Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2325–2340
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A coupled HMM approach to video-realistic speech animation ☆

Lei Xie*, Zhi-Qiang Liu

School of Creative Media, City University of Hong Kong, Hong Kong SAR, China

Received 13 March 2006; received in revised form 4 October 2006; accepted 1 December 2006

Abstract

We propose a coupled hidden Markov model (CHMM) approach to video-realistic speech animation, which realizes realistic facial animations driven by speaker independent continuous speech. Different from hidden Markov model (HMM)-based animation approaches that use a single-state chain, we use CHMMs to explicitly model the subtle characteristics of audio–visual speech, e.g., the asynchrony, temporal dependency (synchrony), and different speech classes between the two modalities. We derive an expectation maximization (EM)-based A/V conversion algorithm for the CHMMs, which converts acoustic speech into decent facial animation parameters. We also present a video-realistic speech animation system. The system transforms the facial animation parameters to a mouth animation sequence, refines the animation with a performance refinement process, and finally stitches the animated mouth with a background facial sequence seamlessly. We have compared the animation performance of the CHMM with the HMMs, the multi-stream HMMs and the factorial HMMs both objectively and subjectively. Results show that the CHMMs achieve superior animation performance. The *ph-vi*-CHMM system, which adopts different state variables (phoneme states and viseme states) in the audio and visual modalities, performs the best. The proposed approach indicates that explicitly modelling audio–visual speech is promising for speech animation.

Keywords: Speech animation; Audio-to-visual conversion; Talking faces; Facial animation; Coupled hidden Markov models (CHMMs)

1. Introduction

Speech-driven talking faces are playing an increasingly indispensable role in multimedia applications such as computer games, online virtual characters, video telephony, and other interactive human–machine interfaces. For example, talking faces can provide visual speech perceptual information for hearing-impaired people to better communicate with machines through lipreading [1]. Recent studies have shown that the trust and attention of humans towards machines can be significantly increased by 30% if humans are interacting with a human face instead of text only [2]. Current Internet videophones can transmit videos, but due to bandwidth limitations and network congestion, facial motion accompanied with audio often appears jerky since many frames are lost during transmissions. Therefore, a video-realistic speech-driven talking face may provide a good alteration.

The essential problem of speech-driven talking faces is speech animation—synthesizing speech-related facial animation from audio. Despite of decades of extensive research, realistic speech animation still remains to be one of the challenging talks due to the variabilities of human speech, mostly commonly the coarticulation phenomenon [3]. Various approaches have been proposed during the last decade, which significantly improve the animation performance. Some approaches use a 3D mesh to define the head shape and map a face texture to the mesh [4–6]. Others realize photo- or video-realistic animation from recorded image sequences of face and render facial movements directly at the image level [6–10].

Despite of different head models, according to the audio/visual conversion method, speech animation can be categorized into *speech classes from audio* and *animation parameters from audio* [6]. In the former approaches, audio is first segmented to a string of speech classes (e.g., phonemes) manually

☆ This work is supported by the Hong Kong RGC CERG project CityU 1247/03E.
* Corresponding author. Tel.: +852 21942209; fax: +852 27887165.
E-mail addresses: xielei21st@gmail.com (L. Xie), zq.liu@cityu.edu.uk (Z.-Q. Liu).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.12.001

or automatically by a speech recognizer. Subsequently these units are mapped simply to lip poses, ignoring dynamic factors such as speech rate and prosody. The latter approaches derive animation parameters directly from speech acoustics, where speech dynamics are preserved. During the last two decades, machine learning methods, such as neural networks [11], Gaussian mixture models (GMMs) and hidden Markov models (HMMs) have been extensively used in the audio/visual conversion.

In GMM-based approaches [12,13], Gaussian mixtures are used to model the probability distribution of audio–visual data. After the GMM is learned using the expectation maximization (EM) algorithm and the audio–visual training data, the visual parameters are mapped analytically from the audio. The *universal* mapping of the GMM ignores the context cues that are inherent in speech. To utilize the context cues, a mapping can be tailored to a specific linguistic unit, e.g., a word. Hence, HMM-based approaches have been recently explored.

To the best of our knowledge, Yamamoto et al. [14] were the first to introduce HMMs into speech animation, which has led the way to some new developments [13,15–20]. Their approach, namely Viterbi single-state approach, trained HMMs from audio data, and aligned the corresponding animation parameters to the HMM states. During the synthesis stage, an optimal HMM state sequence is selected for a novel audio using the Viterbi alignment algorithm [21]; and the visual parameters associated with each state is retrieved. Such approaches produce jerky animations since the predicted visual parameter set for each frame was an average of the Gaussian mixture components associated with the current single state, and it was indirectly related to the current audio input. In some other techniques, e.g., the mixture-based HMM [13] and the remapping HMM in Voice Puppetry [15], the visual output was made dependent not only on the current state, but also the audio input, resulting in improved performance. The mixture-based HMM technique [13] trained joint audio–visual HMMs which encapsulate the synchronization between the two modalities of speech. Recently, Aleksic et al. [20] proposed a correlation-HMM system using MPEG-4 visual features, which integrated independently trained acoustic HMM and visual HMM, allowing for increased flexibility in model topologies.

However, all the above methods heavily rely on the Viterbi algorithm that lacks robustness to noise [17]. If speech is contaminated by ambient noise, the animation quality will suffer greatly [15]. Moreover, the Viterbi sequence is deficient for speech animation in that it represents only a small fraction of the total probability mass, and many other slightly different state sequences potentially have nearly equal likelihoods [22].

Moon et al. [23] proposed a hidden Markov model inversion (HMMI) method for robust speech recognition. Choi et al. [16,17] extended this method to audio–visual domain for speech animation, in which audio and video were jointly modelled by phoneme HMMs. They were able to generate animation parameters directly from the audio input by a conversion algorithm considered as an inversion of EM-based parameter training. In this way, they managed to avoid using the Viterbi algorithm, and made use of all possible state sequences to represent a quite large fraction of the total probability mass. More recently, Fu et al. [22] have demonstrated that the HMMI method outperforms the remapping HMMs [15] and the mixture-based HMMs [13] on a common test bed.

The conventional one-chain HMMs do have limitations in describing audio–visual speech: (1) we know that due to the difference in discrimination abilities, audio speech and visual speech can be categorized to different speech classes—*phonemes* and *visemes* [24]. Therefore, the bimodal speech is better modelled by different atoms explicitly. (2) speech production and perception are inherently coupled processes with both *synchrony* and *asynchrony* between the audio and visual modalities [25]. In previous HMM-based speech animations, the bimodal speech is modelled by a single Markov chain, which cannot reflect the above important facts.

The above two facts are important in audio–visual speech recognition (AVSR) or automatic *lipreading* [26]. These facts also affect the performance of the *inverse* problem (i.e., speech animation) since human perception system is sensitive to artifacts induced by loss of synchrony or asynchrony. Therefore, in order to make animation look more natural, it is necessary to take these facts into consideration. In this paper, we propose a coupled HMM (CHMM) approach to video-realistic speech animation, in which we use CHMMs to model the above characteristics of audio–visual speech.

In the following section, we describe the diagram of our speech animation system. Section 3 presents the AV-CHMMs used in our speech animation system, including our motivations, the model structures and the model training procedure. Section 4 derives the EM-based A/V conversion algorithm for the AV-CHMMs. Section 5 describes the audio-visual front-end. Our facial animation unit is presented in Section 6. Section 7 gives the comparative evaluations both objectively and subjectively. Finally, conclusions and future work are given in Section 8.

2. Speech animation system overview

Fig. 1 shows the block diagram of the proposed speech animation system. The system is composed of two main phases—the AV modelling phase (offline) and the speech-to-video synthesis phase (online). The offline phase is used to model the audio–visual speech as well as learn the correspondences between the two modalities from the AV facial recordings. Given the AV models, the online synthesizer converts acoustic audio to visual parameters (i.e., animation parameters) and synthesizes facial animations from these parameters.

In the AV modelling phase, initially the audio and video processing units extract representative features for audio and video, respectively. The video processing unit also statistically learns an appearance space of mouth articulation. Subsequently, we train AV models (AV-CHMMs) and build up the correspondences between acoustic audio and visual articulation from an audio–visual dataset (the JEWEL dataset) as well as an audio-only speech corpus (the TIMIT corpus). We use two datasets to establish audio–visual correspondences from facial recordings of a subject, and also to learn audio

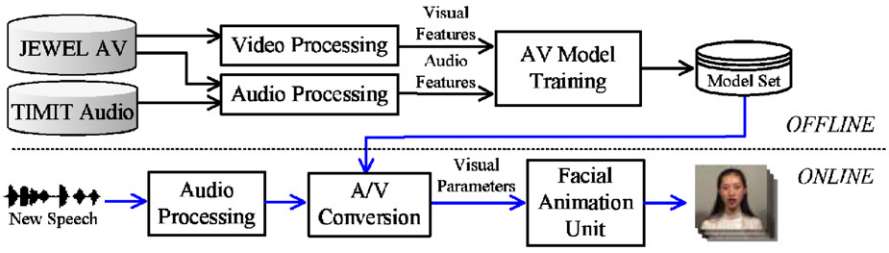

Fig. 1. Diagram of the speech animation system.

