THE PHONEME-LEVEL ARTICULATOR DYNAMICS FOR PRONUNCIATION ANIMATION

Sheng Li[1,2], Lan Wang[1,2] & En Qi[1,2]
[1] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*
[2] *The Chinese University of Hong Kong, Hong Kong, China*
{sheng.li, lan.wang, en.qi}@siat.ac.cn

***Abstract*—**Speech visualization can be extended to a task of pronunciation animation for language learners. In this paper, a three dimensional English articulation database is recorded using Carstens Electro-Magnetic Articulograph (EMA AG500). An HMM-based visual synthesis method for continuous speech is implemented to recover 3D articulatory information. The synthesized articulations are then compared to the EMA recordings for objective evaluation. Using a data-driven 3D talking head, the distinctions between the confusable phonemes can be depicted through both external and internal articulatory movements. The experiments have demonstrated that the HMM-based synthesis with limited training data can achieve the minimum RMS error of less than 2mm. The synthesized articulatory movements can be used for computer assisted pronunciation training.

***Keywords*—** external and internal articulators, pronunciation animation, EMA recordings

I. INTRODUCTION

The development of speech visualization involves a range of explorations and technologies on speech production, synthesis and perception. The task can be extended by implementing a realistic pronunciation animation for language learners in which a transparent 3D talking head is made to present both the external and internal articulator dynamics. With the multimodal information, learners can easily distinguish the articulation differences between confusable phoneme pairs and can improve their pronunciations. The research in [1, 2] introduced the use of 3D talking head in computer-assisted language learning, where the target language was English/Swedish. Previous studies [8, 10, 5, 3, 11, 9] focused on recovering articulatory information from the speech acoustic for visual synthesis. The HMM-based technique has been used for articulatory synthesis [12]. For our purpose, the synthesized movements of articulators should not only fit to the co-articulation of continuous speech, but also reflect the distinctions of phonemes. Therefore, more exact data should be acquired to recover the phoneme level articulatory trajectories.

This paper designed an audio-visual corpus for English pronunciation animation, and then the three-dimensional articulatory movement data was recorded from a native American speaker using the Carstens EMA AG500. Assuming that the articulatory movement extracted from the EMA recording can depict the distinctions among English phonemes, a 3D head model is introduced for pronunciation animation controlled by the recorded EMA articulatory displacement vectors. In particular, we deal with the challenge of synthesizing the articulation by HMM-based synthesis technology. The synthesized articulatory motions can be evaluated by comparing them to the EMA recordings. The experiments of pronunciation animation were conducted on all individual phonemes, and a set of minimal pairs covering the phonemes commonly mis-pronounced by Chinese learners were chosen for implementation.

The rest of this paper is organized as follows: Section 2 introduces the process of articulatory data collection and data processing. In Section 3, HMM-based articulatory synthesis for recovering is implemented. Section 4 presents experimental details and results. And the conclusion and discussions are in the final section.

II. RECORDING AND PROCESSING THE 3D ARTICULAORY MOTIONS

We start from the EMA data collection for phoneme-level articulatory dynamics. The motions of articulators are recorded when a native speaker is reading the prompts. The prompts include the individual phonemes in terms of IPA, each of which is embedded in an example word. Moreover, the prompts involve a set of minimal pairs and a paragraph that are commonly mispronounced by the learners.

In the data collection, the facial and intra-oral articulator movements are recorded using the Carstens EMA AG500 at 200 frames per second. The movements are synchronized with the speech waveform. As previous studies like [6], 3 coils (L_1, L_2, L_3) are placed on the speaker's lips, 1 coil on her jaw, 3 coils (T_1, T_2, T_3) on her tongue to record the internal movements. Since the recorded sensor motion is usually a mixture of head movement and the actual articulator motion. The head motion normalization is required to remove the head movement from the recorded data. So we put 3 coils (H_1, H_2, H_3) on nose bridge and behind the ears for head motion normalization [4]. Fig. 1 shows the coils on the speakers' head and the corresponding points on a generic 3D head.

