DATA-DRIVEN 3D VISUAL PRONUNCIATION OF CHINESE IPA FOR LANGUAGE LEARNING

Jun YU[1, 3]
National Engineering Lab for Speech and Language Information Processing[1]
University of Science & Technology of China
Hefei 230027, China

Aijun LI[2], Fang HU[2], Qiang FANG[2]
Institute of Linguistics[2]
Chinese Academy of Social Sciences
Beijing, 100732, China

Chen JIANG[3, 4], Xian LI[3], Jing YANG[3]
Dept. of Automation[3]
University of Science & Technology of China
Hefei 230027, China

Zeng-fu WANG[1, 3, 4]
Institute of Intelligent Machine[4]
Chinese Academy of Science
Hefei 230031, China

Abstract — In the framework of intelligent aided language learning, a real-time data-driven 3D visual pronunciation system of Chinese IPA is proposed. First, a high quality articulatory speech corpus including speech and 3D articulatory data of lips, tongue and jaw movements is collected through Electro-Magnetic Articulograph; second, the 3D articulatory modeling including shape design and motion synthesis is conducted. The articulatory shape is obtained by designing a precise 3D facial model including internal and external articulators. The articulatory motion synthesis is obtained combining parameterized model and anatomical model. The system can thus illustrate the slight differences among phonemes by synthesizing both internal and external articulatory movements. The perceptual evaluation shows the suitability of the system for instructing language learners to articulate.

Index Terms — Intelligent aided language learning, 3D articulatory modeling, Electro-Magnetic Articulograph recording.

I. INTRODUCTION

Traditional method of phonetic training is teacher-oriented: students follow a teacher to pronounce and the teacher corrects the students' imitations when necessary. This teacher-oriented method has several limitations. First, the training effect is limited by the teacher's degree of professionalism. Second, it blocks the possibility of self-study, which is a significant ability in the second language (L2) learning process (such as Teaching Chinese as a Foreign Language (TCFL) and Teaching English as a Foreign Language (TEFL)), especially for the adult learners. Therefore, computer-aided language learning (CALL) is of great importance, especially to adult L2 learners, and it has been a hot topic since the late 1990s in the research field.

CALL provides a promising assistant method for language learning especially through mobile devices, as it could gradually reduce the dependence on language teachers. In China, a number of research institutions, such as IFLYTEK Laboratory of University of Science and Technology of China, Chinese University of Hong Kong as well as the Institute of Acoustics of Chinese Academy of Sciences (CAS) and the Institute of Automation of CAS, are working on language teaching and evaluation on the basis of speech technologies. Moreover, Shenzhen Institutes of Advanced Technology of CAS has built up a platform for English pronunciation training by applying 3D acoustical animation technology [1-4], which is demonstrated as an efficient way in pronunciation training for L2 learners [5].

The present study aims to build up a phonetic pronunciation system through a data-driven 3D talking head system, with a special focus on training IPA, syllable and word pronunciations for both Chinese students and the students who are learning Chinese as L2 (such as those in Confucius Institutes or other organizations).

The trainings consist of two parts. First is the training on IPA pronunciations, which is based on the vowel and consonant charts for Chinese dialects [6]. Second is the training on pronunciations of monosyllables and vocabularies, which covers all the syllable and word lists in TCFL published by the Ministry of Education of China and China's national linguistics work committee [7][8].

This paper focuses on the Chinese IPA pronunciation training, concerning the articulatory speech corpus collection, and the data-driven 3D articulatory modeling. And the data-driven 3D articulatory modeling consists of two parts: first is articulatory shape design, which is obtained by designing a precise 3D facial model including internal and external articulators; second is articulatory motion synthesis, which is obtained based on articulatory movement recording.

II. ARTICULATORY SPEECH CORPUS OF CHINESE IPA

A. Chinese IPA

International Phonetic Alphabet (IPA) is a useful tool in language learning as well as in language or dialect survey. Currently there are two commonly used sound resources for IPA illustration by Chinese phoneticians. One was recorded by the famous linguist Yuan-ren Chao in 1981. The other was recorded by Mr. Dianfu Zhou in 1985. All recordings were published as parts of a multi-media CD called the Documentaries of Phonetics in China [9].

However, these old recordings are audio only and the sound quality is limited. Therefore, a high quality articulatory speech corpus including 3D articulatory data of lips, tongue and jaw movements as well as audio sounds is collected in current study. This new corpus is used for data training in this study and is also expected to help language learners and researchers in general.