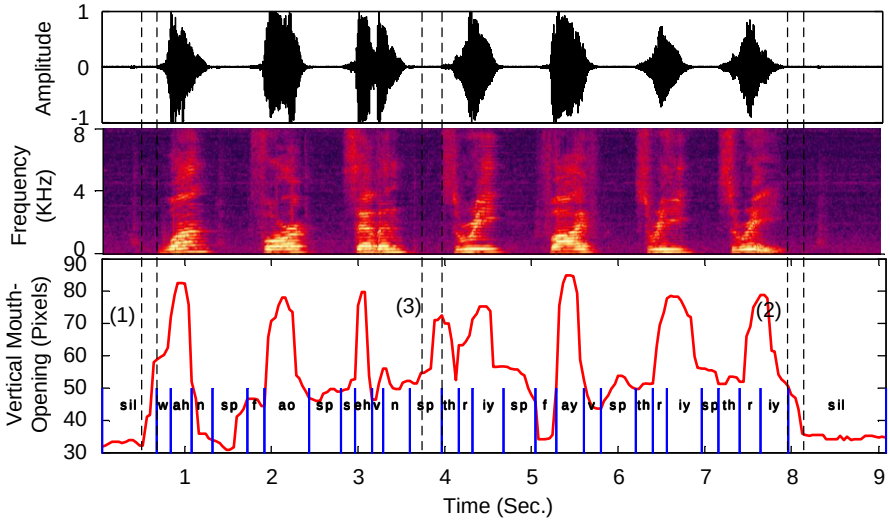

Fig. 2. Example of asynchrony between audio and visual speech. Up: speech waveform; middle: speech spectra; bottom: vertical mouth opening with phoneme labels transcribed from audio.

distributions from various speakers to realize speaker independent animation.

The speech-to-video synthesis phase processes a sequential structure. After feature extraction, a new audio is fed into an A/V converter, resulting in visual parameters. The A/V converter adopts an EM-based conversion algorithm (described in Section 4) which directly generates decent visual parameters frame by frame. Finally, the facial animation unit resembles framewise visual parameters to a mouth image sequence, and stitches the animated mouth with a background facial sequence (described in Section 6).

3. Coupled HMMs for AV modelling

3.1. Characteristics of bimodal speech

3.1.1. Asynchrony and synchrony

Asynchrony arises naturally both in audio–visual speech perception and in speech production. From the speech production point of view, it has been proven that usually visual speech activity precedes the audio signal by as much as 120 ms [27], which is close to the average duration of a phoneme. Lavagetto [28] has shown that visible articulators (i.e., lips, tongue and jaw), during uttering, start and complete their trajectories asynchronously, resulting in both forward and backward coarticulation with respect to the acoustic speech wave. Intuitively this makes much sense, since articulators have to position themselves properly before and after the start and end of an acoustic utterance. This time interval is the well-known voice-on-set time (VOT), which is defined as the time delay between when a consonant is released and when voicing, the vibration of the vocal cords, begins. The VOT is an important cue to the voicing features in perceiving stop consonants, e.g., [p], [t], [k], [n], [m]. Fig. 2 shows an example of the asynchrony between audio speech and visual speech. From the comparison between the speech waveform, the spectrum and the vertical mouth opening scale, we can clearly see that obvious time intervals exist between the audio signal and visual articulation (for example, areas (1), (2) and (3)). On the other hand, audio speech and visual speech are correlated since they are orginated from the same articulation process, and thus need to be synchronized within a time period.

From the speech perception point of view, we know that sound and light travel at quite different speeds, and a 10 m

Table 1
The 13 visemes mapped from 47 phonemes

Place of articulation	Viseme	Phoneme
Silence	*sip*	/sil/, /sp/
Lip rounding-based vowels	*lr1*	/ao, /aa/, /ah/, /er/, /oy/, /aw/, /hh/
	lr2	/uw/, /uh/, /ow/, /em/
	lr3	/ae/, /eh/, /ey/, /ay/
	lr4	/ih/, /iy/, /ax/, /axr/, /ix/
Alveolar semi-vowels	*as*	/l/, /el/, /r/, /y/
Alveolar fricatives	*af*	/s/, /z/
Alveolar	*al*	/t/, /d/, /n/, /en/
Palato Alveolar	*pa*	/sh/, /zh/, /ch/, /jh/
Bilabial	*bi*	/p/, /b/, /m/
Dental	*de*	/th/, /dh/
Labio-dental	*ld*	/f/, /v/
Velar	*ve*	/ng/, /k/, /g/, /w/

distance between the speaker and the listener will introduce roughly a 30 ms delay in the audio channel. In addition, McGrath and SummerLeld [29] found that an audio lead of less than 80 ms or lag of less than 140 ms could not heavily affect the speech perception ability. However, if the audio was delayed by more than 160 ms, the perception ability will be severely affected.

3.1.2. Speech classes for audio and video

We know that *phoneme* is the basic distinct atom of acoustic speech that describes how speech conveys linguistic information. In American English, usually 40–50 phonemes are used according to the dialects. Not all phonemes are visually distinct since human vision system only can observe the visible articulators (such as lips, mouth and teeth) that describe the post formulation of uttering. However, phonemes can be clustered into the so-called *visemes*. Visemes are defined as the smallest visibly distinguishable atoms of speech. There are many acoustic sounds that are visually ambiguous, which are grouped into the same viseme class. For example, the bailable phonemes [p], [b] and [m] are all produced by a visually distinguishable closed mouth; and they fall into one viseme class. There is therefore a many-to-one mapping between phonemes and visemes. Table 1 shows a phoneme-to-viseme mapping table according to the *place* of articulation [30]. By the same token, there are many visemes that are acoustically ambiguous. An example of this can be seen in the acoustic domain when people use so-called phonetic alphabets, e.g., 'B as in boy' or 'D as in Deta' to reduce ambiguity for spelling. These auditory confusions are usually distinguishable in the visual modality. This highlights the bimodal nature of speech, and the fact that to properly understand what is being said information is required from both modalities.

Therefore, it is more appropriate and intuitive to model the audio and visual modalities using phonemes and visemes, respectively. Especially for speech animation which intends to derive visual articulation from acoustic audio, explicitly design of audio–visual interactions are fairly important.

3.2. DBNs for audio–visual speech

In previous HMM-based speech animation approaches, single-stream HMMs are used to model audio–visual speech with tight inter-modal synchronization. This kind of model structure does not accord with the characteristics of audio–visual speech as indicated in Section 3.1, and a richer model structure intuitively produces better speech animation performance. During the last decade, a more general framework, dynamic Bayesian networks (DBNs) [31,32], has emerged as a more powerful and flexible tool to model complex stochastic processes. DBNs generalize HMMs by representing hidden states as *state variables*, and allow the states to have complicated inter-dependencies. The conventional HMM is just a special case with only one state variable in a time slice. Fig. 3(a) shows the repeating structure of HMM described in a DBN framework.

Among DBNs, several model structures such as multi-stream HMMs (MSHMMs) [33], factorial HMMs (FHMMs) [34], and CHMMs [35] have been introduced to model the bimodal speech in AVSR, resulting in improved recognition performance as compared to HMMs. Especially, CHMMs have shown superior performance [36]. Fig. 3(b)–(d) shows their model structures suitable for describing audio–visual speech. However, these richer structures have not been introduced into speech animation yet. In this paper, we propose a CHMM approach to realize video-realistic speech animation.

3.3. Coupled HMMs

The CHMMs can be considered as a generalization of the conventional single-stream HMMs suitable for modelling time series in a large variety of multimedia applications that integrate multiple streams of data. The CHMMs were first introduced by Brand et al. [35] and were successfully used for hand gestures recognition [35], 3D surface inspection [37], speech prosody recognition [38] and AVSR [36].

A CHMM can be seen as a collection of multiple HMM chains coupled through cross-time and cross-chain conditional probabilities. Fig. 3(d) shows the specific structure of a two-chain CHMM in audio–visual speech modelling, namely AV-CHMM, where two hidden Markov chains are incorporated to describe the audio and visual modalities, respectively. There are two variables for each chain in each time frame—the hidden state variable q_t^s, $s \in \{a, v\}$ and the observation variable $\mathbf{o}_t^s$, $s \in \{a, v\}$. A state variable at frame t is dependent on its two parents in previous frame $t-1$, which describes the temporal *coupling* relationship between the two streams. This structure models the *asynchrony* of the audio and visual modalities while still preserving their natural correlation over time (i.e., *synchrony*). This model but also allows us to use different modality *atoms* (in terms of state variables) in audio and video. Intuitively, this model may better capture the interprocess influences between the audio and visual modalities of speech.

