Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 845–856

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Phoneme-level articulatory animation in pronunciation training

Lan Wang [a,*], Hui Chen [c], Sheng Li [a], Helen M. Meng [a,b]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[b] The Chinese University of HongKong, China
[c] Institute of Software, Chinese Academy of Sciences, China

Received 10 December 2010; received in revised form 18 February 2012; accepted 18 February 2012
Available online 13 March 2012

Abstract

Speech visualization is extended to use animated talking heads for computer assisted pronunciation training. In this paper, we design a data-driven 3D talking head system for articulatory animations with synthesized articulator dynamics at the phoneme level. A database of AG500 EMA-recordings of three-dimensional articulatory movements is proposed to explore the distinctions of producing the sounds. Visual synthesis methods are then investigated, including a phoneme-based articulatory model with a modified blending method. A commonly used HMM-based synthesis is also performed with a Maximum Likelihood Parameter Generation algorithm for smoothing. The 3D articulators are then controlled by synthesized articulatory movements, to illustrate both internal and external motions. Experimental results have shown the performances of visual synthesis methods by root mean square errors. A perception test is then presented to evaluate the 3D animations, where a word identification accuracy is 91.6% among 286 tests, and an average realism score is 3.5 (1 = bad to 5 = excellent).

Keywords: Phoneme-based articulatory models; HMM-based visual synthesis; 3D articulatory animation

1. Introduction

Technologies enabling speech visualization with a synthesized talking head have been developed for applications from speech animation to language acquisition for hearing-impaired children (Massaro et al., 2004; Rathinavelu et al., 2007). Earlier studies presented animated lip movements to improve the intelligibility of synthetic speech. Recent investigations have used physiology-based head models with more realistic visual synthesis to enhance speech perception (Tarabalka et al., 2007; Wik and Engwall, 2008; Serrurier and Badin, 2008). Most of the existing efforts in this direction focus on the following aspects: head modeling to design 2D or 3D articulator models (Serrurier and Badin, 2008; Badin et al., 2008), data acquisition to control the synchronized movements of articulators (Fagel et al., 2004; Deng and Neumann, 2008; Badin et al., 2010), and visual synthesis which uses the limited data to synthesize any articulatory motions in speech production (Wang et al., 2009; Ma et al., 2004).

Virtual head models in recent studies have used 2D/3D models of face, lips, tongue and jaw, even velum and nasopharyngeal wall (Serrurier and Badin, 2008). Both facial and intra-oral articulators are modeled on the basis of physiological data, like MRI, CT or X-ray images. Related works in (Fagel et al., 2004; Grauwinkel et al., 2007; Wik and Engwall, 2008) have enabled the visualization of internal articulator dynamics on synthesized audiovisual speech. To control the motion of these articulatory models, various types of data recorded from the real speakers is used. Data acquisition equipments include three-dimensional facial motion capture (Ma et al., 2004; Engwall and mri, 2003) and Electro-Magnetic Articulography (EMA) (Engwall and mri, 2003; Fagel et al., 2004; Tarabalka et al., 2007).

* Corresponding author. Tel.: +86 75586392171.
E-mail addresses: lan.wang@siat.ac.cn (L. Wang), chenhui@iscas.ac.cn (H. Chen), sheng.li@siat.ac.cn (S. Li), hmmeng@se.cuhk.edu.hk (H.M. Meng).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.02.003

In these studies, a motion capture device was used to record the complex movements of lips, jaw and tongue at certain discrete data points, which were mapped to a 3D face model. The EMA was commonly used for audio-visual data recording since the coils can be attached to the intraoral articulators to get the positions of tongue, teeth, jaw, etc.. The video-fluoroscopic images (Murray et al., 1993; Wang et al., 2009; Chen et al., 2010) also provide the data of recording movements of tongue body, lips, teeth, palate and oral cavity in real-time. Because speakers would be exposed to excessive amounts of X-ray radiation during long-term recording, limited data was collected and used for speech visualization.

Most research efforts (Ma et al., 2004; Fagel et al., 2004; Grauwinkel et al., 2007; Tarabalka et al., 2007) defined a viseme or vowel/consonant group as the basic unit for audio-visual synthesis. To animate the movements of a viseme, a set of displacement vectors or parameters were determined for the corresponding vertices of anticipatory and target viseme positions in a 3D synthetic head models. The visualization of articulatory movements were acquired by linear or non-linear combinations of these displacement vectors. The articulatory dominance functions in (Cohen and Massaro, 1993) has been widely used, where the contour was controlled by time offset, duration and magnitude. It is a general form for visual synthesis and our preliminary study (Wang et al., 2009) had explored how to produce the control parameters of the dominance function for any phonetic context. The study in (Guenther et al., 1995) had presented a Directions Into Velocities of Articulators model for speech production, where an orosensory-to-articulatory mapping was proposed but the computation of this mapping is complex. The statistical model (Blackburn and Young, 2000) was also used to predict articulator movements with a set of X-ray trajectories for training. The study in (Dang et al., 2005) developed a carrier model where the articulation was represented by a vocalic movement and a consonantal movement. The linear form of articulation was evaluated by the distribution of tongue tip during the central vowel of VCVCV sequence. For the evaluation, the synthetic articulator movements were compared to the natural motions, the subjects evaluated the intelligibility of the synthesis (Fagel et al., 2004). The work in (Tarabalka et al., 2007; Wik and Engwall, 2008) evaluated the subjects abilities to use the visual information to perform recognition of consonants or sentences with degraded audio. In these works, the subjects gave identification scores for the stimuli under different conditions, including audio information only, audio-visual information combined with tongue and audio-visual information combined with both tongue and face.

Recent developments of speech visualization involve a range of explorations and technologies on speech production, synthesis and perception. The task can be extended to use animated talking heads for computer assisted language learning, as in previous works (Wik and Hjalmarsson, 2009; Wang et al., 2009; Chen et al., 2010). This research investigates the phoneme-level articulatory movements, using visual synthesis methods and a transparent 3D talking head to present both external and internal articulatory animations for pronunciation training. With the multimodal information, language learners would more easily understand and mimic the phonetic sounds that are not present in his/her mother language. For this purpose, we used EMA AG500 to collect three-dimensional visual data, and designed a corpus to explore articulatory distinctions at the phoneme level. A phoneme-based concatenation with smoothing is then presented for visual synthesis, where a blending method is investigated for a good fit. For comparison, an HMM-based visual synthesis is also conducted using a Maximum Likelihood Parameter Generation (MLPG) to smooth the articulatory trajectory. Given the speech input, articulatory movements are synthesized to control the 3D articulator models, where a dynamic displacement-based deformation is applied. In the experiments, we compared two visual synthesis methods to EMA-recorded curves, and then implemented a set of 3D animations to depict the differences among phonemes. We also presented an audio-visual test in which the subjects evaluated the 3D animations using an identification accuracy and a realism score.

The rest of this paper is organized as follows: Section 2 introduces the recording of 3D articulatory motions and EMA data processing, to explore articulatory distinctions of phonemes. In Section 3, two different methods are proposed to synthesize articulatory motions for continuous speech. Section 4 gives a description on developing a data-driven 3D talking head system for articulatory animation. Section 5 presents experimental details and results. The conclusions and discussions are in the final section.

2. Recording the 3D articulatory motions and data processing

2.1. Corpus design and data collection

Studies on phonetics (Lado, 1957) have indicated that the phonemes that are absent from the learner's native language are difficult for learners to pronounce correctly. The phonetics flash animation project in Iowa University (Iowa, xxxx) has illustrated the articulation of the sounds of English/Spanish/German for students of phonetics and foreign language. In this work, we focus on the acquisition of English (i.e. the target secondary language) by learners whose native language is Chinese, so as to develop a 3D talking head to instruct the learners to produce the sounds.

