Visualization of Mandarin Articulation by using a Physiological Articulatory Model

Dian Huang*, Xiyu Wu†, Jianguo Wei*, Hongcui Wang*, Chan Song*, Qingzhi Hou*, and Jianwu Dang*†
*Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China
E-mail: {huang_dian@163.com, jianguo.fr@gmail.com, laurelwind@gmail.com,
songchan_8855@126.com, darcy.hou@gmail.com} Tel: +86-15822883934
†Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: {xiyuwu@jaist.ac.jp, jdang@jaist.ac.jp}

***Abstract*—It is difficult for language learners to produce unfamiliar speech sounds accurately because they may not manipulate articulatory movements precisely by auditory feedback alone. Visual feedback can help identify the errors and promote the learning progress, especially in language learning and speech rehabilitation fields. In this paper, we propose a visualization method for Mandarin phoneme pronunciation using a three-dimensional (3D) articulatory physiological model driven by Chinese Electromagnetic Articulographic (EMA) data. A mapping from EMA data to physiological articulatory model was constructed using three points on the mid-sagittal plane of the tongue. To do so, we analyzed configurations of 30 Chinese phonemes based on an EMA database. At the same time, we designed nearly 150,000 muscle activation patterns and applied them to the physiological model to generate model-based articulatory movements. As the result, we developed a visualized articulation system with 2.5 dimensional and 3D views respectively. The mapping was evaluated using MRI data. It is found that the mean deviation was about 0.21cm for seven vowels.**

I. INTRODUCTION

The studies of pronunciation learning have shown that detailed and accurate error feedback is effective in correcting the errors addressed in the learning process [1] and visualized feedback is playing an important rule. Learners will evaluate their learning through auditory feedback if there is no other feedback available. However, even if the learner can recognize the discrepancy between their utterance and the target speech sounds, it is difficult for them to adjust their articulations. In language learning process, explicit guidance is more effective than implicit introduction [2]. A Computer Assisted Language Learning (CALL) system which contains visual feedback makes the learners easier to correct their articulations by providing a visualized articulatory target. In this study, we put forward a Mandarin phoneme pronunciation visualization method by using Chinese EMA data to drive the 3D articulatory physiological model.

With the development of speech analysis and observation technology, the observation and presentation of pronunciation visualization is becoming much easier. On the one hand, some researchers used two-dimensional (2D) model to visualize articulation. For example, Kaburagi and Honda proposed a 2D model to predict articulator movements for continuous speech based on EMA data [3]. A 2D visual-speech synthesizer was presented to animate the human articulators by Wong et al [4]. LaRocca, *et al.* presented a system which used articulatory information in the form of a side-view of a transparent head to detect spoken segmental errors and provide corrected feedback so that the learner could see articulator placement [5]. As proposed by Eskenazi et al [6], a mid-sagittal 2D model was employed to present immediate corrected articulatory help for each type of possible phonetic or prosodic error made by the students. The accuracy of these methods may be guaranteed but they are not easy to understand. On the other hand, 3D model is used to present articulators movements. For example, Computer Graph (CG) technology is used to construct a 3D model for online Chinese learning [7]. If learners' pronunciation is incorrect, the system will demonstrate the correct articulator organs' movement, as well as wrong pronunciations made by learners themselves [8]. The 3D method above is easy to observe and understand but how to guarantee its accuracy is the biggest challenge. In the previous studies, one can see that in the field of articulation visualization, especially in Mandarin, higher accuracy and intuitive representation cannot be combined well in a visualization method.

In this paper, we constructed a 3D visualization system for Chinese phoneme via jointing the advantage of EMA data's high temporal resolution and 3D physiological model's high space resolution, which was expected to provide intuitive and accurate visualization of articulator movements flexibly. Our visualization system consists of two modules. One is the 3D articulatory model which is extracted from a physiological model, and the other is 2D Chinese EMA data. We applied three points on the tongue in EMA data to select the best-matched mid-sagittal shape of 3D model so that we can get the 3D visualized articulatory organ's movements for each Chinese phoneme. To evaluate the accuracy of our method, we compared the seven vowels' best-matched model data with Chinese MRI data. The result showed that the accuracy of our model is acceptable.

