Evaluation of External and Internal Articulator Dynamics for Pronunciation Learning

Lan Wang*, Hui Chen and JianJun Ouyang

CAS/CUHK ShenZhen Institute of Advanced Integration Technologies, Chinese Academy of Sciences

lan.wang,hui.chen@siat.ac.cn, jj.ouyang@sub.siat.ac.cn

ABSTRACT

In this paper we present a data-driven 3D talking head system using facial video and a X-ray film database for speech research. In order to construct a database recording the three dimensional positions of articulators at phoneme-level, the feature points of articulators were defined and labeled in facial and X-ray images for each English phoneme. Dynamic displacement based deformations were used in three modes to simulate the motions of both external and internal articulators. For continuous speech, the articulatory movements of each phoneme within an utterance were concatenated. A blending function was also employed to smooth the concatenation. In audio-visual test, a set of minimal pairs were used as the stimuli to access the realistic degree of articulatory motions of the 3D talking head. In the experiments where the subjects are native speakers and professional English teachers, a word identification accuracy of 91.1% among 156 tests was obtained.

Index Terms— pronunciation learning, external and internal articulator dynamics, X-ray film

1. INTRODUCTION

Audio-visual speech synthesizer and recognizer have been developed to animate lip movements to improve the intelligibility of synthesized speech and lip reading. More recent studies of speech visualization focused on whether visual information can help in speech perception [8, 5]. In these studies, the articulatory models were established to form a virtual talking head, aiming at enhancing the hearing or language impaired users. The positive feedbacks from subjects [8] encourage us to develop more realistic virtual talking head and apply it to foreign language pronunciation learning. In addition to audio stream, visual information can further help pronunciation learning. Students often observe teachers’ facial and vowel tract movements to understand the manner of pronunciation. In this study, a virtual language tutor using 3D articulatory movement models was developed, in which both the external and internal articulatory movements were simulated with audio stream. Using this combined form of information, the learners can easily distinguish the pronunciations of English phonemes through audio-visual information, and therefore improve their pronunciations.

The related works in [6, 7, 5] have shown the visualization of internal articulator dynamics on synthesized audiovisual speech. The Electro-Magnetic Articulography (EMA) device was commonly used to record the movements of articulators at some discrete points [6, 8]. However, an EMA device is costly and the calibration of the device is complex. The video-fluoroscopic images [2] were also used to record the movements of tongue body, lips, teeth, palate and oral cavity in real-time. Because speakers may be exposed to excessive amounts of X-ray radiation during the long-term recording, fewer resources were used for 3D talking head. However this approach was in principle the most appropriate method to determine the actual tongue shapes when speaking [8]. For accurate automatic visible speech synthesis of facial appearance, the motion capture system [3] was used to record the complex lip movements. The data was mapped to a generic 3D face model. To animate the movements of viseme [6, 7, 8, 5], a set of displacement vectors or parameters were defined for the corresponding vertices of pre- and target viseme positions in the 3D synthetic head models. The visualization of articulatory movements were acquired by linear combinations of these displacement vectors. The evaluation of the audio-visual perception test was conducted in [8, 5], the subjects gave identification scores for the stimuli under different conditions, including audio information only, audio-visual stream combined with tongue and audio-visual information combined with both tongue and face.

In this paper, the articulatory movements of a generic 3D talking head were established. The visual materials are video and videofluoroscopy recording the speech of the native English speakers. Automatic speech recognition (ASR) was used to align the utterance to a sequence of phonemes with phoneme durations. For a phoneme in English inventory, the learners can easily distinguish the pronunciations of English phonemes through audio-visual information, and therefore improve their pronunciations.

The related works in [6, 7, 5] have shown the visualization of internal articulator dynamics on synthesized audiovisual speech. The Electro-Magnetic Articulography (EMA) device was commonly used to record the movements of articulators at some discrete points [6, 8]. However, an EMA device is costly and the calibration of the device is complex. The video-fluoroscopic images [2] were also used to record the movements of tongue body, lips, teeth, palate and oral cavity in real-time. Because speakers may be exposed to excessive amounts of X-ray radiation during the long-term recording, fewer resources were used for 3D talking head. However this approach was in principle the most appropriate method to determine the actual tongue shapes when speaking [8]. For accurate automatic visible speech synthesis of facial appearance, the motion capture system [3] was used to record the complex lip movements. The data was mapped to a generic 3D face model. To animate the movements of viseme [6, 7, 8, 5], a set of displacement vectors or parameters were defined for the corresponding vertices of pre- and target viseme positions in the 3D synthetic head models. The visualization of articulatory movements were acquired by linear combinations of these displacement vectors. The evaluation of the audio-visual perception test was conducted in [8, 5], the subjects gave identification scores for the stimuli under different conditions, including audio information only, audio-visual stream combined with tongue and audio-visual information combined with both tongue and face.