To make discriminations among those sounds that are not present in the Chinese language and easily mis-pronunced by learners, the prompts for the audio-visual data collection are designed to contain two sessions. Session I covers all individual English phonemes in terms of IPA, each of which is followed by an example word as in the Appendix. It consists of a total of 45 phonemes and 45 example words including 3 polysyllabic words (more than

one vowel). Session II consists of 50 groups of minimal pairs (a total of 200 words including 21 polysyllabic words), in order to compare the differences in articulations between two similar phonemes. The average number of occurrences of each phoneme in the prompts is 27, and all phonemes appear more than three times.

To collect the audio-visual data, a native speaker is invited to read the prompts, while articulatory movements are recorded using the EMA AG500 at 200 frames per second. Each recording lasts about 15 seconds, and the movements are synchronized with the speech waveform. To record both external the internal articulatory movements, seven coils are placed on the speaker's face and three coils on her tongue. In particular, three coils are put on the nose and ears $H_{1,2,3}$ for calibration of the head movements (Hoole et al., 2003; Hoole et al., 2010), as shown in Fig. 1 (left side). The coils are put on the right lip corner L_1, the upper lip L_2 and the lower lip L_3. The coils on the tongue are tongue tip, T_1, tongue body T_2 and tongue dorsum T_3. Another coil is attached on the lower incisors for the jaw J_1. The right side of Fig. 1 shows the 3D head models with the corresponding points. Moreover, two cameras are recording the speaker simultaneously from the frontal and profile views for supplementary observation. During recording, the software provided by the manufacture is used to monitor the movements of all coils. The speaker is asked to repeat the prompts when any coil has unusual movements. The head motion normalization is post-processed with the tools provided by the manufacture (Hoole et al., 2003; Hoole et al., 2010), any EMA frame with outlier is discarded.

2.2. Feature state determination

With the use of individual phonemes of Session I, we recover the 3D positions of lips and tongue to illustrate the slight differences among phonemes. The articulator motions are tracked through the successive frames, and the silence frames and speech frames are automatically labeled in a recording by an automatic speech recognition system. The static state is selected from the silence frames to define the starting point of articulatory movement. During data collection, the speaker is asked to keep her lips and tongue still before speaking. The video can help us to observe whether the speaker has lip smacking or swallowing, and then select the silence region without lip motions. However, her tongue is not always at the same position in the silence region. We then use an analysis window (20ms) to segment the silence frames without overlapping. By calculating the mean and variance of the tongue position in a segment, we assume that the tongue in a segment with the smallest variance is relatively stable. Thus the average tongue position of this segment is then defined as the static state. For each EMA recording in Session I, the static states of lips and tongue are given below,

$$L_i(pos_{stat}), \quad i \in 1,2,3,4; \qquad T_j(pos_{stat}), \quad j \in 1,2,3,$$

where pos_{stat} refers to three dimensional position (x,y,z) of coil L_i or T_j. In Fig. 2, the silence segments in a recording are plotted where the solid line of tongue is the selected static state, and the dash lines are the other average positions of tongue. The selected static states have a small variance, compared to the tongue positions while speaking. Since there is no coil put on the left lip corner L_4, its three-dimensional position is estimated by assuming that the line between H_2 and H_3 is parallel to the line between L_1 and L_4, and the distance between L_1 and L_2 is the same as that between L_4 and L_2.

The feature state is defined as the peak position of an individual phoneme, which should refer to the characteristics of producing this sound. For instance, the peak position of the phoneme /æ/ should be selected with the maximally opened mouth, while the tongue is also at its lowest point. Fig. 3 depicts the lips and tongue positions of individual vowels /ɛ/ versus /æ/. It shows the profile view of lips and tongue at the feature state, in contrast to that of the static state.

Fig. 4 indicates the Z-axis movements of the upper and lower lips along with the time scale, where the feature state is chosen at t_{p_k}. By checking the articulatory movements of all individual phonemes, we confirmed that the EMA recordings can illustrate the distinctions between the confusable phonemes, like /ɛ/ and /æ/. Given the large number of frames in the EMA-recordings, the feature state of each

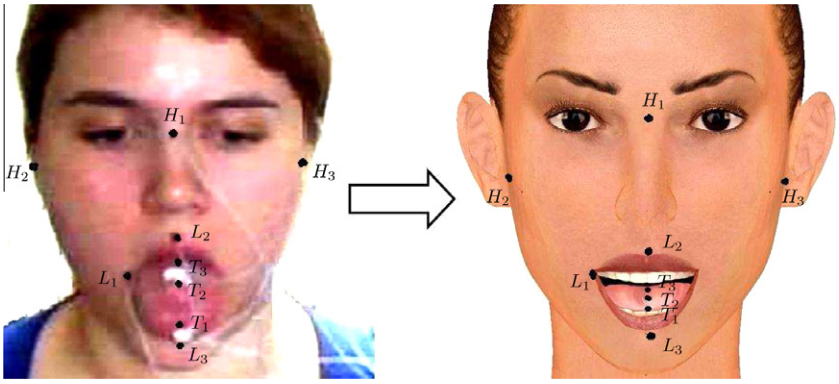

Fig. 1. The EMA coils on the speaker face and the corresponding points on a 3D head model.

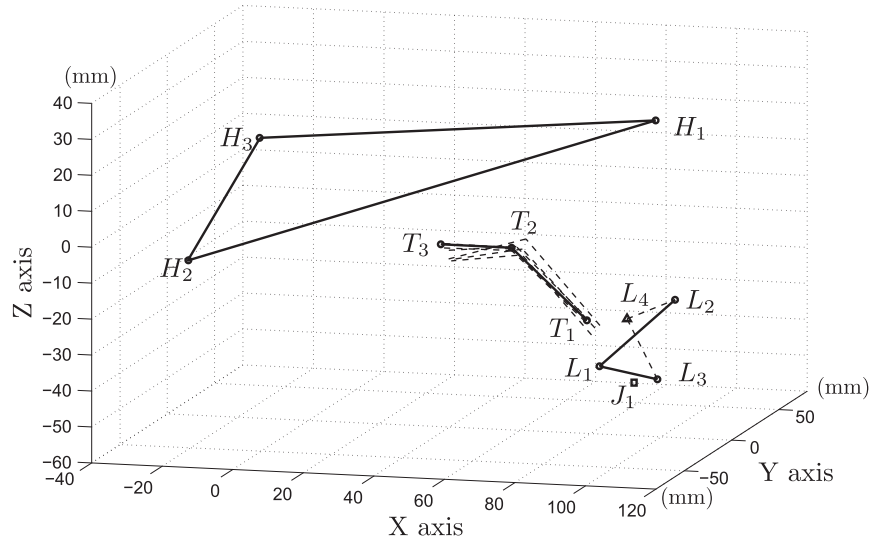

Fig. 2. The variance of the tongue positions while keeping the reference coils static.

