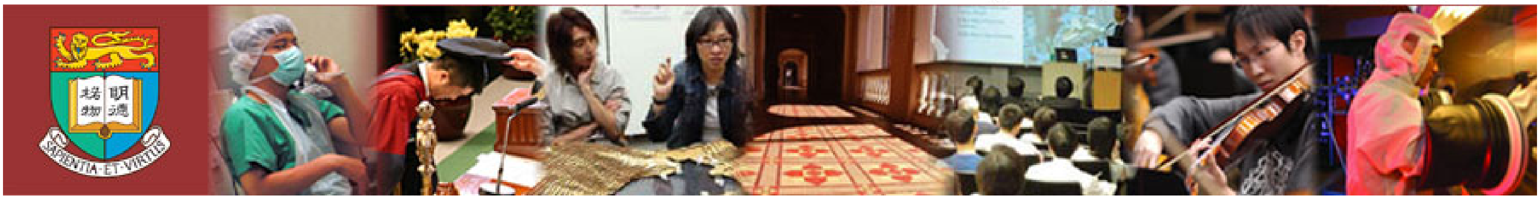


Proceeding of the IEEE
International Conference on Information and Automation
Yinchuan, China, August 2013

An Interactive Speech Training System with Virtual Reality Articulation for Mandarin-speaking Hearing Impaired Children

Speech training system with VR articulation

Xiaoqian Liu, Nan Yan, Lan Wang
Laboratory for Ambient Intelligence & Multimodal System
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences/The Chinese University of Hong Kong
Shenzhen, Nostate 518000, China
E-mail: {liu.xq, nan.yan, lan.wang}@siat.ac.cn

Xueling Wu
QingQing Speech Rehabilitation Centre
Shenzhen, Nostate 518000, China
E-mail: wuxueling666@163.com

Manwa L. Ng
Speech Science Laboratory, Division of Speech and Hearing Sciences
The University of Hong Kong
Hong Kong, China
Email: manwa@hku.hk

***Abstract* - The present project involved the development of a novel interactive speech training system based on virtual reality articulation and examination of the efficacy of the system for hearing impaired (HI) children. Twenty meaningful Mandarin words were presented to the HI children via a 3-D talking head during articulation training. Electromagnetic Articulography (EMA) and graphic transform technology were used to depict movements of various articulators. In addition, speech corpuses were organized in listening and speaking training modules of the system to help improve language skills of the HI children. Accuracy of virtual reality articulatory movement was evaluated through a series of experiments. Finally, a pilot test was performed to train two HI children using the system. Preliminary results showed improvement in speech production by the HI children, and the system was recognized as acceptable and interesting for children. It can be concluded that the training system is effective and valid in articulation training for HI children.**

Index Terms - hearing impaired; speech training; interactive; articulatory tutor; virtual reality

I. INTRODUCTION

Hearing impaired (HI) children often lack necessary phonetic skills in daily communication. Resmi et al. found that infant and childhood deafness had immense impact on communication, manifested mainly in delayed speech and language development [1]. Engwall and his colleagues also pointed out that children with hearing loss showed deficits in acoustic speech targets with which to imitate and compare their own speech [2]. However, hearing impaired children can learn speech by relying on visual cues of phonetic features. A recent study showed that speech perception based on audio-visual feedback appeared to be superior to visual- or auditory-only perception [3]. Visualization might help children conceptualize the place of inner articulators and control of movements during speech production [3]. This visual feedback can be enabled using three-dimensional virtual reality (VR) articulatory movement during speech trainings.

On the other hand, studies demonstrated that the mirror neuron system is mainly connected to hand, mouth and foot actions [4-5]. Speech production might be influenced by such mirror neuron system. Franceschini et al. found that the role of human mirror neuron system in language falls into two different classes [6]. One relates to articulated speech, in which perception of linguistically relevant sounds appears to depend on previous experience in producing those sounds. The other aspect of mirror neuron system and language is that brain area for specific motor execution is involved in the understanding language describing such action. Motor imagery induced by motor observation, imitation and execution may enhance language rehabilitation of those with articulatory disorders, such as HI children. Therefore, the importance of imitation in

-This Research is supported by National Natural Science Foundation of China (NSFC 61135003, NSFC 90920002), Guangdong Innovative Research Team Program (No. 201001D0104648280).

