Realistic Visual Speech Synthesis Based on Hybrid Concatenation Method

Jianhua Tao, Member, IEEE, Le Xin, and Panrong Yin

Abstract—This paper presents a realistic visual speech synthesis based on the hybrid concatenation method. Unlike previous methods based on phoneme level unit selection or hidden Markov model (HMM), etc., the hybrid concatenation method uses a frame level-based unit selection method combined with a fused HMM, and is able to generate more expressive and stable facial animations. The fused HMM can be used to explicitly model the loose synchronization of tightly coupled streams, with much better results than a normal HMM for audiovisual mapping. After fused HMM is created, facial animation is generated via the unit selection method at the frame level by using the fused HMM output probabilities. To accelerate the computing efficiency of the unit selection on a large corpus, this paper also proposes a two-layer Viterbi search method in which only the subsets that have been selected in the first layer are further checked in the second layer. Using this idea, the system has been successfully integrated into real-time applications. Furthermore, the paper also proposes a mapping method to generate emotional facial expressions from neutral facial expressions based on Gaussian mixture models (GMMs). Final experiments prove that the method described can neutral facial expressions based on Gaussian mixture models (GMMs). Final experiments prove that the method described can

Index Terms—Fused hidden Markov model (HMM), inversion, speech-driven facial animation, unit concatenation, visual speech synthesis.

I. INTRODUCTION

W

With the intensive requirements of human–computer interaction (HCI), realistic visual speech synthesis has become a very popular subject in both research and flourishing industry domains. It aims to construct a sequence of corresponding face movements given an incoming audio stream input. However, there are some basic problems with this process, including the expressiveness of facial animation and computing complexity. Although many techniques have been tried to solve these problems, most of them simply focus on lip movements; they include the rule-based method [10], [15], [17], [18], vector quantization (VQ) method [11], [19], neural network (NN) method [9], [12], multilayer preceptrons (MLPs) [12], K nearest neighbor (KNN) [17], Gaussian mixture model (GMM) [27], and hidden Markov model (HMM) [11], [20], [23].

The rule-based method [10], [18] and VQ method [11], [19] are two direct and easily realized methods. However, their results are usually inaccurate and discontinuous due to limited rules and codebooks. The GMM method is normally used for learning the relationship between audio and visual, and predicting visual parameters such as mapping functions from audio input [9]. Although the GMM method has the merits of moderate sample amounts and smooth synthesized results, it gets easily bogged down into problems of over smoothing for direct visual parameters prediction.

While HMM has been widely used in speech recognition, it has also been used to build a phoneme recognition model, which directly maps recognized phonemes to lip shapes using a static viseme model [11]. Bregler [7] extended this work by reordering existing visual frames based on recognized phonemes. All these methods require the use of a speech recognition engine which not only limits the system quality by recognition error rates, but also limits the system in real time applications due to very low speed.

To increase the expressiveness and speed of such methods, some researchers have applied the unit selection method to reorder or concatenate recorded audiovisual units to form new visual sequences. For instance, Cosatto [6] and Tao [2] present a method that selects corresponding visual frames according to the distance between new audio tracks and stored audio tracks, and then concatenates the candidates to form the smoothest sequence. Although this method gives a relatively faster speed, it simplifies audiovisual mapping to a process of one-to-one mapping. Thus, it is hard to obtain stable facial animation output, although some smoothing or morphing methods might be used.

Although some work has been done on facial animation from speech, the naturalness and expressiveness of the synthesized results are still open questions. In this paper, we present a real time realistic audiovisual speech synthesis system based on a hybrid concatenation method. Similar to the traditional unit selection method, the hybrid concatenation method uses existing facial deformation information in the training corpus to produce a natural realization of visual speech synthesis. However, unlike traditional methods, which are based on phoneme or frame units, our method uses subsequences as the basic unit for concatenation.

For the selection of the best subsequences for synthesis, we define the target costs of all candidates by a fused HMM, which
can be used to model the loose synchronization nature of two tightly coupled audio and visual streams. In fused HMM, the joint probabilistic distribution of the novel audio input and the existing visual deformation is a better valuation of the target cost formulated from a probabilistic viewpoint. Further, the audio and facial parameter error between two selected subsequences is a probabilistic valuation of the concatenation cost.