B. Training material of Chinese IPA pronunciation

Training on Chinese IPA pronunciation consists of tones, consonants, vowels and diphthongs.

1) Tones (25 training inventories): Level tones, rising, falling tones, rising-falling tones, and falling-rising tones.

2) Consonants (306 training inventories): The consonants are listed according to place and manner of articulation. The majority of the consonants are pronounced together with /a, i, u/, in the form of "CVCV, CVC" as in /papa, pap/; Palatals are pronounced together with /ia, i, iu/ in the form of "CVCV, CVC"; Retroflex sounds are pronounced together with /a, ɻ, u/ in the form of "CVCV, CVC".

3) Monophthongs (40 training inventories): Rounded and unrounded monophthongs are produced in pairs. Apical vowels and syllabic consonants /m, n, ŋ, l/ are also included.

4) Diphthongs and others (117 training inventories): Combinations of possible vowel elements; sequences of ten common vowels with /i, u, y, m, n,ŋ, p, t, k, ʔ/ respectively; ten common vowels in nasalized and rhotacized versions.

C. Articulatory data recording

The Electro-Magnetic Articulograph (EMA, the Carstens AG500 system) was used to trace the 3D trajectories of articulatory movements. Four dialect experts (2 males and 2 females) from the Institute of Linguistics, Chinese Academy of Social Sciences (CASS) participated in the IPA recording. The sensors were attached onto the positions depicted in Fig. 1 as Tongue Rear (TR), Tongue Blade (TB), Tongue Tip (TT), Lower Incisor (LI), Lower Lip (LL) and Upper Lip (UL). Another 3 sensors were glued to the bridge of nose (NOSE), the back of left ear (LE), and the back of right ear (RE) respectively. Those points are relatively stable during the recording, so they were used as reference to remove head movements of the speaker.

The articulatory data were sampled at 200 Hz. The audio data were recorded simultaneously with the articulatory data at a sampling rate of 22 KHz.

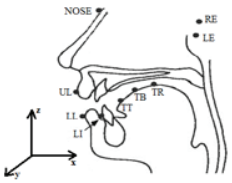


Fig. 1. Sensor attachment. X-axis corresponds to the posterior-anterior dimension and z-axis denotes the inferior-superior dimension.

III. Data-driven 3D Articulatory Modeling

The data-driven 3D articulatory modeling includes two parts, namely shape design and motion synthesis.

The first part is the design of a precise 3D facial model (Fig. 2(a)-(d)) according to the requirements of controllability and expressiveness. The model consists of skin, eyes, teeth, tongue, oral cavity and skeleton, and thus is sufficient for highly realistic facial animation and articulatory animation.

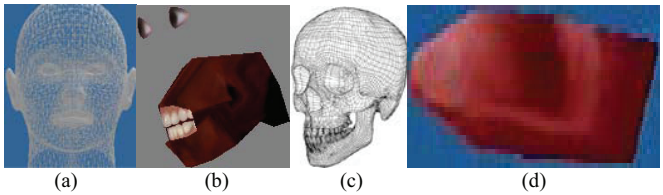


Fig. 2. (a) Skin; (b) Organs; (c) Skeleton; (d) Tongue.

The second part is the mechanism for 3D articulatory animation on the basis of EMA recordings. The mechanism consists of 3 components, namely tongue animation, lip animation, and the synchronicity between speech and animation.

IV. Speech Synchronized 3D Articulatory Animation

The movement of the articulators changes the shape of oral cavity and thus plays a vital role in speech production. The upper articulators are generally passive during speech production. In particular, the up-down rotation degree of the jaw is computed with the trajectory of LI, and we will focus on the tongue and lips in the discussion on the basis of the sampled EMA data.

Another issue for a highly realistic 3D visual pronunciation training system is the synchronization of articulatory animation and the audio speech.

A. Tongue Animation

The realistic 3D articulatory animation requires a deformable tongue model, which consists of the geometry that defines the surface of the tongue, a parameterization that defines a space of possible tongue shapes, and a mapping that deforms the geometry based on the values of parameters.

First, the geometry should meet the requirements of realistic shape and fast deformation. These requirements were addressed by defining a Non-Uniform Rational B-Spline (NURBS) surface.