The following paper is organized as below: Part Ⅱ and Ⅲ describe the details of the construction of Chinese EMA database and 3D model based database. Part Ⅳ introduces the method for building up the mapping, and part Ⅴ shows the

proposed system. Finally, part Ⅵ will give a summary.

II. Processing of a Multichannel Chinese EMA Database

In this part, we first introduce the Chinese EMA database employed in this study.

A. Corpus Design

A corpus was designed to cover 30 Chinese phonemes which was presented in table Ⅰ, and recorded the articulatory data of continuous utterances from a native Mandarin speaker using the EMA system [9]. The corpus and speaker information are listed in table Ⅱ.

TABLE Ⅰ
CHINESE PHONEME LIST

Category	Phoneme	Number
Vowel	[a] [o] [ɤ] [i] [u] [y] [ə]	7
Consonant	[t] [t‘] [n] [l] [p] [p‘] [m] [f] [k] [k‘] [x] [tɕ] [tɕ‘] [ɕ] ts] [ts‘] [s] [tʂ] [tʂ‘] [ʂ] [ʐ] [j] [w]	23

TABLE Ⅱ
CORPUS DESIGN

Speak	Sentence	Syllable	Duration
Female, Age 24, Beijing, China	374	7726	37mins

B. Data Labeling

In this EMA database, the movements were synchronized to the speech waveform. So we labeled the speech in phoneme level and extracted the corresponding movement data. After label and extraction, all the included data are 10647 phonemes. The first ten phonemes with highest occurrence are listed in table Ⅲ.

TABLE Ⅲ
PHONEME FREQUENCY LIST

Phoneme	Frequency	Phoneme	Frequency
[i]	1155	[t]	593
[j]	901	[tɕ]	464
[u]	667	[ʂ]	462
[l]	663	[tʂ]	411
[ɤ]	658	[w]	411

C. Data Processing

In EMA recording process, two sensors were placed on the speaker’s lips, one coil on the jaw, three sensors on the tongue surface to record the internal movements, and three reference sensors on Nose Bridge and skull behind the ears respectively. After labeling, we got a central point of each phoneme segment by finding out the minimum velocity frame based on the 3 sensors on the tongue. Then, by means of checking the tongue movements’ distribution of each phoneme, we excluded some incorrect points for the phonemes according to the deviation from the gathering center.

As a result, we adopted acoustic and articulation method to screen EMA data. The procedure is as follow: (1) calculate the LPC coefficient of speech sound for the phonemes and transform the FFT to get their envelopes; (2) for a given phoneme, if its deviation from the mean envelop is more than two times of the standard deviation, the phoneme would be excluded; (3) repeat step (1) and (2) twice; (4) calculate the deviation from the mean articulation position for each pronunciation based on the three tongue sensors, by using the same approach as that for acoustics method. Fig. 1 shows the data distribution before and after the filtering. One can see that the isolated points were removed by this processing.

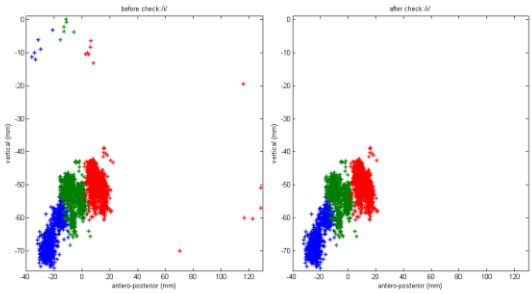

Fig. 1 EMA data preprocessing. Left figure shows the original articulation distribution for phoneme [i], and right figure presents the filtered distribution.

III. Construction of Simulation Database Using 3D Model

In this study, we applied a 3D continuum physiological model as the visualization platform shown in Fig. 2, which was constructed based on the previous work [10] by means of ArtiSynth [11], a 3D biomechanical simulation toolkit.