After the target costs are obtained by fused HMM, the optimal subsequence selection can be performed with a Viterbi search algorithm. However, compared with phoneme-based unit selection, the amount of subsequences will be tens or hundreds times greater. To obtain real-time visual speech synthesis on a large audiovisual bimodal corpus, we propose a two-layer framework. Only the subsets that were selected in the first layer will be further checked for existing subsequences in the second layer.

Finally, this paper proposes a GMM-based mapping method to investigate the correlation between neutral facial expressions and emotional facial expressions. Due to the uncertainty between facial expressiveness and audio inputs, the method keeps both the expressiveness and stability of the emotion outputs at the frame level. With this model, the emotional facial animation is generated by a combination of the GMM and audiovisual unit selection methods.

The rest of this paper is organized as follows. Section II introduces the audiovisual corpus used for our work. The audio and visual parameters are also defined. In Section III, fused HMM is introduced to model the loose synchronization nature of audio and visual streams explicitly, and fused HMM is integrated into the unit selection framework. The concatenation cost is also defined. The corpus is also classified into a series of subclasses by considering both visual configurations and audio observations. The two layer Viterbi search framework is then introduced to efficiently obtain optimal subsequence selection. Section IV proposes a method to generate emotional facial expressions by mapping the neutral facial parameters to emotional parameters. In Section V, the validity and effectiveness of the proposed method are proved via extensive experiments and discussion. Finally, we conclude our work and discuss future work in Section VI.

II. AUDIOVISUAL CORPUS

Herein, we have designed two kinds of corpora for our research. One includes one-hour audiovisual data of 700 phoneme-balanced utterances with a neutral speaking style. All Mandarin di-phones are covered in this corpora. The other was designed for emotional facial animation. It includes 300 utterances, and each utterance was repeated with five different emotion states, i.e., neutralness, happiness, sadness, anger, and surprise, for a total of 1500 expressive utterances.

To obtain powerful facial movement representation, 50 markers compatible with FDPs in MPEG-4 were selected to encode the face shape, as shown in Fig. 1. These markers were acquired by a commercial motion capture system using eight digital cameras (1024 × 768 pixels) and 75 frames per second (FPS). The 3-D trajectories of the facial animation markers and their accompanying time-aligned audio samples were recorded simultaneously in real time.

<table>
<thead>
<tr>
<th>FAP#</th>
<th>FAP name</th>
<th>FAP#</th>
<th>FAP name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>open_jaw</td>
<td>38</td>
<td>squeeze_r_eyebrow</td>
</tr>
<tr>
<td>14</td>
<td>thrust_jaw</td>
<td>39</td>
<td>puf_f_1_cheek</td>
</tr>
<tr>
<td>15</td>
<td>shift_jaw</td>
<td>40</td>
<td>puf_f_cheek</td>
</tr>
<tr>
<td>16</td>
<td>push_b_lip</td>
<td>41</td>
<td>lift_1_cheek</td>
</tr>
<tr>
<td>17</td>
<td>push_t_lip</td>
<td>42</td>
<td>lift_r_cheek</td>
</tr>
<tr>
<td>19</td>
<td>close_t_l_eyelid</td>
<td>51</td>
<td>lower_t_midlip_o</td>
</tr>
<tr>
<td>20</td>
<td>close_t_r_eyelid</td>
<td>52</td>
<td>raise_b_midlip_o</td>
</tr>
<tr>
<td>21</td>
<td>close_b_l_eyelid</td>
<td>53</td>
<td>stretch_t_cornerlip_o</td>
</tr>
<tr>
<td>22</td>
<td>close_b_r_eyelid</td>
<td>54</td>
<td>stretch_r_cornerlip_o</td>
</tr>
<tr>
<td>31</td>
<td>raise_l_i_eyebrow</td>
<td>55</td>
<td>lower_t_lip_o</td>
</tr>
<tr>
<td>32</td>
<td>raise_r_i_eyebrow</td>
<td>56</td>
<td>lower_t_lip_o</td>
</tr>
<tr>
<td>33</td>
<td>raise_l_m_eyebrow</td>
<td>57</td>
<td>raise_b_lip_o</td>
</tr>
<tr>
<td>34</td>
<td>raise_r_m_eyebrow</td>
<td>58</td>
<td>raise_b_lip_o</td>
</tr>
<tr>
<td>35</td>
<td>raise_l_o_eyebrow</td>
<td>59</td>
<td>raise_l_cornerlip_o</td>
</tr>
<tr>
<td>36</td>
<td>raise_r_o_eyebrow</td>
<td>60</td>
<td>raise_r_cornerlip_o</td>
</tr>
<tr>
<td>37</td>
<td>squeeze_t_eyebrow</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All audio data was recorded by both a close-talk microphone (AKG C420) and a large membrane microphone (RODE K2). The audio data from the large membrane microphone was used for the speech analysis, while the data from the close-talk microphone was used for the time alignment between the audio and visual channels. A laryngograph was also used to support the detection of pitch pulses.