The conditional probability distributions (CPDs) associated with variables for each frame describe the following

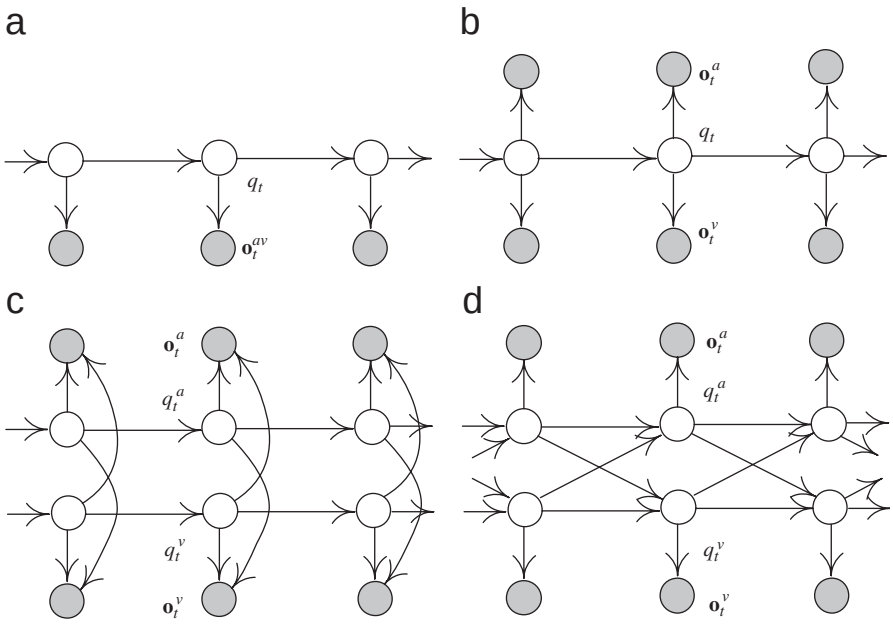

Fig. 3. DBN models for audio–visual speech (a: HMM, b: MSHMM, c: FHMM and d: CHMM).

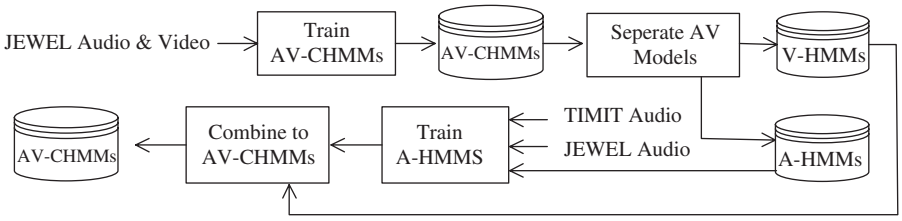

Fig. 4. Training procedure of AV-CHMMs.

probabilities:

- $P(\mathbf{o}_t^s|q_t^s), s \in \{a, v\}$: observation probability and
- $P(q_t^s|\mathbf{q}_{t-1}), s \in \{a, v\}$: state transition probability,

where the 'meta-state' $\mathbf{q}_{t-1}=[q_{t-1}^a, q_{t-1}^v]$, and for initialization, $P(q_1^s|\mathbf{q}_0) = P(q_1^s)$. The state transition probability is described by a 3D table; and a continuous Gaussian mixture is associated with each q_t^s:

$$P(\mathbf{o}_t^s|q_t^s) = \sum_{k=1}^{K} w_{q_t^s k} \mathcal{N}(\mathbf{o}_t^s, \mu_{q_t^s k}, \Sigma_{q_t^s k}). \quad (1)$$

Correspondingly, the joint probability distribution (JPD) among this DBN structure is

$$P(\mathbf{O}^{av}|\mathbf{q}) = \prod_t \prod_s P(q_t^s|\mathbf{q}_{t-1}) P(\mathbf{o}_t^s|q_t^s). \quad (2)$$

The training of the model parameters (i.e., CPDs) can be performed using the EM algorithm. Although the topology of a CHMM resembles that of an ordinary HMM, the EM training of ordinary HMMs is not directly applicable. We derive the EM algorithm for CHMM parameter training in Appendix A.

3.4. Training procedure for AV-CHMMs

Since our objective is to realize speech animation driven by speaker independent continuous speech, we need AV facial recordings as well as a large acoustic speech corpus. We use the JEWEL audio–visual dataset together with the TIMIT speech corpus to build the AV-CHMMs. Fig. 4 shows the diagram of the training procedure.

The JEWEL audio–visual dataset [39] contains 524 recordings of one female speaker uttering sentences from the TIMIT corpus. The training set is composed of 2 SA sentences and 450 SX sentences, and the testing set contains 50 SI sentences. Another 22 SI sentences are used as a small validation set. These sentences have a good coverage of English phonetic contexts. In the dataset, the speaker's head-and-shoulder front view against a white background is shot by a digital video cameracorder in a studio environment, where synchronized audio and video are recorded. For each of the sentences, the dataset provides a

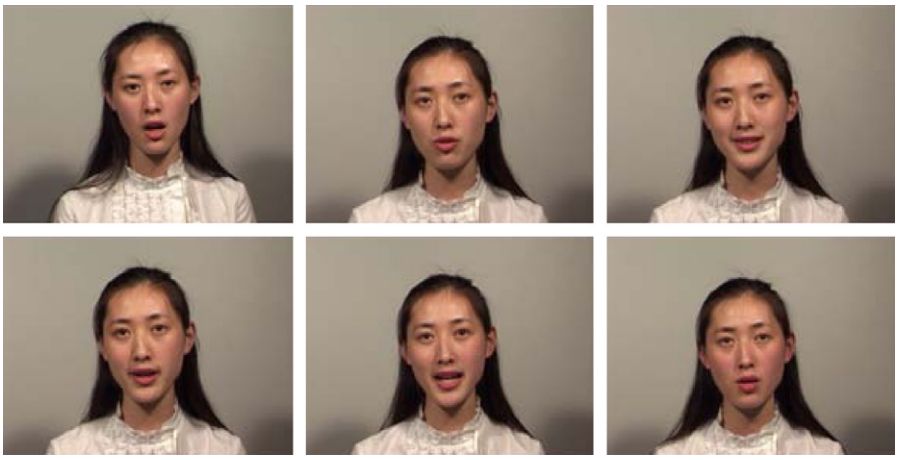

Fig. 5. Some snapshots from the JEWEL AV dataset.

speech waveform in Microsoft WAV format, a word-level transcription file, and an MPEG-2 AVI video file time synchronized with the speech waveform. Audio is acquired at a rate of 16 Hz, and video is recorded at 25 frames/s in PAL format. In total, the audio–visual recordings are about 2500 s in duration. Fig. 5 shows some snapshots from the dataset. To realize speech animation, the mouth region of interest (ROI) is tracked by a method described in Ref. [40].

Firstly, we use time-synchronized joint audio–visual features extracted from the JEWEL training set (452 sentences) to train the AV-CHMMs. This process learns the correspondences between the acoustic speech and visual articulation, and most importantly, transition probabilities between the audio states and visual states. The EM algorithm derived in Appendix A is used to train the AV-CHMMs.

Secondly, we use the acoustic features extracted from the TIMIT training set (4620 utterances from 630 speakers) as well as the JEWEL training set to train the audio observation distributions (pdfs) extensively, while keeping the visual observation distributions (pdfs) and transition probabilities intact. To do this, single-stream audio-HMMs (A-HMMs) and visual-HMMs (V-HMMs) are separated from the AV-CHMMs. The A- and V-HMMs directly adopt the observation emission probabilities from the corresponding streams of AV-CHMMs, and the new state transition matrices are extracted from the 3D tables of AV-CHMMs. Subsequently, we train the A-HMMs using the large training set (the JEWEL audio and TIMIT audio) via the EM algorithm for conventional HMMs [21]. This process intends to achieve good distribution estimations of acoustic signal from abundant audio samples collected from numerous speakers. Finally, the A-HMMs and V-HMMs are re-combined to AV-CHMMs, with newly trained audio distributions, former visual observation distributions and the transition probabilities from the AV-CHMMs.

4. EM-based A/V conversion on AV-CHMMs

Given the trained AV models, a simple, common A/V conversion approach is to derive sub-phonemic transcriptions from a novel audio via the Viterbi algorithm, and the visual Gaussian mean associated with the current state label is used as the visual parameter vector of the current frame. As indicated in Section 1, this approach has a major defect in that the facial animation performance relies heavily on the Viterbi state sequence which is not robust to acoustic degradation, e.g., additive noise. Choi's HMMI approach [17] has managed to avoid the Viterbi algorithm, which uses maximum likelihood (ML) criterion to generate visual parameters and catches a large fraction of the total probability mass by considering all the HMM states at each time slice.

4.1. The algorithm

Based on Choi's work [17], we derive the specific A/V conversion algorithm for the AV-CHMMs under the ML criterion. As the inversion of ML-based model parameter training, the conversion algorithm searches for the missing visual parameters (observations) by maximizing the likelihood of visual parameters given the trained CHMMs and the audio input. We use the EM algorithm to solve the ML problem. According to the EM algorithm [41], the optimal visual parameters $\widehat{\mathbf{O}}^v$ can be found by iteratively maximizing the auxiliary function $Q(\lambda, \lambda, \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'})$, i.e.,

$$\widehat{\mathbf{O}}^v = \arg \max_{\mathbf{O}^{v'} \in \mathcal{O}^v} Q(\lambda, \lambda, \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'}), \tag{3}$$

where $\mathbf{O}^v$ and $\mathbf{O}^{v'}$ denote the old and new visual parameter sequences in the visual parameter space $\mathcal{O}^v$, respectively.