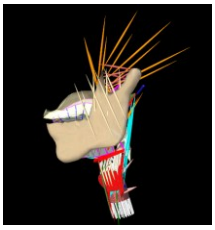

Fig. 2 Three-dimensional continuum physiological model we used.

A. Model Simulation

In order to build an articulation database from model simulation, 149,275 muscle activation patterns were designed to cover the possible articulation based on the method used by Fang [12]. Each muscle activation pattern is a combination of 18 muscle control parameters. By changing the size of the muscle force, we can control the movement of the model, especially the tongue’s movement. Then, we applied these activation patterns to model to obtain articulatory simulation.

B. Extract Tongue Shapes in the Mid-Sagittal Plane

After the model simulation was finished, we extracted the tongue’s mid-sagittal plane data of each simulation to prepare for the mapping construction. In the mid-sagittal plane, 11 points were used to present the tongue (see Fig. 3).

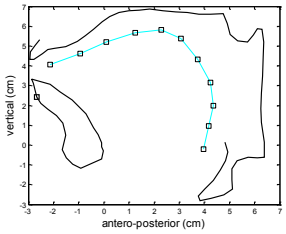

Fig. 3 Model mid-sagittal plane.

IV. Mapping Between EMA Data and Model Simulation

We employed Chinese EMA data obtained in Section II to drive a 3D model which was extracted from a physiological model. Fig. 4 shows the fusion procedure. The mid-sagittal configuration of the tongue was extracted from the simulation database constructed in Section III. By matching the three points' data on the tongue of Chinese EMA data with tongue's mid-sagittal data of 3D model, the mapping function was obtained.

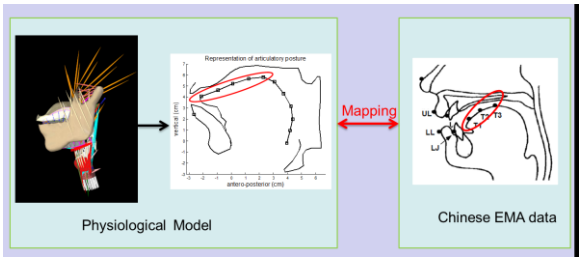

Fig. 4 Procedure of data fusion.

A. Coordinate System Adjustment

In this paper, we projected the EMA coordinate system to 3D model. The hard palate was selected as the reference. Thin-plate spline (TPS) warping method [13] was used to project EMA data on the model in order to make palates in EMA data and 3D model data consistent.

B. Calculation of Mapping Function

After the adjustment of coordinate system, we could project EMA phoneme pronunciation data on 3D model's coordinate system. Allowing for the fact that points on the mid-sagittal tongue plane are too sparse, we adopted Lagrange interpolation method to insert 10 points between every two original tongue points. Since the EMA data of Chinese phoneme [ɤ] show best matching to the rest position of 3-D model, it is used for calibration. The best matching points for the first, second, third point of tongue EMA data are 9th, 27th, 43rd points on the expanded 3D model tongue trajectory respectively, where their mean deviation was 0.09cm. Fig. 5 shows original distribution and distribution after normalization. Therefore, we can use this mapping relationship to find the best match 3D model for other phonemes.

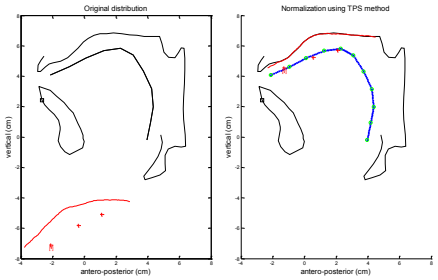

Fig. 5 Left figure: original distribution of EMA data and model. Right figure: matching between Chinese phoneme [ɤ] and 3-D model after TPS normalization.