In this paper, we used MPEG-4 standard facial animation parameters (FAPs) for the visual parameters. To reduce the noise for audiovisual mapping, we removed all frames related to eye blinks. The parameters related to the tongue, nose, head movement, ears, eyeballs, pupils, and inner lips, which are not captured by the motion capture system, were not used. Then, after coordinate normalization and affine transformation, 31 FAPs (see Table I) related to lip and facial deformation were extracted.

For the audio parameters, both vocal parameters and prosody features were included. Vocal parameters form a major factor for lip and jaw movement, while prosody parameters may enhance the facial expressiveness of the whole facial area [3]. Although many acoustic parameters, such as the linear prediction coefficient (LPC), formant, and linear spectrum pair (LSP), have been used to denote vocal features, the Mel-frequency cepstral coefficient (MFCC) has been proven as the most efficient parameter...
in speech recognition and perception. With this knowledge, we chose to use it in our research. As for prosody features, we could only use logarithmic F0s and logarithmic energies because our research is based on the frame-unit level.

Then, for each audiovisual utterance, we can denote the visual parameters by a sequence of facial parameter vectors, shown as
\[ (v_1, v_2, \ldots, v_{t-1}, v_t, \ldots) \]  
and describe the audio parameters by
\[ (a_1, a_2, \ldots, a_{t-1}, a_t, \ldots). \]

III. AUDIOVISUAL UNIT SELECTION BASED ON FUSED HMM

A. Basic Unit Sets for Audiovisual Unit Selection

Unlike the phoneme-based unit selection system, our work is based on the frame level. As context-dependent acoustic features have been intensively applied in speech recognition and synthesis for many years, with subsequent significant increases in recognition and synthesis quality, we also try to integrate context information by associating each visual frame with the corresponding acoustic features from past \( N \) and future \( N \) frames. In our real application, \( N = 7 \) was considered for a 150-ms audio block. This forms a kind of tradeoff between a stable description of coarticulation and the requirement of real-time implementation. As shown in Fig. 2, by means of a tapped-delay line incorporating coarticulation into the audiovisual mapping, training examples are formed by a series of subsequences.

In Fig. 2, \( A_k \) and \( V_k \) denote the input audio and visual sequences with \( N \) frames, respectively,
\[ A_k = (a_{t-7}, a_{t-6}, \ldots, a_t, \ldots, a_{t+6}, a_{t+7}) \]  
\[ V_k = (v_{t-7}, v_{t-6}, \ldots, v_t, \ldots, v_{t+6}, v_{t+7}) \]  

B. Fused HMM-Based Target Cost

With the audiovisual data being aligned by subsequences, audiovisual mapping becomes a many-to-many mapping issue. In [4], Pan et al. proposed a fused HMM that had successful performance for object identification in audiovisual data [4], [5]. The model constructs a new structure linking the two component HMMs, which is optimal according to the maximum entropy principle and the maximum mutual information (MMI) criterion. Similar to the coupled HMM method, fused HMM consists of two separated component HMMs. However, unlike the Coupled HMM’s graphical structure in which the connections between the two component HMMs are completed by subsequent hidden states, the connections in the fused HMM’s graphical structure are between the hidden state node and the observation node of different HMMs. This has the advantage of reaching a better balance between model complexity and performance than other existing model fusion methods, such as the coupled HMM [8], [24] and mixed-memory HMM [21] methods.

