

# TONGUE SHAPE SYNTHESIS BASED ON ACTIVE SHAPE MODEL

Chan Song<sup>1</sup>, Jianguo Wei<sup>1</sup>, Qiang Fang<sup>2\*</sup>, Shen Liu<sup>1</sup>, Yuguang Wang<sup>1</sup>, Jianwu Dang<sup>1,3</sup>

<sup>1</sup> School of Computer Science and Technology, Tianjin University,  
92 Weijin Road, Nankai District, Tianjin 300072, China

<sup>2</sup> Phonetics Lab., Institute of Linguistics, Chinese Academy of Social Sciences, China

<sup>3</sup> School of Information Science, Japan Advanced Institute of Science and Technology, Japan  
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

## ABSTRACT

Nowadays magnetic resonance imaging (MRI) technique has been widely used in speech production research since it acquires high spatial resolution data of vocal tract shape without any known harm of radiation. However, it would be time consuming and expensive to establish an overall articulatory database using MRI technique due to its low temporal resolution and the large expense of the MRI equipment. In this study, we propose a method to interpolate tongue shapes between static vowels to acquire dynamic tongue shapes. Firstly, a set of parameters is extracted to control tongue shape based on Active Shape Model (ASM). Then, control parameters are interpolated to synthesize dynamic tongue shapes from static vowels' articulation. To evaluate the method, a set of key points were chosen from both the MRI images and the synthesize tongue shapes. Results suggested that the dynamic properties of these key points from the synthesized tongue shapes resemble those of the actual dynamic tongue shapes.

**Index Terms**—MRI, Active shape model, PCA

## 1. INTRODUCTION

Speech is the vocalized form of human communication. In order to study the mechanism of speech production, several techniques have been used to acquire movement of speech organs. The earliest used technology and still being used by many researches in speech production is X-ray (Fant, 1970)[1]. The main disadvantage of X-ray is the radiation, which is not safe to subject. Besides, the contours of X-ray images are very vague, which makes it difficult for automatic edge detecting.

Electromagnetic articulography (EMA) is also a widely used technology in speech production (Perkell et al., 1992; Dang et al., 2009)[2][3]. Due to its high temporal resolution, it is good for tracking the movements of speech organs. But it's not suitable for extracting area functions because it can only provide a few sensors' location.

Magnetic resonance imaging (MRI) is widely used in

speech production field and it is so far the best choice of acquiring high spatial resolution data of vocal tract shape. Baer (1991) firstly used MRI to measure the detailed vocal tract shape[4]. However, it would be time consuming and too expensive to establish an overall articulatory database using MRI technique due to its bad temporal resolution and the large expense of the MRI equipment.

This study aims to synthesize dynamic tongue shapes from static tongue shapes. Active Shape Model (ASM), which is proposed by Cootes and Taylor in 1995, describes the image shape of object of interest by obtaining a statistical shape model in examples from a training set (Cootes, 1995)[5]. By using ASM, we can train the tongue shapes and then presentate a tongue shape through only a parameter vector. At the same time, it is easier to interpolate in figures than in pictures for our experiment. We trained a set of parameters to control the tongue shape based on Active Shape Model (ASM) and then interpolated control parameters to synthesize dynamic tongue shapes. To evaluate our method, a set of key points were chosen from both the MRI images and the synthesize tongue shapes.

This paper is organized as follows: the second section introduces the proposed method for interpolating tongue shapes; the third section describes our experiment and results analysis and then follows the conclusion in section 4.

## 2. METHODS

### 2.1 MRI database of Mandarin speech production

The Chinese Mandarin MRI database is the first systematic 3D vocal tract database for Chinese. It covers static 3D MRI volumetric images of 10 static vowels and 10 sustained consonants, dynamic 2D data of 54 syllables and 15 retroflex, and two set of dynamic 3D data of 5 syllables. 6 male and 2 female Chinese native speakers participated in the experiment. It contains more than 100,000 images and can provide considerable information of vocal tract for analyzing the detailed morphological characteristics of Chinese Mandarin. The MRI images of a male subject are used in this study.

\*Corresponding author

## 2.2 Active Shape Model

Active Shape Model(ASM) has been proved great promising for automatically tracking objects from images. ASM was developed by Cootes and Taylor in 1995, which is a statistical point distribution model(PDM)[5]. The shapes are constrained by the PDM (point distribution model) Statistical Shape Model to vary only in ways seen in a training set of labeled examples. The shape of an object is represented by a set of points (controlled by the shape model). The ASM algorithm aims to match the model to a new image. The technique has been widely used to analyze images of faces[6], mechanical assemblies and medical images (in both 2D and 3D).

