Multimed Tools Appl (2016) 75:5247–5263
DOI 10.1007/s11042-015-3040-4

# Multi-modal recording and modeling of vocal tract movements

**Jianguo Wei[1] · Song Wang[1] · Wenhuan Lu[1] · Qingzhi Hou[2] · Qiang Fang[3] · Jianwu Dang[2,4]**

Received: 21 April 2015 / Revised: 6 October 2015 / Accepted: 21 October 2015 /
Published online: 14 December 2015


**Abstract** The complexity of vocal tract movement causes the difficult to record whole information of vocal tract during speech. Dynamic articulation has been acquired by implementing a variety of instruments, each of which has its advantages and shortcomings. However, the measurement of vocal tract movements is a difficult task to accomplish using one type of recording technique, and this has led to the simultaneous application of multiple instruments. Thus, we used an ultrasound system in combination with the electromagnetic articulography (EMA) system to record the multi-modality movement of the tongue. Data of the vocal tract movements were obtained by the ultrasound-based speech recording system developed by us, with which ultrasound images and synchronized audio signals are recorded synchronously. The EMA system is also used for the simultaneous collection of articulatory data with the audio. The EMA and ultrasound data were registered and matched to the same audio signal, after which these two sets of data were fused for each time point. In addition, a method for vocal tract shape reconstruction and modeling is proposed for the ultrasound dataset by using an active shape model. The averaged reconstruction error does not exceed 1.26 mm.

**Keywords** Ultrasound images · EMA · Vocal tract · Alignment

✉ Wenhuan Lu
wenhuan@tju.edu.cn

[1] School of Computer Software, Tianjin University, Tianjin, China

[2] Tianjin Key Laboratory of Cognitive Computation and Application, Tianjin University, Tianjin, China

[3] Phonetics Laboratory, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China

[4] Japan Advanced Institute of Science and Technology, Nomi, Japan

## 1 Introduction

Vocal tract (VT) recording and modeling has been the subject of investigation of many research groups. However, the complexity of the inner structure of the VT is such that the acquisition of sound generated in the VT remains a challenge to us. A variety of instruments have been implemented to record dynamic articulations during speech, each of which has its own advantages and disadvantages. However, none of these instruments has the ability to record data containing all the information of the articulators. Alongside X-ray, computed tomography (CT), and magnetic resonance imaging (MRI), ultrasound imaging is one of the four major medical imaging techniques [17, 18]. Each of these four instrumental techniques has its particular advantages and disadvantages in respect to recording articulation. Ultrasound imaging technology, which is widely used in clinics, has the advantages of being convenient, safe, fast, and offering real-time scan results. However, due to the particularity of the imaging mechanism, ultrasound images are noted for their extensive speckle noise and provide limited information of the subject's articulator [2]. On the other hand, although electromagnetic articulography (EMA) [6, 13] data contain the precise location of sensor information, they lack complete information of the surface of the tongue.

Therefore, in our experiment, the EMA and ultrasound systems are used simultaneously as a complementary pair to record tongue movement as the ultrasound images provide the EMA data with a complete tongue contour while the EMA data offer additional key point information to the images such as information about the upper and lower tooth, lips, and tongue tip. Information about these parts plays an important role in complementing the information pertaining to the tongue and finding the relationship between different ultrasound image frames. The ultrasound images and the synchronized audio were obtained by a portable ultrasound system with the data collection software that was developed by our team. The EMA system was used to collect the flesh-point information of speech articulation in synchrony with the audio. Then, the ultrasound images and the EMA data were registered and matched by using the audio stream. High speech cameras were also used to collect facial information. In total there are four modalities of data sources that were synchronized and recorded together. The ultrasound images and the EMA data for each time point were also integrated spatially.

The multi-modalities articulatory data can be applied for vocal tract visualization, speech training, and silent speech recognition applications. After recording the multi-modality data, an active shape model-based approach was proposed to model the articulatory data.

The paper is organized as followed. Section 2 introduces the acquisition system, including the hardware and software system. The analytical process and the procedure of integrating the ultrasound and EMA data are described in Section 3. The active shape model-based tongue shape reconstruction approach is presented in Section 4. The conclusion is given in Section 5.

## 2 Data acquisition system

### 2.1 Brief description of the system

The acquisition system mainly includes four parts, i.e., the portable ultrasound system for collecting ultrasound images, EMA system for collecting EMA data, and audio system for

collecting the synchronizing audio signal for both of these systems. The main components of the system are shown in Fig. 1.

Details of the equipment are as follows:

- Portable ultrasound system: a Terason T3000 ultrasound system with a 8MC3 micro-convex transducer and a stand for the transducer
- WAVE system: the WAVE system of Northern Digital Inc. with 8 channels is an instrument for collecting EMA data. The WAVE system contains: a field generator, mounting arm, system control unit, system interface unit, micro-sensor, and audio synchronization cable.
- Audio system: the audio interface Roland Octa-Capture UA-1010, a Studio Project CS5 condenser microphone, two 6.5 mm to 3.5 mm audio adapters for connecting the audio interface and the laptops
- Helmet and stand: in order to stabilize the ultrasound probe, we developed a helmet-based stand and a magic arm-based stand for stabilization during data recording [15]. The helmet is constructed of a special material usually used in head surgery, to reduce its weight. The magic stand and helmet stand can be used for different situations and purposes. The helmet-based probe stand is able to stabilize the ultrasound probe to a greater extent than the magic arm-based stand. However, it could affect the quality of the recorded facial information. In this work, we selected the magic arm-based stand for experiments.