978-1-4799-1334-3/13/$31.00 ©2013 IEEE

language study should be emphasized. In the present study, children's imitation was encouraged by watching and imitating virtual reality articulatory movement.

Traditional speech rehabilitation training attempts to display the articulatory movement and location during the syllable pronunciation by teacher's example. However, the movement of inner articulatory such as tongue cannot show to HI children accurately. In recent years, audio-visual speech training systems with 3D articulatory tutors were developed to fill in the gaps of traditional training. Rathinavelu et al. developed a computer-aided speech training system for the deaf persons. MRI data were used in the system to model the movement of articulators and pronunciation sequences. The pronunciation sequences were then used to train hearing impaired patients [7-8]. Another text-to-audiovisual speech synthesizer developed using visemes that contained a set of images spanning a large range of mouth shapes corresponding to each phoneme was also developed by colleagues at MIT [9]. Panasonic Speech Training Technology Lab worked out a speech training system in which several types of instrumentally measured articulatory and acoustic data (namely palatography, nasal vibration, airflow and presence or absence of voicing) were integrated and presented in both technical and motivating game format. They also invented another system that allowed a student to enter any utterance to be learned and instructed the articulatory model to realize the utterance. The system measured a student's production and evaluated it against the parameters of standard utterance for its similarity [10-11]. Olle Bälter et al. also developed an interesting ARTUR system for language learning and HI rehabilitation. The system included a 3D animated head to illustrate the process of speech production [12]. Meanwhile, the pronunciation process of the speaker was recorded by the system and compared with standard pronunciations. Similar work has been done in France and Taiwan [13-14] where computer animation, sound and video signals were combined to teach sound production to deaf children. After children have listened to and repeated a particular phoneme or word, speech recognition system compared the speech with standard data in order to provide feedback.

Although systems and training methods already exist, a 3D articulatory tutor that focuses on Mandarin speaking and HI children rehabilitation is still lacking. There are currently over 20,000,000 HI children who need effective language training in China, making the design of a computer-aided Chinese speech training system especially urgent [15]. Moreover, Mandarin Chinese consists of enormous coarticulation. Synthesizing higher levels of articulatory movements, such as word or sentence, is a big challenge. Since articulatory movements could change vastly in continuous speech compared with single phonemes, movement data associated with continuous should be obtained and applied to 3D talking heads for smoother and more accurate VR articulation.

The aim of the present study was to develop an interactive speech training system with VR articulatory movement for Mandarin speaking training. The three-dimensional vocal tract articulatory model allows a realistic movement of jaw, tongue and lips. Through training with this system, HI children could imitate the articulatory movements presented in the VR talking heads. The system also included listening and speaking training modules for HI children who need to practice daily words and confusable words, so that their language skills could be improved.

II. System Design

The requirements of such a system should include:

- Providing lessons with suitable difficulty for children.
- Providing an interface that is easily accepted and welcomed by children.
- Carrying out speech recognition and comparison module that is flexible and accepts speech input just like a human being.
- Motivating children to practice more and guide them to correct themselves.
- Providing data that can be updated from time to time and that can be selected by user freely.