NURBS patches result in a smooth surface that can be quickly deformed by displacing the control points. Based on the tongue geometry introduced in Fig. 2, we created a NURBS surface capable of producing realistic tongue shapes by the fitting procedure. The fitting procedure has been widely used in computer graphic community [10]. This NURBS surface is composed of a 16×7 grid of bi-cubic patches over 70 control points. It is capable of approximating any non-intersecting tongue shape except for a pronounced medial groove with steep walls. With the addition of 20 control points and a 16×3 grid, the groove is also approximated, and then a 3D deformable tongue model (Fig. 3) is obtained.

Fig. 3. 3D deformable tongue model.

Second, we developed a tongue parameterization to abstract the specification of the tongue shape away from the actual geometry so that it is easier to specify a particular tongue shape.

The concept of parameterizing a model is to reduce the degrees of freedom while maintaining an adequate level of control. It is preferred to have a parameterization with just a handful of parameters, as it is onerous to specify each vertex separately.

As stated in [11], tongue shape is systematically related to tongue position since the tongue volume can only be redistributed rather than being changed. This led us to believe that a few number of parameters would be sufficient to specify tongue shapes, and we found that the recording positions of TR, TB and TT (squares in Fig. 3) are sufficient to deform the tongue model into any shape necessary for the visual representation of speech. Therefore, these x, y, z coordinates of TR, TB and TT are defined as the parameters for the shape of tongue model.

Third, a motion equation of NURBS control points is defined as the mapping between the deformation of geometry and the values of parameters. At the same time, a simple and fast method is embedded to the equation for approximating volume preservation.

The control point $C_{i,j}$ are arranged in a 9×10 grid with the 9 control points along column i and located along the tongue from front to back, and with the 10 control points along column j and located along the tongue from top to down.

The new control point $C'_{i,j}$ after movement are calculated as:

$$C'_{i,j} = C_{i,j} + \sum_{k=1}^{9} \beta_{k,j} p_k \alpha_{k,i} \tag{1}$$

Where $C_{i,j}$ is the old control point, $\beta_{k,j}$ is the weight for row j of the *kth* parameter p_k and $\alpha_{k,i}$ is the weight for column i of the *kth* parameter p_k. There are separate weights for x, y, z, and the criteria of designing them are: reducing the tongue width when the tip is extended, and expanding the tongue during retraction. For example, when the value of $\alpha_{k,i}$ for z is enlarged, the value of $\beta_{k,j}$ for x is reduced simultaneously. Based on these criteria, volume preservation is simulated.

As a result, when the recording positions of TR, TB, TT are given, they are set as the parameters ($p_1, \cdots, p_9$) of NURBS surface, and the positions of control points are obtained by equation (1), then the positions of vertices are obtained by NURBS interpolation.

B. Lip Animation

The lips are muscular hydrostats in that they are composed entirely of soft tissues and move under local deformation. Given the recording positions of LL, UL, constrained deformations are applied to the adjacent points according to the muscular model [12][13]. The model divides facial muscles into linear muscle, sheet muscle and sphincter muscle based on the motion characteristics of facial muscles [14] (Fig. 4). The first two muscles are used for tension, while the last one is used for shrinkage. The model uses the directional characteristics of muscle motion to build a vector model which is independent of lower skeleton structure, and each vector has its action area which changes in accordance with the distance from attachment point. Therefore, the vertices of 3D facial model are controlled to produce facial animation through the motion of vector model. For our lip animation, linear muscle and sphincter muscle are used.

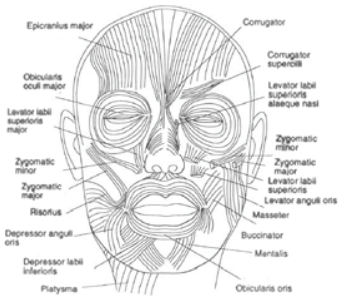
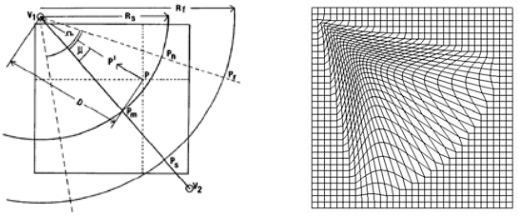


Fig. 4. Distribution of facial muscles [12].