## 2.2 Hardware setup of the acquisition system

During experiments, it is unavoidable that the subject's head will move. Thus, we need a helmet to fix the probe to the subject's chin. Ultrasound-based systems, which have been developed to date, include ESPCI [5], HATS [16], or Palatron [12]. Our acquisition system used to employ a helmet for stabilizing the ultrasound probe, but this approach caused the subject to become tired after prolonged experimental recording sessions. Thus, we replaced the helmet by a stand with a plier-like instrument to stabilize the probe, as shown in Fig. 2.

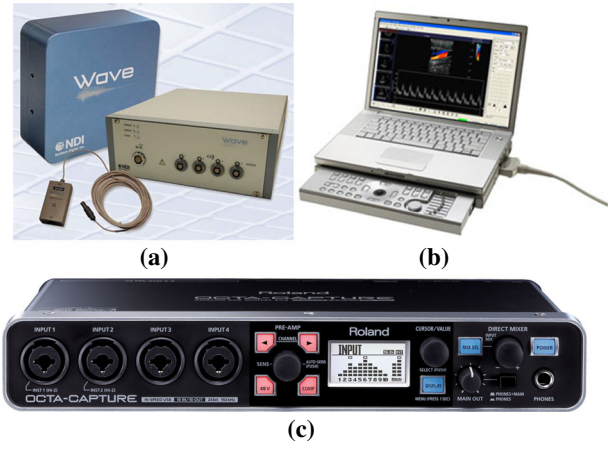



**Fig. 1** **a** EMA system **b** ultrasound system **c** Roland Octa-Capture UA-1010

**Fig. 2** Photograph showing data being recorded with the acquisition system

### 2.3 Software part of the acquisition system

We attempt to obtain three sets of data from the experiments: ultrasound images, EMA data, and data synchronized with audio signals. The ultrasound image acquisition program was developed based on the SDK of Terason Ultrasound System to which selected new functions have been added, such as file storage, information display, and synchronizations between imaging and audio. The audio and image streams are processed in parallel using multithreading programming techniques [15]. A timestamp is attached to each image to facilitate synchronization with the audio signal and also the EMA data.

The EMA data includes the trajectories of the flash points and audio files provided by the EMA system.

## 3 Vocal tract movement data acquisition

### 3.1 Description of experiment

Firstly, we set up the EMA system and place the magnetic field generator in position. Then, we start attaching sensors to the subject. All the sensors are cleaned with alcohol and pasted to 11 points on the subject (left ear, right ear, nose, tongue tip of mid-sagittal plane, tongue blade of mid-sagittal plane, a point between the tongue blade and tongue dorsum, tongue dorsum of mid-sagittal plane, upper teeth, lower teeth, upper lip, and lower lip) to collect articulatory data. Two more sensors are attached to the ultrasound probe as references. The sensor positions on the tongue and those pasted to the probe are shown in Fig. 3. Once the sensors are attached to the subject, their status is detected by the software together with the location or connection. Adjustment needed to ensure all sensors are functioning. Secondly, after the status of all sensors status is checked, the ultrasound system is started to check whether the ultrasound probe is working properly. Then, the subject positions his head on the ultrasound probe, which is attached with glue to improve the quality of the images, by adjusting their head slightly. Finally, guided by the live images on the monitor, the image parameters (including image size, image depth,

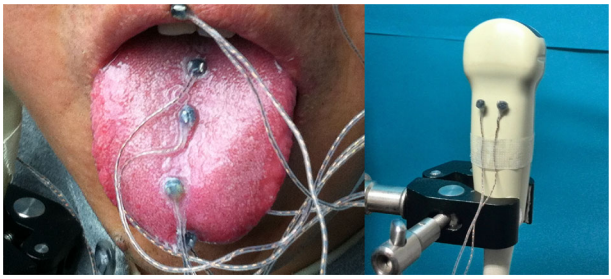


**Fig. 3** **a** Sensors positions on the tongue and **b** sensors on the probe

gain, compression ratio, noise suppression, etc.) are adjusted to ensure that clear tongue contours are observed with a more stable and higher acquisition frequency.

After finishing the preparation, the recording starts and the data is recorded sentence by sentence. A beep sound prompts subject to start reading.

The corpus is a Chinese speech database designed for speech synthesis applications (the database is named Corpus of speech synthesis of the National "863" Project) [11]. For this experiment, 350 sentences were selected from the corpus in which about 8–15 s are spent reading each sentence at normal reading speed. The whole recording process lasts approximately one hour.

### 3.2 Structure of collected data

The ultrasound system collects ultrasound images and synchronizes audio files. The images are in bitmap, 8-bit gray scale, with a resolution of 640 × 480 with a name, including a timestamp, given by the system clock. The frame frequency of the image stream is 60 fps. The audio file has a size of 16 bits and the sampling rate is 44.1 kHz.

The EMA system generates '.raw' and '.wav' files after data collection in which it stores the articulatory data and speech sound, respectively. In '.raw' files, the sensor trajectory and the corresponding audio alignment is done through the EMA equipment. The sampling frequency for the sensor movement is set to 100 Hz by default. As with the ultrasound system, the audio file is recorded with a sampling rate of 22,050 Hz.

### 3.3 Date analysis and fusion

#### *3.3.1 Synchronization*

Two groups of audio files are recorded during a single acquisition procedure which is synchronized with the ultrasound images and EMA data, respectively. The two audio signals are generated by the same microphone and is transferred from the audio interface to the two computers; hence, theoretically the two audio files should be identical and have the same length. Although the impedance difference between the soundcards of the different computers causes an amplitude difference, the basic features of the two audio files are similar. After we locate the beep sound position and use it as a reference, the two sets of audio files can be completely co-registered. With the two audio files, the ultrasound images and EMA data can be aligned on the time axis. Figure 4 shows the waveform of the audio files for a certain sentence in Chinese for both data sets.