C. Discussion and Evaluation

Statistical analysis shows that the mean deviation for all 30 phonemes between original EMA data and best matched 3D model is 0.16cm for the three match points. In the EMA data, only three points on the fontal and dorsal parts of the tongue can be observed. Once tongue's front part is fixed, how about the tongue's back part? Whether the tongue's back part will move without constraint or not? In this paper, two experiments were conducted for evaluation.

A) Investigation of the tongue's back part

At first, based on five Chinese people's MRI data, we investigated the mean deviation relationship between the tongue's front and back parts. When one produced seven vowels, the mean deviation of the tongue's front part and that of back part were calculated respectively. Fig. 6 displays the relationship between tongue's front and back parts.

The 3D data were extracted from the model simulation. There are constraints on the physiological model, like volume. So, if the tongue's front part is moving, its back part will change with it possibly. Therefore, a 0.6cm mean deviation perturbation range for tongue's front part was set to select the best several models to investigate the movement of tongue's back part. The statistics show that the mean deviation for the tongue's back part is 0.63cm when that for the tongue's front part is 0.6cm. The above data are almost consistent with the observation from MRI data (see Fig.6), proving that using three points on the tongue front part as projection is feasible.

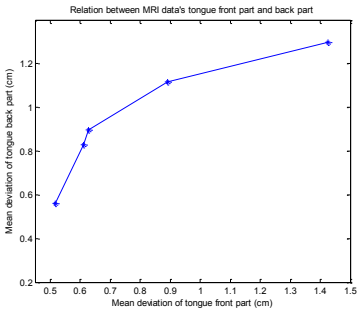

Fig. 6 The relationship between tongue front and back parts.

B) Compare with MRI data

To confirm the method's accuracy, the tongue's mid-sagittal data of the best-matched 3D model was compared with MRI data, based on seven vowel phonemes normalized by using TPS [13] method. The MRI data is presented in Fig. 7.

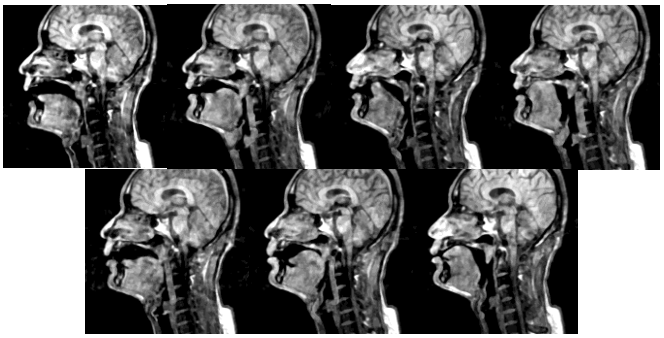

Fig. 7 Chinese MRI data for phonemes [a] [ɤ] [ə] [i] [o] [u] [y] respectively.

The comparison indicates the mean deviation between the vowel's best-matched 3D model and MRI data is 0.21cm which is acceptable.

V. Visualization Presenting System

Base on the above work, we built a visualization system according to the procedure shown in Fig. 8.

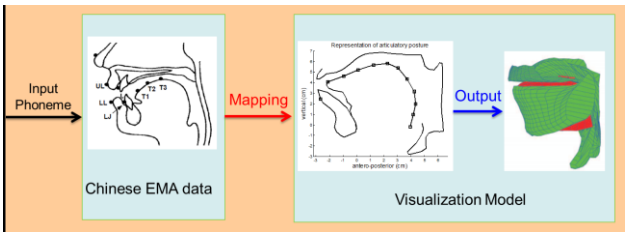

Fig. 8 Visualization system work procedure.

When a phoneme is required to visualize its articulatory movement, the system will first judge whether the phoneme belongs to the Chinese 30 phonemes or not. If not, an error warning will be shown to remind users. Otherwise, the system will depend on the tongue's EMA data of the input phoneme and mapping relationship between EMA data and 3D model data. And then the system will search our model simulation database to find the best-matched mid-sagittal shape, and the corresponding 3D model simulation movement will be presented via a Matlab visualize program soon afterwards.