Fused HMM has been proven as a good model in obtaining the possibility of training pairs being fused; however, it has only been used for object identification thus far. To use it for audiovisual mapping with the unit selection method, we classify all visual parameters into several visual clusters by the k-means method and choose a four-state right-left HMM model for each cluster. The visual cluster represents the continuous deformation of the face shape. Based on the time synchronization between audio and visual representation in the audiovisual corpus, the sequences for each clustered visual feature also have corresponding audio frames. Then, for each cluster sequence, we also train a three-state right-left HMM model for the audio data with MFCCs. The best hidden state sequences of the audio component HMMs are found using the Viterbi algorithm, while a GMM is fitted on the visual frame data for each estimated hidden state. This procedure is described as follows.

Given the observed audiovisual parameters \( A_k \), \( V_k \), and their corresponding HMM, the fused HMM was proposed to construct a structure linking the component HMMs together by giving optimal estimation of the joint probability \( p(A_k, V_k) \). Taking advantage of the fact that the data from a single sensor can be individually modeled by an HMM, and according to the maximum entropy principle and the MMI criterion, the fusion model yields the following two structures, as shown by [4]
\[ p^{(1)}(A_k; V_k) = p(A_k)p(V_k|\tilde{U}^a) \]  
\[ p^{(2)}(V_k; A_k) = p(V_k)p(A_k|\tilde{U}^v) \]  
where the most possible hidden state sequences \( \tilde{U}^a \) and \( \tilde{U}^v \) are estimated by the Viterbi algorithm.

The two structures shown in (5) and (6) are totally different. In (5), \( \tilde{U}^a \) is asked to be reliably estimated, while in (6) \( \tilde{U}^v \) has to be exactly determined. Previous studies [4], [5] have proven that the first structure will generate more stable results in bi-modal object recognition because the hidden states of the speech HMM can be estimated more reliably. It is intuitive to use the
same experience for audiovisual mapping, and thus in our study structure (5) is preferred.

Then, the training process will include the following three steps.

1) Two individual HMMs consisting of a visual component HMM and an audio component HMM are trained independently by the EM algorithm [34].

2) The best hidden state sequences of the audio component HMMs are found by using the Viterbi algorithm.

3) The coupling parameters are determined.

Then, each trained fused HMM corresponding to a subsequence cluster will contain an audio HMM and the audio HMM state related coupling parameters. The coupling parameter represents the conditional probability distribution of visual observation in the visual component HMM by the given states of the audio component HMM. Unlike the coupling parameter that was based on discrete observations in [4], we extended it to continuous observation by the following mixture Gaussian function

\[
b_j(V_t) = \sum_{k=1}^{K} c_{jk} N(V_t | \mu_{jk}, \Sigma_{jk}), \quad 1 \leq j \leq N.
\]  

where the mixture Gaussian \(b_j(V_t)\) is the visual observation in the audio state \(j\) of the audio component HMM, and \(K\) is the amount of Gaussians. The parameters are estimated with the expectation–maximization (EM) algorithm [34].

The framework for using fused HMM as the target cost for audiovisual unit selection is then shown in Fig. 3.

For each audio input, we first use audio HMM of each cluster and the Viterbi algorithm to get the best HMM state alignment. Then, we use these aligned audio HMM states and the visual cluster centers to get the conditional probabilities by (7). From the conditional probabilities, we will find the best visual cluster for the input audio. We further use aligned audio HMM states corresponding to the selected best visual cluster and formula (7) to get the new conditional probabilities for all visual subsequences within the selected cluster. These new conditional probabilities are considered as target costs for visual subsequences by giving audio input. We then concatenate selected visual subsequences together to compose new facial animation parameters.

**C. Fused HMMs in Two-Layer Framework**

If we do not control the amount of clusters, we will have a very large number of audiovisual candidates when compared to phoneme-based units. To reduce the computing complexity, therefore, we use a two-layer framework by classifying the corpus into a series of subsets by considering both visual and audio configurations. This two-layer framework is performed by the following steps.

In the first layer, we only classify all audiovisual subsequences into 20 clusters. Each cluster center represents the repertoire of facial specification. Furthermore, each cluster is classified into subclusters by the k-means method. These subclusters constitute the second layer. Then, we can train more fused HMMs for subclusters below the representative fused HMM.

In unit selection procedure, we use fused HMMs of the first layer to select the best cluster. Then, all fused HMMs of the second layer within the selected cluster will be further checked to find the best subcluster. The target costs will only be calculated for the visual subsequences within this selected subcluster.