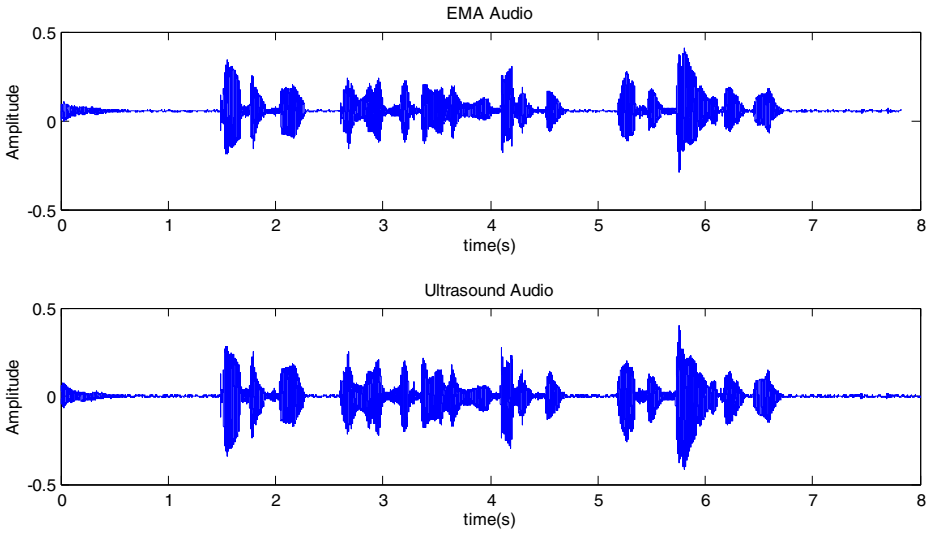



**Fig. 4** Signals of two different audio files corresponding to the EMA data (*top*) and ultrasound data (*bottom*)

### *3.3.2 Fusion of multi-modal data in the space*

As described above, the EMA system collects articulatory data and stores it in '.raw' files which contain 5D information including the position and the rotation for each sensor at each timestamp. The ultrasound system collects ultrasound images of the mid-sagittal plane of the tongue in 640 × 480 bitmap image format which can be considered as a 2D matrix. In order to fuse the two different kinds of data together, we need to find the relationship between the two coordinate systems and perform a coordinate transformation to incorporate them into the same coordinate system.

1) Finding the mid-sagittal plane:

We have to locate the mid-sagittal plane so as to fuse the EMA data and ultrasound images together. The mid-sagittal plane for the EMA data can be determined by using the three points (left ear, right ear, and nose) as the reference after compensating for the movement of the head. We specify the points on the left ear, right ear, and nose as point A, B, and C, respectively. We obtain the mid-sagittal plane of the head by using the three points A, B, and C. Points A and B form the vector $\overrightarrow{AB}$, which is perpendicular to the mid-sagittal plane, whereas point C is located on the plane.

We obtain the three-dimensional (3D) coordinates of the points A, B, and C from the EMA data with which we can obtain the equation of the mid-sagittal plane:

$$ax + by + cz + d = 0 \tag{1}$$

where a, b, c, and d are constants, determined by the coordinates of the points A, B, and C.

The intersection point of the plane and the connection line of the two ears (vector $\overrightarrow{AB}$) is set as the origin of the plane-coordinate system. The direction from the origin to the point of the

nose (point C) is the direction of the X-axis. The Y-axis is perpendicular to the X-axis, pointing down from the nose to the tongue.

To convert the 3D EMA data to 2D data, we project the other EMA points onto the mid-sagittal plane and obtain the plane coordinates. After given a point $\overrightarrow{P}(X, Y, Z)$, we can obtain its projection point $\overrightarrow{P}_0(X_0, Y_0, Z_0)$ on the mid-sagittal plane. $\overrightarrow{P_0}$ is calculated by the following formula.

$$\overrightarrow{P}_0 = \overrightarrow{P} + [at, bt, ct] \tag{2}$$

Here, *a*, *b*, *and c* are constant in Eq. (1) and *t* is defined by:

$$t = -(aX + bY + cZ + d) \Big/ \left(a^2 + b^2 + c^2\right) \tag{3}$$

where *X*, *Y*, *and Z* are the coordinates of point $\overrightarrow{P}$ and *a*, *b*, *c*, and *d* are constant in Eq. (1).

2) Transforming the coordinates:

After determining the mid-sagittal plane of EMA data and transferring the points onto it, we can say, the EMA and ultrasound coordinates are in the same plane. Simple coordinate transformation is used here by performing three steps: scaling, rotation, and translation.

Once the point $\overrightarrow{P}_0$ on the EMA coordinate is known, we can use the coordinate transformation as follows to obtain the new coordinate $\overrightarrow{P}_u$ in the ultrasound coordinate system.

$$\overrightarrow{P}_u = s \times \left(\overrightarrow{P}_0 \times r\right) + t \tag{4}$$

where *s*, *r*, and *t* are parameters corresponding to the values of scaling, rotation, and translation, respectively, *s* is a constant that represents enlargement or reduction during the coordinate transformation, *r* is the degree representing the angle of rotation, and *t* is a vector representing the distance and direction of translation.