Fig.2(a) is the EMA data of single vowels projected on sagittal plate, and we made a reference to the standard English tongue position chart shown in Fig.2(b). In Fig.2(a), every point is an displacement of tongue tip at that frame, and the data is distinctive enough to let each cluster of points belongs to a single vowel. We can see from Fig.2(a) that the EMA data can be a support and an addition to the abstract

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.13

tongue position chart. So the EMA data integrated with 3D talking head is convincing for computer-assisted pronunciation training.

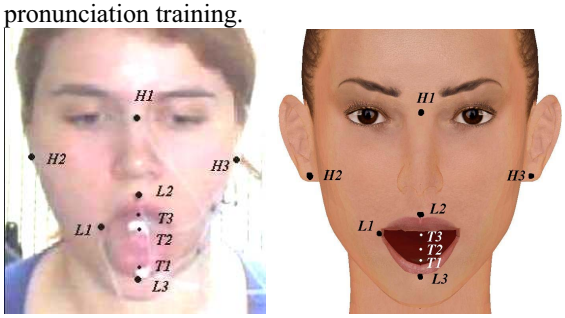

Fig. 1 The coils on the speakers' head and the corresponding points on the 3D head (the coil on the jaw is not shown in this figure)

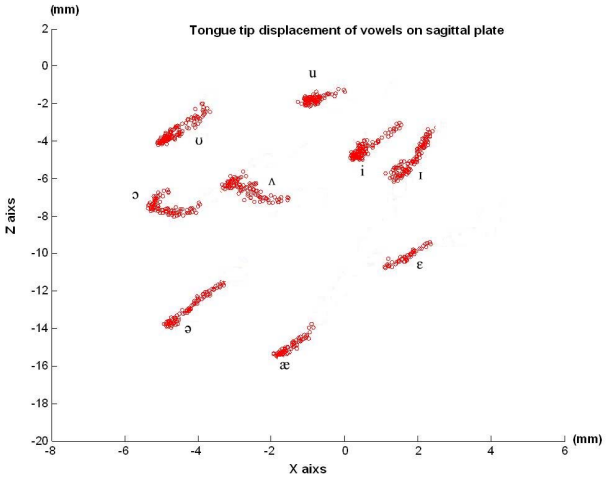

Fig.2(a) Tongue tip displacement of vowels on sagittal plate

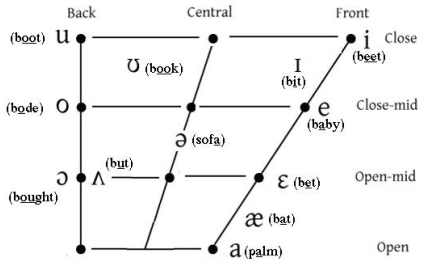

Fig.2(b) standard tongue position chart of English

III. ARTICULATORY SYNTHESIS

When the articulation data of the words were not included in the recording prompts, we must synthesize them. To synthesize the word level continuous articulation, the co-articulation phenomenon should be taken into consideration. For this reason, the synthesis is not simply a concatenation task and our study needs a relatively complicated synthesis mechanism.

The HMM-based synthesis technology, which belongs to the data driven method category, is fully investigated for speech synthesis in these 10 years. It was first introduced in audio-visual synthesis by Japanese researchers [12].

The HMM based synthesis includes three main stages: 1).the feature extraction stage, 2).the training stage, 3).the synthesis stage.

Suppose we have an observed parameter vector sequence O. Let

$$O = [o'_1, o'_2, ..., o'_T]' \quad (1)$$

The structure of the observation vector of the t th frame ($1 \le t \le T$) o_t will be organized like:

$$o_t = [c'_t, \Delta c'_t, \Delta^2 c'_t]' \quad (2)$$

where c_t is the n dimensional static feature of the t th frame ($1 \le t \le T$) of the static feature sequence

$$C = [c'_1, c'_2, ..., c'_T]' \quad (3)$$

Δc_t and $\Delta^2 c_t$ are dynamic features ($1 \le t \le T$):

$$\Delta c_t = 0.5(c_{t+1} - c_{t-1}),\ \Delta^2 c_t = c_{t+1} - 2c_t + c_{t-1} \quad (4)$$

In the feature extraction stage, the features for the articulatory models are displacement vectors corresponding to each coil at the lips, tongue and jaw. Both static features and dynamic features (deltas and delta-deltas) can be used to parameterize the articulatory movements.