An ASM describes the image shape of object of interest by obtaining a statistical shape model in examples from a training set. ASM minimize the difference between the synthesized image from the model and an unseen image by tuning the model parameters, when it is applied to image interpretation or segmentation[7].

The ASM was built in following steps. Before implementing ASM, tongue contours on the mid-sagittal MRI images are manually annotated for both static vowels and vowel-vowel (VV) sequence. All this manually labeled images make up the training set. Figure 1 shows the manually marked tongue contours. In the training set, each tongue contour is described by  $n$  evenly distributed points. We define  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in}, y_{i1}, y_{i2}, \dots, y_{in})$  as the  $i$ -th contour, where  $(x_k, y_k)$  is the coordinates of the  $k$ -th point.

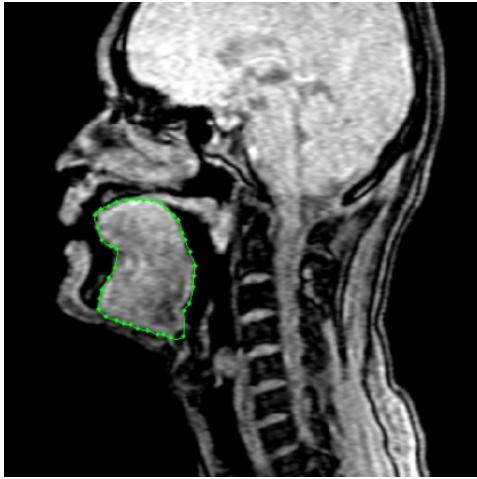


Figure 1 The marked tongue contour on mid-sagittal plane.

Firstly, We calculate the covariance matrix of the adjusted shape vectors. The covariance matrix is defined as follows:

$$S = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (1)$$

Where  $\bar{\mathbf{x}}$  is the mean shape of all the vectors in the training set[5].

Secondly, we calculate the eigenvalue sequence of  $S = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6)$  Where  $\lambda_i \geq \lambda_{i+1}$ , and  $\lambda_i$ ,  $i=1, 2, \dots, 6$ ; We choose the first  $t$  eigenvalues under the following conditions

$$\frac{\sum_{i=1}^t \lambda_i}{\sum_{i=1}^m \lambda_i} \geq 90\% \quad (2)$$

Then we calculate the corresponding eigenvector of the first  $t \lambda_i$  to make up  $\mathbf{P}$ , recording as  $\mathbf{P} = (p_1, p_2)$ .

After we get the mean shape vector  $\bar{\mathbf{x}}$  and the eigenvector  $P$  of the training set, our tongue shape model can be expressed as  $x = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}$ . So we can get an unique  $\mathbf{b}$  vector of a certain shape through the known model,  $\mathbf{b} = (b_1, b_2, b_3, b_4, b_5, b_6)$ ; At the same time, if we have a known  $\mathbf{b}$ , we can also get an unique tongue shape.

## 3. EXPERIMENTS AND RESULTS

### 3.1 Initialization of the MRI pictures and tongue labeling

In our experiment, the training set totally contains 90 mid-sagittal MRI figures, which was composed of different static and dynamic articulation tongue shape. A total 40 points represent each contour of the training set. Because the dynamic tongue coordinate values were different from the static ones and ordinary method can mix the jaw movement characteristics in the tongue deformation, initialization process, like translation, became very important. We chose a certain fixed point of the lower jaw as origin to establish a new coordinate system. In this standard system, we marked totally 90 tongue shapes, every contours was replaced by a shape vector  $\mathbf{x}$  of 40 points. The example is followed:

### 3.2 PCA analysis of the training set

After we get the training set, we conduct the PCA analysis to get the basic parameters for following model and calculate the mean shape. The result was shown in Table 1.

Table1. The contribution of first six factors

$\lambda$	Eigenvalue	Percentage	Accumulated percentage
$\lambda_1$	3669.78	62.79%	62.79%
$\lambda_2$	618.95	10.59%	73.38%
$\lambda_3$	532.26	9.11%	82.49%
$\lambda_4$	286.79	4.91%	87.40%
$\lambda_5$	194.49	3.33%	90.73%
$\lambda_6$	135.92	2.33%	93.05%

We select the first six eigenvalues to be the main influence factors, the corresponding accumulated contribution rate is 93%, as shown in the Table 1. Then, calculating the corresponding  $p_i$  ( $i=1,2,\dots,6$ ) of  $\lambda_i$ , and we get one of the parameters  $\mathbf{P} = (p_1, p_2, \dots, p_6)$  of ASM. The tongue ASM is build at last.