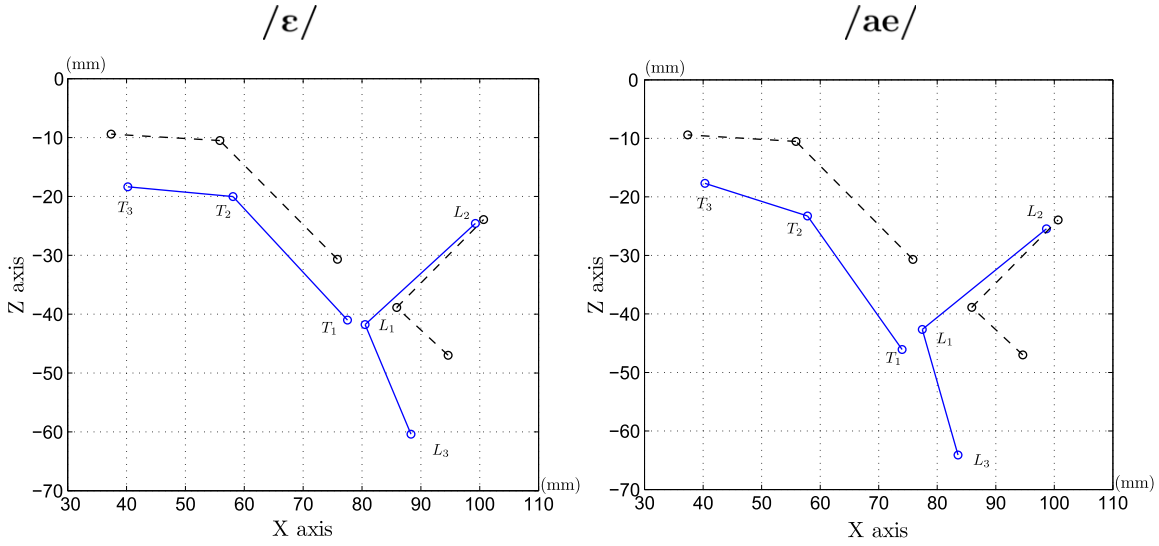

Fig. 3. The static state (dash lines) and feature state (solid lines) of the vowels */ɛ/* and */æ/*.

phoneme should be selected automatically. By searching through all EMA frames of a phoneme, the state that has maximal Euclidean distance from the static state is selected as the feature state. Since each EMA recording has its static state, we calculate the feature state for every individual phoneme in this recording as below,

$$t_{p_k} = \arg\max_{t_p}\Big\{\sqrt{\sum_i [L_i(pos,t_p) - L_i(pos_{stat})]^2} + \sqrt{\sum_j [T_j(pos,t_p) - T_j(pos_{stat})]^2}\Big\}, \quad (1)$$

where $t_{p_s} \leqslant t_p \leqslant t_{p_e}$ refers to the phoneme p starts at t_{p_s} and ends at t_{p_e}, and the feature state of this phoneme is determined at t_{p_k}.

For diphthongs, stops and liquids, one feature state may not represent articulatory characteristics. According to Eq. (1), the feature state of **/b/** is selected at $t_{p_{k1}}$, which has the maximal open mouth as shown in Fig. 5. However, another important feature to produce this sound should be at $t_{p_{k2}}$, where the lips brought together to obstruct the oral cavity. So, we manually chose the successive frames to determine the feature state, and then compute the local maximum using Eq. (1) within this interval. The selected feature states for all individual phonemes are illustrated and checked, in order to represent the distinctions of this phoneme.

3. Audio-visual synthesis based on the phoneme-level articulatory movements

3.1. Phoneme-based articulatory models with smoothing

To synthesize three-dimensional articulatory motions for natural speech, a straightforward way is to concatenate

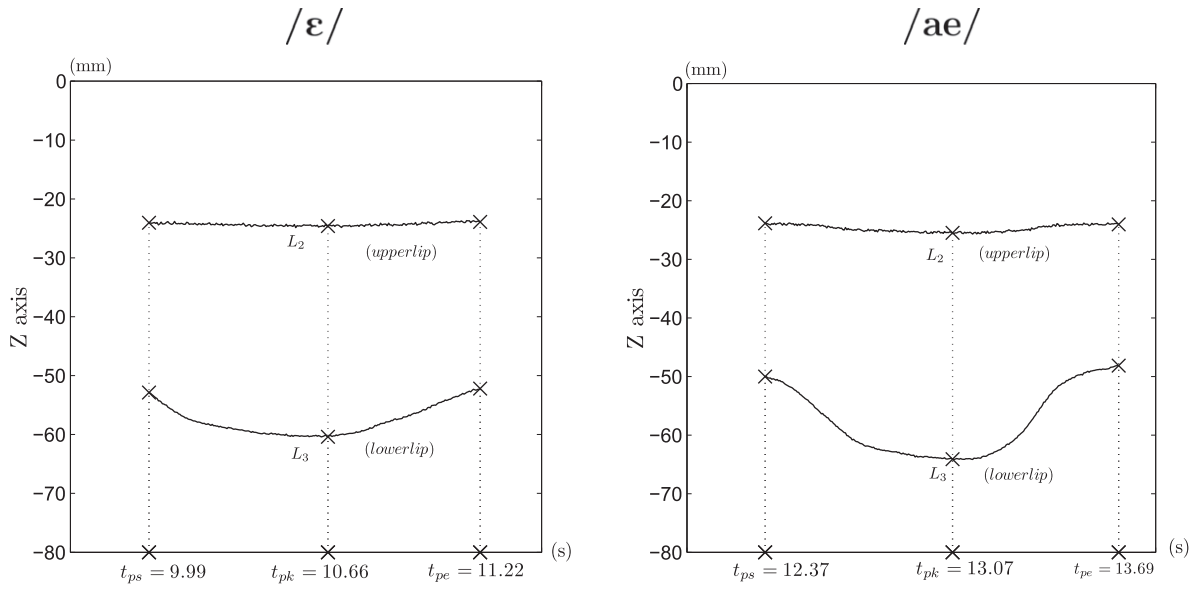

Fig. 4. The feature state of /ɛ/ or /æ/ is selected at t_{p_k}.

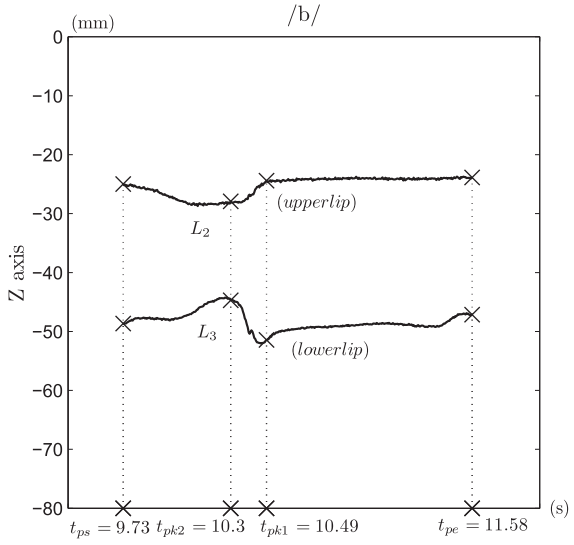

Fig. 5. The feature states of /b/ are selected at $t_{p_{k1}}$ and $t_{p_{k2}}$.

the movements of individual phonemes to form continuous motions, based on the durations of phonemes in continuous speech. These durations are obtained by performing forced-alignment of each waveform against its phonemic transcription, using an automatic speech recognition system. Since the duration of an individual phoneme usually differs from that in continuous speech, time warping is required. However, the articulatory movements of individual phonemes cannot be connected directly, since the displacement vectors of phonemes in continuous speech have different scales.