In the model training stage, the monophone models are trained with the left-to-right HMM topology, each of which has three states. State clustering is not used, because we need to illustrate the distinctions between the confusable phonemes. More details of the first two stages will be discussed in the experiment section.

The synthesis can be achieved by simply concatenating the states together if state duration is provided. The state duration determines how many frames will be generated for each state. The state duration can be provided by an automatic speech recognizer (ASR). Since the output is a distribution, the position at each frame of time is stochastic. We use the MLPG algorithm, a parameter generation algorithm based on the Maximum Likelihood criterion [11] integrated in HTS[7], can find the most likely sequence of generated parameters.

According to condition (2)(3)(4), o_t can be rewritten in the form like:

$$o_t = w_t \bullet \hat{C} \quad (5)$$

w_t can be defined as follows:

$$w_t = \begin{bmatrix} 0_{1,1} \ldots & 0_{1,t-1} & I_{1,t} & 0_{1,t+1} \ldots 0_{1,T} \\ 0_{2,1} \ldots -\frac{1}{2} I_{2,t-1} & 0_{2,t} & \frac{1}{2} I_{2,t+1} & \ldots 0_{2,T} \\ 0_{3,1} \ldots & I_{3,t-1} & -2I_{3,t} & I_{3,t+1} \ldots 0_{3,T} \end{bmatrix}, 1 < t < T \quad (6)$$

where w_t is a $(3 \times n)$-by-$(T \times n)$ matrix, I is a n-by-n identity matrix and 0 is a n-by-n zero matrix. w_1 and w_T have their specific forms as defined in [11].

The predicted static feature vector $\hat{C}$ can be derived as the following forms, let

$$W = [w'_1, w'_2, \ldots, w'_T]' \quad (7)$$

then

$$\hat{C} = (W'\Sigma^{-1}W)^{-1}W'\Sigma^{-1}\mu \quad (8)$$

where μ and Σ are the mean vector and covariance matrix of articulatory HMMs. Other detailed deduction can be found in [11].

IV. EXPERIMENTS AND EVALUATIONS

In the experiment, we compared EMA recordings to our HMM-based synthesis output where root mean square (RMS) error can be used to evaluate the synthesized curves. Moreover, a data-driven 3D talking head system was developed to load the synthesized data for word level continuous speech, aiming to depict the distinctions between confusable phonemes for Chinese learners of English.

A. Experimental Setup

The word level data are used in this experiment. 200 words were used for training an HMM-based articulatory model. The other 45 words were used for testing and evaluation.

The acoustic waveforms were recorded simultaneously while the EMA articulatory movements were being recorded. The phone boundaries have been first labeled by a forced alignment procedure using an acoustic model trained from TIMIT, and then followed with manual corrections.

To synchronize with the acoustic features, the EMA data in the train set for HMM-based synthesis did down sample rate from 200 frames per second to 100 frames per second. From all of the EMA data, 6 coils with their articulatory displacement on X-axis (back to front direction) and Z-axis (bottom to up direction) were selected. They are upper lip (L_2), lower lip (L_3), tongue tip (T_1), tongue body (T_2), tongue dorsum (T_3), and jaw (J_1). And 1 coil with its articulatory displacement on Y-axis (right to left direction) also has been chosen. It is the right edge of the lip (L_1).

The features were extracted from the displacement of each coil on each axis using a 25ms window at a rate of 100 frames per second, to form the 39 dimension feature vectors (13^{th} order together with their deltas and delta-deltas).

We use HTK to train models initialized with flat-start and remaining the model at monophone stage. The output model can be considered an average articulatory displacement model. It has 3 emitting states, left-to-right topology and single Gaussian mixture per state.

B. Synthesis and evaluations

In HMM-based synthesis, we connect state level articulatory HMMs and estimate the optimal trajectory. All these were made according to the state-level forced-alignment output transcriptions using an acoustic model well trained from TIMIT database.