### 3.3 Interpolation synthesis

From the MRI database, we get not only the total 14 dynamic tongue contour varies when the subject pronounce continuous /ier/, but also the data of isolated sustained articulation of /i/ and /er/. The 14 images of dynamic trail were treated as the evaluation set. In the control parameters getting from images of static /i/ and /er/, we conduct interpolation procedure. The interpolation equation is polynomial equation, whose control parameters are obtained from dynamic tongue shapes. Then, we use the static shape images as starting and ending points, between which we get 12 synthesized control parameters by interpolation. The synthesized parameters are applied to synthesis 12 tongue shapes corresponding to mid-sagittal 12 tongue shapes of dynamic /ier/. We evaluate the synthesized shapes by comparing the difference betewwn synthesis tongue shapes and dynamic ones. As shown in figure 2, we choose three key points on the back of the tongue to see the tendency. The result was shown in figure 3 and figure 4, where the green ones present our synthesis data, the blue ones present original tongue shape deformation tendency.



Figure 2. Key points (denoted by blue points) selected from the tongue contour.

### 3.4 Performance evaluation

The results shown in figure 3 and figure 4 show the similar tongue movement trajectories between the synthesized tongue images and original tongue movement, when the people pronounce continuous /ier/. Thus, this demonstrates our method is feasible in synthesizing dynamic articulation. The main difference root in that we interpolate in the static articulations, but the evaluating set were dynamic articulations. There are differences between static /i/ and /er/ and dynamic ones. On the other hand, the PCA analysis removes some little efficient factors. The pictures we synthesis with only six factors and with all the factors must be different. The second reason of differences is inevitable. Both the two error is shown in figure 5. The first panel show the difference between static /i/ and dynamic /i/, while the second panel show the difference between static /er/ and dynamic /er/. The third and forth panels show the synthesized dynamic /i/ and /er/ comparing with original dynamic /i/ and /er/.

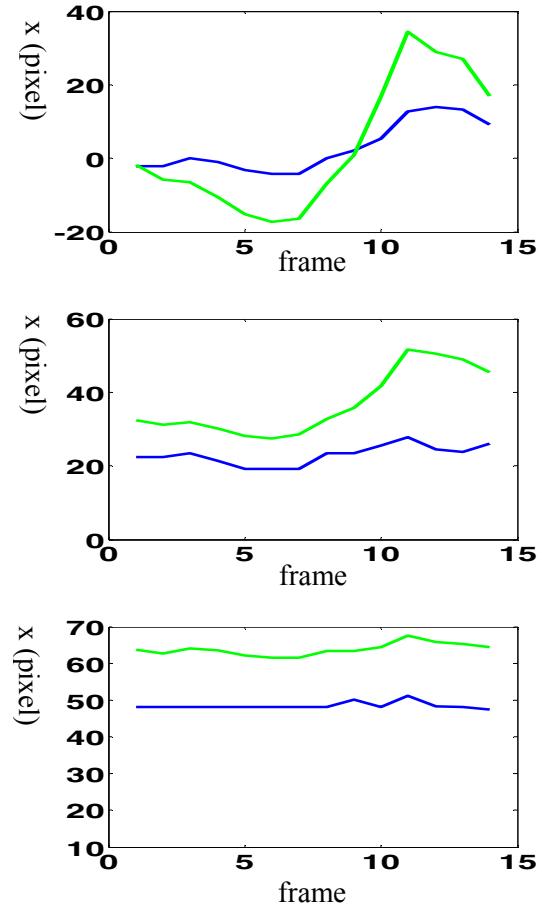


Figure 3. The x coordinate value tendency comparison of our synthesis data and original dynamic tongue shape, the first picture were the x value comparison of tongue tip, the second one was the second points near the tongue tip and the third one was the point furthest to the tongue tip.