Given the audio-visual observation sequence, $\mathbf{O}^{av} = [\mathbf{O}^a, \mathbf{O}^v]$, the audio visual state sequence, $\mathbf{q} = [\mathbf{q}^a, \mathbf{q}^v]$, and a well-trained AV model set λ, according to the Markov property of independent relationships between variables, the *complete-data* likelihood, or the JPD (see Eq. (2)), can be formed as

$$P(\mathbf{O}^a, \mathbf{O}^v, \mathbf{q}|\lambda) = \prod_{t=1}^{T} [P(q_t^a|\mathbf{q}_{t-1}) \times P(q_t^v|\mathbf{q}_{t-1}) P(\mathbf{o}_t^a|q_t^a) P(\mathbf{o}_t^v|q_t^v)]. \tag{4}$$

Input : audio observation sequence $\mathbf{O}^a = [\mathbf{o}_t^a]_{t=1}^T$,
trained CHMM parameter set
$\forall$ *state* $i^s, \lambda_{i^s} = \{\mu_{i^s k}, \Sigma_{i^s k}, w_{i^s k}, P(i^s|\mathbf{j})\}, s \in \{a, v\}, k = 1, \cdots, K$.
Output : optimal visual parameter sequence $\widehat{\mathbf{O}}^v = [\hat{\mathbf{o}}_t^v]_{t=1}^T$.
Initialize: $\mathbf{O}^{v,(0)} = [\sum_k w_{\theta_t^v k} \mu_{\theta_t^v k}]_{t=1}^T$, where $[\theta_t^v]_{t=1}^T$ is the Viterbi state sequence from a speech recognition engine.

```
begin
    l = 0, where l is the iteration number;
    repeat
        l++;
        for (t = 1; t ≤ T; t++) do
            E-Step: ∀ q_t, compute α_t^(l)(q_t) and β_t^(l)(q_t) using {O^a, O^{v,(l−1)}} and
            the frontier algorithm;
                ∀ q_t^s, compute γ_t^(l)(q_t^s, k) using Eq. (A.15);
            M-Step: estimate o_t^{v,(l)} using Eq. (9);
        end
    until ||O^{v,(l)} − O^{v,(l−1)}|| < ξ;
end
```

Fig. 6. The EM-based A/V conversion algorithm on AV-CHMMs.

The auxiliary function can be further expressed as

$$\begin{aligned} Q(\lambda, \lambda, \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'}) &= \sum_{\mathbf{q}} P(\mathbf{O}^a, \mathbf{O}^v, \mathbf{q}|\lambda) \log P(\mathbf{O}^a, \mathbf{O}^{v'}, \mathbf{q}|\lambda) \\ &= \sum_{\mathbf{q}} P(\mathbf{O}^a, \mathbf{O}^v, \mathbf{q}|\lambda) \left\{ \sum_{t=1}^T \log P(q_t^a|\mathbf{q}_{t-1}) \right. \\ &\quad + \sum_{t=1}^T \log P(q_t^v|\mathbf{q}_{t-1}) \\ &\quad + \sum_{t=1}^T \log P(\mathbf{o}_t^a|q_t^a) \\ &\quad \left. + \sum_{t=1}^T \log P(\mathbf{o}_t^{v'}|q_t^v) \right\}. \end{aligned} \tag{5}$$

By taking the derivative of $Q(\lambda, \lambda, \mathbf{O}^v, \mathbf{O}^v, \mathbf{O}^{v'})$ with respect to $o_{t,i}^{v'}$ (the ith coefficient of $\mathbf{o}_t^{v'}$) and equaling it to be zero, we get

$$\begin{aligned} \frac{\partial Q}{\partial o_{t,i}^{v'}} &= \sum_{\mathbf{q}} P(\mathbf{O}^a, \mathbf{O}^v, \mathbf{q}|\lambda) \frac{\partial}{\partial o_{t,i}^{v'}} [\log P(\mathbf{o}_t^{v'}|q_t^v)] \\ &= \sum_{\mathbf{q}_t} P(\mathbf{O}^a, \mathbf{O}^v, \mathbf{q}_t|\lambda) \frac{\partial}{\partial o_{t,i}^{v'}} [\log P(\mathbf{o}_t^{v'}|q_t^v)] = 0, \end{aligned} \tag{6}$$

where $\mathbf{q}_t$ denotes the possible value vector of state variables at time t. The derivative $\partial \log P(\mathbf{o}_t^{v'}|q_t^v)/\partial o_{t,i}^{v'}$ is calculated by differentiating Eq. (1):

$$\begin{aligned} \frac{\partial \log P(\mathbf{o}_t^{v'}|q_t^v)}{\partial o_{t,i}^{v'}} &= \sum_{k=1}^K w_{q_t^v k} (2\pi)^{-P^v/2} |\Sigma_{q_t^v k}|^{-1/2} \\ &\quad \times \left[\sum_{j=1}^{P^v} \sigma_{q_t^v k}(i, j)(\mu_{q_t^v k}(j) - o_{t,j}^{v'}) \right], \end{aligned} \tag{7}$$

where P^v is the dimensionality of $\mathbf{o}_t^{v'}$, and $\mathbf{o}_t^{v'} = [o_{t,i}^{v'}]_{i=1}^{P^v}$. $\mu_{q_t^v k}(j)$ is the jth coefficient of $\mu_{q_t^v k}$; $\sigma_{q_t^v k}(i, j)$ denotes the (i, j)th element of the inverse covariance matrix $\Sigma_{q_t^v k}^{-1}$. If the covariance matrix is diagonal, Eq. (7) can be simplified to

$$\begin{aligned} \frac{\partial \log P(\mathbf{o}_t^{v'}|q_t^v)}{\partial o_{t,i}^{v'}} &= \sum_{k=1}^k w_{q_t^v k} (2\pi)^{-P^v/2} |\Sigma_{q_t^v k}|^{-1/2} \cdot \sigma_{q_t^v k}(i, i) \\ &\quad \times (\mu_{q_t^v k}(i) - o_{t,i}^{v'}). \end{aligned} \tag{8}$$

Using Eqs. (6) and (8), we get

$$o_{t,i}^{v'} = \frac{\sum_{\mathbf{q}_t} \sum_k \gamma_t(q_t^v, k) w_{q_t^v k} \sigma_{q_t^v k}(i, i) \mu_{q_t^v k}(i)}{\sum_{\mathbf{q}_t} \sum_k \gamma_t(q_t^v, k) w_{q_t^v k} \sigma_{q_t^v k}(i, i)}, \tag{9}$$

where the state occupation probability $\gamma_t(q_t^v, k)$ is calculated using Eq. (A.15) in Appendix A.

Similarly for MSHMMs and FHMMs shown in Fig. 3, the reestimation formulas can be derived as

$$o_{t,i}^{v'} = \frac{\sum_{q_t} \sum_k \gamma_t(q_t, k) w_{q_t k} \sigma_{q_t k}^v(i, i) \mu_{q_t k}^v(i)}{\sum_{q_t} \sum_k \gamma_t(q_t, k) w_{q_t k} \sigma_{q_t k}^v(i, i)}, \tag{10}$$

and

$$o_{t,i}^{v'} = \frac{\sum_{\mathbf{q}_t} \sum_k \gamma_t(\mathbf{q}_t, k) w_{\mathbf{q}_t k} \sigma_{\mathbf{q}_t k}^v(i, i) \mu_{\mathbf{q}_t k}^v(i)}{\sum_{\mathbf{q}_t} \sum_k \gamma_t(\mathbf{q}_t, k) w_{\mathbf{q}_t k} \sigma_{\mathbf{q}_t k}^v(i, i)}. \tag{11}$$

Fig. 6 summarizes the EM-based A/V conversion algorithm on AV-CHMMs. We use the Gaussian mixture centers of each states from the Viterbi sequence as the initial visual parameter values at the start of the prediction iterations. The Viterbi state sequence is obtained from an audio-only speech recognition engine.

4.2. Discussions

As pointed out in Ref. [17], the conversion algorithm moves the visual parameters $\mathbf{o}_t^{v'}$ to maximize the likelihood of these

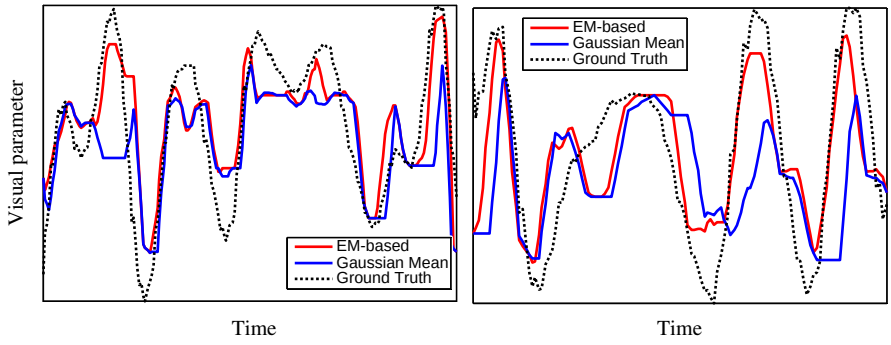

Fig. 7. Predicted visual parameter trajectories vs. the ground truth.

visual parameters for fixed mixture means, which retains the original distributions of visual parameters. The weighted averages of Gaussian mixture means of *all possible* states are used as the predicted visual parameters in each time frame t. Compared with the Viterbi single-state approach which considers the probability distribution of a single state at each time slice, the visual parameters predicted by the EM-based algorithm represent quite a large fraction of the total probability mass, making the estimation more robust to speech degradations. Even when the speech is not contaminated, the EM-based approach also shows superior performance due to the ML criterion. Fig. 7 shows two snippets of the predicted visual parameter time trajectories (the EM-based and Viterbi single-state methods) compared with the ground truth (actual parameters extracted from original facial image sequences). It illustrates the advantages of the EM-based method over the Viterbi single-state method, where the curve predicted by the EM-based approach match the actual parameters more closely.