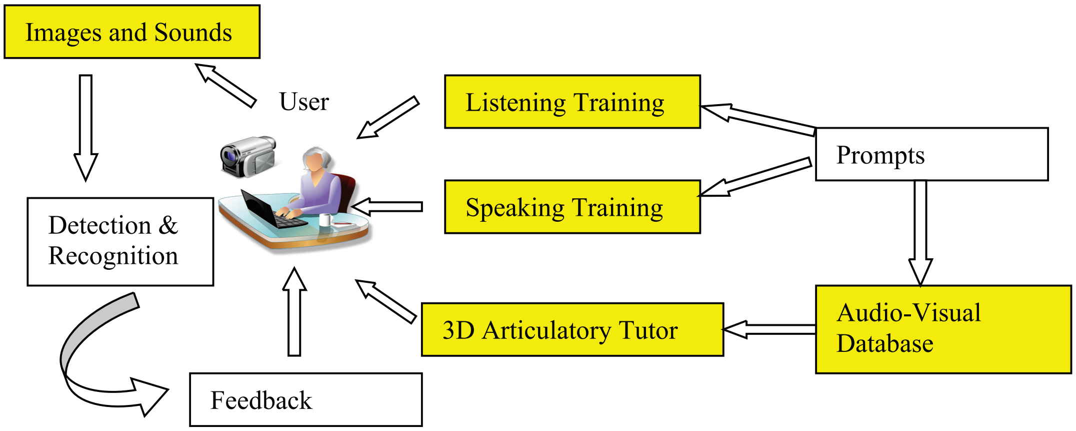


Figure 1. The structure of interactive speech training system

Figure1 illustrates the structure of interactive speech training system. The system contains three main modules: Listening Training Module, 3D Talking Head Articulatory Tutor, and Speaking Training Module. Three-dimensional articulatory data is stored in audio-visual data base and could be selected by user. Once a group of pronunciation videos is selected, the videos could be viewed in 3D Articulatory Tutor's interface repeatedly, until user has finished learning. The prompts of Listen Training and Speaking Training Module could also be selected in line with user's requirements. The sounds of speaker could be automatically detected and recognized by backstage recognition system and speaker could receive feedback instantly. Through training with the system, HI children could simulate standard articulation movements and practice with words in daily life repeatedly.

A. 3D Articulatory Tutor

Children of HI face the difficulty in distinguishing between the phonemes with articulatory movements. Therefore our system specifies on developing a three-dimensional articulatory model based on a set of vocal tract geometrical data, which is acquired by Electro-Magnetic Articulograpthy (EMA). The EMA device is commonly used for audio-visual data recording since the sensors can be attached to the intraoral articulators to get the positions of articulator such as tongue, teeth, and jaw [16-17]. To explore the movement trajectory of main articulators, ten EMA sensors were used, with 4 on the lips, 3 on the tongue. A set of three-dimensional articulatory model was then developed using EMA data with the method mentioned in [16].

Articulatory movements could be synthesized using phoneme level EMA data. However, directly connecting each phoneme didn't receive satisfactory results, because coarticulation is common in Chinese pronunciation and it could have immense influence on articulatory movements. Therefore, Cohen-Massaro Coarticulation Model was used to achieve high-quality synthesized words [18]. Cohen-Massaro Coarticulation Modeling contains two steps.

1): Synthesization. The difference between key frame and static frame is calculated as the displacement vector of each phoneme. The displacement vector of phoneme p is labeled as R_p and the movement range of phoneme p is calculated as (1).

$$\alpha_p = \begin{cases} \dfrac{\max_p\left|R_p\right| - \left|R_p\right|}{\max_p\left|R_p\right|} & \left|R_p\right| \neq \max_p\left|R_p\right| \\ 1 & \left|R_p\right| = \max_p\left|R_p\right| \end{cases} \tag{1}$$

Then, an exponential function is used to simulate pronunciation movements of single phoneme. The function of phoneme p is defined as (2). In the equation c is a constant and we make it 2 after the experiments. θ_d and θ_g are also constants which represent the speed of growth and decline.

$$D_p(\tau) = \begin{cases} \alpha_p \bullet e^{-\theta_d\left|\tau\right|^c}, \tau \leq 0 \\ \alpha_p \bullet e^{-\theta_g\left|\tau\right|^c}, \tau > 0 \end{cases} \quad \tau = t_{pk} - t \tag{2}$$