In this paper, the articulatory movements of a generic 3D talking head were established. The visual materials are video and videofluoroscopy recording the speech of the native English speakers. Automatic speech recognition (ASR) was used to align the utterance to a sequence of phonemes with phoneme durations. For a phoneme in English inventory, the corresponding key images were chosen from videos, which have shown the significant differences of mouth shapes or tongue positions among phonemes. Three dimensional po-
sition data of feature points on lips, jaw, tongue, teeth were defined and calculated, then mapped to a generic 3D head model. A data-driven 3D talking head simulated the co-articulation, by concatenating the phoneme-level articulatory movements in accordance with the phoneme sequence of continuous speech. The blending function was derived from the position data of the articulatory feature points to smooth the motions. The audio-visual tests were performed to access whether the external and internal articulator dynamics can illustrate the differences between confusable English phonemes, and help the language learners to improve their pronunciations. A set of minimal pairs were used as the stimuli, and two groups of subjects were asked to listen and then watch the audio-visual representations. In performance evaluation the subjects wrote down the words that the 3D talking head uttered and scored the realistic degrees of the animations.

The rest of this paper is organized as follows: Section 2 introduces the collection and data processing of video and X-ray film. Section 3 describes articulatory deformation of a 3D head model at the phoneme level. In Section 4, the blending function is derived and used with phoneme-level articulator dynamics to form the data-driven 3D talking head. Section 5 presents the experimental details and results. The conclusions and discussions are in Section 6.

2. THE DATABASE RECORDING THE ARTICULATORY MOVEMENTS

To record both external and internal articulation motions, two types of video streams were collected and processed. A native American speaker was invited to speak isolated phonemes, words, and continuous sentences respectively, facial articulatory movements were recorded synchronously by video cameras from frontal and profile views. Internal articulatory motions were segmented from the existed X-ray film. To define the corresponding frames at phoneme level, an HMM-based speech recognition system was used for the audio of all video streams. Using American English as the target language, the acoustic models were built with TIMIT corpus recorded from native American English speakers. The forced-alignment was performed to generate the phoneme-level sequence with phoneme boundary for each utterance in the video, and the image frames of phonemes were defined accordingly.

The phoneme-viseme mapping used in previous studies can’t represent both face and tongue movements, since the phonemes in a viseme class may have the similar mouth shapes but different tongue positions, like /th/ versus /s/. In the paper, both external and internal articulatory motions were tracked at phoneme-level. The peak frames of facial and internal images were selected synchronously to represent the most significant articulatory movements for each phoneme. Tongue segmentation in the X-ray images were manually processed by a medical graduate. The feature points were defined and labeled in the reference images as shown in Figure 1. Selected feature points of facial appearance on frontal and profile views of the speaker include eight points on the lips, nose tip, and chin. The internal articulators feature points include seven points on the tongue, two points on the upper and lower teeth respectively. Accordingly, the feature positions on the 3D head model were evaluated and adjusted via the registration of all above images. The 2-D trajectories at two views were used to recover the 3D positions of face. All the facial and tongue feature positions were then mapped to a generic 3D head model. Since the motions of articulation were determined by the images in the video streams, the calibration algorithm was developed to remove the head movements. A database recording the 3D positions of articulation at phoneme-level was then constructed, where the relaxed state of articulators was defined as the starting point of the articulatory movements.