In the HMMI method, audio–visual speech is modelled using a single Markov chain with tight inter-modal synchronization, where the audio and visual data are described by joint observation distributions of same speech classes and lack inter-modal interactions. This is not consistent with the facts of speech production and perception. In contrast, the AV-CHMMs are able to model the bimodal speech more accurately by mimicking the interprocess influences between the audio and visual modalities of speech. For example, audio and visual data are described by separate observation distributions of separate speech classes, and audio visual interactions are modelled by inter-modal-class transitions. Consequently, the weighted averages of Gaussian means (Eq. (9)) may become more appropriate than that of the HMMI.

5. Audio visual front-end

Prior to AV modelling, front-end processing is performed to achieve representative features of audio and visual speech (see Fig. 1). The speech signal, sampled at 16 kHz mono, is processed in frames of 25 ms with a 15 ms overlapping (rate = 100 Hz). We first pre-emphasize speech frames with an FIR filter ($H(z) = 1 - az^{-1}, a = 0.97$), and weight them with a Hamming window to avoid spectral distortions. After pre-processing, we extract Mel Frequency Cepstral Coefficients (MFCCs) [42] as the acoustic features. Each acoustic feature vector consists of 12 MFCCs, the energy term, and the corresponding velocity and acceleration derivatives. The dimensionality of acoustic feature vector is 39 for each frame.

Since we are targeting video-realistic speech animation, we use one of the most effective feature extraction methods—principal components analysis (PCA) [43] to get the visual features that capture mouth appearance in a low dimension. In the PCA implementation, the correlation matrix $\mathbf{R}$ of mouth images is first computed. Subsequently, $\mathbf{R}$ is diagonalized as $\mathbf{R} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$, where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_d]$, and $\mathbf{a}_i$ is the eigenvectors of $\mathbf{R}$, $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{R}$. The eigenvectors are called *eigenmouths* or *eigenlips*, representing the statistical basis of mouth appearance space.

A mouth image $\mathbf{I}$ can be approximately represented by a linear combination of the r ($r \ll d$) most significant eigen mouths $\mathbf{a}_i$ ($i = 1, \ldots, r$), that is,

$$\mathbf{I} = \bar{\mathbf{I}} + \tilde{\mathbf{A}}\mathbf{g}, \tag{12}$$

where $\tilde{\mathbf{A}} = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ and $\bar{\mathbf{I}}$ is the mean mouth image. The weighting coefficient vector $\mathbf{g}$ is the visual feature vector that we look for, which can be computed by

$$\mathbf{g} = \tilde{\mathbf{A}}^{\mathrm{T}}(\mathbf{I} - \bar{\mathbf{I}}). \tag{13}$$

Fig. 8 depicts the 30 most significant eigenmouth images of the red channel calculated on the 980 representative mouth images from the JEWEL AV dataset, which preserves 96.7% of the statistical variance of mouth appearance. Totally, we use a set of 90 visual features (30 for each color channel) for each frame, and the feature vector is up-sampled to 100 frames/s to meet the audio feature rate (100 Hz).

6. Facial animation unit

The facial animation unit first smoothes the predicted visual parameters by a moving average filter (3 frames wide) to remove jitters, and then augments the fine details of the mouth appearance using a *performance refinement* process. Subsequently, mouth animation is generated from the visual parameters by the PCA expansion process [43]. Finally, we overlay

Fig. 8. The 30 most significant eigenmouth images of the red channel calculated on the JEWEL AV dataset. The images are ordered in descending significance from left to right and top to bottom.

the synthesized mouth animation onto a background sequence which contains natural head and eye movements.

6.1. Performance refinement

Although the PCA-based visual parameters have already represented the most significant statistical variances of the speech-related mouth appearance, the mouth images resembled by PCA expansion still lack fine details due to the low dimensionality of the visual parameters (90 dimensions). Introducing more visual parameters will result in more prediction errors. Therefore, we propose a performance refinement process to improve the realism of the animation. We select a set of 500 typical normalized mouth images from the JEWEL dataset, which covers almost all possible articulation-related mouth appearances. We then save their full-dimension PCA coefficients (980 dimensions) of each color channel (R,G and B) to a candidate set $\mathbf{G}$:

$$\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{500}\}, \quad \mathbf{g} = [\mathbf{g}^r, \mathbf{g}^d], \tag{14}$$

where $r = 1{:}30$, $d = 31{:}980$ and $\mathbf{g}$ denotes the PCA coefficient vector for a color channel. We choose the visual parameters $\hat{\mathbf{g}}_t^d$ as the fine detail augments of frame t using the following MSE criterion:

$$\hat{\mathbf{g}}_t^d = \arg\min_{\mathbf{g}_j^d} ||\hat{\mathbf{o}}_t^v - \mathbf{g}_j^r||, \quad j = 1, 2, \ldots, 500, \tag{15}$$

where $\hat{\mathbf{o}}_t^v$ denotes the estimated visual parameters. Therefore, the full-dimension visual parameter vector will be $\mathbf{g}_t = [\hat{\mathbf{o}}_t^v, \hat{\mathbf{g}}_t^d]$. The mouth image $\mathbf{I}_t$ at frame t for a specific color channel is resembled by the PCA expansion [43]:

$$\mathbf{I}_t = \bar{\mathbf{I}}_t + \mathbf{A}\mathbf{g}_t. \tag{16}$$

Fig. 9 shows some snapshots of the synthesized mouth images before (up) and after (bottom) the performance refinement process. Appearance details have been rewritten after the refinement. Finally, we add Gaussian noise to the synthesized image to regain the camera image sensing noise. The noise is estimated from the original facial images.

6.2. Overlaying onto a background sequence

To realize facial animation, we overlay the synthesized mouth animation onto a background sequence with natural head and eye movements. Since our system exhibits only movement in the mouth region, 'zombie'-like artifacts can be easily detected if we directly stitch the animated mouth with the background facial sequence. Therefore, we add natural jaw (lower face) movements by a jaw selection process. Since we find that the jaw downward action is approximately in proportion to the energy term of the acoustic signal, we associate an appropriate jaw mask from a jaw candidates set to each synthesized mouth image according to this proportion relationship. The jaw candidates set contains lower face image masks with different jaw movements selected from the JEWEL dataset. The synthesized mouth, the corresponding jaw and the face background video snippet are stitched together by the Poisson cloning technique [44] according to the manually labelled stitching positions.

7. Experiments

7.1. Experiment setup

We have compared the CHMMs with three models—HMMs, MSHMMs and FHMMs in performance of speech animation. The tested systems are summarized in Table 2, where $C(\bullet)$ denotes the cardinality of state variables. The systems named '*ph*-$*$' adopt only phoneme states for both audio and visual modalities, while the systems named '*ph-vi*-$*$' adopt phonemes state for audio and viseme states for video. Each phoneme (viseme) is modelled by five states, and the 13 visemes are

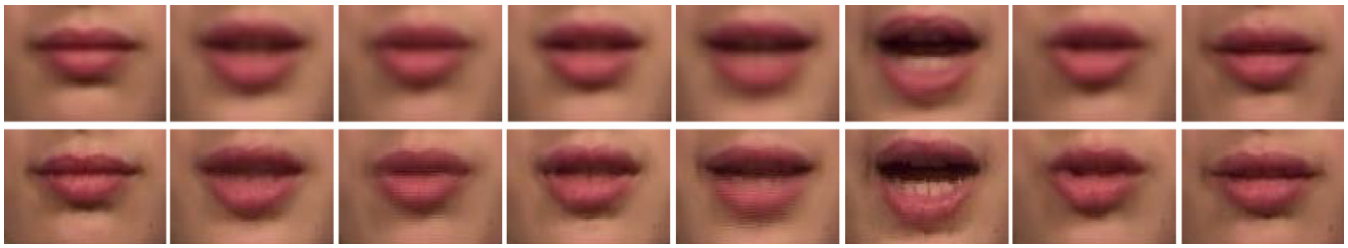

Fig. 9. Snapshots of the synthesized mouth images before (up) and after (bottom) the performance refinement process.

