SPEAKER-INDEPENDENT LIPS AND TONGUE VISUALIZATION OF VOWELS

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

This paper proposes a scheme of speech-driven lips and tongue animation synthesis in a speaker-independent manner. Directional relative displacement (DRD) features are proposed based on the Electromagnetic Articulograph (EMA) data to describe human's lips and tongue movements, which are more stable across different speakers than the raw EMA data. Multi speakers' acoustic-articulatory data of vowels are used to learn the acoustic-to-articulatory inversion mapping. We build 2D geometric models of lips and tongue for visualization. With the trained mapping and the geometric models, visualization of lips and tongue movements from acoustic signal of vowels uttered by arbitrary speaker is realized. The experimental results demonstrate that the animations we synthesized are effective aids in helping people identifying vowels.

Index Terms— speech visualization, articulatory models, electromagnetic articulograph, acoustic-to-articulatory inversion

1. INTRODUCTION

The visual information of lips and tongue's movements can enhance speech perception, especially under noisy environment or when one or more talkers have hearing disorder [1]. They are also effective hearing aids in second language learning and teaching hearing impaired people how to speak [2]. Therefore, visualization of lips and tongue movements with acoustic signal would have many potential applications, which is a technology that integrates visualization method and acoustic-to-articulatory inversion mapping algorithm.

Speech visualization technology with virtual articulatory models has been developed by many researchers. Two-dimensional (2D) or three dimensional (3D) articulatory mesh models of lips, tongue and jaw are generally used in prior woks [3-5]. To control the motion of these articulatory models, various types of data recorded from the real speaker are used, such as 3D motion capture data and Electromagnetic Articulograph (EMA) data. Most of prior works extract control parameters from the 3D coordinates of sensors attached on speaker's organs to control their virtual models [4-6]. The reconstruction of articulatory movements from acoustic signal is considered as a difficulty and ill-posed problem for the highly nonlinearity and the "one-to-many" nature. Many corpus-based methods have been proposed such as codebook model [7], mixture density network [8], and Gaussian Mixture Model (GMM) based mapping [9]. All these approaches use single speaker's acoustic-articulatory data to learn the mapping between acoustic space and articulatory space. Ghosh and Narayanan [10] proposed a subject-independent acoustic-to-articulatory inversion method use only one speaker's data to learn the mapping. However, the acoustic-articulatory data from single speaker is limited, and the widely adoption of EMA device makes it possible to obtain multi speakers' EMA data.

To control virtual articulatory models using multi speakers' data is the aim of our work. For this purpose, the coordinates of EMA coils are not suitable features to represent articulatory movements, because the shape and size of human articulator varies across speakers and virtual models, and EMA coils may shift or be reattached during recording procedure. We propose directional relative displacement (DRD) features as representation of articulatory space, which are more stable across different speakers than the raw EMA data. GMM-based mapping method is adopted to construct acoustic-to-articulatory inversion mapping, and multi speakers' acoustic-articulatory data are used to learn and test the mapping. We built 2D geometric models of lips and tongue with B-spline curves for visualization, which performs well in showing animation for its clarity. The estimated DRD features are then used to drive the models in real-time. The experiment results show that the 2D animations we synthesized are helpful for vowels identification.

The rest of this paper is organized as follows. Section 2 gives the details of our acoustic-articulatory database. Section 3 describes the DRD features. In Section 4, geometric models of lips and tongue are described. In Section 5, The GMM-based mapping method is reviewed. The experiments and results are shown in Section 6. We summarize this paper in Section 7.

2. DATA ACQUISITION

In this study, we built an acoustic-articulatory database which includes six speakers' data. EMA device was used to capture the movements of lips and tongue, and acoustic signal was recorded synchronously by a uni-directional microphone. The EMA sam-

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

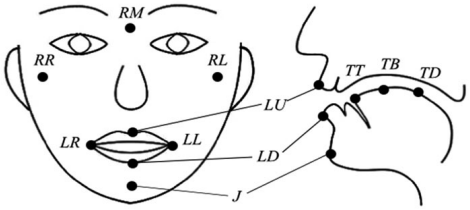

Fig. 1. EMA coil positions.

pling frequency was set to 100Hz, and the audio recording sampling frequency was 22.05 kHz. Three male and three female adult Chinese speakers participated in the recording procedure. None of the speakers was professional broadcaster, and their utterance may have different accents even they were requested to pronounce as accurate as they can. Speakers were asked to keep their mouth closed and keep their tongue close to palate when they were not uttering.