Fig. 5 is the diagram of linear muscle, and illustrates the effect of muscle extraction. V_1 is the fixed endpoint; V_2 is the removable endpoint; P is an arbitrary point in the action area; P' is the point after movement of P; D is the distance between P and V_1; Ω is the biggest angle in the action area; μ is the angle between V_1P and V_1V_2. Therefore, after the movement of linear muscle, the displacement from P to P' is:

$$\overline{PP'} = K \cdot \cos(\mu) \cdot h(D) \cdot \overline{V_1P} \Big/ \left\| \overline{V_1P} \right\| \tag{2}$$

K is constant, $h(D) = \begin{cases} \cos(1 - D/R_s), & D \le R_s \\ \cos((D - R_s)/(R_f - R_s)), & R_s \le D \le R_f \end{cases}$.



(a) (b)

Fig. 5. Linear muscle. (a) Diagram. (b) The effect of muscle extraction.

Fig. 6 is the diagram of sphincter muscle, and illustrates the effect of muscle extraction. v_i and u_i are vertical coordinate and horizontal coordinate of point x_i; a' and b' are long radius and short radius of elliptical action area; O is center of gravity. Therefore, after the movement of sphincter muscle, the displacement from O to x_i is:

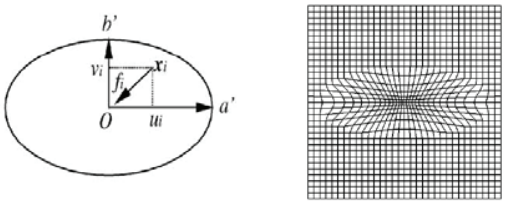
$$f_i = \alpha\theta(r_i)(O - x_i)/\|O - x_i\| \quad (3)$$

α is elastic parameter, and $\theta(r_i)$ is defined as:

$$\theta(r_i) = \cos\left((1 - r_i)\cdot\pi/2\right),\ 0 \le r_i \le 1 \quad (4)$$

$$\theta(r_i) = \cos\left(\left((r_i - 1)/(R - 1)\right)\cdot\pi/2\right),\ 1 \le r_i < R \quad (5)$$

$r_i = \sqrt{u_i^2 a'^2 + v_i^2 b'^2}\,/\,(a'b')$ is a weighted distance; R is a threshold, beyond which $\theta(r_i)$ is zero.



(a) (b)

Fig. 6. Sphincter muscle. (a) Diagram. (b) The effect of muscle extraction.

In the 3D facial model, a sphincter muscle is defined in each of the upper lip and the lower lip, and a linear muscle is defined in each of the left lip corner and the right lip corner. Fig. 7 illustrates two action areas of sphincter muscles in the lips.

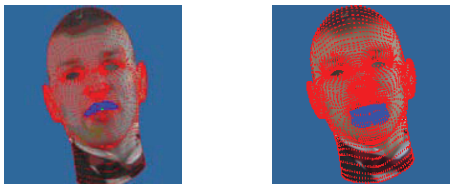


Fig. 7. Action areas (blue points) of sphincter muscles in the lips.

C. Establishing the synchronicity between speech and Animation

The speech and animation should be synchronized in phoneme level, otherwise, it will introduce error to language learners in pronunciation training.

For the speech input, phonemes and durations are extracted from speech by speech recognition engine, then the 3D articulatory animation corresponding to each phoneme are concatenated to form continuous animation, and are played with speech simultaneously according to phonemes durations.

V. EXPERIMENTS

Experiments are conducted under such configuration: AMD Athlon (tm) II X4 640 3.01GHz, memory 2GB, NVIDIA GT200.

A. 3D Facial Model Adaptation

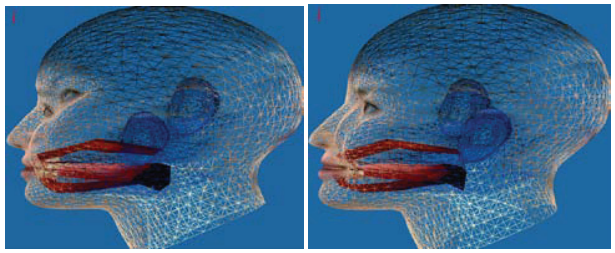
As the foundation of 3D animation, 3D facial model adaptation is applied by a deforming-based approach [15]. Fig. 8 is the results on two images of a foreign man and a Chinese man.