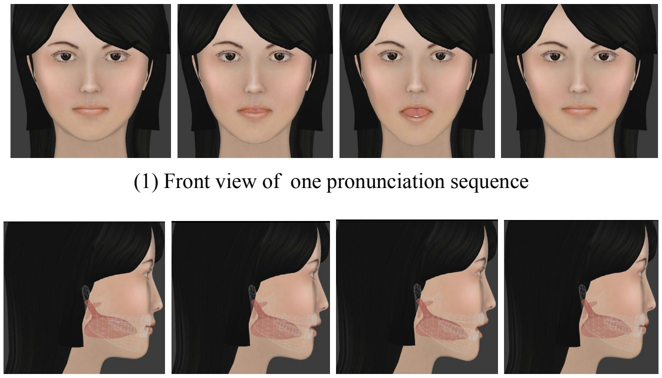
θ_d and θ_g are decided by (3). The labels t_{ps} , t_{pk} and t_{pe} respectively represents the time of starting frame, key frame and ending frame in phoneme p . The label ε is also a constant which represents the local minimum of two exponential functions and we make it 0.22 after experiments.

$$\begin{cases} \alpha_p \bullet e^{-\theta_d\left|t_{pk} - t_{pe}\right|} = \varepsilon \\ \alpha_p \bullet e^{-\theta_g\left|t_{pk} - t_{ps}\right|} = \varepsilon \end{cases} \tag{3}$$

2): Smooth. A blending function proposed in paper [15] is used to smooth between pronunciation movements of two adjacent phonemes and the function can be viewed as (4).

$$\widehat{F}_w(t) = \sum_{p=1}^{N} R_p \bullet D_p(t - t_p) \tag{4}$$

To realize the VR talking head, the 3D articulatory model can be combined with a human face, creating a more familiar environment. Dirichlet Free-Form Transformation algorithm (DFFD) was also utilized to transform VR articulation movement [19]. Totally, the combination of consonants and simple vowels helped us build 20 meaningful words. The key frames of the exampled pronunciation demos show in Figure 2.



(1) Front view of one pronunciation sequence

(2) Side view of one pronunciation sequence

Figure 2. 3D Pronunciation Demos of /ɑː/

Figure 3 illustrates an interface for 3D articulatory tutor in our system. When a child is trained using the system, he (or she) is asked to read following the tutor first. Then the recording of his (or her) pronunciation is processed by backstage recognition system and the key frames of child's lips could be viewed in the interface. By comparing the key frames with standard pronunciation demos, children may find their mistakes in pronunciation and learn quickly.

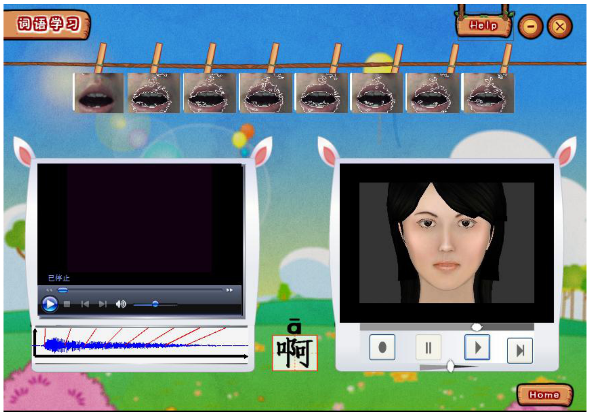


Figure 3. 3D Articulatory Tutor

B. *Listening Training Module*

The listening training module includes 18 groups of ambiguous words, with three words in each group. When a child hears the sound of one word, he/she is asked to select the word he/she perceives and the system gives him/her correct or incorrect feedback. If the child makes a mistake, the system plays the sound of the word again to help the user get a deeper understanding. Figure 4 shows the interface of listening session.