The EMA coil positions are shown in Fig. 1. We use eleven coils: coils *RR* (reference right)*, RM* (reference middle) and *RL* (reference left) are reference coils, they were used to record the rigid movement of the head; coils *LR* (lip right), *LU* (lip up), *LL* (lip left) and *LD* (lip down) were used to record the movement of lips; coils *TT* (tongue tip), *TB* (tongue body) and *TD* (tongue dorsum) were used to record the movement of tongue; coil *J* (jaw) was used to record the rotational movement of jaw. Coils *TT*, *TB* and *TD* were glued on symmetric line of tongue's upper surface, and other coils were glued on face.

The corpus consisted of 6 simple vowels of Mandarin: a, o, e, i ,u, ü; and 29 compound vowels of Mandarin: ai, ei, ao, ou, ia, ie, iao, iou, ua, uo, uai, uei, üe, an, en, ang, eng, ong, ian, in, iang, ing, iong, uan, uen, uang, ueng, üan, ün (denoted by Chinese pinyin). Each term was uttered twice by each speaker and each speech session lasts for 3 seconds with about 1 second silence period at the beginning and the end of the speech session.

3. FEATURE EXTRACTION

3.1. Acoustic feature extraction

The original speech signal were divided into acoustic frames, the frame length and shift were 20ms and 10ms, respectively. Thus, the acoustic frame rate was 100Hz, which is the same as the sampling rate of EMA data. The RMS amplitude and 16-order Line Spectral Pairs (LSPs) were adopted as acoustic features. The features were extracted by Speech Signal Processing Toolkit (SPTK)[11].

3.2 Articulatory feature extraction

We propose directional relative displacement (DRD) features as representation of the lips and tongue's movements because they are more stable across different speakers than the raw EMA data. The DRD features are explained as follows. For each frame, we calculate each EMA coil's displacement, which is the Euclidean distance of the coil's position to its initial position. A relative displacement is the ratio of an EMA coil displacement to a normalization unit, and a DRD feature is the projection of a relative displacement on

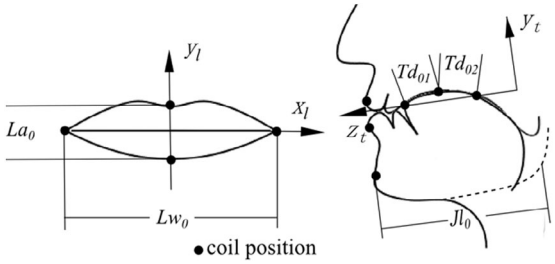

Fig. 2. Initial positions of EMA coils and definition of main directions and normalization units for DRD features.

one of the main directions. Initial positions of EMA coils are obtained from silence period of a speech session where speaker's mouth is closed and tongue is kept close to palate. See Fig. 2 for the initial position of coils and definition of main directions and normalization units. Table 1 shows the details of DRD features. The subscript 0 denotes initial position of each coil， $X_t - X_0$ denotes the displacement vector of coil X at *t*-th frame. Lw_0 is the distance between right lip coil and left lip coil, and La_0 is the distance between upper lip coil and lower lip coil at initial position. Td_0 is defined by $Td_0 = (Td_{01} + Td_{02})/2$, where Td_{01} is the Euclidean distance between TT_0 and TB_0 , Td_{02} is the Euclidean distance between TB_0 and TD_0 . Jl_0 is the length of the speaker's under jaw. $Lw_0/2$, Td_0 and Jl_0 are defined as normalization units for lips, tongue and jaw, respectively. Vectors $\vec{x}_l$, $\vec{y}_l$ and $\vec{z}_l$ ($\vec{z}_l = \vec{x}_l \times \vec{y}_l$) are orthonormal vectors for lip space, the direction of $\vec{x}_l$ is form LR_0 to LL_0 , and the direction of $\vec{y}_l$ is from LD_0 to the LU_0 ; vectors $\vec{y}_t$ and $\vec{z}_t$ are orthonormal vectors for tongue space, the direction of $\vec{z}_t$ is from TD_0 to TT_0 , and $\vec{y}_t$ is perpendicular to $\vec{z}_t$ in mid-sagittal plane. The directions of those orthonormal vectors are defined as main directions. As we see, different speakers have different initial parameters Lw_0 , La_0 , Jl_0 and Td_0 , and the main directions also change across speakers. We recalculate these parameters at silence period of each speech session. By introducing the initial parameters, we reduce the effects that are caused by the speakers' variability in lips and tongue shapes and EMA coils' position shift. Because of the symmetry of lips, we use *Lw* (lip width) instead of the DRD features of *LR* and *LL* on x_l directions. *La* (Lip aperture) is the combination of DRD features of *LU* and *LD* on y_l directions. *Ja* (jaw angle) is approximation of the angle of jaw's rotational movement.