**D. Concatenation Cost**

With the unit selection based on the target cost, the facial animation can be further optimized by the concatenation cost, which can be calculated by the facial parameter errors between two selected sequences. To enhance the smoothing, the delta and delta-delta parameters are also used. The concatenation cost of the facial parameters is then calculated by

\[C^c_{t,n} = u_{t,1}^v S(v_{t-1}, v_t) + w_{t,1}^v S(v_{t-1}, v'_t) + w_{t,2}^v S(v''_{t-1}, v''_t).\]  

Here, \(S(v_{t-1}, v_t)\), \(S(v'_{t-1}, v'_t)\) and \(S(v''_{t-1}, v''_t)\) denote the Euclidean distance of facial parameters, their delta parameters, and their delta-delta parameters, respectively, between visual frame \(t-1\) and \(t\). \(t\) and \(t-1\) represent the two frames between two selected sequences. \(u_{t,1}^v\), \(w_{t,1}^v\), and \(w_{t,2}^v\) are the weights of these distances. Normally, they are all just set to 1.

From another point of view that continuous visual data always accompanied by continuous audio, it is believed that the result of unit selection will be very stable if the selected audiovisual data comes from the same utterance with continuous audio alignment. Thus, it is also very useful to use the concatenation errors of the audio frames between two selected sequences, which tries to limit the selected units in the same utterance as much as possible. The audio concatenation cost is denoted by

\[C^a_{t,n} = u_{t,1}^a S(a_{t-1}, a_t) + w_{t,1}^a S(a'_{t-1}, a'_t) + w_{t,2}^a S(a''_{t-1}, a''_t)\]  

where \(S(a_{t-1}, a_t)\), \(S(a'_{t-1}, a'_t)\), and \(S(a''_{t-1}, a''_t)\) denote the Euclidean distance of the audio parameters, their delta parameters, and their delta-delta parameters, respectively, between audio frame \(t-1\) and \(t\). \(u_{t,1}^a\), \(w_{t,1}^a\), and \(w_{t,2}^a\) are the weights of these distances. Just as the weights in the visual distances calculation, these weights are also set to 1 here.

With these target and concatenation costs, the whole system framework can be described as shown in Fig. 4, where the \(U_{t,n}\)
denotes one candidate of a visual subsequence selected from the corpora at time $t$.

The final output of the audiovisual sequences will be generated with a Viterbi search among all audiovisual candidates.

IV. EMOTION GENERATION BY THE MAPPING METHOD

To investigate emotional facial animation generation, we built an additional audiovisual corpus containing four emotion states (happiness, sadness, anger, and surprise) for each utterance and which shares the same size and same text as the neutral utterances mentioned in Section II.

In the emotional audiovisual data, we found most of the facial expressions are different to those in the neutral state. Even in the mouth area, lip movements were also affected by the emotion states. However, emotional facial animation is not simply the combination of static emotion facial expressions and speech related neutral facial actions. Fig. 5 shows the trajectory examples of FAP#33 (vertical displacement of left middle eyebrow), FAP#51 (horizontal displacement of right outer-lip corner) and FAP#60 (vertical displacement of right outer-lip corner) between the neutral state and the emotional states.

With this analysis result of uncertainty between facial expressiveness and the audio, any results based solely on the unit selection method will be very noisy. In particular, because our unit selection method is based on frame levels, to get more smoothed results, we use a neutral-to-emotion mapping method to generate emotional facial animation based on the neutral audiovisual unit selection results. Since the styles of expressive representation vary among people, it is not reasonable to simply classify the emotion expressions into a certain series of patterns; therefore, we use GMMs to model the probability distribution of the neutral-emotion deformation vectors. Each emotion category corresponds to one GMM.

In the training set, these neutral vector sequences and emotional vectors are aligned with dynamic time warping (DTW) [28]. Then, we concatenate the neutral facial deformation features with the emotional facial deformation features to compose the joint feature vector: $Z_{ik} = [X_k, Y_{ik}]^T$, where $i \in [0, 1, 2, 3, 4, 5]$, means different emotions, $k = 1, \ldots, N$, denotes the different training sample. Therefore, the joint probability distribution of the neutral-emotion deformation vectors is modeled by the GMM

$$P(Z) = \sum_{q=1}^{Q} w_q N(Z|\mu_q, \Sigma_q), \sum_{q=1}^{Q} w_q = 1 \quad (10)$$

where $N(Z|\mu_q, \Sigma_q)$ is the Gaussian distributed density component, $\mu_q$ and $\Sigma_q$ are the qth mean vector and qth covariance matrix, respectively, $w_q$ is the mixture weight, and $Q$ is the number of Gaussian functions in the GMM.