As proposed in (Cohen and Massaro, 1993), an exponential function was used to approximate the motion of an individual phoneme, and a blending function was required to smooth the articulatory motions of preceding and succeeding phonemes. A general form of dominance function $D_p(t)$ for the phoneme p is given as below,

$$D_p(\tau) = \begin{cases} \alpha_p \cdot e^{-\theta_d|\tau|^c}, & if \quad \tau \leqslant 0; \qquad \tau = t_{p_k} - t \\ \alpha_p \cdot e^{-\theta_g|\tau|^c}, & if \quad \tau > 0. \end{cases} \tag{2}$$

For any axis of any feature point (coil) at the lip or tongue, t_{p_k} refers to the occurrence of feature state of the pth phoneme. In this study, the coefficients are calculated based on the displacement of determined articulatory feature states. Thus, the exponential growth θ_g and decay constant θ_d in Eq. (2) is defined as below,

$$\begin{cases} \alpha_p \cdot e^{-\theta_d|t_{p_k} - t_{p_e}|^c} = \epsilon \\ \alpha_p \cdot e^{-\theta_g|t_{p_k} - t_{p_s}|^c} = \epsilon \end{cases} \tag{3}$$

where ϵ is a local minimum (in the experiments, $\epsilon = 0.22$) to define the overlap between two exponential functions, and t_{p_s} and t_{p_e} are the starting/ending time of the pth phoneme. Moreover, α_p determines the magnitude of the pth phoneme in continuous speech (Wang et al., 2009),

$$\alpha_p = \begin{cases} \frac{max_p|R_p| - |R_p|}{max_p|R_p|}, & if \quad |R_p| \neq max_p|R_p| \\ 1.0, & if \quad |R_p| = max_p|R_p| \end{cases} \tag{4}$$

where R_p is the displacement of the feature state of the pth phoneme compared to the static state. In the above definition, the phoneme with maximal displacement will play a significant role in a sequence, such as a vowel in a CVC structure word. However, for a phoneme sequence with more vowels (for instance, VCV and VCVC words), the dominance of the second vowel is greatly reduced if using Eq. (4). So a more general form of α_p is given as below,

$$\alpha_p = \frac{|R_p|}{max_p|R_p|}. \tag{5}$$

Given the dominance function, the deformation curve for smoothing was expressed as in (Cohen and Massaro, 1993; Wang et al., 2009),

$$F_w(t) = \frac{\sum_{p=1}^{N} R_p \cdot D_p(t - t_p)}{\sum_{p=1}^{N} D_p(t - t_p)} \tag{6}$$

In most studies, the dominance function with $c = 1$ is used, and the normalization term in the blending function is designed to guarantee each phoneme can reach the target feature value at t_{p_k}.

For continuous speech, it is observed that the principle articulator of most consonants in the articulation fails to reach its target position due to the lack of sufficient time. We then investigate a dominance function with $c = 2$ (as shown in Fig. 6) and a modified blending function, aiming to obtain a good fit. The work of (King, 2001) had suggested that the dominance function with $c = 2$ is convex on each side of apex, which can prevent the synthesized curves from a discontinuity in the 1st order derivative. Discontinuities may result in unrealistic and visible artifacts in the animation. Since the dominance function with $c = 2$ performs as a combination of two Gaussians, the blending function presented in this study is a sum of weighted Gaussians,

$$\widehat{F}_w(t) = \sum_{p=1}^{N} R_p \cdot D_p(t - t_p). \tag{7}$$

The above definition does not guarantee each phoneme in continuous speech can reach its target value, only the dominant phoneme of articulation which has both sufficient time and largest displacement will reach the target. The synthesized curves using different articulatory models are compared to the recorded movements to examine the performance.

3.2. The HMM-based articulatory synthesis

The statistical model based speech synthesis has been used successfully, which pushed the development of HMM-based articulatory inversion. Works of (Youssef et al., 2009; Tamura et al., 1999; Zhang et al., 2008) provided the implementations to prove that the jointly trained acoustic and articulatory models can give low RMS errors compared to the real data, with enough data for training.

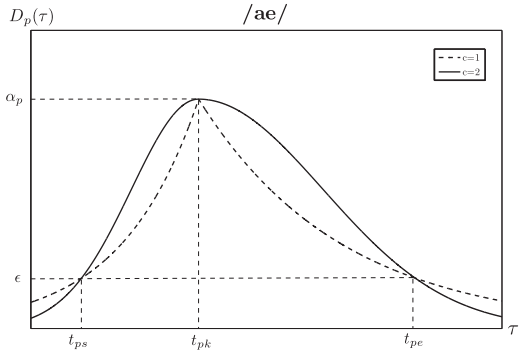

Fig. 6. The dominance functions of an individual phoneme.

The HMM based synthesis includes three main stages: feature extraction stage, training stage and synthesizing stage. In the feature extraction stage, the features for the articulatory models are displacement vectors corresponding to each coils at the lips, tongue and jaw. The coefficients and its delta and delta-delta coefficients can be used to parameterize articulatory movements. So the structure of the observation vector of the tth frame will be organized in forms like:

$$o_t = [\mathbf{c}_t', \Delta\mathbf{c}_t', \Delta^2\mathbf{c}_t']' \tag{8}$$

where $\mathbf{c}_t$ is the n dimensional static feature, $\Delta\mathbf{c}_t$ and $\Delta^2\mathbf{c}_t$ are the dynamic features. $\Delta\mathbf{c}_t$ and $\Delta^2\mathbf{c}_t$ are defined as in (Ling et al., 2008),

$$\Delta\mathbf{c}_t = 0.5 \cdot (\mathbf{c}_{t+1} - \mathbf{c}_{t-1}) \tag{9}$$

$$\Delta^2\mathbf{c}_t = \mathbf{c}_{t+1} - 2 \cdot \mathbf{c}_t + \mathbf{c}_{t-1} \tag{10}$$

In the model training stage, the monophone models are trained with the left-to-right topology, each of which has three emitting states. The conventional Expectation-Maximization (EM) algorithm is used with the Maximum Likelihood criterion. Clustering (Youssef et al., 2009) is not used, because we need to illustrate the distinctions between the confusable phonemes. The synthesis can be achieved by simply concatenating the states together if state duration is provided. The state sequence can be provided by an automatic speech recognizer (ASR). Since the output is a distribution, the position at each frame of time is stochastic.

MLPG, the parameter generation algorithm based on the Maximum Likelihood criterion (Tokuda et al., 1995), can give out an optimized smoothed articulatory trajectory. In the MLPG algorithm, the predicted static feature vector $\mathbf{c}$ can be derived as the following forms,

$$\left[\mathbf{W}'\Sigma^{-1}\mathbf{W}\right]\mathbf{c} = \mathbf{W}'\Sigma^{-1}\mu \tag{11}$$

where μ and Σ are the mean vector and covariance matrix associated with the specific mixture of the state at time t, $\mathbf{W}$ is the weight matrix. Thus we can obtain an updated trajectory $\mathbf{c}$.

4. The 3D articulatory animation system

This section introduces a data driven three-dimensional talking head system, in which the EMA-recorded data is used to synthesize articulatory movements, and then control the 3D articulator dynamics. The whole system consists of three modules: the static 3D head models, animation of the 3D head models and continuous speech control. For any speech input, the system would present the phoneme-level distinctions of articulatory motions.

The static three-dimensional head models are established based on the templates from MRI images. Articulators including the lips, jaw, teeth, palate, alveolar ridge, tongue, velum and pharynx have been recreated. Overall the whole three-dimensional head model is made of 28,656 triangles (Chen et al., 2010). The coordinates of

the static head model are defined in relation to the referenced MRI images. The origin is the nose tip, the X-axis is parallel with eyeline, Z-axis is at the cross-sectional view and perpendicular to X-axis, while Y-axis is perpendicular to XOZ plane. The EMA data is registered to the static head model via affine transformation. Accordingly feature points in the 3D head model are identified manually as in Fig. 1.