3. THE PHONEME-LEVEL ARTICULATORY DEFORMATIONS

To achieve natural-looking articulator dynamics and improve the phoneme-level intelligibility, a physiological head model was established based on the templates of MRI slices. Overall the whole three dimensional head model is made of 28656 triangles, including articulators of upper and lower lips, jaw, upper and lower teeth, palate, oral cavity and tongue [9]. Simulating the articulator dynamics is a complex process, requiring the synchronization of all articulators. The phoneme-level articulator deformation is based on the above database recording the movements of articulators. For the feature positions of external or internal articulators, three dimensional displacements were obtained by calculating the differences between the feature positions of the key image and that of the relaxed state image. In this study, the phoneme-level deformation was performed with the feature positions of articulators, rather than the use of the motion parameters, such as lip width and height, jaw height, tongue tip height [6]. Three modes of articulator deformation were defined according to the physiological properties of articulators.

The lips and tongue are both muscular tissues, and move under local deformation. The feature points on both the lips and tongue then moved under constrained displacements for...
each phoneme, transferring deformations were then applied to the adjacent points on facial skin or tongue. The animation process of jaw, chin skin and linked lower teeth is mainly controlled in six degrees of freedom. The vertices related to the skull were defined with three orientation angles of yaw, pitch, roll and three positions of the longitudinal, vertical and lateral axis. The motion mode of the skull linked upper teeth and facial skin was set as fixed constraint. The movements of the fixed part can be used to simulate the head motions when speaking [9].

4. THE BLENDING FUNCTION FOR CONTINUOUS SPEECH

The data-driven 3D talking head was based on the phoneme-level articulator dynamics to simulate the articulatory movements of continuous speech. The ASR system was used for phoneme segmentation, and the phoneme sequence with duration information was generated by running forced-alignment on each utterance. The phoneme-level deformation was implemented within the phoneme duration using the displacement-based articulator dynamics method described in Section 3. All deformations of the phonemes in a sequence were then concatenated to form the articulatory movements of continuous speech. Hereby, a blending function was derived to smooth the articulatory motions of precedent and succeed phonemes.

The dominance function introduced in [1] was used with modification based on the database recording the feature positions of articulators.

\[
D_{ip}(\tau) = \begin{cases} 
\alpha_{ip} \cdot e^{-\theta_d|\tau|}, & \text{if } \tau \leq 0 \\
\alpha_{ip} \cdot e^{-\theta_g|\tau|}, & \text{if } \tau > 0 
\end{cases} \tag{1}
\]

Here \( ip \) is the parameter \( p \) of \( i \)th phoneme in the sequence, \( \alpha_{ip} \) is given by

\[
\alpha_{ip} = \left\{ \begin{array}{ll}
\frac{\text{Max}_i |R_{ip}|}{\text{Max}_i |R_{ip}|} & \text{if } |R_{ip}| \neq \text{Max}_i |R_{ip}| \\
1.0 & \text{if } |R_{ip}| = \text{Max}_i |R_{ip}|
\end{array} \right. \tag{2}
\]

where \( R_{ip} \) is the peak position of the \( i \)th phoneme in a sequence. The \( \theta_g \) and \( \theta_d \) in Eq. (1) are growth and decay constants of exponential function. \( \tau \) is given by \( \tau = t_{e_i} - t \) with \( t_{e_i} \) is the peak time of articulatory motion of the \( i \)th phoneme. The exponential growth and decay constants can be obtained by the following derivation,

\[
\begin{align*}
\alpha_{ip} \cdot e^{-\theta_d(t_{e_i} - t_i)} &= \epsilon \\
\alpha_{ip} \cdot e^{-\theta_g(t_{e_i} - t_i)} &= \epsilon
\end{align*}
\]

where \( \epsilon \) is a given a value 0.05 in the experiments, \( t_{e_i} \) and \( t_i \) are the starting/ending time of the \( i \)th phoneme. Given the dominance function, the deformation parameter curve was expressed by

\[
F_{Np}(t) = \frac{\sum_{i=1}^{N} R_{ip} \cdot D_{ip}(t - t_i)}{\sum_{i=1}^{N} D_{ip}(t - t_i)} \tag{3}
\]

The dominance function derived using Eq. (1-2) and the deformation parameter curve were illustrated for the lower lip point of the word “hat” with three phonemes /\h/ ae /t/.

5. EXPERIMENTS

For Chinese learners of English, some confusable phonemes are commonly mispronounced. Hence, a set of minimal pairs were chosen to evaluate the 3D talking head. A total of 12 words contain the phonemes missed from Chinese inventory /æl/ and /\h/, the confusable phonemes of labiodental fricative /v/ and bilabial glide /w/, lingua-dental fricative /th/ and lingua-alveolar fricative /s/, etc. The audio-visual perception test attempts to access whether the animations of the 3D talking head are correct and realistic.