Fig.3 shows an example word pair (bait and bite) of the EMA true displacement trajectories of lower lip on Z axis, in comparison with the synthesized curves. The synthesized trajectory seems to have more disturbances which may lead to unsmoothness. RMS errors were computed to assess the precision of the synthesis methods. We calculated 13 RMS errors, to represent the 7 coils on 3 axis directions in Fig.4. From the RMS values, we can have an overall idea of the synthesized data. The RMS values of J1 (jaw coil) and T3 (tongue dorsum coil) in direction from back to front is relatively high. This may caused by distortions of the training EMA data in this direction. It is clear that the accuracy of our HMM-based synthesis still has enough room for improvement. This may due to the limited training data for HMM-based synthesis.

The data-driven 3D talking head was then controlled by the synthesized data, illustrating in Fig.5. Using the transparent mode, both external and internal articulatory movements can be well depicted. In these figures, we can find the synthesized result of common CVC word structures with single vowels (bat) is acceptable. But when our synthesis encountered with diphthong (boy), the result may be a little confusing.

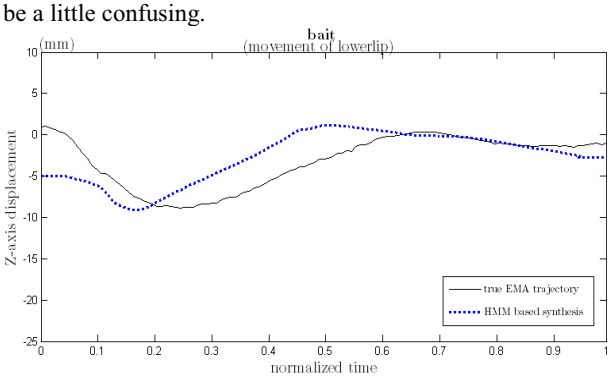

Fig.3 (a) comparison of the synthesized and true displacement curves of tongue tip when pronouncing 'bait'

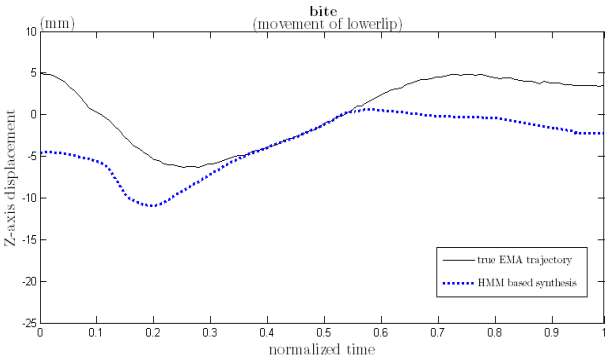

Fig.3 (b) comparison of the synthesized and true displacement curves of lower lip when pronouncing 'bite'

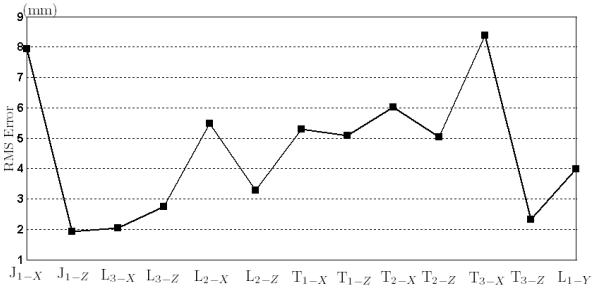

Fig.4 the RMS errors of the HMM based synthesis methods

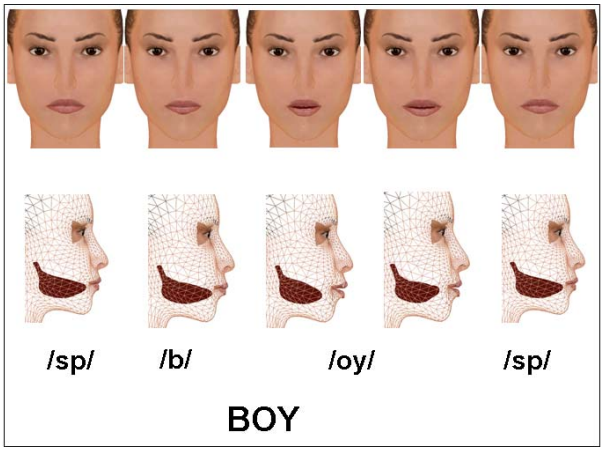

Fig.5(b) 3D system implemented with synthesize data (word structure including diphthong)