Phoneme-level articulator dynamics are the basis of the 3D articulatory animation system. According to the anatomy, the motions of articulators are divided into muscular soft tissue deformations of lips and tongue; rotational up-down movements of chin; and relatively fixed parts (Chen et al., 2010). The lips and tongue are both muscular hydrostats in that they are composed entirely of soft tissue and move under local deformation. Given that feature points ($L_{1,2,3,4}$ and $T_{1,2,3}$) move under displacements of each phoneme in the inventory, constrained deformations are applied to the adjacent points of facial skin or tongue. Up-down rotation of the jaw affects the animation process of jaw, lower teeth, the linked chin skin, and tongue body. The degree of the rotation is computed with the displacement of the feature point on the lower incisors (J_1). The skull and upper teeth remain still while speaking, given that the head does not move.

Unlike other audio-visual synthesis, the 3D articulatory animation system in this paper is worked for pronunciation training with the given speech input. In this case, an automatic speech recognition system is used to segment the speech input into a sequence of phonemes with a time boundary. To apply the phoneme-based articulatory model, the duration of each individual phoneme is linearly mapped to its segmentation in the phoneme sequence. The co-articulator blending algorithm is then used to generate the movement contours. The HMM-based method is also based on the segmentation by ASR to synthesize the curves. Accordingly, the synthesized curve is sampled at 25 frames per second, and the displacement vectors are used for the deformations of each feature point on 3D articulators.

5. Experiments

In order to demonstrate the ability of phoneme-based and HMM-based models to synthesize articulatory motions, we compare it with the recorded EMA data. With the experiments, we aim at showing the performances of 3D articulatory animations at the phoneme level, and an audio-visual test is performed on the animations of 11 minimal pairs (22 words) selected from Session I.

5.1. Experiments on phoneme-based articulatory models with smoothing

For this method, the EMA-recorded articulatory positions of 45 individual phonemes were segmented and processed. The static state and feature states were automatically determined according to the strategy in Subsection 2.2. An automatic speech recognition system was used for phone-level segmentation, where the HMM parameters were trained with TIMIT corpus. The Perceptual Linear Predictive (PLP) features with its delta and acceleration were extracted and trained the cross-word triphone HMMs under ML criterion. Each speech HMM has the left-to-right topology with three emitting states and each state has 12 Gaussians. The test set is 45 example words of Session I including 42 single syllabic words and 3 poly-syllabic words (**butter**, **about**, **azure**).

We first illustrate the use of different exponential functions in the blending function. When setting $c = 1$ or $c = 2$, the dominance functions and the corresponding blending functions are drawn in comparison with the EMA-recording of continuous speech. Fig. 7 shows the synthetic movements of the Z-axis of the lower lip for the word "**bat**", in comparison to the EMA-recorded curve $E_w(t)$. In the above figure, $b_{1,2}$ and $t_{1,2}$ refer to two feature states selected for stops, and *sp* refers to the speech pause at the beginning and ending of a word.

The Root Mean Square (RMS) errors were used as the objective measure to evaluate the performance of synthetic articulatory motions. As in Table 1, the average RMS errors are listed for different forms of the blending functions. In particular, $F_w(t)$ refers to the definition in Eq. (6) with $c = 1$ and normalization term, while the form without normalization term has also been tested. $\widehat{F}_w(t)$ is given by Eq. (7) using $c = 2$ and without normalization term, for comparison, the form with normalization term is also tested. It is found that the curve of the blending function $\widehat{F}_w(t)$ is closer to the EMA-recorded movement than that of $F_w(t)$ and obtains a lower RMS error. In the following experiments, $\widehat{F}_w(t)$ without normalization term is used to smooth the concatenation of the phonemes in the sequence of continuous speech.

Table 2 shows the average RMS errors of synthesized movements by the use of different forms of α_p, where α_1 refers to the definition in Eq. (4) and α_2 is that of Eq. (5). Usually, the blending function using α_1 has a good approximation to the EMA-recorded curve for single syllabic words. However, the synthetic curve using α_2 performs better than the former for poly-syllabic words, like CVCV/VCVC/VCV words. Since these complex structure words have two more vowels, the definition of α_1 will degrade the dominance of another vowel. Therefore, we apply α_1 for the words with only one vowel, while α_2 for the words with more than one vowel, and the overall RMS errors of $\widehat{F}_w(t)$ is **3.46 mm**.

5.2. Experiments on HMM-based synthesis

To train the HMM-based articulatory model, the EMA data of Session II was used. The same test set in Subsection 5.1 was also applied. A total of 200 words were used for training the articulatory HMMs, and there is no overlap between the training and testing data. The time information

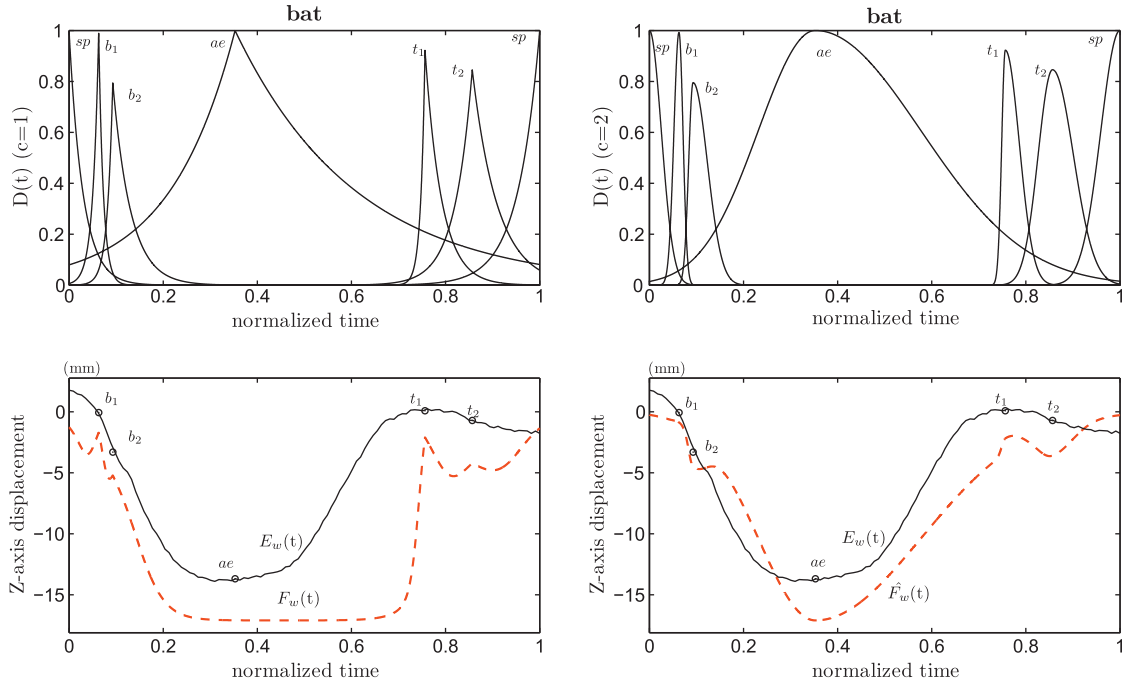

Fig. 7. Comparison of the synthetic movements of lower lip versus the EMA-recorded curve for the word "**bat**".

Table 1
The average RMS errors for different forms of the blending functions.