Table 2
The tested systems

System	Model structure	Conversion algorithm
ph-HMM	q_t describes phoneme state, $C(q_t) = 47 \times 5$	HMMI [16]
ph-MSHMM		
ph-CHMM	$q_t^s, s \in \{a, v\}$ describes phoneme state, $C(q_t^s) = 47 \times 5$	EM-based A/V conversion
ph-FHMM		
ph-vi-CHMM	q_t^a describes phoneme state, $C(q_t^a) = 47 \times 5$,	
ph-vi-FHMM	q_t^v describes viseme state, $C(q_t^v) = 13 \times 5$	

mapped from the 47 phonemes using Table 1. We performed an iterative mixture splitting scheme [42] to achieve the optimal Gaussian mixture numbers (i.e., K) using the JEWEL validation data set. All the systems were built using the same diagram illustrated in Fig. 1, except that the AV model training and the A/V conversion were carried out, respectively, for the four kinds of AV models. The *ph*-HMM system uses the HMMI [17] as the A/V conversion algorithm, while the other systems use the EM-based A/V conversion algorithm described in Section 4.

After empirical testing on the JEWEL validation data set, we chose the following constraints on state transitions to model the causality in speech generation and to decrease the model computation complexity. For HMMs and MSHMMs,

$$P(q_t|q_{t-1}) = 0 \quad \text{if } q_t \notin \{q_{t-1}, q_{t-1} + 1\}. \tag{17}$$

For FHMMs,

$$P(q_t^s|q_{t-1}^s) = 0 \quad \text{if } q_t^s \notin \{q_{t-1}^s, q_{t-1}^s + 1\}. \tag{18}$$

The above two constraints were set to ensure the state non-skip policy. For CHMMs, a further constraint on audio–visual asynchrony and synchrony was imposed:

$$P(q_t^s|\mathbf{q}_{t-1}) = 0 \quad \text{if } \begin{cases} q_t^s \notin \{q_{t-1}^s, q_{t-1}^s + 1\}, \\ |q_t^s - q_t^{s'}| \geqslant 2, \quad s' \neq s, \end{cases} \tag{19}$$

which ensured that only one-state asynchrony relationship was allowed between the audio and visual modalities.

Fig. 10 shows the convergence property of the EM-based A/V conversion algorithm with the above constraints. These curves were calculated using the *ph-vi*-CHMM system. Usually, the algorithm converges to a local minimum within a very few iterations.

7.2. Objective evaluations

We objectively evaluated the prediction performance on the JEWEL testing set using two quantitative measurements: the average mean square error (AMSE) ε and the average correlation coefficient (ACC) ρ, defined by

$$\varepsilon = \frac{1}{T} \sum_{t=1}^{T} ||\hat{\mathbf{o}}_t^v - \mathbf{o}_t^v||, \tag{20}$$

$$\rho = \frac{1}{T \cdot P^v} \sum_{t=1}^{T} \sum_{i=1}^{P^v} \frac{(o_{t,i}^v - \mu_{o_i^v})(\hat{o}_{t,i}^v - \mu_{\hat{o}_i^v})}{\sigma_{o_i^v} \sigma_{\hat{o}_i^v}}, \tag{21}$$

where $\mathbf{o}_t^v$ and $\hat{\mathbf{o}}_t^v$ denote the actual and predicted visual parameter vectors; $o_{t,i}^v$ and $\hat{o}_{t,i}^v$ are their ith coefficients, respectively. μ and σ are their mean and standard deviation. T is the total number of frames in the testing set. Table 3 shows the evaluation results.

From Table 3 we can see that the *ph*-MSHMM system achieves similar performance with the *ph*-HMM system. The two systems using FHMMs (*ph*-FHMM and *ph-vi*-FHMM) give improved performance compared with the *ph*-HMM and *ph*-MSHMM system. However, the introduction of visemes does not show great improvement for FHMMs. The CHMMs outperform all the other models tested. Especially, the *ph-vi*-CHMM system shows superior performance in predicting visual parameters with the lowest AMSE of 6.911 and the highest ACC of 0.696.

The FHMMs do not lead to good performance probably because the modelling advantage offered by FHMMs can only become evident if less correlated features are used. However, audio speech and visual speech are highly correlated since they are originated from the same articulation source. Moreover, independent state variables cannot embed appropriate (a)synchrony relationships between the audio and video. Instead, the

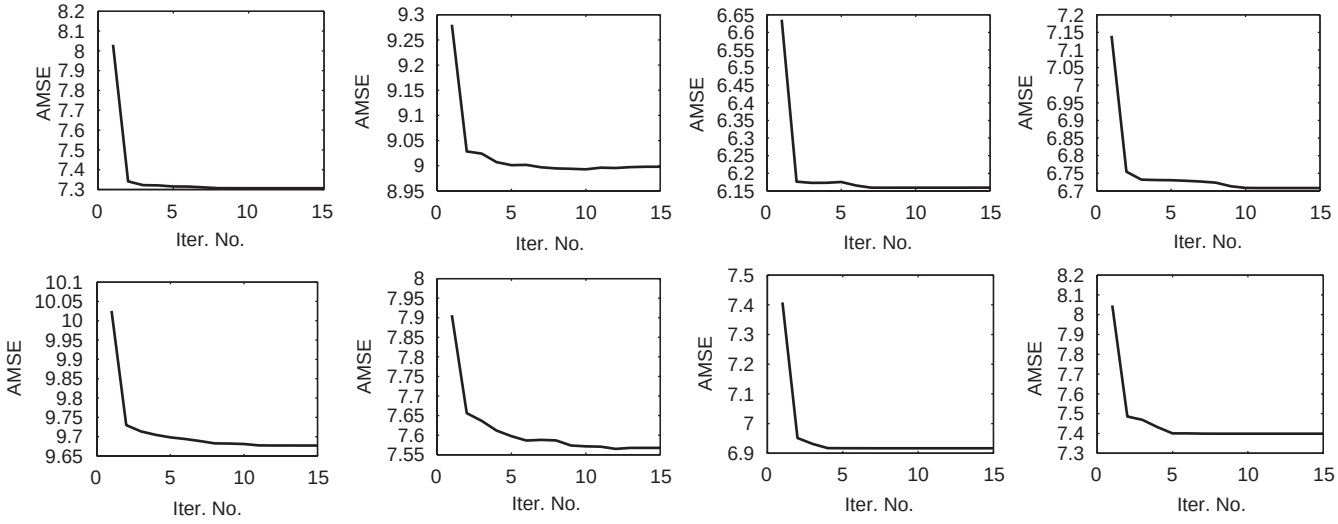

Fig. 10. Iteration curves of the EM-based A/V conversion algorithm for eight utterances from the JEWEL testing set.

Table 3
Objective evaluation results in terms of average MSE (ε) and average correlation coefficient (ρ)

System	AMSE (ε)	ACC (ρ)
ph-HMM	8.043	0.581
ph-MSHMM	8.050	0.576
ph-FHMM	7.778	0.593
ph-vi-FHMM	7.679	0.608
ph-CHMM	7.239	0.630
ph-vi-CHMM	6.911	0.696

CHMMs encapsulate the coupling relationship and the (a)synchrony between the two building blocks of speech by an appropriate structure and the imposed constraints, leading to better predicting performance. Fig. 11 shows some examples of the predicted and original trajectories for the first visual parameter (the first PCA coefficient for the red channel) for the six testing sentences: *(1) maybe it is taking longer to get things squared away than the bankers expected*; *(2) the staff deserves a lot of credit working down here under real obstacles*; *(3) to create such a lamp, order a wired pedestal from any lamp shop*; *(4) some observers speculated that this might be his revenge on his home town*; *(5) nobody really expects to evacuate*; *(6) why do we need bigger and better bombs*? These trajectories were generated by the *ph-vi*-CHMM system using the EM-based A/V conversion algorithm. We can clearly observe that the predicted visual parameters match the original ones very well.

7.3. Subjective evaluations

Since speech animation is to provide a natural human–machine communication method, subjective evaluations that human observers provide feedback on animation qualities are more appropriate than objective measurements. We first performed experiments on the JEWEL testing set (50 sentences) to evaluate the speaker dependent animation performance, and then carried out experiments on an AV subjective testing set collected from the Internet containing speech utterances from various speakers to evaluate the speaker independent animation performance. In both subjective evaluations, we synthesized the mouth video from audio using the predicted visual parameters (90 PCA coefficients in frames) as well as their detail augments achieved from the performance refinement process described in Section 6.1. As a benchmark, we also cropped the mouth sequences from the original JEWEL testing videos, and enrolled them in the speaker-dependent test.

We made only evaluations on the synthesized mouth region. Cosatto et al. [6] has suggested that for any subjective tests, it is important to separate the different factors influencing the perceived quality of speech. If the evaluations are performed on the whole face, the results might be affected by the motions and appearances from other facial parts other than the articulated mouth. Therefore, we eliminated other factors such as head movements and eye blinks, which enabled us to focus on the quality of articulation.

The AV subjective testing set contains 30 AV snippets with lengths from 15 s to 100 s containing contents about news reports, distinguished speeches, weather reports and interviews, etc. The synthesized mouth videos were overlaid onto the original videos as the subjective evaluation set (Fig. 12), which simulated a potential application for the hearing-impaired people perceiving speech by lipreading when watching Internet AV contents. A realistic, auxiliary, animated mouth can help hearing-impaired people to better understand what linguistic information the AV contents convey.