(a) (b) (c) (d) (e) (f)

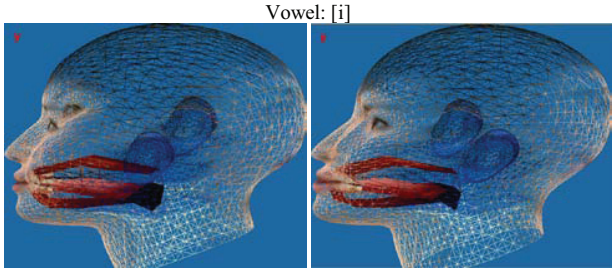
Fig. 8. (a) Carphone image. (b) Front of adaptation result on (a). (c) Profile of (b). (d) Jun YU image. (e) Front of adaptation result on (d). (f) Profile of (e).

B. Speech Synchronized 3D Articulatory Animation

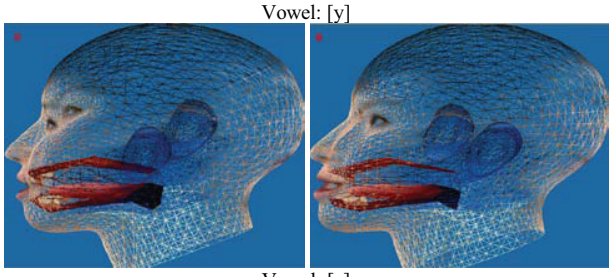
According to the articulatory speech corpus, we test the reality of articulators with transparent skin. As shown in Fig. 9, we can see that the articulators have a highly realistic appearance, and their movements cooperate well with the speech.



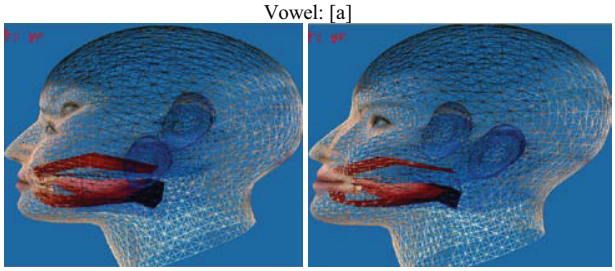
Vowel: [i]



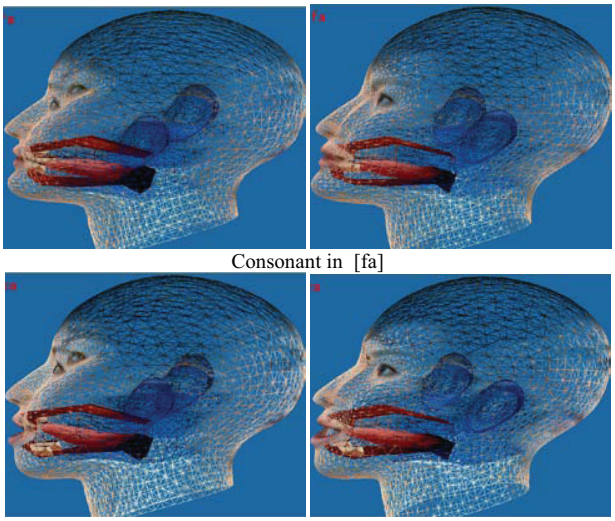
Vowel: [y]



Vowel: [a]



Retroflexed vowel [yr]



Consonant in [fa]

Consonant in [tsa]

Fig. 9. 3D articulatory animation with transparent skin.

To evaluate the ability of real-time, table 1 shows the average time each frame takes when producing 3D articulatory animation. As it shows, the system can run in real time.

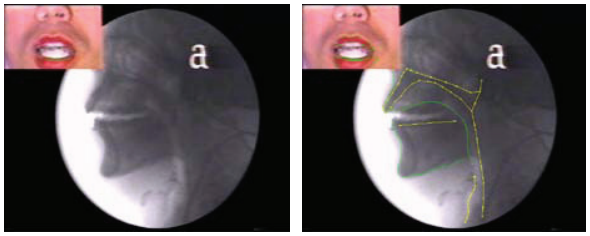
Table 1. Average time each frame takes when producing 3D articulatory animation.

	Average time each frame takes
3D articulatory animation	0.035s.

C. Perceptual Evaluation

The evaluation of 3D visual pronunciation has not been more formal than observing how the 3D articulatory animation fits to its corresponding movement of the real articulator during pronunciation. The problem to an objective evaluation lies in the difficulty to get the absolute truth of real 3D articulatory motion, not to mention that EMA data points are too sparse due to the limited number of coils. Magnetic resonance imaging (MRI) and X-ray imaging provide possible solutions. But MRI has the drawbacks of low accuracy and poor real-time capability. X-ray technique has the unique advantage of real-time capability, and is in principle the most appropriate method to determine the actual tongue shapes during speaking. However, it is difficult to identify individual articulators on the videofluoroscopic images of X-ray, and it is harmful to the speaker's health.