Test set	$F_w(t)$ (mm)		$\widehat{F}_w(t)$ (mm)	
	With norm.	Without norm.	With norm.	Without norm.
Single syll. words	3.55	3.54	3.65	3.50
Poly-syll. words	2.98	3.25	3.04	3.22
Overall	3.51	3.52	3.61	3.48

Table 2
The average RMS errors for different forms of α_p in the blending functions.

Test set	$F_w(t)$ (mm)		$\widehat{F}_w(t)$ (mm)	
	α_1	α_2	α_1	α_2
Single syll. words	3.55	3.53	**3.50**	3.51
Poly-syll. words	2.98	2.94	3.22	**2.98**

can be extracted given the speech input. The phone level boundaries were automatically determined by the ASR as mentioned in the above section. To synchronize with the waveforms, the EMA data must be down sampled. The articulatory feature vectors comprise of the displacements on the X-axis (back to front direction) and Z-axis (bottom to up direction) of six main coils on the lip(L_2, L_3), tongue (T_1, T_2, T_3) and lower incisor (J_1). Moreover, the displacement on the Y-axis (right to left direction) of the right edge of the lip (L_1) is also used. These displacements together with their deltas and delta-deltas between frames form a 39 dimensional feature vectors.

The monophone HMMs were then trained with the left-to-right topology, where each state has a single Gaussian mixture. When synthesizing, connecting the related phone level articulatory HMMs and estimating the optimal trajectory were performed according to the state sequence generated by forced-alignment. After concatenation, MLPG algorithm was applied to smooth the trajectory.

Furthermore, we compared phoneme-based articulatory models with smoothing and HMM-based synthesis.

Fig. 8 shows the EMA true displacement trajectories of the lower lip and tongue tip, in comparison with the synthesized curves from two different synthesis methods. In particular, Fig. 9 shows the synthesized movements for standard CVC word, in contrast to three poly-syllabic words. It was observed that phoneme-based articulatory models $\hat{F}_w(t)$ can outperform the HMM-based synthesis $\widehat{H}_w(t)$, which has lower RMS errors compared to the EMA true displacement trajectories $E_w(t)$.

It can be seen from Fig. 10 that the phoneme-based motion model with smoothing deals well with low RMS errors. The average RMS for phoneme-based method is **3.46 mm**, while it is **4.67 mm** for HMM-based synthesis. One concern is that the estimation of the articulatory HMMs may not be exact due to the limited training data, and the acoustic models and articulator models were not trained together. This may result in the worse performance of HMM-based method in the experiments.

5.3. Experiments on the articulatory animation system

For Chinese learners of English, some phonemes are commonly mispronounced as in Table 3, a set of minimal

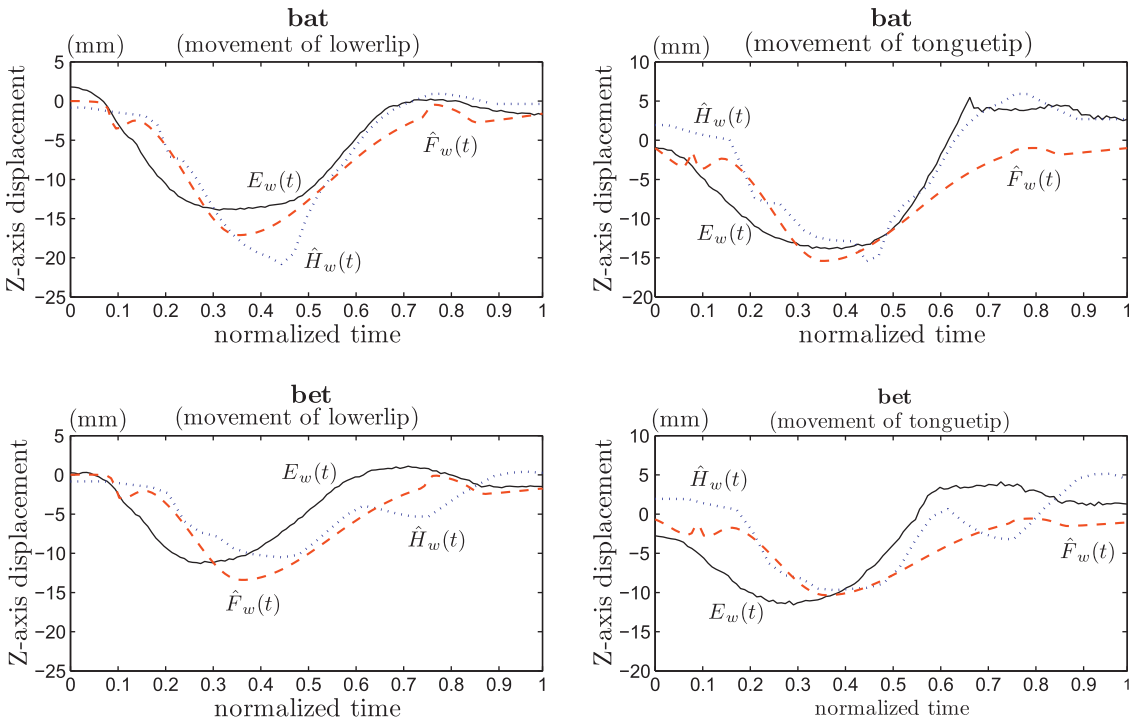

Fig. 8. Comparison of two synthetic movements versus the EMA-recorded curve for a minimal pair.

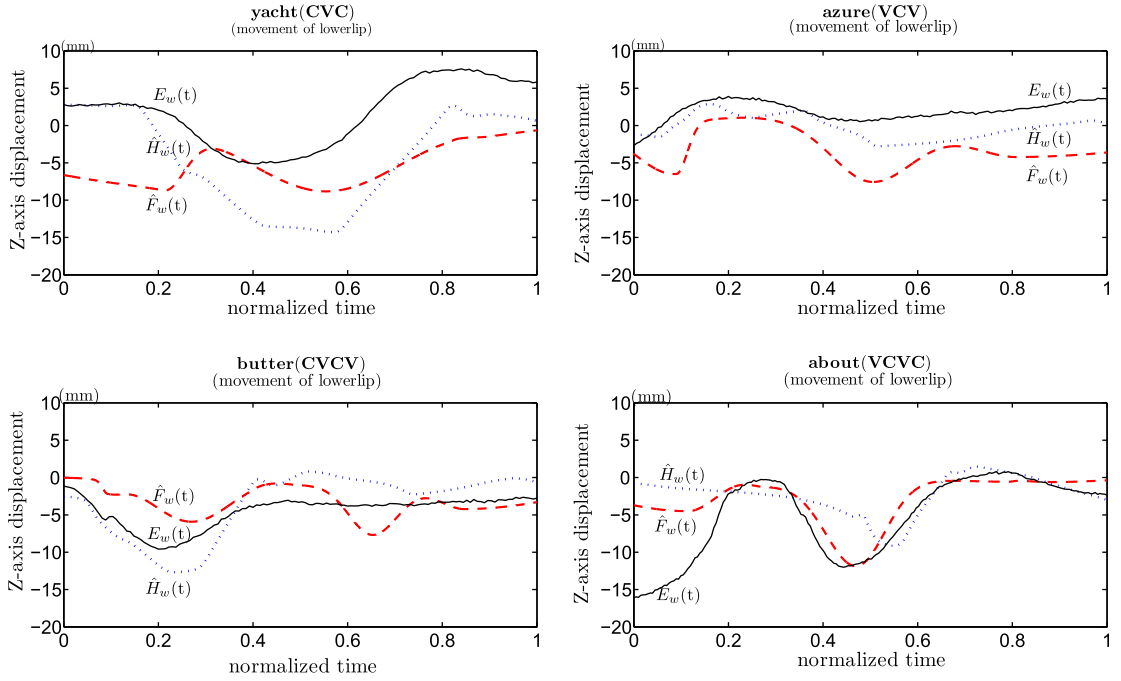

Fig. 9. Comparison of two synthetic movements versus the EMA-recorded curve for a set of complex words.

pairs were selected to present the synthetic talking head. In a total of 22 words, the phonemes not presented in Chinese language were included, like /ɛ/ and /æ/, while the confusable phonemes were also shown, such as labio-dental fricative /v/ and bilabial glide /w/, apico-dental fricative /θ/ and apico-alveolar fricative /s/. With the 3D talking head system presented in Section 4, the animations of these articulations are implemented. In Fig. 11, the deformations on both external and internal articulators are displayed at the feature states. Clearly, the 3D head can reveal the distinctions of the articulator motions at the phoneme level.