Among the three parameters above, the scaling value *s* is a certain value which can be decided in the following way. The unit length is 1 mm in the EMA system. In order to calculate the value of the scaling parameter, we need to know the unit length of the coordinate system of the ultrasound image which corresponds to 1 pixel in an image. When we configure the parameters of the ultrasound image, we set the depth (the height of the sector) as 8 cm as is shown in Fig.5. This enables us to calculate the unit length of the ultrasound image, which is 0.223 mm, and the scaling parameter *s* is 4.4875 in the coordinate transformation.

3) Locating the reference image and adjusting the other images automatically:

Firstly, a reference image corresponding to the silent posture is selected and then the parameters of translation and rotation are manually selected to match the EMA data to the ultrasound image. For the same sentence shown in Fig. 4, the reference image is chosen and the mapping result is shown in Fig. 6. The rotation degree is $-20^{\circ}$ and the translations along the X- and Y-axes are −10 and 90 pixels, respectively.

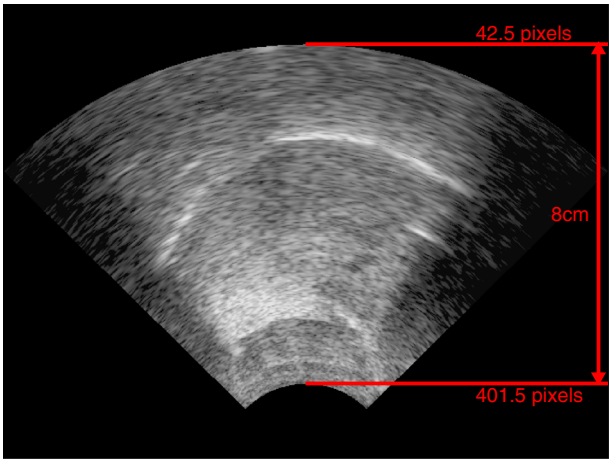



**Fig. 5** Scan depth of ultrasound signal for the ultrasound imaging system as indicated by the distance between the two *red lines*

After selecting the reference image, we obtain the position of the probe as a reference. For the other images, the translation of the probe means the translation of the image. Thus, the position of each sensor is determined by subtracting the translation of the probe as a

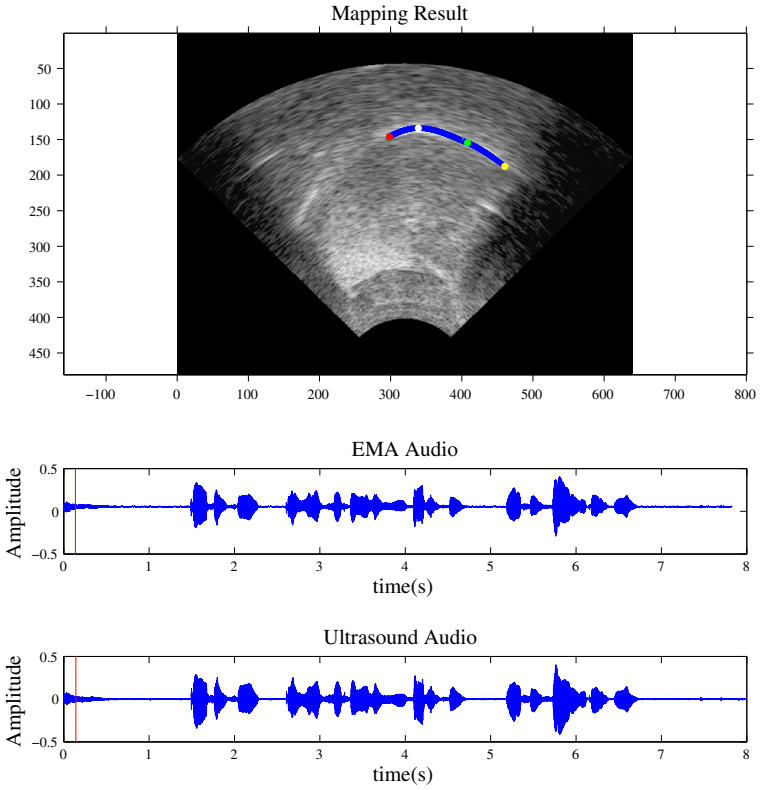



**Fig. 6** Reference image and mapping result. The *red*, *white*, *green*, and *yellow* points represent the four sensors pasted on the tongue. The *blue line* is the result of interpolation between the four points

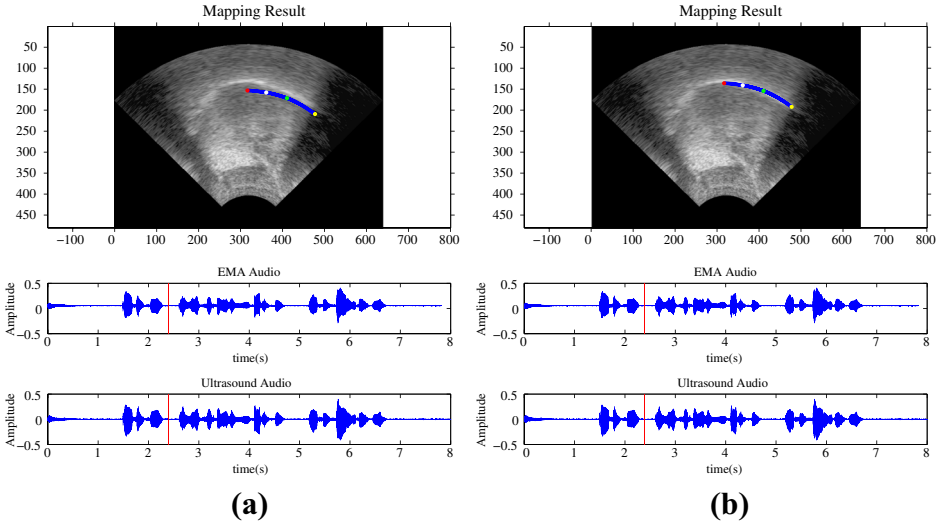



**Fig. 7** Mapping result **a** without and **b** with the complementary process

complement. This approach enables us to reduce the effect of the moving probe. Figure 7 shows the result before and after the complementary process, whereas Fig. 8 shows the mapping result of the sentence as a function of time.