In this study, the available videofluoroscopic images of X-ray for Mandarin Chinese [16] are used as the baseline for comparison. The articulatory shapes on the videofluoroscopic images of X-ray were detected by active appearance model (AAM) [17]. AAM is a computer vision algorithm for matching a statistical model of object shape and appearance to a new image. They are built during a training phase. A set of images, together with coordinates of landmarks that appear in all of the images, is provided to the training supervisor. The algorithm uses the difference between the current estimate of appearance and the target image to drive an optimization process. By least square techniques, the results can match to new images very swiftly. Our AAM is built on 113 images during the training, and the initial position is set by manual adaptation. In practice, due to the limit of training database, articulatory shapes may not be accurate enough, so manual intervention has to be added. Fig. 10 shows the detection result of articulatory shape on one frame of the videofluoroscopic images of X-ray during pronunciation.



(a) (b)

Fig. 10. (a) One frame of the videofluoroscopic images of X-ray when pronouncing vowel [a], (b) The articulatory shape detection result of (a).

Users' reactions interacting with the 3D visual pronunciation system are evaluated based on the above baseline. It is a between-subject experiment. There are 34 participants. They all speak Mandarin Chinese but are from different dialectal regions. The distributions of their age, gender and dialect background are shown in table 2. The main goal of the evaluation is to decide if the 3D articulatory animation is consistent with the corresponding real articulatory movement in the videofluoroscopic images.

Table 2. Distributions of participants.

Construct	Profiles	Distribution
Age	**1.** Younger than 20. **2**. Between 20 and 30. **3**. Older than 30.	8/18/8
Gender	**1**. Male. **2**. Female.	20/14
Dialect background	**1**. Northeastern. **2**. Beijing. **3**. Zhongyuan. **4**. Southwestern. **5**. Xiang. **6**. Gan. **7**. Minnan. **8**. Cantonese. **9**. Jilu. **10**. Guiliu Pian. **11**. Wu.	2/3/7/3/2/3/2/4/3/1/4

The first stage is a questionnaire survey. Table 3 shows the constructs and questions of the survey related to the system performance. The first construct, i.e. expressiveness, verifies if the articulators can display the corresponding visual signature of speech, or have any unusual movement, such as tongue tied, etc. The second construct, i.e. coherence, verifies if the dynamics of articulatory animation are comparable to those of articulatory movements in the videofluoroscopic images of X-ray. The third construct, i.e. appearance, verifies if the appearance of articulators is realistic, which is important in motivating/inspiring learners during language learning. The answers to these questions are given from 'absolutely disagree' to 'totally agree' on a ten-point scale. In order to determine if these constructs have a significant relationship, a Cronbach's alpha test [18] is carried out between questions for each construct. Cronbach's alpha is a coefficient of reliability, and is commonly used as a measure of

internal consistency or reliability of psychometric test score for a psychometric instrument. Typically, an alpha of 0.7 or greater is considered acceptable in psychological experiments.

Table 3. Cronbach's alpha results of questionnaire.

Construct	Question	Cronbach's alpha
Expressiveness	Articulatory movements look natural.	0.756
Coherence	Articulatory animation is coherent with the movement of real articulator.	0.787
Appearance	I like articulatory appearance.	0.734

As shown in table 3, the alpha values suggest that the created groups of questions refer to the same topic. In other words, the questionnaire is suitable for the evaluation.

In the second stage, the 3D visual pronunciation system performed 3D articulatory animation of several IPA pronunciations, and the participants were shown the corresponding articulatory movements in the videofluoroscopic images of X-ray simultaneously. The participants were then asked to fill in the questionnaire.

Table 4 is the result of mean scores after evaluation. The maximum is 10, and the minimum is 0. The fact that the scores for all the constructs are greater than 7.5 suggests that the 3D articulatory animation is consistent with the corresponding real articulatory movement in the videofluoroscopic images of X-ray.

Table 4. The mean score of evaluation.