A group of 10 relatively inexperienced viewers were involved to rank the performance of the speech animation in terms of naturalness of the mouth matching the audio accompanied. We used a 5-point assessment, where 5 means ‘excellent’, and 1 means ‘bad’. To get impartial evaluations, the videos (synthesized and real) were randomly named and mixed together prior

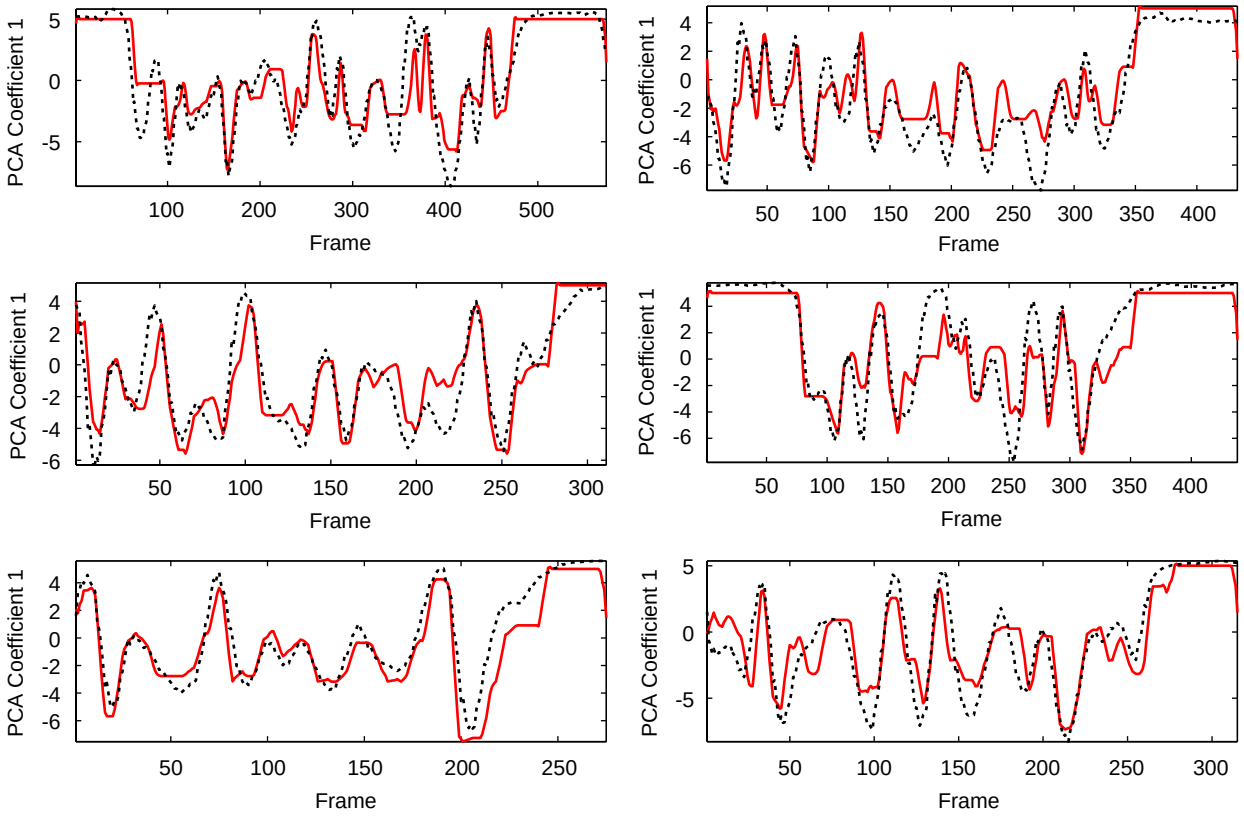

Fig. 11. Predicted (solid line) vs. actual (dotted line) trajectories for the first coefficient of visual parameters for six testing utterances.

Fig. 12. Snapshots of the synthesized speech animation overlaid onto the videos from the AV subjective testing set.

to the testing session. We used four specific criteria (smoothness, closure, protrusion, turning) as well as the overall performance in the assessments. The four criteria are known as the most important factors that viewers are sensitive to. Tables 4 and 5 summarize the mean opinion score (MOS) calculated on the two testing sets, respectively.

The subjective evaluation results in Tables 4 and 5 clearly show that, among all the tested models, the *ph-vi*-CHMM system demonstrates the best performances in both speaker dependent and independent tests with relative overall MOS improvements of 22.47% (speaker dependent) and 22.37% (speaker independent) as compared to the *ph*-HMM system. The *ph*-MSHMM system exhibits no obvious improvement. This is because the MSHMM structure in Fig. 3(b) still uses a single-state variable to synchronously model the audio and video, and the only difference is that the audio and visual observations

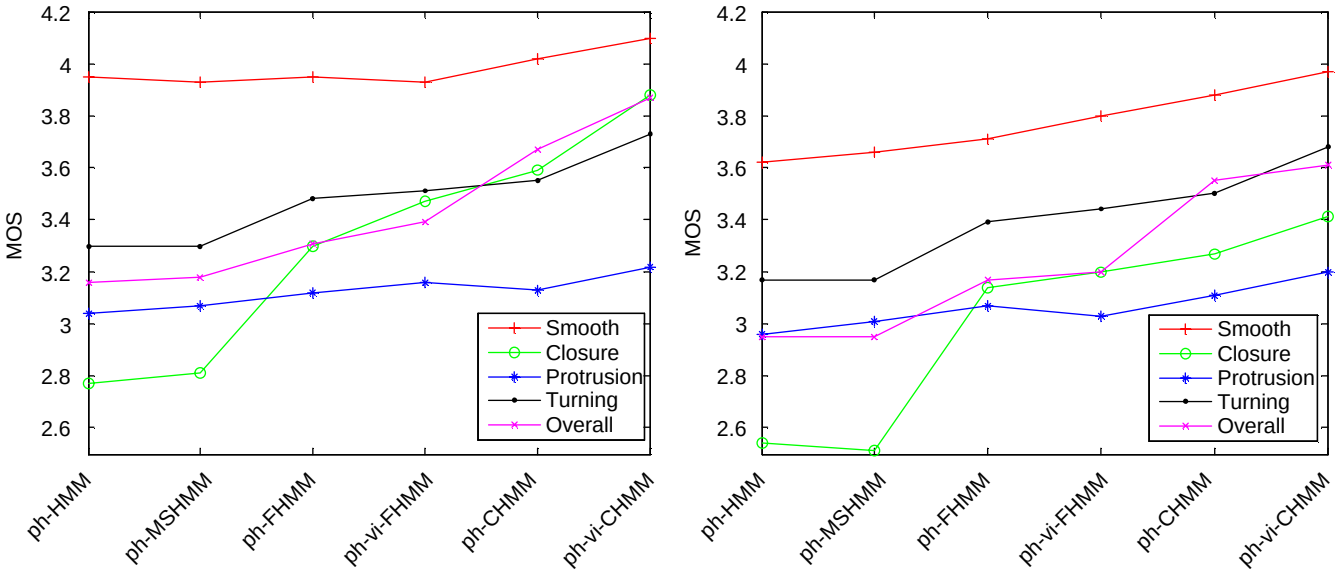

Fig. 13. MOS curves. Left: speaker-dependent test, right: speaker-independent test.

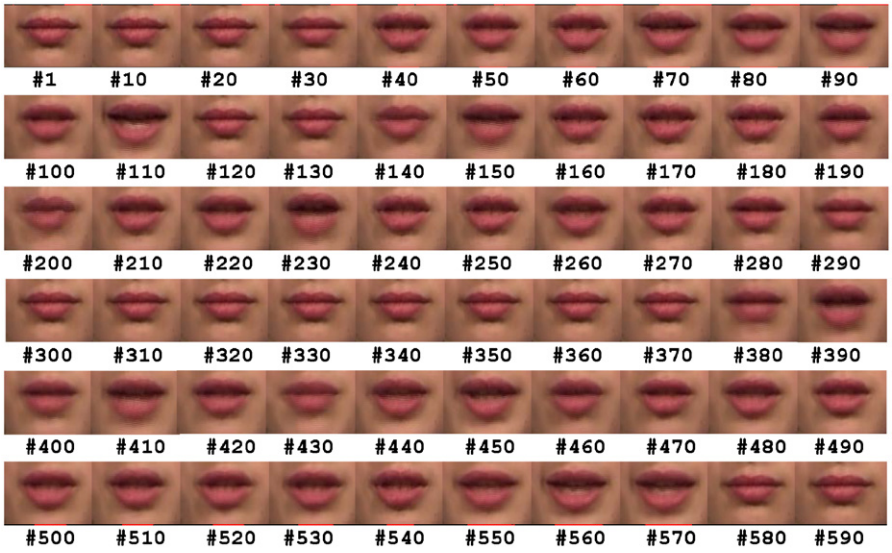

Fig. 14. Snapshots from the synthesized mouth animation for the utterance 'This is Tony Blair, the prime minister of United Kingdom. I'm glad to be able to speak to you today'.

are modelled by separate distributions. The FHMMs are also capable of effectively improving the subjective performance, but they are not as good as the CHMMs. This is understandable since the FHMMs lack inter-modal coupling dependencies. Generally speaking, introducing asynchrony does help a lot in the animation performance, while using different speech atoms for the two speech modalities can achieve further improvements. Comparing Table 4 with Table 5, we can observe that the performance of speaker-independent test is worse than that of speaker-dependent test. This is due to the following two major reasons: (1) speaker-dependent models can catch a specific subject's acoustic distributions more accurately than speaker independent models. (2) Many sentences in the AV subjective test set have much longer durations than that of the JEWEL testing set, and the defects can be easily detected within a relative long time.