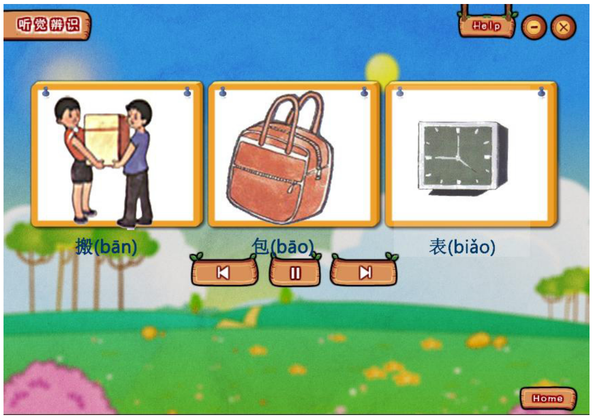


Figure 4. The interface of listening scenario

C. *Speaking Training Module*

To encourage the HI child speaking, the speaking training module was designed. This module includes 4 groups of scenarios, respectively titled as Garden, Zoo, Home and Classroom. In each scenario, some cartoon objects which belong to the category are presented on the screen. When a child sees the highlighted object, he/she needs to say the name of mentioned item. The system will give him/her the correct or incorrect feedback according to automatic backstage speech recognition results. If the child finishes all items in the scenario, the system will give child a positive feedback such as “Good! You speak well.” to encourage HI child speaking. Figure 5 shows an example of speaking training module scenario.

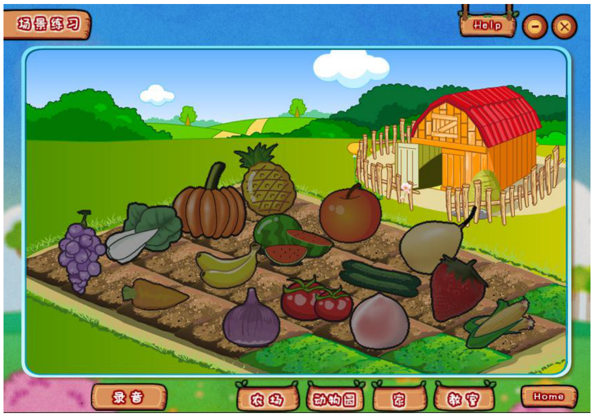


Figure 5. The interface of speaking scenario

III. Interviews

To evaluate the accuracy of VR articulation, two evaluations were conducted before training experiment.

A. *Difference between Confusable Words*

Twenty words used in the system were divided into 10 pairs and 10 individuals with outstanding Mandarin were asked to distinguish within each pair. For example, to test the difference between /ci/ and /tʃ i/, each person was told that this pair of videos includes /ci/ and /tʃ i/ before watching the videos and the participants should identify the two words after watching them. The overall identification accuracy is listed in Table 1. The average identification rate reach 90%, indicating that the tutor could be distinguished between confusable words accurately.

B. *Correctness of Single Word*

Another 8 native Mandarin-speaking participants were recruited to rank each word with a score between “1” and “5”, with “1” for “Poor”, “2” for “General”, “3” for Good, “4” for “Better”, and “5” for “Outstanding”. The average score of each word is listed in Table 2. The average score of all 3D talking heads is higher than 4.0 and these results support that this 3D articulatory tutor could represent the words selected correctly.

IV. Pilot Experiments And Results

In order to investigate the efficiency and acceptability of the present system for HI children training, a pilot experiment was performed. One male and one female HI child (both were five years old) were recruited in this experiment from Shenzhen QingQing Speech Rehabilitation Centre. Both of them were deaf and were using hearing aid or cochlear implant. They also had dysarthria and language disorders. Both children had one year of experience of speech training but little experience of computer-aid speech training system. Yet, they showed interest to training with the VR system.

The primary training procedure was given for two weeks. They were trained for at least forty minutes per day, five days per week, and a total 10 training sessions. In each training session, three special training programs were performed for HI children, including five sets of listening training, ten sets of articulation training, and two sets of speaking training. Two speech therapists administered the sessions and helped the HI