### 3.4 Validation

Verification of the precision of the method requires us to quantify the distance between the EMA mapping results and the real contour of the tongue. A tool named EdgeTrak is used here for extracting the contour of the tongue semi-automatically [10]. Although EdgeTrak needs several manually selected contours as a reference, which will undoubtedly introduce errors, it is a preliminary and a convincible method for validating the mapping results.

In total, 365 ultrasound images are used for calculating the average error using the following equation:

$$Err_{ave} = \frac{\sum_{i=1}^{k} \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}}{k} \tag{5}$$

where $(x_c, y_c)$ and $(x_i, y_i)$are the coordinates of the reference points on the labeled contour of the EMA points respectively, $k$ represents the total number of points, and the average error is 1.8 mm.

### 3.5 Date set description

In this study, we recorded data for three subjects. One of the subjects recorded 350 sentences, which included a one-hour dataset. The other two subjects each recorded 100 sentences that included 20–30 min of data. The corpus we used is a Chinese speech database designed for speech synthesis applications as described in the previous section. The data was preprocessed by performing de-noising and data cleaning. The annotation was conducted manually.

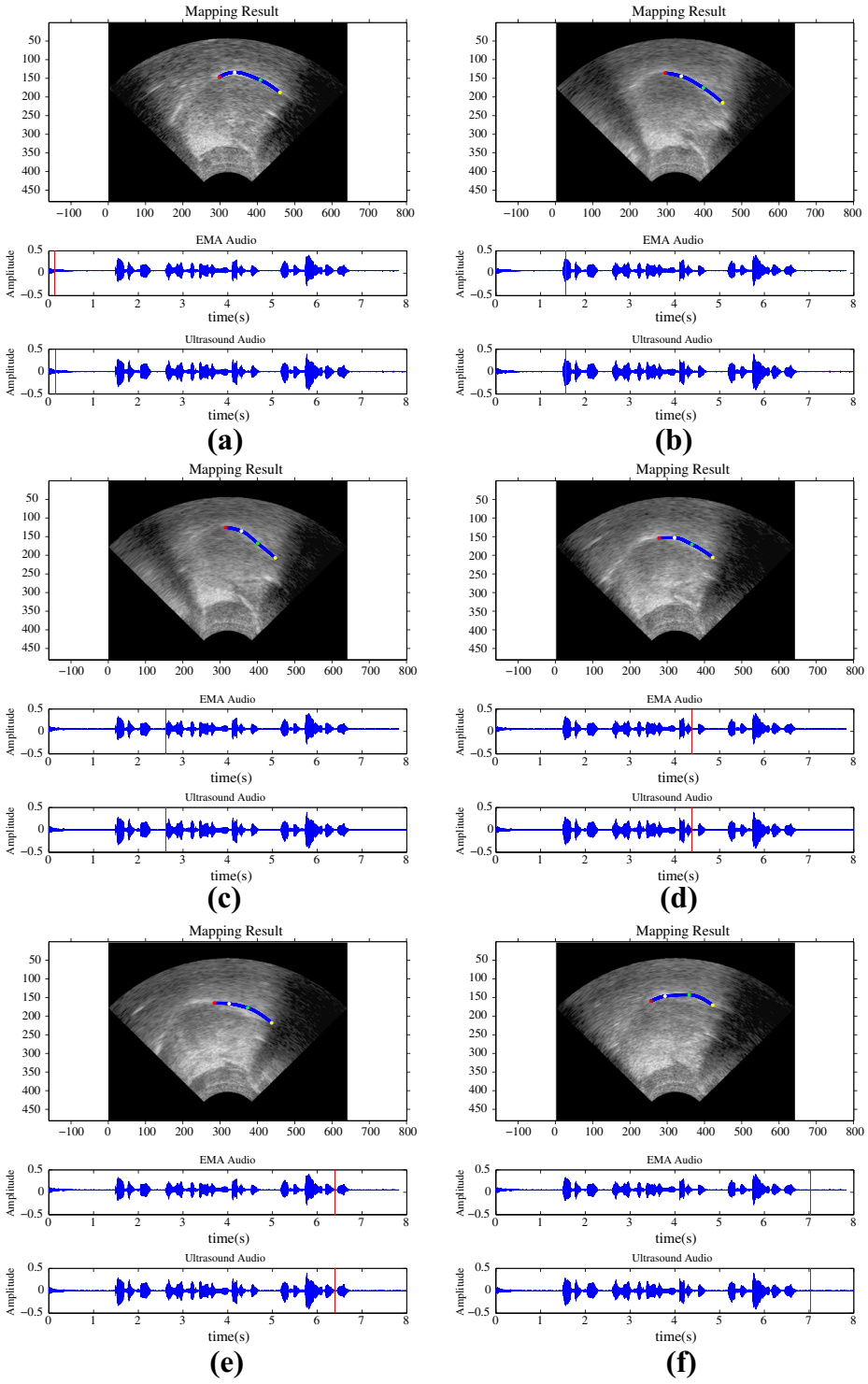



**Fig. 8** Mapping results of the sentence at different times

## 4 Modeling the tongue movements

### 4.1 Active shape model

Active Shape Model (ASM) has been successfully used to automatically track objects from images. ASM was proposed by Cootes and Taylor in 1995, as a statistical point distribution model (PDM) [4]. The shapes of the object are represented by a set of points (controlled by the shape model). The ASM algorithm is targeted at matching the model to an unseen image. This approach has been widely used to analyze images of faces [9], mechanical assemblies, and medical images (in both 2D and 3D).