Generally speaking, the lips and tongue move slowly and smoothly, and the articulatory feature sequences should be low-pass, however, the original EMA data contain high frequency components [12]. Therefore, we smooth the DRD feature sequences with low-pass filter. Cutoff frequency of the filter was set to 15Hz.

4. LIPS AND TONGUE MODEL

We build 2D lips and tongue geometric models with B-spline curves, which will also be called curve models in the rest of this paper. The models consist of front lip model and lateral model. These models are abbreviated compared to 2D or 3D mesh models, but can perform well in showing animation for its clarity. As is shown in Fig. 3(a), the front lip model is constructed by four curves, two for upper lip and two for lower lip. There are four

Table 1. Definition of DRD features.

Feature	Definition
Lw (lip width)	$\|LL_t - LR_t\| / (Lw_0 / 2)$
La (lip aperture)	$((LU_t - LD_t) \cdot \vec{y}_l - La_0) / (Lw_0 / 2)$
LU_z	$(LU_t - LU_0) \cdot \vec{z}_l / (Lw_0 / 2)$
LD_z	$(LD_t - LD_0) \cdot \vec{z}_l / (Lw_0 / 2)$
Jw (jaw angle)	$\|J_t - J_0\| / Jl_0$
TT_y	$(TT_t - TT_0) \cdot \vec{y}_t / Td_0$
TT_z	$(TT_t - TT_0) \cdot \vec{z}_t / Td_0$
TB_y	$(TB_t - TB_0) \cdot \vec{y}_t / Td_0$
TB_z	$(TB_t - TB_0) \cdot \vec{z}_t / Td_0$
TD_y	$(TD_t - TD_0) \cdot \vec{y}_t / Td_0$
TD_z	$(TD_t - TD_0) \cdot \vec{z}_t / Td_0$

points that control these lip curves, corresponding to four EMA coils glued on the lips. Fig. 3(b) shows the lateral model, in which we use five curves to represent the surface of upper lip, lower lip and jaw, hard palate, lower teeth, and tongue in mid-sagittal plane, respectively. Each key point on those curves corresponds to the EMA coil on the corresponding position. Lip thickness is kept stable. The lower teeth curve will rotate rigidly with the jaw key point, and the hard palate curve is fixed. The front lip model is capable to show the deformation of lips in front view while the lateral model is capable to show the lips open and close movements, lips extension, rotational up and down movements of jaw and deformations of tongue in the mid-sagittal plane. The deformable curves between the key points are interpolated by B-spline interpolation algorithm. Control points of B-spline are obtained by reverse the key points so that the B-spline will go through the key points.

Because of the correspondence of the model key points and EMA coils, we define the main directions and normalization units of models using the same method as shown in Fig. 2 treating model key points as EMA coils. Initial parameters of the models are decided by a real speaker's organs (lips in the front model are scaled larger on its size). The key points' positions in each frame can be calculated by an inverse process of the DRD features extraction method. Therefore, we can reconstruct the movement of lips and tongue through DRD features.

5. GMM-BASED MAPPING

We apply the GMM-based method to the inversion mapping. This method has been adopted for acoustic-to-articulatory inversion mapping in [9]. Let x and y denote acoustic feature vector and articulatory feature vector (DRD feature vector in our case), respectively. In this method, a GMM on joint probability $P(x, y \mid \lambda)$ is trained with the EM algorithm at the beginning, and the mapping function from an acoustic feature vector to an articulatory feature vector is

$$\hat{y}_t = \sum_{m=1}^{M} P(m \mid x_t, \lambda) E_{m,t}^{(y)} \tag{1}$$

$$P(m \mid x_t, \lambda) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})} \tag{2}$$

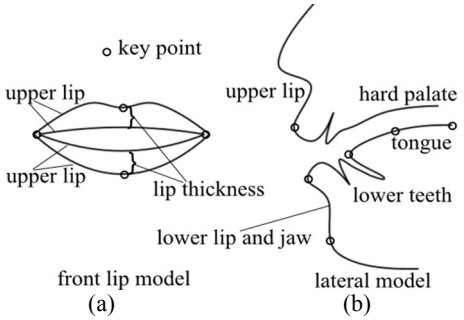

Fig. 3. Lips and tongue models consist of B-spline curves.