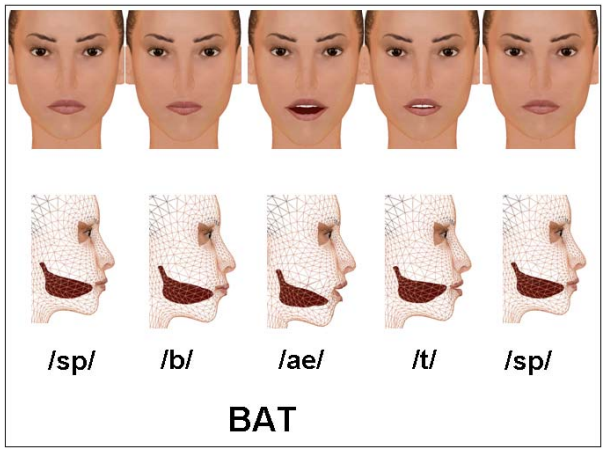

Fig.5(b) 3D system implemented with synthesized data (common CVC word structure only including single vowels)

V. CONCLUSIONS AND FUTURE WORKS

This study discussed the feasibility of an EMA data driven 3D talking head for second language computer-assisted pronunciation training. A HMM based synthesis method is used to recover some articulatory movements. And the synthesized data has proved acceptable in most common cases.

In our future work, we will collect more EMA data to improve our synthesis result. Also this research can give us enough suggestion when we begin our study in Chinese articulator dynamics.

ACKNOWLEDGEMENT

The work is supported by National Science Foundation of China (NSFC 90920002) and the Knowledge Innovative Project of CAS (No. KJCZ-YW-617).

REFERENCES

[1] L. Wang, H. Chen, JJ. Ouyang, “Evaluation of External and Internal Articulator Dynamics for Pronunciation Learning”, In proceedings of Interspeech 2009, pp. 2247-2250, Brisbane,UK, 2009

[2] P. Wik and Hjalmarsson, “Embodied Conversational Agents in Computer Assisted Language Learning”, in Speech Communication, 51:1024-1037, 2009

[3] B. Youssef, A., P. Badin, G. Bailly and P. Heracleous, “Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models”, Interspeech 2009, Brighton, UK, 2009

[4] C. Kroos, "Using sensor orientation information for computational head stabilisation in 3D Electromagnetic Articulography (EMA)". In Proceedings from Interspeech 2009, pp. 776-779, Brighton, UK, 2009

[5] L. Zhang, S. Renals, ”Acoustic-Articulatory Modeling With the Trajectory HMM”, IEEE Signal Processing Letters, 2008, vol. 15, pp. 245-248

[6] Y. Tarabalka, P. Badin, F. Elisei and G. Bailly, "Can You Read Tongue Movements? Evaluation of The Contribution of Tongue Display to Speech Understanding", in Proceedings of ASSISTH 2007, France, pp. 187-190, November 2007.

[7] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The hmm-based speech synthesis system (hts) version 2.0,” 6th ISCA Workshop on Speech Synthesis (SSW6), vol. 6, pp.294–299, Agosto 2007.

[8] G. Hofer, H. Shimodaira, and J. Yamagishi. “Lip motion synthesis using a context dependent trajectory hidden Markov model”. Poster at SCA 2007, 2007.

[9] K. Richmond. “A trajectory mixture density network for the acoustic-articulatory inversion mapping”. In Proc. Interspeech, Pittsburgh, USA, September 2006.

[10] H. Zen, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features”, Ph.D. dissertation, Nagoya Institute of Technology, 2006.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, 2000. “Speech parameter generation algorithms for HMM-based speech synthesis”. ICASSP 2000, vol. 3. pp. 1315–1318.

[12] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM", EUROSPEECH, Budapest, 1999.