An ASM describes the image shape of the object of interest by obtaining a statistical shape model in examples from a training set. ASM minimizes 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 [1, 3, 8, 14, 19, 20, 21]. The vocal tract shapes obtained from articulatory images can be applied to acoustic simulations [7].

The ASM was built in the following steps of our study [3]. Before implementing ASM, the contours on the mid-sagittal ultrasound images of the tongue shape were semi-automatically annotated for both static vowels and vowel-vowel (VV) sequences by the tool EdgeTrak [10]. In order to find the relationship between different frames, the EMA points are used as identical points in different ultrasound images which also play the role of segmenting the tongue contour into a small piece from the tongue dorsum to the tongue tip as shown in Fig. 9. All those labeled images were adopted to form the training set. In the training set, $n$ evenly distributed points were used to describe each contour of the tongue where $n = 41$. We define $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$ as the i-th contour, where $(x_k, y_k)$ are the coordinates of the k-th point.

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

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - x\right)\left(x_i - x\right)^T \tag{6}$$

where $x$ is the mean shape of all the vectors in the training set [4].

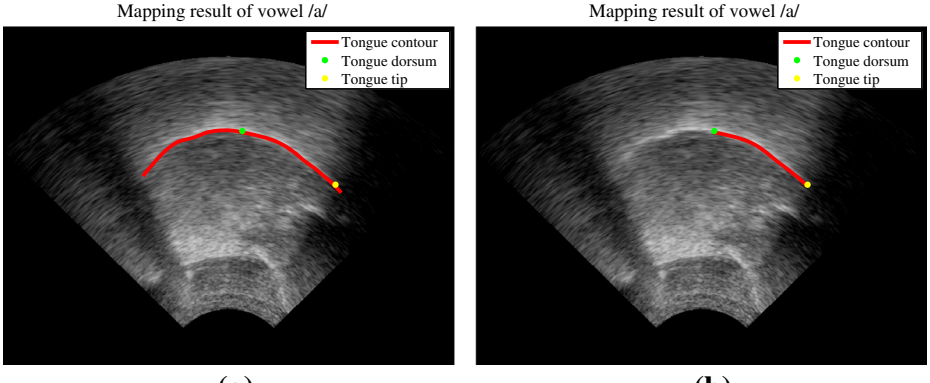



**Fig. 9** Tongue contour **a** before and **b** after segmentation by the EMA points (tongue tip and tongue dorsum)

**Table 1** Contribution of the first two factors

| λ | Eigenvalue | Percentage | Accumulated percentage |
|---|---|---|---|
| $\lambda_1$ | 32,438.24 | 71.32 % | 71.32 % |
| $\lambda_2$ | 11,046.71 | 24.29 % | 95.61 % |

Secondly, we calculate the eigenvalue sequence of $\mathbf{S} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_i \geq \lambda_{i+1}$, and where $i = 1,2,\ldots n\text{-}1$. We choose the first $t$ eigenvalues under the following conditions

$$\sum_{i=1}^{t} \lambda_i \Big/ \sum_{i=1}^{m} \lambda_i \geq 90\% \tag{7}$$

Then, we calculate the corresponding eigenvector of the first $t$ $\lambda_i$ to form **P**, recorded as $\mathbf{P} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_t})$.

After we obtain the mean shape vector $x$ and the eigenvector $P$ of the training set, our tongue shape model can be expressed as$x = x + \mathbf{Pb}$. Thus, we can obtain a **b** vector of a certain shape through the known model, $\mathbf{b} = (\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_t})$. At the same time, if we have a known **b**, we can also obtain a certain tongue shape.

## 4.2 Experiments and results

### *4.2.1 Initialization of the ultrasound images and tongue labeling*

In our experiment, the training set contains a total of 145 mid-sagittal ultrasound figures of/a/,/i/and/u/, which include dynamic tongue shapes in articulation. A total of 41 points were used to represent each contour of the training set.

### *4.2.2 PCA analysis of the training data*

After processing the training set, we conducted the PCA analysis to extract the parameters of the following model and calculate the mean shape of the training set. The result is presented in Table 1.

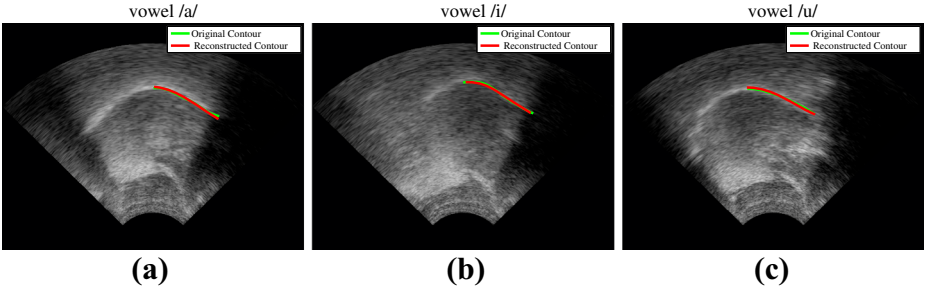



**Fig. 10** Original and reconstructed tongue contour of vowel (a)/a/(b)/i/(c)/u/

We choose the first two eigenvalues to be the main factors; the accumulated contribution rate reached 95.61 %, as shown in the Table 1. Then, we calculated the corresponding $\mathbf{p}_i$ ($i$ = 1,2), and obtained one of the parameters $\mathbf{P} = (\mathbf{p_1},\mathbf{p_2})$ of the ASM of our dataset. The tongue ASM was built by these coefficients.

