, Σ^B can be explained as weighted summation of the covariance matrix of all landmarks for source and that for target, respectively. The weight is $m_i^A \Big/ \sum_{i=1}^{N} m_i^A$ for source and $m_i^B \Big/ \sum_{i=1}^{N} m_i^B$ for target, which ensures that the more frames the phoneme *i* has, the more important the covariance matrix of *i* will be. Furthermore, any phoneme which contains less than 10 frames is removed from our landmark set. Now the objective function (6) can be written as

$$J_\lambda(u) = \omega \sum_{i=1}^{n} \| q_i - u(p_i) \|^2 + \lambda J(u) \tag{12}$$

and the solution for the TPSA problem is

$$\begin{cases} \left(K + \dfrac{\lambda}{\omega} I \right) W + P(a_0, a_1, a_2)^{\mathsf{T}} = v \\ P^T W = 0 \end{cases} \tag{13}$$

Equation (3), and (13) are the TPSA morph method we use. We set $\lambda = 1$ so that the smoothness will only be decided by ω. If ω is large, we obtain a solution with good approximation behavior, and if ω is small, we obtain a very smooth morph. Figure 3 shows the morph result with this method. The male speakers' landmarks are transformed to close to the females' with a smooth transformation. Figure 4 is the mapping result of a section of TT trajectory in the mid-sagittal plane. The male speaker's data is converted to be similar to the female speakers' after TPSA morph.

3. PROPOSED METHOD

It is obvious that the more accurate the landmarks are the better result the TPSA method will get. The calculation of landmarks is highly affected by the accuracy of phonetic segmentation and the quality of the articulatory data. In the EMA data record procedure, the EMA sensors may shift its position or be reattached, which may lead to inconsistent landmarks. Therefore, the robustness of TPSA method is weak. The reason of having this disadvantage is that the TPSA morph only considers the phonetic information. The dynamics of the tongue movement and the acoustic information are not modeled. To take them into consideration, we combine the TPSA morph with codebook mapping. We will discuss the codebook mapping and then the combination strategy.

Codebook mapping method has been used in non-parallel training for voice conversion. In this paper, the codebook unit we use is the time frames. The codebook can be trained by the LBG algorithm with all target training frame vectors if the amount of training vectors is large. The frame vectors consist of both articulatory features and acoustic parameters, which can be written as $(Aco^{\mathsf{T}}, Art^{\mathsf{T}})^{\mathsf{T}}$, where *Aco* represents the acoustic parameters and *Art* represents the articulatory features (which are EMA data). We choose 12-order MFCCs and log energy as acoustic parameters. For a given source vector, we use the nearest neighbor in the codebook as the codebook mapping result. The distances are measured by Euclidean distances of the source vectors and the codebook vectors. Each component of *Aco* and *Art* are z-score normalized before the learning of codebook and calculation of distance. Generally speaking, the tongue moves smoothly so we smooth the mapping result of EMA data with low-pass filter (LPF). The cutoff frequency is set to 10Hz according to the study of Ghosh and Narayanan [14]. Two problems need to be solved when using the codebook mapping method. The first one is that the EMA sensors' coordinate system is different if the data was recorded with different initial settings. Even though we normalize each component of EMA data, the rotation and non-rigid transformation are not considered, which may lead to incorrect distances. The second problem is that when the target training data is insufficient, it may lack some patterns, for example, there may be some phonemes that not appear in the target training data. In this case, the performance of codebook mapping method will decline.

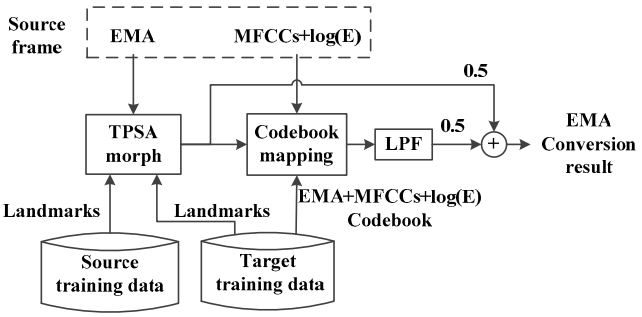

Figure 5. Workflow diagram of the proposed conversion method.