After the GMMs are trained with the training data by the EM method [34], the optimal estimate of emotional facial deformation ($Y_{ik}$) given by neutral facial deformation ($X_k$) can be obtained according to the transform function of conditional expectation

$$Y_{ik} = E[Y_{ik}/X_k] = \sum_{q=1}^{Q} p_q(X_k) \left[ \mu_q^Y + \Sigma_q^Y X_k \Sigma_q^{-1} (X_k - \mu_q^X) \right]^{-1} (X_k - \mu_q^X) \quad (11)$$

where $p_q(X_k)$ is the probability that the given neutral observation belongs to the mixture component

$$p_q(X_k) = \frac{w_q N(X_k; \mu_q, \Sigma_q)}{\sum_{p=1}^{Q} w_p N(X_k; \mu_p, \Sigma_p)} \quad (12)$$

V. EXPERIMENTS AND DISCUSSION

When the speech is input, the system will calculate MFCCs and prosody parameters, respectively. Once the neutral FAP stream is synthesized by the fused HMM-based concatenation model, the optional GMM will be used to predict the emotional FAP stream based on the neutral FAP stream. To optimize the unit selection results and reduce noise in the output, a Bezier curve [29] is also used to smooth the output FAP sequences.
Fig. 6. Comparisons with synthesized FAP stream results with the corresponding recorded data frame by frame. The dash line denotes recorded FAP stream and the real line denotes synthesized FAP stream. (a) FAP#51 in the validation set. (b) FAP#60 in the validation set. (c) FAP#51 in the test set. (d) FAP#60 in the test set.

Of the total audiovisual corpus (including the emotional corpus), 80% is used for training, 10% is used for validation and the rest is used for testing. The performance of the method mentioned in the paper is be measured by the normalized mean square error (MSE) $\varepsilon$ and average correlation coefficient (ACC) $\rho$

$$\varepsilon = \frac{1}{T} \frac{1}{\sigma^2} \sum_{t=1}^{T} (\hat{v}_t - v_t)^2$$  \hspace{1cm} (13)

$$\rho = \frac{1}{T} \sum_{t=1}^{T} \frac{(\hat{v}_t - \mu_v)(\hat{v}_t - \mu_v)}{\sigma_v \sigma_v}$$  \hspace{1cm} (14)

where $v_t$ and $\hat{v}_t$ denote the recorded and the synthesized individual FAP streams, respectively, and $\sigma_v$ and $\sigma_\theta$ are the self-correlation coefficients of $v_t$ and $\hat{v}_t$, respectively. $T$ is the total number of frames in the database, and $\mu_v$ and $\mu_\theta$ are the corresponding means of all $v_t$ and $\hat{v}_t$, respectively.

A. Experiment on Neutral Audiovisual Corpus

1) Results: The experimental results for the fused HMM-based concatenation model on neutral audiovisual corpus are shown in Figs. 6 and 7.

In Fig. 6, we compare the smoothed synthesized stream of a lip related FAP (FAP#51, FAP#60) with a recorded FAP stream from a validation set and test set. In Fig. 6(a) and (b), the two curves (smoothed synthesized sequence and recorded sequence) are very close because the validation speech input is more likely to identify the complete sequence of the same sentence from the corpus. In Fig. 6(c) and (d), the input speech is more different from that in the training process. Although the two curves are not very close, their slopes are similar in most cases.

Fig. 6 shows that our method obtains good performance of the synthesized results. Although the test sentences show smaller correlation coefficients, they still can synthesize FAP shapes similar to those recorded in the corpus.

The normalized mean square errors and average correlation coefficients for all FAPs over the whole testing set are shown in Fig. 7.

In Fig. 7, we can see that the normalized mean square errors range from 0.08 to 0.25, and that the average correlation coefficients range from 0.58 to 0.88, indicating that we obtained very good synthesized results. From Fig. 7(a) and (b), we find that FAPs #3, #14–#17 and #51–#60 contain smaller error distributions and higher correlations than the others, indicating that lip-related FAPs have the smallest error distribution, while the eyebrow area and cheek areas contain relatively large prediction errors. These results confirm that the movement of the lips area is more highly correlated with audio information. The FAPs #19–#22 also show relatively high prediction errors, indicating that the prediction results of eyelid movement are also not very stable, although we tried to reduce the influence from eye blinks at the beginning of the experiment.