Construct	Question	Mean score
Expressiveness	Articulatory movements look natural.	8.23
Coherence	Articulatory animation is coherent with the movement of real articulator.	8.57
Appearance	I like articulatory appearance.	7.94

D. Comparison with Other Visual Pronunciation System

Compared with current visual pronunciation systems [1][2][3][4], our work has following characteristics: *1)* it is the first visual pronunciation system about Chinese; *2)* a combination of parameterized model, i.e. NURBS interpolation, and anatomical model, i.e. muscular model, can produce realistic articulatory animation while maintaining the real-time ability.

VI. CONCLUSION

A real-time data-driven 3D visual pronunciation system of Chinese IPA is proposed. First, a high quality articulatory speech corpus is collected; second, the 3D articulatory modeling including shape design and motion synthesis is conducted. The system can thus illustrate the slight differences among phonemes by synthesizing both internal and external articulatory movements. The perceptual evaluation suggests that it is suitable for the instruction purpose in language (articulation) learning.

In further steps, the ability of collision processing among tongue, teeth and oral cavity is to be added, and the volume preservation of tongue is to be improved. What's more, the system is to be transplanted to embedded mobile platforms such as smart phone.

ACKNOWLEDGEMENT

This research is supported by the Innovation Project of CASS, the National Natural Science Foundation of China (No. 61303150), the Fundamental Research Funds for the Central Universities (No. WK2100100020), the China Post Doctoral Science Foundation (No. 2012M521248), and the STP of Anhui (No. 11010202192).

VII. REFERENCES

[1] Li, Sheng, Wang Lan, Qi En. "The phoneme-level articulator dynamics for pronunciation animation," IEEE International Conference on Asian Language Processing (IALP) 2011, pp. 283-286.

[2] Lan Wang, Hui Chen, Sheng Li, et al, "Phoneme-level articulatory animation in pronunciation training," Speech Communication, 54, 2012, pp. 845-856.

[3] Hui Chen, Lan Wang, Wenxi Liu, et al, "combined x-ray and facial videos for phoneme-level articulator dynamics," The Visual Computing, 26 6-8, 2010, pp. 283-286.

[4] Wang Lan, Chen Hui, Ouyang Jianjun. "Evaluation of external and internal articulator dynamics for pronunciation learning," InterSpeech, 2009, pp. 2247-2250.

[5] Pierre Badin, Yuliya Tarabalka, Fredric Elisei, Gerard Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," Speech Communication, 52, 2010, pp. 493–503.

[6] Institute of Linguistics of CASS, "The character tables for Chinese dialect survey," Beijing: The commercial Press, 1981.

[7] China National Office for Teaching Chinese as a Foreign Language ed, "The graded Chinese syllables, characters and words for the application of teaching Chinese to the speakers of other language: interpretation for the national standard," BLCUP, 2010.

[8] Ministry of Education of PRC and China's national linguistics work committee. "The graded Chinese syllables, characters and words for the application of teaching Chinese to the speakers of other language." BLCUP, 2010.

[9] Institute of Linguistics, CASS. (ed.), "Documentaries of phonetics in china (CD-Rom)," Beijing: Social Sciences Academic Press.

[10] Djordje Brujic, Iain Ainsworth, Mihailo Ristic, "Fast and accurate NURBS fitting for reverse engineering," International Journal of Manufacture technology, 54 5-8, 2011, pp. 691-700.

[11] Scott A K, Richard E P, "A 3D parametric tongue model for animated speech," Journal of Visualization and Computer Animation, 12, 2001, pp. 107-115.

[12] K. Waters, "A muscle model for animating three dimensional facial expression," Computer Graphics, 22(4), 1987, pp. 17-24.

[13] Marcos S, Bermejo JGG, Zalama E, "A realistic facial animation suitable for human-robot interfacing," International Conference on Intelligent Robots and Systems (ICIRS) 2008, pp. 3810-3815.

[14] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," Palo Alto, CA, USA: Psychologists Press, 1978.

[15] Hu Yuankui, Zheng Ying, Wang Zengfu, "Reconstruction of 3D face from a single 2D image for face recognition," IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 217-222.

[16] Bao Huaiqiao and Yang Lili (edits), "X-ray recoding data for Chinese Syllables," 1963.

[17] I. Matthews, J. Xiao, and S. Baker, "2D vs. 3D deformable face models: Representational power, construction, and real-time fitting," International journal of computer vision, 75(1), 2007, pp. 93-113.

[18] S Marcos, J Gomez-Garcia-Bermejo, "A realistic, virtual head for human-computer interaction," Interacting with Computers, 22, 2010, pp. 176-192.

[This paper was published in O-COCOSDA,2013]