Form the description of the TPSA morph and the codebook mapping we can see that these two methods are mutually complementary. If we use the TPSA method before the codebook mapping, the effect caused by difference of coordinate system and the non-rigid transformation will be reduced, as we can see in Figure 4, the TPSA results are more suitable for the calculation of distance than the original EMA data. For the second problem of the codebook mapping, we need to decide whether to trust the codebook mapping result if the distance between the TPSA result and its' nearest neighbor in codebook is large. The large distance can be caused by the TPSA error or the absence of similar pattern in the target training data. To compromise the errors caused by both of the reasons, we can use the mean value of the TPSA result and its codebook mapping result as the conversion result. In order to perform the algorithm in a uniform way, we apply this strategy to all source frames, because if the distance is very small, it doesn't matter which result to use. Based on the above analyses, we propose our combination method. Figure 5 is the workflow diagram of the proposed conversion method. Here, the articulatory features are EMA data, and the acoustic parameters are 12-order MFCCs and log energy (log(E)). The source EMA data will be spatial morphed by TPSA, and then be fed to the codebook mapping along with the source acoustic parameters, LPF is used to smooth the output EMA data of codebook mapping, the conversion result is the mean value

of the TPSA output and the result of codebook mapping. In this strategy, the conversion result consist of the real target data modeled in the codebook and the movement patterns of the source data preserved by TPSA morph. This method takes advantage of the phonetic information by using landmarks obtained with phonetic segmentation. The codebook mapping performance is improved with TPSA as preprocessing and the data-sparsity problem is alleviated by using the mean value of the TPSA result and the codebook mapping result.

4. EXPERIMENTS

4.1. Data

In our experiments, both training data and test data were from the MOCHA-TIMIT database. This database contains two speakers' data, one female (fsew0) and one male (mask0). The British TIMIT sentences were uttered by each speaker in this database. We only use the EMA data of TT, TB and TD in this research, those data were given by x and y coordinates on the mid-sagittal plane. The EMA raw data was sampled at 500Hz, we down sampled them to 100Hz and smoothed them with 10Hz low-pass filter. The acoustic data were parameterized with 12-order MFCCs and log energy, the frame length and shift were 20ms and 10ms, respectively. All silence and breath sections were removed. From the 460 pairs of utterances, 92 utterances of fsew0 (file number ended with 3 and 8) and 92 utterances of mask0 (file number ended with 0 and 5) were used as non-parallel training data. The rest of the utterances (file number ended with 1, 2, 4, 6, 7 and 9) were test data. Both male to female and female to male experiments were conducted. The experiments were performed with different amount of training data of target speaker (varies from 10 seconds to 150 seconds), and all the training data of the source speaker were used in the experiments (254 seconds for the female speaker and 202 seconds for the male speaker in all). The codebooks were learned from the target training data with LBG algorithm and the size of the codebooks was set to 512. The label files included in the database were used for the phonetic segmentation of training data to calculate the landmarks. Test data were time aligned with dynamic time warping algorithm based on distance of acoustic features.

4.2. Evaluations

The results are evaluated by root mean square errors (RMSE) and correlation coefficients. All the evaluation are presented by the average results of the six components (x and y coordinates of three EMA sensors). Figure 6 shows the evaluation for the male to female conversion and Figure 7 shows that for the female to male conversion. CB denotes the codebook mapping method with original EMA data and acoustic parameters, all features are normalized for the codebook mapping. TPSA-CB-Rep denotes the codebook mapping method with TPSA mapping as preprocessing, in this method the TPSA mapping result is replaced by its nearest neighbor in the codebook. TPSA-CB-Mean denotes the proposed method, in which the conversion output is equivalent to the mean value of TPSA mapping result and the TPSA-CB-Rep result.

As we discussed in section 3, the TPSA method turn out to be unstable. It works well in the male to female conversion task, but yields poor results in the female to male task. The abnormal performance reduction in Figure 7 around 80 seconds training data is probably caused by a re-attachment of TD sensor of mask0 data in the 40th utterance. Even though, in both of the tasks, the TPSA-CB-Rep yields lower RMSEs than the CB method, which indicate that the combination with TPSA can improve the performance of CB despite the instability of TPSA. The correlation coefficients of CB and TPSA-CB-Rep decrease sharply with the decreasing of target data when the amount of target training data is small, which is the disadvantage of CB method. This phenomenon is caused by the lack of motion patterns in the target training data. The TPSA-CB-Mean method yields the lowest RMSEs and highest correlation coefficients in both of the tasks regardless of the amount of target training data. It maintains relatively higher correlation coefficients even when there are only a few utterances of target speaker. The results indicate that the proposed method has the advantages of both the TPSA morph and the codebook mapping.