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)-1} (x_t - \mu_m^{(x)}) \tag{3}$$

where the $\hat{y}_t$ is the predicted articulatory feature vector and x_t is the acoustic feature vector in frame t. The total number of mixtures is *M*. A set of model parameters λ consists of weights, mean vectors and covariance matrices. The weight of the *m*-th mixture component is w_m. The vectors $\mu_m^{(x)}$ and $\mu_m^{(y)}$ denote the mean vectors of the *m*-th mixture for x and y, respectively. The matrices $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yx)}$ denote the covariance matrix of the *m*-th mixture for x and the cross-covariance matrix of the *m*-th mixture for x and y, respectively.

6. EXPERIMENTS AND RESULTS

6.1. Objective evaluation

We use 393 pairs of acoustic-articulatory parallel data from our database in the experiments. Silence frames were removed from the database using a threshold-based silence detection method after the DRD features were calculated. Three male speakers and three female speakers are represented by M1, M2, M3 and F1, F2, F3, respectively. There are 9 different male-female groups. Each time, we chose one group's data as test data and use the other four speakers' data to train the GMM on joint probability, to make sure that both training data and test data were gender balanced. Full covariance matrixes were used in our experiments. The number of Gaussian mixtures was varied from 1 to 32 (1 2 4 8 16 32). We smooth the estimated DRD feature sequences with the same filter that used for input DRD features (cutoff frequency is 15Hz). Fig. 4 shows the mean RMS error of estimated and measured DRD features as a function of number of Gaussian mixtures, Fig. 5 shows the estimated and measured DRD features trajectories for a compound vowel uttered by speaker M1.

6.2. Subjective evaluation

The inversion mapping is performed frame by frame in our scheme, which makes 100 DRD feature vectors per second. However, it is not necessary to use such a high frame rate in animation. We divided the estimated DRD feature sequences into animation frame, the frame length was set to 9 samples and the frame shift was set to 2 samples. For each animation frame, we use the mean value of the 9 samples to drive the models. The tongue's upper surface curve would go cross the palate curve if features TT_y, TB_y and TD_y

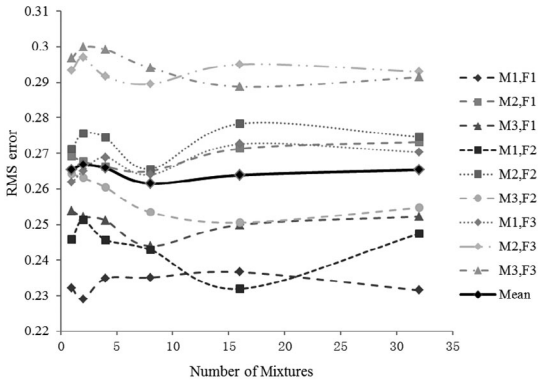

Fig. 4. Mean RMS error between measured and estimated DRD features as a function of number of Gaussian mixtures. The labels show the speakers whose data were used for test while the others were used to train the GMM on joint probability.

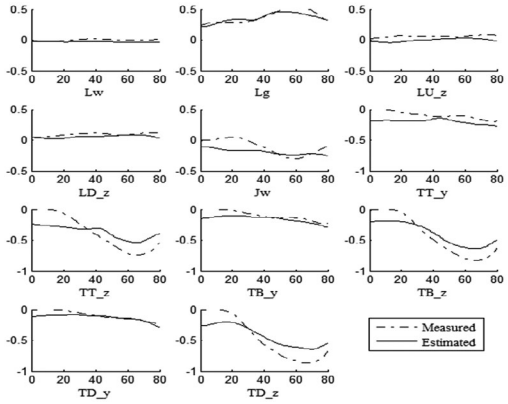

Fig. 5. A sample of measured and estimated trajectory of DRD features of a compound vowel.

are positive according the definition of DRD features. Therefore, we set the three features to zero whenever the estimated values are positive (which happens occasionally).

In the subjective evaluation test, we use the mapping model trained by the data of speaker M2, M3, F2, and F2, and the acoustic signal of M1 and F1 were used to drive the models. The number of Gaussian mixtures was set to 8. Two groups of compound vowel were used in our test, and each of them consisted of 4 confusable compound vowels. We show the synthesized animations driven by one group of compound vowels uttered by one test speaker to 9 subjects without audio each time, and let the subjects to label the animations with 4 optional pinyin symbols (each symbol may use only once). The subjects have not been trained on lip or tongue reading. Fig. 6 shows a series of frames that driven by the word "ia" uttered by a M1.