TABLE I. IDENTIFICATION ACCURACY BETWEEN CONFUSABLE WORDS

Pair of Words	Identification Rate	Pair of Words	Identification Rate	Pair of Words	Identification Rate
/ɑː/ vs. /ɔː/	90%	/ləː/ vs. /nəː/	85%	/ʒiː/ vs. /tʃ iː/	85%
/əː/ vs. /iː/	100%	/gəː/ vs. /dəː/	80%	/ziː/ vs. /zhiː/	100%
/uː/ vs. /yuː/	90%	/xiː/ vs. /qiː/	80%	/ciː/ vs. /tʃ iː/	100%
				/siː/ vs. /ʃ iː/	90%

TABLE II. AVERAGE SCORE OF EACH SINGLE WORD

Word	Average Score	Word	Average Score	Word	Average Score	Word	Average Score
/ɑː/	4.1	/yuː/	4.4	/xiː/	4.4	/ciː/	4.0
/ɔː/	4.4	/gəː/	4.1	/qiː/	4.1	/siː/	3.1
/əː/	3.8	/dəː/	4.0	/jiː/	4.1	/dʒiː/	4.1
/iː/	3.8	/ləː/	4.0	/ʒiː/	3.8	/tʃ iː/	4.4
/uː/	4.0	/nəː/	4.8	/ziː/	4.0	/ʃ iː/	3.8

TABLE III. SCORE OF SPEECH INTELLIGIBILITY AND USER EXPERIENCE

	Score of SI		Most incorrect syllables		User experience		
	Pre-training	*Post-training*	*Pre-training*	*Post-training*	*Enjoyment*	*Motivation*	*Acceptable*
S1	64	77	/nəː/,/ciː/,/siː/,/ʃ iː/,/jiː/, /xiː/, /dʒiː/	/ji:/, /xiː/, /dʒiː/	4	5	4
S2	61	72	/tʃ iː/, /ʃ iː/, /ʒiː/, /jiː/, /siː/	/ʒiː/, /jiː/, /si:/	4	4	3

children perform these trainings. The system logged each subject's training session, including the time for each exercise and each pronunciation. Before and after the training, their speech performance was tested and re-tested using the speech intelligibility (SI) measures in [20] to evaluate the efficiency of present system for HI children rehabilitation. Furthermore, user experience evaluation was executed after 5 training sessions. The user experience evaluation included three criteria: enjoyment, motivation and acceptable. Children were asked to rate these items using a five-point scale.

Table 3 shows the results of pre-training, post-training SI measures and user experience evaluation. The speech intelligibility measures referred to the number of correct pronunciation out of 100 standard syllables, which consisted of bilabial, alveolar, velars, retroflex and front palatals consonants, and compound vowels. Results of SI measures showed that a clear improvement in speech intelligibility in both children after two weeks training, especially on bilabial, alveolar and retroflex consonant productions. However, improvement in front palatals was not seen. The results of user experience evaluation indicated that the HI children had strong enjoyment (mean = 4) and motivation (mean = 4.5) in daily training using the VR system, with medium acceptability (mean = 3.5) for the initial training procedure.

V. CONCLUSION

The primary goal of this study was to develop an effective Mandarin speech training system for hearing impaired children. In this system, a novel 3D articulatory tutor was developed based on continuous EMA data. Both the pronunciation accuracy of 3D articulatory tutor for each single word and the distinction between confusable words were then evaluated. The results showed that our articulatory model can well illustrate the trajectory of articulators during production of ambiguous and confusing words. Though the number of children attending the training using this system was relatively small, it is clear that the system helped enhanced their language skills after a period of training. Also, the interface and feedback of system were characterized by cartoons and therefore the system was favored and could easily be accepted by children.

However, there are limitations in the system at the present form. Existing articulatory tutors mainly focuses on the movements of lips, jaw, teeth and tongue. Aerodynamic spread which is important in some pronunciations is not being incorporated in the current model. Therefore, in future developments representation of airflow could be emphasized. Moreover, the training corpuses in this system are not sufficient for mass application. More 3D articulatory demos and corpuses should be included.

ACKNOWLEDGMENT

I hereby express my heartfelt gratitude for Miss Yu of QingQing Speech Rehabilitation Centre, who gave us a lot of meaningful suggestions about our system. Also, many thanks for our workmates Hongna Zheng, Yun Zhu, who spent plenty of time on the synthesis and improvement of 3D Articulatory Talking Head.