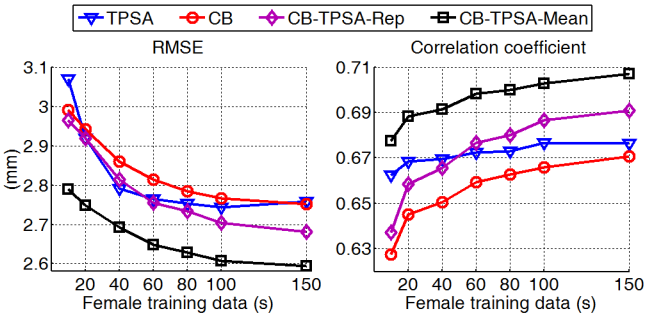

Figure 6. Evaluation of conversion results, male to female.

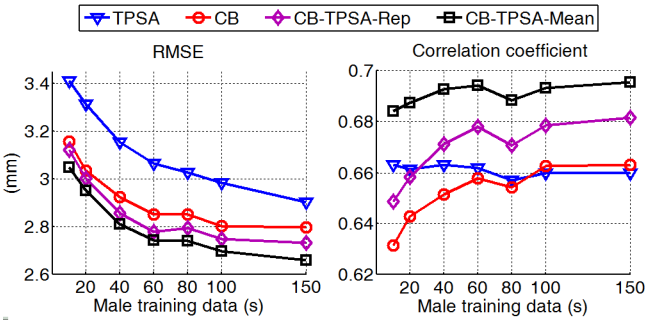

Figure 7. Evaluation of conversion results, female to male.

5. CONCLUSIONS

In this paper, we propose a tongue shape conversion method for non-parallel training, which combines codebook mapping with TPSA morph. It has the advantages of both methods by considering the spatial configuration and the acoustic parameters. The landmarks of TPSA are extracted from the EMA data with phonetic segmentations, and we also propose a method to evaluate the degree of certainty of landmarks. Our method can be used in the conversion task for EMA-like articulatory data and we will apply this method to other articulators in our future study. This method will be directly used in our work on audio-visual speech synthesis. We will adapt the articulatory data to a virtual articulatory model using the conversion method with a few prerecorded utterances of the model speaker, so that the data can be visualized by the virtual model. Another application is to integrate databases. There are articulatory databases that contain only utterances of phonemes or syllables such as the database adopted in [4]. The method we proposed will be used to solve the data sparsity problem by convert more data to adapt a given speaker.

7. REFERENCES

[1] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.

[2] K. Richmond, P. Hoole, and S. King, "Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus," in *Proc. Interspeech 2011*, Prague, Czech Republic, pp. 1505-1508, 2011.

[3] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data," in *Articulated Motion and Deformable Objects*. vol. 5098, ed: Springer Berlin Heidelberg, 2008, pp. 132-143.

[4] H. Li, M. Yang, and J. Tao, "Speaker-independent lips and tongue visualization of vowels," in *Proc. ICASSP 2013*, Vancouver,Canada, pp. 8106-8110, 2013.

[5] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology,* vol. 6, pp. 201-251, 1989.

[6] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. ICASSP 2008*, Las Vegas, U.S.A., pp. 4605-4608, 2008.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP 1988*, New York, U.S.A., pp. 655-658, 1988.

[8] L. M. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. Eurospeech 1997*, Rhodes, Greece, pp. 1347-1350, 1997.

[9] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," *Speech Communication,* vol. 48, pp. 1677-1690, 2006.

[10] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *The Journal of the Acoustical Society of America,* vol. 104, pp. 2426-2437, 1998.

[11] J. Wei and J. Dang, "Morphological normalization of vocal tract shape," in *Proc. ICASSP 2010*, Dallas, U.S.A, pp. 4186-4189, 2010.

[12] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 11, pp. 567-585, 1989.

[13] K. Rohr*, et al.*, "Point-based elastic registration of medical image data using approximating thin-plate splines," in *Proc. Visualization in Biomedical Computing*, pp. 297-306, 1996.

[14] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America,* vol. 128, pp. 2162-2172, 2010.