Table 2 shows the ratios for each compound vowel be identified correctly by 9 subjects. The accuracies are much higher than random guess, except "ua" uttered by M1, which indicate that the animations are greatly helpful for identification of vowels when the acoustic signal is absent. Our test result agrees with the work by Badin [1] which also indicates that the visual information of lips and tongue can significantly improve the speech identification accuracy when the acoustic signal is absent. The results also prove

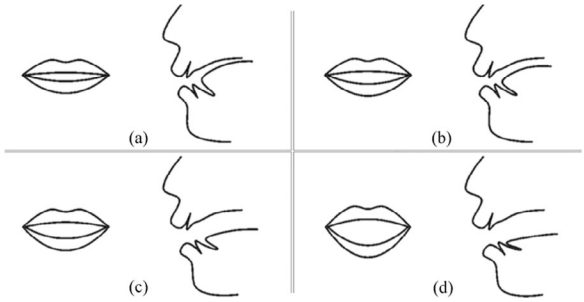

Fig. 6. Four frames of synthesized animation that driven by "ia" uttered by M1.

Table 2. The identification accuracy of compound vowels by 9 subjects given only synthesized animations.

		Test group 1			
Speaker	ai	ia	ao	ua	Mean
M1	0.56	0.44	0.89	0.22	0.53
F1	0.78	0.67	0.56	0.78	0.69
		Test group 2			
Speaker	ei	ie	ou	uo	Mean
M1	0.67	0.44	0.56	0.56	0.56
F1	0.78	0.56	0.67	1.00	0.75

that the curve model animations are similar with the movements of real human's lips and tongue movements, so that we can identify them with our life experiences.

7. CONCLUSIONS

This paper presents an experimental study on speech visualization of lips and tongue. We propose DRD features as a bridge between different speakers' EMA raw data and our curve models. Multi speakers' parallel data are used to learn the GMM-based mapping, which makes the mapping speaker-independent. We can synthesize movements from speech signal of vowels uttered by an arbitrary speaker with this method, and the 2D animations we synthesized prove to be helpful for vowels identification, even though the B-spline curves are not capable to show many details of the non-linear deformation of lips and tongue. A larger database which consists of more speakers' data is needed to make the performance of our system more stable, and more systematic tests are needed to evaluate the performance. It is also part of our future work to expand our scheme to syllable and continuous speech visualization for hearing aids.

8. RELATION TO PRIOR WORK

The work presented in this paper has focused on the visualization of articulators from acoustic signal frame by frame, while the work by Wang et al [6] visualize articulators in phoneme level. We use more than one speaker's EMA data to learn the mapping, which is different from Ghosh and Narayanan [10] who propose a subject-independent acoustic-to-articulatory inversion method. The present study is also related to articulatory model control with EMA data [1, 4]. The DRD features we proposed are inspired by the definition of face animation parameters (FAPs) [13].

9. REFERENCES

[1] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication,* vol. 52, pp. 493-503, 2010.

[2] D. W. Massaro and J. Light, "Using visible speech to train perception and production of speech for individuals with hearing loss," *Journal of Speech, Language and Hearing Research,* vol. 47, pp. 304-320, 2004.

[3] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis — Determination, adjustment, evaluation," *Speech Communication,* vol. 44, pp. 141-154, 2004.

[4] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication,* vol. 41, pp. 303-329, 2003.

[5] M. Kaihui, T. Jianhua, c. Jianfeng, and Y. Minghao, "Real-time speech-driven lip synchronization," in *Universal Communication Symposium (IUCS), 2010 4th International*, Beijing, China, pp. 378-382, 2010.

[6] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication,* vol. 54, pp. 845-856, 2012.

[7] S. Suzuki, T. Okadome, and M. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *Proc. ICSLP*, Sydney, Australia, pp. 2251-2254, 1998.

[8] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. ICSLP*, Pittsburgh,USA, pp. 577–580, 2006.

[9] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. ICSLP*, Jeju, Korea, pp. 1129–1132, 2004.

[10] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proc. ICASSP*, Prague, Czech Republic, pp. 4624-4627, 2011.

[11] S. Imai*, et al.*, "Speech signal processing toolkit (SPTK), Version 3.5," http://sp-tk.sourceforge.net, 2011.

[12] 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.

[13] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication,* vol. 15, pp. 387-421, 2000.