REFERENCES

[1] K. Resmi, S. Kumar, H. K. Sardana and R. Chhabra, "Graphical Speech Training system for hearing impaired," Proc. International Conference on Image Information Processing (ICIIP 2011), Nov. 2011, pp. 1-6.

[2] O. Engwall, O. Bälter, A.M. Öster and H. Kjellström, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," Behaviour & Information Technology, vol. 25, Feb. 2006, pp. 353-365.

[3] P. Badin, A. Ben Youssef, G. Bailly, F. Elisei and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," Actes de SLATE, 2010, pp. 1-10.

[4] G. Buccino, A. Solodkin, and S.L. Small, "Functions of the mirror neuron system: implications for neurorehabilitation," Cognitive and behavioral neurology, vol. 19, Mar. 2006, pp. 55-63.

[5] S.L. Small, G. Buccino, and A. Solodkin, "The mirror neuron system and treatment of stroke," Developmental Psychobiology, vol. 54, Apr. 2012, pp. 293-310.

[6] M. Franceschini et .al, "Mirror neurons: action observation treatment as a tool in stroke rehabilitation," European journal of physical and rehabilitation medicine, vol. 46, Sep. 2010, pp. 517-523.

[7] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," Proc. 4th Internation Conference on Universal Access in Human Computer Interaction, vol. 4554, Jul. 2007, pp. 786-794.

[8] A. Rathinavelu and G. Yuvaraj, "Data Visualization Model for Speech Articulators," Proc. AICERA, 2011, pp. 155-159.

[9] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," International Journal of Computer Vision, vol. 38, Jun. 2000, pp. 45-47.

[10] H. Javkin, N. Antonanzasbarroso, A. Das, D. Zerkle, Y. Yamada, N. Murata, H. Levitt, and K. Youdelman, "A Motivation-Sustaining Articulatory/Acoustic Speech Training System for Profoundly Deaf-Chlidren," Icassp-93: 1993 Ieee International Conference on Acoustics, Speech, and Signal Processing, vol. 1-5, 1993, pp. A145-A148.

[11] H. R. G. Javkin et al. "Synthesis-based speech training system," United States Patent 5340316, 1994.

[12] O. Bälter, O. Engwall, A.M. Öster, and H. Kjellström, "Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction," Proc. 7th international ACM SIGACCESS conference on Computers and accessibility, Oct. 2005, pp. 36-43.

[13] E. Rooney, F. Carraro, W. Dempsey, et .al. "HARP: an autonomous speech rehabilitation system for hearing-impaired people," Proc. Third International Conference on Spoken Language Processing, Sep. 1994, pp. 2019-2022.

[14] M.L. Hsiao, P.T. Li, P.Y. Lin, S.T. Tang, T.C. Lee, and S.T. Young. "A computer based software for hearing impaired children's speech training and learning between teacher and parents in Taiwan," Proc. 23rd Annual International Conference of the IEEE, vol. 2, 2001, pp. 1457-1459.

[15] Xibin Sun, Zhandong Guo, "Evaluation report on the rehabilitation of hearing and speech abilities of deaf children," Modern Rehabilitation, vol. 3, Oct. 1999, pp. 1288-1291.

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

[17] H. Chen, L. Wang, W. Liu, P.A. Heng, "Combined X-ray and facial videos for phoneme-level articulator dynamics." The Visual Computer, vol. 26, Apr. 2010, pp. 477-486.

[18] M. Cohen, D.W. Massaro, "Modeling articulation in synthetic visual speech," Models Technique in Computer Animation, no. 92, 1993, pp. 139-156.

[19] L. Moccozet, and N.M. Thalmann, "Dirichlet free-form deformations and their application to hand simulation," Proc. Computer Animation, Jun. 1997, pp. 93-102.

[20] Xibin Sun, Evaluation criteria and methods of hearing impaired children. Beijing: Sanchen video press, 2009.