2) Comparison and Discussion: Finally, a MPEG-4 3-D facial animation engine was used to assess our synthesized FAP streams qualitatively. The animation model displays at a frame rate of 30 fps. Fig. 8 shows some frames of synthesized neutral facial animation.
To compare these results with previous research, here we make a detailed analysis with reference to the results of other studies. Such comparisons are not easy to perform because there is no common audiovisual corpus used for evaluation, and most other previous research only investigated the lips area. However, we still can make some comparisons with respect to quality, speed, and stability.

a) Compared with other unit selection methods: Previous unit selection methods for visual speech synthesis were based on direct visual data selection according to the distance between new audio and the stored audio [6], [7]. They worked on the basic assumption that audiovisual mapping is a one-to-one mapping process. We have previously tried a similar, direct unit selection method and shown the error distribution in [2] based on the same audiovisual corpus used in this paper. A comparison of the error distributions between these two studies shows that our new method in this paper generates more stable facial parameter outputs than direct unit selection methods; however, the expressiveness of the facial expression remains similar.

b) Compared with HMM-based methods: Traditional HMM-based visual speech synthesis is based on phonemic representation and direct mapping from recognized phonemes. With audio input, the method must use a speech recognition engine [11]. To our knowledge, the speech recognition engine can only work on whole utterances. Further, it normally has a longer time delay than methods based on frames, making visual speech synthesis very slow. In addition, the results are also deeply influenced by speech recognition results. The method used in our study, on the other hand, has a very short time delay of only seven frames (about 150 ms) if Viterbi search is not used. While we do perform a Viterbi search for every ten subsequences that contain a time of 1.5 s, the whole system runs with a time delay of 1.8 s. Thus, the time delay of the system is nearly the same as the window size set for the Viterbi search. This delay can be adjusted according to the requirements of different applications.

c) Compared with other visual parameter prediction methods: Other training methods for audiovisual mapping include VQ [11], [19], NN [9], MLPs [12], KNN [17], and GMM [27]. Among these, the VQ method has a problem in stability similar to that in the previously introduced unit selection methods. Due to the limitations of their algorithms, most of these methods can only process the visual parameters in the lips area. While we tried to extend these algorithms for entire facial expressions, we often encountered an over-smoothing problem that also occurs in GMM-based voice conversion systems [30] for facial expressions. This problem makes visual speech synthesis results less expressive than recorded video or the results of unit selection. The method we introduce here provides a very good balance between parameter prediction and unit selection. It also guarantees both expressiveness and stability of the synthesis results.

B. Experiment on Emotional Audiovisual Corpus

1) Results: The results of a further experiment on GMM-based emotional facial expression generation for the emotional audiovisual corpus are shown in Figs. 9 and 10 and Table II. From Fig. 7(a) and (b), we find that FAP#33, FAP#51, and FAP#60 are three typical FAPs in the lips and eyebrow areas that denote the major prediction errors in facial expression. To simplify the results for comparison, we only list these three FAPs in the figure and table.

In Fig. 9, the synthesized emotional trajectories show good performance following the trend of recorded trajectories. It is noted that the ends of synthesized trajectories do not well estimate facial deformation; this is mainly because the energy is very low during the final part of speech, and thus the features of the end parts of two aligned sequences do not strictly correspond.

The average correlation coefficients for different emotions are shown in Table II, from which we can see that estimates for FAP#33 are generally better than those of other points on the lips. Eyebrow movements are observed to be mainly affected by expressions, while lip movements are not only influenced by the speech content, but also extended with emotional exhibition. The trajectories of FAP#51 and FAP#60 are determined by both

![Fig. 9. Synthesized expressive trajectories of (a) FAP#33, (b) FAP#51, and (c) FAP#60 under happy state of “jiu4 shi4 xia4 yu3 ye3 qu4”.