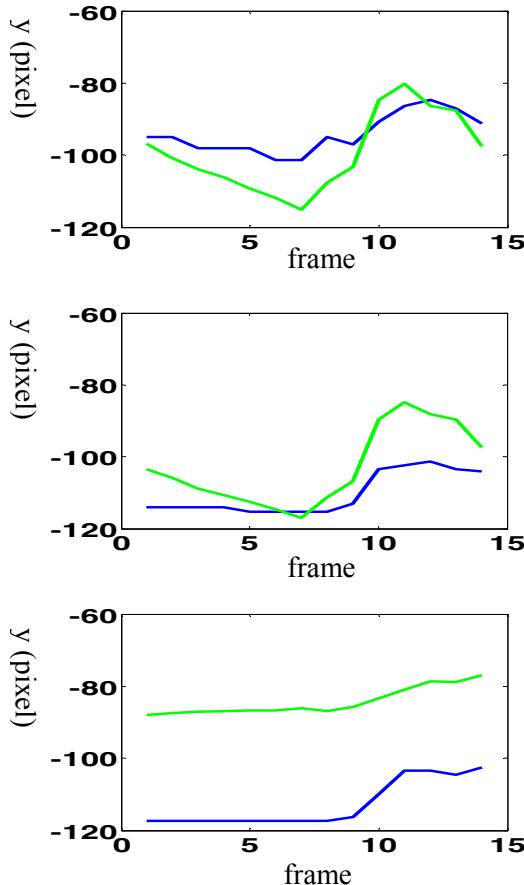


Figure 4. The y coordinate value tendency comparison of our synthesis data and original dynamic tongue shape, the point order is the same as figure3

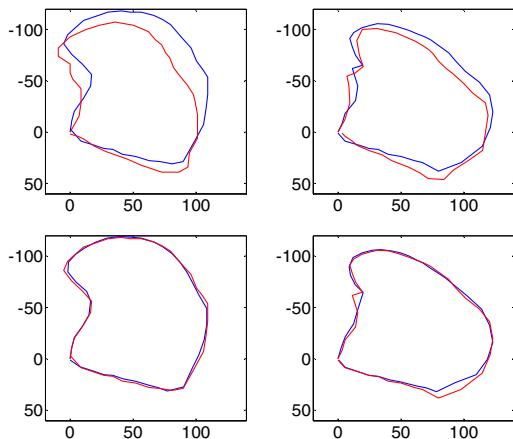


Fig.5. the first row shows the first error we explained in our paper, where the red lines present the static pronounce tongue shape, while the blue lines present the dynamic ones. The second row shows the second one, where the red shows the result of six factors.

#### 4. CONCLUSIONS

In this study, we proposed a method to synthesize dynamic

vocal tract shapes. We trained a set of parameters to control tongue shape based on ASM and then interpolated control parameters to synthesize dynamic tongue shapes. To evaluate our method, a comparison was carried out between the key points choose from the synthesize tongue shapes and the actual MRI dynamic images. The dynamic properties of the synthesized tongue shapes quite resembled the actual tongue movement. We can conclude that our method is an feasible way to get dynamic tongue shapes From static data. In the future, we will try to improve the accuracy of out method.

#### 5. ACKNOWLEDGMENT

This work is sponsored by the national natural science foundation of china under contract No. 61175016and contract No. 61233009.This study was also supported in part by Microsoft Research Asia under contract No. FY11-RES-OPP-008.

#### 6. REFERENCES

- [1] Fant, G. (1970). Acoustic Theory of Speech Production (Mouton, The Hague, Paris), Chap. 1.4, pp. 63–90.
- [2] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta,I. & Jackson, M. (1992). "Electro-magnetic mid-sagittal articulometer (EMMA) systems for transducing speech articulatory movements". *J. Acoustical Soc. America*, 92, 3078-3096.
- [3] Dang, J., Lu, X. (2009, 9) "INVESTIGATION OF THE RELATION BETWEEN SPEECH PRODUCTION AND PERCEPTION BASED ON A VOWEL STUDY," *The Festschrift in honor of Prof. Wu*.
- [4] Baer, T., and J. C. Gore et al. , "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *The Journal of the Acoustical Society of America* , 90(2), 799-828, 1991.
- [5] Cootes T.F., Taylor C.J., Cooper D.H. and Graham J., "Active shape models - their training and application," *Computer Vision and Image Understanding*, (61): 38–59, 1995.
- [6] Skjalg Lepsoy, and Sergio.Cuinga, "Conversion of articulatory parameters into active shape model coefficients for lip motion representation and synthesis", *Signal Processing: Image Communication* 13,pp.209-225,1998.
- [7] Hans C. van Assen, Mikhail G.Danilouchkine, Alejandro F.Frangi, Sebastian Ordas, Jos Jnberg, .M.Westenberg, Johan H.C.Reiber, Boudewijn P.F.Lelieveldt."SPSM: A 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data", *Medical Image Analysis* 10,pp.286-303,2006.
- [8] S.Avila-Garcia. Maria, N.C.John,, and I.D.Robert, "Extracting Tongue Shape Dynamics frovm Magnetic Resonance Image Sequences," *World Academy of Science, Engineering and Technology* 2,2005.
- [9] Bram.van.Ginneken, Alejandro.F.Frangi, Joes.J.Staal, Bart.M.ter.Haar,Romeny, and Max.A.Viergever," Active Shape Model Segmentation With Optimal Features", *IEEE TRANSACTIONS ON MEDICAL IMAGING*, VOL. 21, NO. 8, AUGUST 2002