Consequently, we performed an audio-visual perception test to access the correctness of the animations of the 3D talking head. The participants were three native speakers

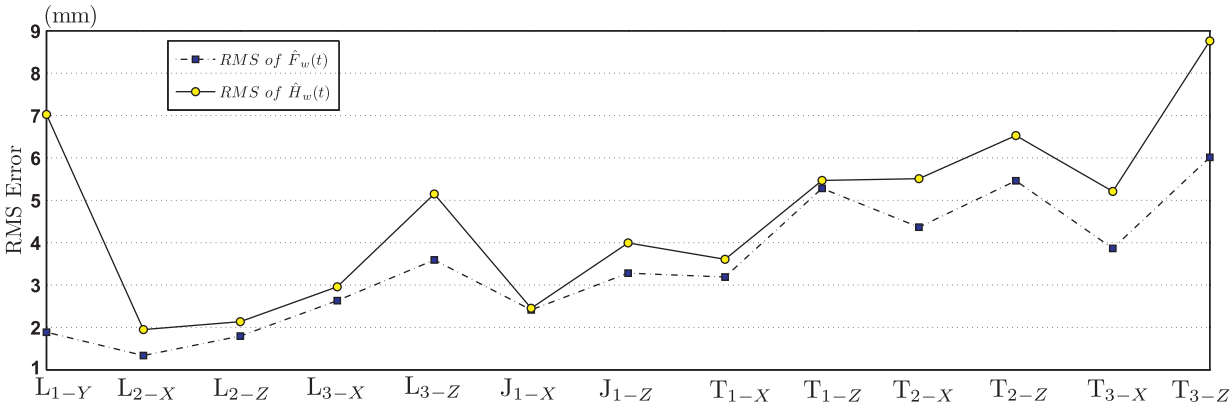

Fig. 10. The RMS errors for all example words on specific coils and coordinates.

Table 3
The identification rates and realism scores of 3D articulatory animations for minimal pairs.

Confusable phonemes	Minimal pairs					
	Word	Identi. rate (%)	Realism score	Word	Identi. rate (%)	Realism score
/æ/ vs./ɛ/	**bat**	84.6	3.5	**bet**	84.6	3.6
/au/ vs. /ɔ:/	**house**	92.3	4.0	**horse**	92.3	3.8
/v/ vs. /w/	**vine**	76.9	3.0	**wine**	76.9	2.9
/θ/ vs. /s/	**thing**	92.3	3.4	**sing**	92.3	3.1
/ð/ vs. /f/	**they**	100	3.4	**fay**	100	3.8
/n/ vs. /m/	**night**	100	3.3	**might**	100	3.4
/l/ vs. /r/	**lay**	84.6	3.2	**ray**	84.6	3.0
/ai/ vs. /ɔi/	**buy**	100	3.4	**boy**	100	3.1
/tʃ/ vs. /dʒ/	**choke**	92.3	3.2	**joke**	92.3	3.2
/s/ vs. /ʃ/	**sea**	84.6	3.2	**she**	84.6	3.1
/iə/ vs. /ɛə/	**beer**	100	3.2	**bear**	100	3.3

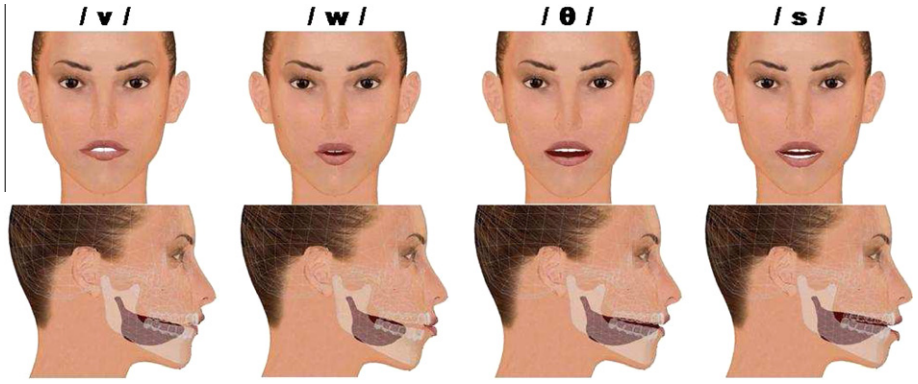

Fig. 11. The 3D talking head presents for individual phonemes with facial and profile (transparent) views.

and ten English teachers in one university. In the test, the audio streams of one minimal pair were played firstly, and then the animations were shown in which two words of one minimal pair appeared in a random order, as shown in Fig. 12. The subjects were asked to identify which animation corresponded to the word. The identification accuracy refers to the ratio of the number of correctly recognized animations and the total number of animations. All the stimuli were presented in two conditions, the visualization with front face (F), the visualization with transparent face and tongue (FT). The subjects also scored the degree of realism of the animation, when the correct label of the animation was shown to them. The score ranges from 1 to 5, 1 for bad, 2 for poor, 3 is fair, 4 for good and 5 for excellence.

The identification rate and realism score of animation of each word are averaged and listed in Table 3, the overall identification accuracy is 91.6% among 286 tests. Most of the subjects could clearly point out the differences of the 3D articulator dynamics between the confusable phonemes. Although the average realism score is over 3.5, the 3D animations of "**vine**" and "**wine**", "**sea**" and "**she**"

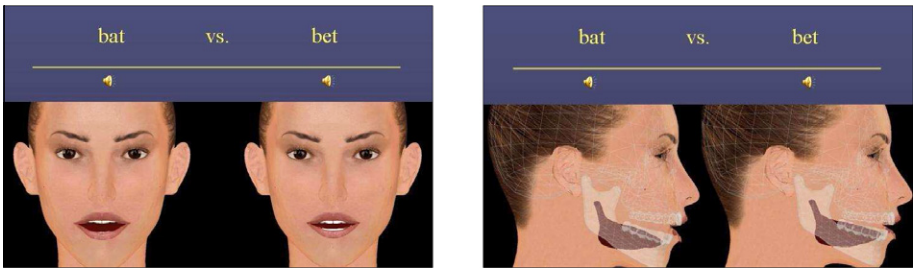

Fig. 12. The 3D articulatory animations in the perception test.

Table 4
The individual phonemes and the example words in Session I.