Fig. 10. Comparisons with synthesized emotional results with the corresponding recorded video data. (a) Neutral state. (b) Surprise state. (c) Happy state.

| TABLE II AVERAGE CORRELATION COEFFICIENTS FOR DIFFERENT EXPRESSIONS |
|----------------|----------------|----------------|
| Correlation Coefficients | FAP#33 | FAP#51 | FAP#60 |
| Happiness             | 0.729    | 0.725  | 0.701   |
| Surprise              | 0.737    | 0.656  | 0.688   |
| Anger                 | 0.810    | 0.731  | 0.726   |
| Sadness               | 0.712    | 0.623  | 0.664   |

FAP#60 are three typical FAPs in the lips and eyebrow areas that denote the major prediction errors in facial expression. To simplify the results for comparison, we only list these three FAPs in the figure and table. The average correlation coefficients for different emotions are shown in Table II, from which we can see that estimates for FAP#33 are generally better than those of other points on the lips. Eyebrow movements are observed to be mainly affected by expressions, while lip movements are not only influenced by the speech content, but also extended with emotional exhibition. The trajectories of FAP#51 and FAP#60 are determined by both
the content and expression, and thus they have smaller correlation coefficients.

2) Difficulties of Emotion Generation: Though plenty of analyses have been performed on the audiovisual distributions among emotions [26], [31]–[33], these emotions have not actually been clearly defined via perception. Even for the same emotion, there are still various expression methods. In audio features, for example, one speaker may increase F0 jitter for “happiness” rather than increasing the overall pitch level. The locations of sentence stress in “anger” utterances can also vary according to differences in content and linguistic emphasis. In most cases, the stress is located in the word that the speaker wants to emphasize. In visual features, the facial expressions vary greatly according to the emotion level and speaker’s habits. These various methods of emotional expression increase the difficulty of emotion generation, because the audio features and visual features can be widely distributed. To solve this problem, we simplified the emotional audiovisual corpus into four basic emotion states. Speakers were asked to practice several times before their speech and facial expressions were recorded. Although this kind of performed data limits the naturalness of the final results, it generates very good expressiveness and makes the training procedure of the system very stable. Our experiments on the output proved that this kind of method can provide very good visual speech synthesis results compared to other methods.

VI. CONCLUSION

In this paper, we have introduced a novel method of visual speech synthesis based on a combined method of the fused HMM and the unit selection models. By modeling two coupling streams explicitly in the fused HMM model, the unit selection technique was used to obtain the visual counterparts given novel audio inputs. When implemented in different clusters identified by two-time k-means clustering, a relative short time delay and realistic synthesized facial animation were obtained.

We also analyzed the correlation between neutral facial deformation and expressive facial deformation and used GMM to model the joint probability distribution. By employing a hierarchical structure, the mapping process could be broken down into two steps. First-layer output gives visual speech synthesis results based on the neutral audiovisual corpus, while second-layer output gives emotional facial expression with neutral-emotion mapping models.

We tried using different voices in our system. The synthesized results well matched the voices, and the face animation seemed very natural. Future studies should investigate how dynamic expressive movement is related to prosody features. Further, FAP sequences also need to be aligned with better strategies, and more appropriate parameters to smooth synthesized trajectories need to be identified.

ACKNOWLEDGMENT

The authors would like to thank the students of NLPR for their cooperation with the experiments.

REFERENCES


Jianhua Tao (M’98) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include human–computer interaction, emotional information processing, and speech synthesis and recognition. He has published more than 60 papers in major journals and proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, Interspeech, ICCV, ICIP, ICME, ICPR, ACII, etc.

Dr. Tao received several awards from important conferences including Eurospeech2001. In 2006, he was elected as Vice-Chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing (SIG-CSLP), and Executive Committee member of the HUMAINE association. Currently, he also serves as chair or program committee member for several major conferences, including ACII, ISCSLP, and ICPR.

Le Xin received the M.S. degree in automatic control from Xi’an Jiaotong University, Xi’an, China, in 2003. He is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include computer vision, computer graphics, pattern recognition, and digital image processing and analysis. He has published papers in conference proceedings, including ICCV, ICIP, and ACII.

Panrong Yin received the M.S. degree in pattern recognition from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2006. She is currently working at the Speech Group, National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. Her current research interests include audiovisual information processing, computer vision, computer graphics, pattern recognition, and speech processing and analysis. She has published three papers in the Proceedings of ACII and the Journal of Information and Computational Science.