### 4.3 Synthesizing tongue shape by using ASM

The tongue shapes were synthesized by using the ASM model described in the last section. The results are shown in Fig. 10, in which the reconstructed tongue shapes are compared with the original annotated tongue shapes of three isolated Chinese vowels/a/,/i/, and/u/. Thus, these results denoted that our approach is feasible for synthesizing articulation. The main cause of the differences between synthesized and real shapes is that only the first two components have been adopted in this ASM modeling procedure. The averaged error calculated following equation (5) over 40 points along the tongue contour is 1.26 mm, where k = 40 × 145 indicates the total number of sample points along the contours.

## 5 Conclusions

This paper introduces our acquisition system for observing ultrasound images and EMA data, which we analyzed and combined into a single dataset. The data combination procedure involved synchronization of these two datasets using the audio files from each set as reference to determine the alignment image with both the EMA information and ultrasound image. A dataset that included ultrasound and EMA data recorded by three subjects was built by using this data recording system and data fusion protocol.

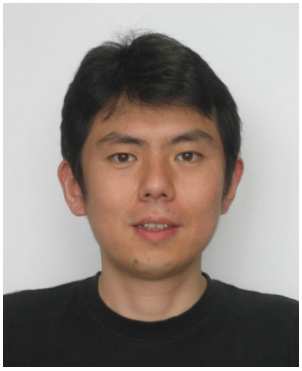
We also proposed a method to synthesize the shapes of the vocal tract by using the recorded dataset. We trained a set of parameters of the ASM-based model to control the deformation of the shape of the tongue, thereby facilitating the determination of the relationship between different frames. Furthermore, we realized the synthesis of tongue shapes by interpolating the control parameters of the ASM-based model. Finally, we evaluated our method by carrying out a comparison between the synthesized and real tongue shapes. The results indicated that our method has the capability to reconstruct tongue shapes with errors not exceeding 1.26 mm, indicating that the system could be applied for vocal tact visualization in the future.

**Acknowledgments** This work was supported by the National Natural Science Foundation (NSFC) of China (No. 61,175,016), as well as a 973 project (No. 2013CB329305), and the National Natural Science Foundation (NSFC) of China (No. 61,304,250).

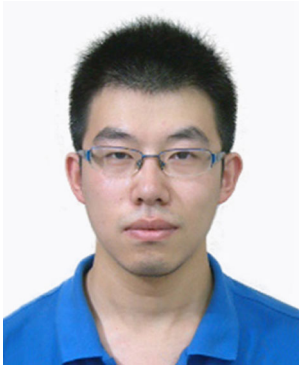
## References

1. Avila-Garcia MS, Carter NJ, Damper RI (2005) Extracting tongue shape dynamics frovm magnetic resonance image sequences. World Acad Sci Eng Technol 2

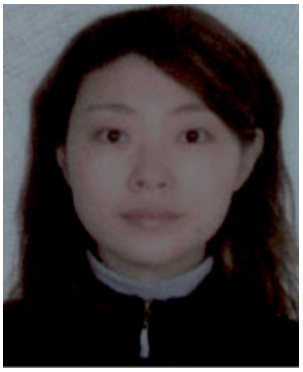
2. Boisvert J, Gobbi D, Vikal S, Rohling R, Fichtinger G, Abolmaesumi P (2008) An open-source solution for interactive acquisition, processing and transfer of interventional ultrasound images, in [workshop on systems and architectures for computer assisted interventions], Miccai 2008
3. Chan Song, Jianguo Wei, Qiang Fang, Shen Liu, Yuguang Wang, Jianwu Dang. Tongue shape synthesis based on active shape model, ISCSLIP, pp 383–386, Dec.2012, Hongkong
4. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models - their training and application. Comput Vis Image Underst 61:38–59
5. Florescu V-M, Crevier-Buchman L, Denby B, Hueber T, Colazo-Simon A, Pillot-Loiseau C, Roussel P, Gendrot C, Quattrochi S (2010) Silent vs vocalized articulation for a portable ultrasound-based silent speech interface, Proceedings of Interspeech (Makuari, Japan), pp. 450–453.
6. Hoole P, Nguyen N (1999) Electromagnetic articulography in coarticulation research, in Hardcastle, W.H., Hewlitt, N. Eds., Coarticulation: Theory, data and techniques, pp. 260–269, Cambridge University Press, 1999.
7. Jianguo Wei, Song Wang, Qingzhi Hou, Jianwu Dang (2015) Generalized finite difference time domain method and its application to acoustics. mathematical problems in engineering, vol. 2015, Article ID 640305, 13 pages
8. Lee Hung LIEW, Beong Yong LEE, Yin Chai WANG, WaiShiang CHEAH (2013) Aerial images rectification using Non-parametric approach. J Converg 4(2):15–21
9. Lepsoy S, Cuiinga S (1998) Conversion of articulatory parameters into active shape model coefficients for lip motion representation and synthesis. Signal Process Image Commun 13:209–225
10. Li M, Kambhamettu C, Stone M (2005) Automatic contour tracking in ultrasound images[J]. Clinical Linguistics & phonetics 19(6–7):545–554
11. Li A, Yin Z, Wang T, Fang Q, Hu F (2004) RASC863 - a Chinese speech corpus with four regional accents. ICSLT-o-COCOSDA, New Delhi, India
12. Mielke J, Baker A, Archangeli D, Racy S (2005) Palatron: a technique for aligning ultrasound images of the tongue and palate. In siddiqi, D., tucker, B.V. (eds.). Coyote Papers 14:97–108
13. Perkell J, Cohen M, Svirsky M, Matthies M, Garabieta I, Jackson M (1992) Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. J Acoust Soc Am 92:3078–3096
14. Shahabi C, Kim SH, Nocera L, Constantinou G, Lu Y, Cai Y, Medioni G, Nevatia R, Banaei-Kashani F (2014) Janus - Multi source event detection and collection system for effective surveillance of criminal activity. J Inf Process Syst 10(1):1–22
15. Song Wang, Shen Liu, Jianguo Wei, Qiang Fang, Jianwu Dang (2012) Reconstruction of vocal track based on multi-source image information, ISCSLP, pp 393–399, Hongkong
16. Stone M, Davis E (1995) A head and transducer support system for making ultrasound images of tongue/jaw movement. J Acoust Soc Am 98(6):3107–3112
17. Stone M, Sonies B, Shawker T, Weiss G, Nadel L (1983) Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. J Phon 11:207–218
18. Thomas H, Elie-Laurent B, Gérard C, Bruce D, Gérard D, Maureen S (2010) Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Comm 52(4):288–300
19. van Assen HC, Danilouchkine MG, Frangi AF, Ordas S, Westenberg JJM, Reiber JHC, Lelieveldt BPF (2006) SPSM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data. Med Image Anal 10:286–303
20. van Ginneken B, Frangi AF, Staal JJ, ter Haar Romeny BM, Viergever MA (2002) Active shape model segmentation with optimal features. IEEE Trans Med Imaging 21(8)
21. Verma P, Singh R, Singh A (2013) A framework to integrate speech based interface for blind web users on the websites of public interest human-centric computing and. Inf Sci 3:21