Fig. 9 is visualization for Chinese phoneme [o] in 2.5-dimensional and 3D views respectively.

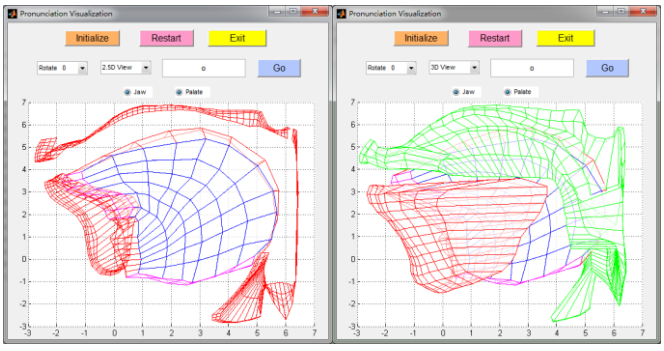

Fig. 9 Visualization for Chinese phoneme [o].

VI. Summary and Conclusions

In this paper, a Chinese phoneme pronunciation visualization method was proposed by combining Chinese EMA data and 3D articulatory physiological model. Comparing with MRI data, it is showed that our method provides an understandable and accurate description of articulator movements for Chinese phonemes. Furthermore, a visualization system is constructed based on the proposed method. For a practical learning-aid system, an inverse estimation modular is necessary. Besides, in order to extend the visualization method to words level, we need to attach great importance on co-articulation problem. These issues are remained for future study.

Acknowledgment

This work is supported in part by the National Basic Research Program of China (No. 2013CB329301), and in part by the national natural science foundation of China under contract No. 61233009 and No.6117501.

References

[1] A. Neri, C. Cucchiarini, and H. Strik, "ASR corrective feedback on pronunciation: does it really work?," *Proc. ISCA Interspeech.* Pittsburgh, PA, pp. 1982-1985, 2006.

[2] N. C. Ellis, *Implicit and Explicit Learning of Languages,* Cambridge University Press, 2000.

[3] T. Kaburagi and M. Honda, "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes", *J. Acoust. Sot. Amer.* 99 (5) (1996) 3154-3170.

[4] K..H. Wong, W. K. Leung, W. K. Lo and H. Meng, "Development of an articulatory visual-speech synthesizer to support language learning," *ISCSLP2010*, Tainan, Taiwan, 2010.

[5] S. LaRocca, S. Bellinger, and T. Potter, "Voice-interactive German homework at the US military academy," *Proc. ISCA ITRWINSTiL2000.* Dundee, Scotland, pp. 26-30, 2000.

[6] M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, "The Native Accent pronunciation tutor: measuring success in the real world," *SLaTE Workshopon Speech and Language Technology in Education.* pp. 124-127, 2007.

[7] LLabs, "MyCT", http://www.myet.com/MyETWeb/

[8] Y. Iribe, S. Manosavan, K. Katsurada, R. Hayashi, C. Zhu, et al., "Improvement of animated articulatory gesture extracted from speech for pronunciation training," *ICASSP2012*. Kyoto, Japan, 2012.

[9] F. Hu, "An acoustic and articulatory analysis of vowels in Ningbo Chinese." *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003.

[10] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Amer.,* vol. 115, no. 2, pp. 853–870, 2004.

[11] S. Fels, F. Vogt, K. V. D. Doel, J. Lloyd and O. Guenter, "Artisynth: Towards realizing anextensible, portable 3D articulatory speech synthesizer." *International Workshop on Auditory Visual Speech Processing*, pages 119–124, 2005.

[12] Q. Fang, J. Dang, "Physiological Articulatory Model for Investigating Speech Production: modeling and Control", VDM Verlag Press, 2009.

[13] J. Wei, J. Dang, "Vocal tract normalization in articulatory space using thin-plate spline method". *J. Acoust. Soc. Amer.*, 123(5): 3885-3885, 2008.