The participants in performance evaluation were divided into two groups. The first group contained three native speakers and ten English teachers in China, and the second were university students who had little knowledge on phonetics. In the test, the audio streams of one minimal pair were played firstly, and then the animations were shown in which two words of one minimal pair appeared in a random order. The subjects were asked to identify which animation corresponded to the word. All the stimuli were presented in two conditions, the visualization with front face (F), the visualization with transparent face and tongue (FT). The subjects also scored the realistic degree of the animation, where 1 is for bad, 2 for poor, 3 is fair, 4 for good and 5 results in excellence.

In the first group, the word identification accuracy is 91.1% among 156 tests. Most of subjects can clearly point out the differences of the 3D articulator dynamics between the confusable phonemes. The realistic scores of animation of each word were averaged and listed in Table 1. Most of words have the realistic scores over 3, however, the 3D animations of “vine” and “wine” are not realistic enough to
For the articulations of “thing” and “sing”, the 3D talking head presents with facial and profile (transparent) views.

<table>
<thead>
<tr>
<th>Confusable Phonemes</th>
<th>Minimal Pairs</th>
<th>word</th>
<th>score</th>
<th>word</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ae/ vs. /eh/</td>
<td>hat</td>
<td>3.1</td>
<td></td>
<td>het</td>
<td>3.0</td>
</tr>
<tr>
<td>/aw/ vs. /ao/</td>
<td>house</td>
<td>4.0</td>
<td></td>
<td>horse</td>
<td>3.8</td>
</tr>
<tr>
<td>/v/ vs. /w/</td>
<td>vine</td>
<td>3.0</td>
<td></td>
<td>wine</td>
<td>2.9</td>
</tr>
<tr>
<td>/th/ vs. /s/</td>
<td>thing</td>
<td>3.4</td>
<td></td>
<td>sing</td>
<td>3.1</td>
</tr>
<tr>
<td>/dh/ vs. /l/</td>
<td>they</td>
<td>3.4</td>
<td></td>
<td>fay</td>
<td>3.8</td>
</tr>
<tr>
<td>/n/ vs. /m/</td>
<td>night</td>
<td>3.3</td>
<td></td>
<td>might</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 1. The realistic scores of minimal pairs

Fig. 3. For the articulations of “thing” and “sing”, the 3D talking head presents with facial and profile (transparent) views.

This study investigated a data-driven 3D talking head system based on the data from the facial videos and X-ray films. We aim at illustrating the phoneme-level differences of English pronunciation through realistic articulatory motions of the 3D models. A database for such purpose was built up, recording three dimensional feature positions calculated based on the key frames of each English phoneme. The displacement-based deformations for both external and internal articulators were simulated. A blending function was developed to smooth the concatenation of the phoneme-level articulatory motions for continuous speech. In audio-visual test, six minimal pairs were used to cover the confusable phonemes commonly mispronounced by Chinese learners of English. The word identification accuracy and the realistic scores from the native speakers and English teachers concluded that over 90% of the articulatory motions can illustrate the differences between the confusable phonemes.

EMA data will be collected in future research for contrast experiments to possibly improve the intelligibility of the 3D talking head system. More audio-visual tests would be conducted for the subjects from the middle schools. The study progress assisted with audio-visual information will be tracked for weeks. It is interesting to access whether a 3D talking head can help students to understand and improve their English pronunciations.

6. CONCLUSIONS

This study investigated a data-driven 3D talking head system based on the data from the facial videos and X-ray films. We aim at illustrating the phoneme-level differences of English pronunciation through realistic articulatory motions of the 3D models. A database for such purpose was built up, recording three dimensional feature positions calculated based on the key frames of each English phoneme. The displacement-based deformations for both external and internal articulators were simulated. A blending function was developed to smooth the concatenation of the phoneme-level articulatory motions for continuous speech. In audio-visual test, six minimal pairs were used to cover the confusable phonemes commonly mispronounced by Chinese learners of English. The word identification accuracy and the realistic scores from the native speakers and English teachers concluded that over 90% of the articulatory motions can illustrate the differences between the confusable phonemes.

7. REFERENCES