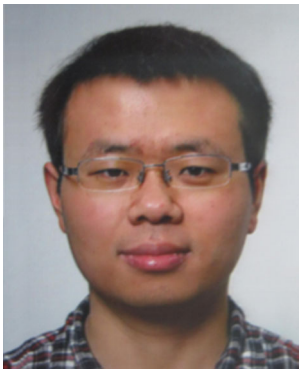
**Jianguo Wei** received the B.S. degree from Jilin University, China, in 1995, the M.S. degree from Tianjin University, Tianjin, in 2002 and the Ph.D. from Japan Advanced Institute of Science and Technology, Japan in 2007. From 2007 to 2008, he was a research fellow in CNRS/ENST. Since 2010, he has been an Associate Professor in School of Computer Software and School of Computer Science and Technology, Tianjin University. His research interests include Speech Production, Multi-modal speech processing and Speech watermarking. He is a member of IEEE, International Speech Communication Association (ISCA) and a member of Acoustic Society of Japan (ASJ).



**Song Wang** currently is a Research assistant of School of Computer Science and Technology, Tianjin University, China. He received his B.S. and M.S. degree from Tianjin University, China. His research interests are mainly related with speech production, articulatory visualization and acoustic modeling.



**Wenhuan Lu** is an associate professor at School of Computer Software in Tianjin University, P. R. China. She received her Ph.D. Degree at Japan Advanced Institute of Science and Technology (JAIST), Japan and she has also worked as postdoctoral researcher at JAIST; afterwards, she joined School of Computer Software, Tianjin University, P. R. China. She is working on knowledge modeling & knowledge representation with semantic technology applied to knowledge-based system. She is currently the Principle Investigator for several national grants in China, including National High-Technology R&D Program of China (863 Program), National Natural Science Foundation of China, and Scientific Research Foundation from Ministry of Education of China, etc. She has also (co)authored more than 20 academic articles and won the "Best paper award" in IJCSS, 2011.



**Qingzhi Hou** Lecture of Tianjin Uniersity, received his Ph.D. in applied mathematics from department of mathematics and computer science of Eindhoven University of Technology, Netherlands. He was a visiting research fellow in Nottingham University, United Kingdom and Shenzhen Research Institute of City University of Hong Kong. Now he is a lecture at school of computer science and technology of Tianjin University. His main research interests are meshless numerical methods, computer fluid simulation, visualization of flow field and its application in water conservancy, Ocean Engineering, etc.

**Qiang Fang** currently is an associate professor of Chinese Academy of Social Sciences. He received his B.S. degree from Nanjing University of Science and Technology, China, in 2001, the M.S. degree from Chinese Academy of Social Sciences, in 2004 and the PH.D degree from Japan Advanced Institute of Science and Technology, Japan in 2008. His research interests include speech production and speech modeling.

**Jianwu Dang** received the B.E. and M.S. degrees from Tsinghua Univ., China in 1982 and 1984, respectively. He worked for Tianjin University as a lecture from 1984 to 1988. He was awarded the Ph.D. Eng. from Shizuoka Univ., Japan in 1992. Dr. Dang worked for ATR human Information Processing Lab., Japan from 1992 to 2001. He joined the University of Waterloo, Canada, as a visiting scholar for one year in 1998. He has been with the Japan advanced institute of science and technology since 2001, where he is a professor. He joined the Institute of Communication Parlee, Center of National Research Scientific, France, as a research scientist the first class for one year in 2002. His research interests are in all of the fields of speech science, especially in speech production. He is a member of the IEEE, ASA, ASJ and IEICE.