Phoneme/word		Phoneme/word		Phoneme/word		Phoneme/word		Phoneme/word	
/iː/	beet	/ɪ/	bit	/eɪ/	bait	/ɛ/	bet	/æ/	bat
/əː/	bird	/ər/	butter	/ə/	about	/ʌ/	but	/uː/	boot
/u/	book	/əu/	boat	/ɔːr/	horse	/ɔ/	bott	/ai/	bite
/au/	house	/ɔi/	boy	/aː/	bar	/ɛə/	bear	/iə/	beer
/juː/	you	/p/	pea	/b/	bee	/t/	tea	/d/	day
/k/	key	/g/	gay	/m/	might	/n/	night	/ŋ/	sing
/l/	lay	/θ/	thing	/ð/	they	/s/	sea	/z/	zone
/ʃ/	she	/ʒ/	azure	/f/	fay	/v/	vine	/h/	hay
/tʃ/	choke	/dʒ/	joke	/w/	wine	/j/	yacht	/r/	ray

are not good enough to identify correctly. All subjects gave comments that the FT condition had importance to identify the words. The HMM-based synthesized curves are also applied to the animation, the perception tests show no significant difference between the two synthesis methods.

6. Conclusions

We have investigated the use of EMA-recorded visual information to capture phoneme-level articulatory distinctions. 3D articulatory movements of phonemes, words and sentences were recorded and processed, then the feature states of individual phonemes were determined to illustrate differences among phonemes. Thus, we developed the phoneme-based articulatory motion model with smoothing, in which different forms of dominance function and blending functions were presented and compared. This visual synthesis is then feasible with a very small amount of EMA data, which can reduce the cost of implementing a data-driven 3D animation system.

For comparison, the HMM-based method was also performed, since it has been commonly used for visual synthesis. Experimental results have shown that the phoneme-based motion model with smoothing obtained a better performance compared with the HMM-based synthesis on phone/word level. And most of synthetic articulatory movements achieved lower RMS errors to the EMA-recorded curves. Due to the limited EMA data for training, the acoustic models and articulator models are not trained together, which may degrade the accuracy of the HMM-based synthesis. So a further investigation should be conducted by designing a large corpus and collecting EMA data from different speakers.

A data-driven 3D articulatory animation system was designed in a way that the physiological head models were controlled by synthetic movements, given the speech input. Rather than making animations at the viseme level to improve the intelligibility of audio-visual synthesis, this work illustrated more slight differences among phonemes, so as to instruct language learners to articulate. In the perception test, the subjects evaluated the 3D animations with a high identification rate. It is worth investigating animations of longer utterances rather than minimal pairs of words, in order to study the effect of co-articulation on the deformation of the articulators.

Acknowledgements

Our work is supported by National Nature Science Foundation of China (NSFC 61135003, NSFC 90920002), National Fundamental Research Grant of Science and Technology (973 Project: 2009CB320804), and The Knowledge Innovation Program of the Chinese Academy of Sciences (KJCXZ-YW-617).

Appendix A. The Session I of EMA corpus includes phonemes and example words according to IPA (see Table 4).

References

Massaro, D.E., Light, J., 2004. Using visible speech to train perception and production of speech for individuals with hearing loss. Journal of Speech, Language and Hearing Research 47, 304–320.

Rathinavelu, A.,Thiagarajan, H., Rajkumar, A., 2007. Three dimensional articulator models for speech acquisition by children with hearing loss.

In: Universal Access in Human Computer Interaction, HCII 2007, pp. 786–794.

Tarabalka, Y., Badin, P., Elisei, F., Bailly, G., 2007. Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding. In: Proceedings of ASSISTH 2007, pp. 187–190.

Wik, P., Engwall, O., 2008. Looking at tongues – can it help in speech perception. In: Proceedings FONETIK 2008 Sweden, pp. 57–61.

Serrurier, A., Badin, P., 2008. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on mri and ct data. Journal of Acoustic Society of American 123 (4), 2335–2355.

Badin, P., Elisei, F., Bailly, Y., Tarabalka, C., 2008. An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data. Proceedings of V Conference on Articulated Motion and Deformable Objects, 132–143.

Fagel, S., Clemens, C., 2004. An articulation model for audio-visual speech synthesis – determination, adjustment, evaluation. Speech Communication 44, 141–154.

Deng, A., Neumann, U., 2008. Expressive speech animation synthesis with phoneme-level controls. Computer Graphics 27 (8), 2096–2113.

Badin, P., Tarabalka, Y., Elisei, F., Bailly, G., 2010. Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Communication 52 (6), 493–503.

Wang, L., Chen, H., Ouyang, J.J., 2009. Evaluation of external and internal articulator dynamics for pronunciation learning. In: Proceedings of InterSpeech 2009, pp. 2247–2250.

Ma, J., Cole, R., Pellom, W., Ward, B., 2004. Accurate automatic visible speech synthesis of arbitrary 3d models based on concatenation of diviseme motion capture data. Computer Animation and Virtual Worlds 15, 485–500.

Grauwinkel, K., Dewitt, B., Fagel, S., 2007. Visualization of internal articulator dynamics and its intelligibility in synthetic audio-visual speech. Proceedings of International Congress on Phonetic Sciences, 2173–2176.

Engwall, O., mri, Combining, 2003. ema & epg in a three-dimensional tongue model. Speech Communication 41 (2–3), 303–329.

Murray, N., Kirk, K.I., Schum, L., 1993. Making typically obscured articulatory activity available to speech readers by means of videofluoroscopy. NCVS Status and Progress Report (4), 41–63.

Chen, H., Wang, L., Heng, P.A., Liu, W.X., 2010. Combine X-ray and facial videos for phoneme-level articulator dynamics. The Visual Computer 26 (1), 477–486.

Cohen, M., Massaro, D.W., 1993. Modeling coarticulation in synthetic visual speech. In: Thalmann, N.M., Thalmann, D. (Eds.), Springer-Verlag, pp. 139–156.

Guenther, F., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review 102, 594–621.

Blackburn, S.C., Young, S., 2000. A self-learning predictive model of articulator movements during speech production. Journal of Acoustical Society of America 107 (3), 1659–1670.

Dang, J.W., Wei, J., Suzuki, T., Perrier, P., 2005. Investigation and modeling of coarticulation during speech. In: Proceedings of InterSpeech 2005, pp. 1025–1028.

Wik, P., Hjalmarsson, A., 2009. Embodied conversational agents in computer assisted language learning. Speech Communication 51, 1024–1037.

Lado, R., 1957. Linguistics Across Cultures: Applied Linguistics for Language Teachers. Ann Arbor.

Iowa. <http://www.uiowa.edu/ acadtech/phonetics/>.

Hoole, P., Zierdt, A., Geng, C., 2003. Beyond 2d in articulatory data acquisition and analysis. Proceedings of the International Conference of Phonetic Sciences XV, 265–268.

Hoole, P., Zierdt, A., 2010. Five-Dimensional Articulography. In: Speech Motor Control: New Developments in Basic and Applied Research. Oxford University Press, pp. 331–349 (Chapter 20).

King, S., 2001. A Facial Model and Animation Techniques for Animated Speech, The Ohio State University.

Youssef, B., Badin, P., Bailly, G., Heracleous, C., 2009. Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden markov models. Proceedings of Interspeech, 2255–2258.

Tamura, M., Kondo, S., Masuko, T., Kobayashi, T., 1999. Text-to-audio-visual speech synthesis based on parameter generation from hmm. In: Proceedings of EuroSpeech 1999, pp. 959–962.

Zhang, L., Renal, S., 2008. Acoustic-articulatory modelling with the trajectory hmm. Signal Processing Letters, IEEE 15, 245–248.

Ling, Z.H., Richmond, K., Yamagishi, J., Wang, R.H., 2008. Articulatory control of hmm-based parametric speech synthesis driven by phonetic knowledge. In: Proceedings of InterSpeech 2008, pp. 573–576.

Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from hmm using dynamic features. In: Proceedings of ICASSP 1995, pp. 660–663.