Fig. 13 plots the MOS curves for different criteria. We can observe that richer structures for audio–visual speech have the greatest influence on the closure criterion, but have little influence on the smoothness and protrusion criteria. It unveils that the improvement of animation performance when using richer structures (like *ph-vi*-CHMMs) mainly comes from good prediction of closures.

The subjective results indicate that the CHMMs are more promising for speech animation. Fig. 14 demonstrates a mouth animation sequence synthesized by the *ph-vi*-CHMM system

Table 4
Subjective evaluation results on the JEWEL testing set

System	MOS					
	S	C	P	T	O	%RI
ph-HMM	3.95	2.77	3.04	3.30	3.16	—
ph-MSHMM	3.93	2.81	3.07	3.30	3.18	0.63
ph-FHMM	3.95	3.30	3.12	3.48	3.31	4.75
ph-vi-FHMM	3.93	3.47	3.16	3.51	3.39	7.28
ph-CHMM	4.02	3.59	3.13	3.55	3.67	16.14
ph-vi-CHMM	4.10	3.88	3.22	3.73	3.87	22.47
Real Recordings	4.76	4.70	4.51	4.80	4.71	49.05

S:smoothness, C:closure, P:protrusion, T:turning, O:overall and %RI:relative percentage MOS improvement as compared to *ph*-HMM.

Table 5
Subjective evaluation results on the AV subjective testing set

	MOS					
System	S	C	P	T	O	%RI
ph-HMM	3.62	2.54	2.96	3.17	2.95	—
ph-MSHMM	3.66	2.51	3.01	3.17	2.95	0
ph-FHMM	3.71	3.14	3.07	3.39	3.17	7.46
ph-vi-FHMM	3.80	3.20	3.03	3.44	3.20	8.47
ph-CHMM	3.88	3.27	3.11	3.50	3.55	20.34
ph-vi-CHMM	3.97	3.41	3.20	3.68	3.61	22.37

S: smoothness, C: closure, P: protrusion, T: turning, O: overall %RI: relative percentage MOS improvement as compared to *ph*-HMM.

for the first 6 s of an AV snippet named 'DL-Blair.avi' from the AV subjective testing set, where the mouth snapshots and their frame numbers are shown on the top and bottom, respectively.

8. Conclusions and future work

In this paper, we have proposed a CHMM approach to video-realistic speech animation. Motivated by the subtle relationships between audio speech and mouth movement, we use the CHMMs to explicitly model the synchrony, asynchrony, temporal coupling and different speech classes between the audio speech and visual speech. The CHMMs use two Markov chains to model the audio–visual asynchrony, while still preserving the natural correlations (i.e., synchrony) through inter-modal dependencies.

We have derived an EM-based A/V conversion algorithm on the CHMMs. Given an audio input and the trained AV-CHMMs, the algorithm outputs the optimal visual parameters (animation parameters) by maximizing the likelihood of these parameters under the ML criterion. We have proposed a facial animation system which learns AV-CHMMs and a mouth appearance space from AV recordings of a female subject as well as the TIMIT speech corpus. Based on the EM-based A/V conversion algorithm and the performance refinement process, the system is able to convert speaker-independent continuous speech to video-realistic facial animation.

We have compared the CHMMs with HMMs, MSHMMs, and FHMMs both objectively and subjectively. Evaluation results show that the CHMM system using phonemes and visemes (*ph-vi*-CHMM) demonstrates superior performance among all the tested systems. The promising result indicates that to animate speech more naturally, it is necessary to explicitly describe the temporal relationships between the two building blocks of speech, namely audio modality and visual modality.

Since explicit modelling of speech is more robust to speech degradation and training–testing mismatch, we are currently trying to realize speech animation under adverse acoustic conditions, e.g., ambient noise and microphone mismatch.

Appendix A. Parameter training for CHMMs

We derive the EM parameter training for CHMMs. Since the initial parameters are essential in achieving good model estimates, we use the Viterbi algorithm in the E Step to get reasonable initial values of model parameters [36].

E Step: The forward probability

$$\alpha_t(\mathbf{q}_t) = P(\mathbf{q}_t|\mathbf{o}^{av}_{1:t}) \tag{A.1}$$

and the backward probability

$$\beta_t(\mathbf{q}_t) = P(\mathbf{o}^{av}_{t+1:T}|\mathbf{q}_t) \tag{A.2}$$

are computed using the frontier algorithm [32] that is a general inference algorithm for DBNs. Note that the forward probability is defined differently with the conventional HMM.

The frontier algorithm sweeps a Markov blanket across the model, first forward then backward, ensuring the frontier (denoted by $\mathscr{F}$) d-separates the past (the left of the frontier denoted by $\mathscr{L}$) from the future (the right of the frontier denoted by $\mathscr{R}$) [32].

Forward pass: The frontier initially contains all the nodes in frame $t-1$:

$$\mathscr{F}_{t-1,0} = \alpha_{t-1}(\mathbf{q}_{t-1}) = P(\mathbf{q}_{t-1}|\mathbf{o}^{av}_{1:t-1}). \tag{A.3}$$

First we add nodes $\mathbf{q}_t$ to the frontier since all their parents are already in the frontier. To do this, we multiply in their CPDs $P(\mathbf{q}_t|\mathbf{q}_{t-1})$

$$\begin{aligned}\mathscr{F}_{t-1,1} &= P(\mathbf{q}_{t-1:t}|\mathbf{o}^{av}_{1:t-1}) \\ &= P(\mathbf{q}_t|\mathbf{q}_{t-1}) \cdot \mathscr{F}_{t-1,0},\end{aligned} \tag{A.4}$$

where $P(\mathbf{q}_t|\mathbf{q}_{t-1}) = \prod_s P(q^s_t|\mathbf{q}_{t-1})$. Second we remove $\mathbf{q}_{t-1}$ by marginalizing it out because all their children are in the frontier

$$\mathscr{F}_{t-1,2} = P(\mathbf{q}_t|\mathbf{o}^{av}_{1:t-1}) = \sum_{\mathbf{q}_{t-1}} \mathscr{F}_{t-1,1}. \tag{A.5}$$

Finally we add $\mathbf{o}^{av}_t$ to the frontier since all their parents are already in the frontier:

$$\begin{aligned}\mathscr{F}_{t-1,3} &= P(\mathbf{q}_t|\mathbf{o}^{av}_{1:t}) \\ &= P(\mathbf{o}^{av}_t|\mathbf{q}_t) \cdot \mathscr{F}_{t-1,2} \\ &= \mathscr{F}_{t,0} = \alpha_t(\mathbf{q}_t),\end{aligned} \tag{A.6}$$

where $P(\mathbf{o}^{av}_t|\mathbf{q}_t) = \prod_s P(\mathbf{o}^s_t|q^s_t)$.

Backward pass: The backward pass advances the frontier from frame t to $t-1$ by adding and removing nodes in the opposite order of the forward pass. The frontier initially contains all the nodes in frame t:

$$\mathscr{F}_{t,0} = \beta_t(\mathbf{q}_t) = P(\mathbf{o}_{t+1:T}^{av}|\mathbf{q}_t). \tag{A.7}$$

First we remove $\mathbf{o}_t^{av}$:

$$\mathscr{F}_{t,1} = P(\mathbf{o}_{t:T}^{av}|\mathbf{q}_t) = P(\mathbf{o}_t^{av}|\mathbf{q}_t) \cdot \mathscr{F}_{t,0}. \tag{A.8}$$

Second we add $\mathbf{q}_{t-1}$:

$$\mathscr{F}_{t,2} = P(\mathbf{o}_{t:T}^{av}|\mathbf{q}_{t-1:t}) = P(\mathbf{o}_{t:T}^{av}|\mathbf{q}_t) = \mathscr{F}_{t,1}. \tag{A.9}$$

Then we remove $\mathbf{q}_t$:

$$\begin{aligned}\mathscr{F}_{t,3} &= P(\mathbf{o}_{t:T}^{av}|\mathbf{q}_{t-1}) \\ &= \sum_{\mathbf{q}_t} P(\mathbf{q}_t, \mathbf{o}_{t:T}^{av}|\mathbf{q}_{t-1}) \\ &= \sum_{\mathbf{q}_t} P(\mathbf{q}_t|\mathbf{q}_{t-1}) P(\mathbf{o}_{t:T}^{av}|\mathbf{q}_{t-1:t}) \\ &= \sum_{\mathbf{q}_t} P(\mathbf{q}_t|\mathbf{q}_{t-1}) \cdot \mathscr{F}_{t,2} \\ &= F_{t-1,0} = \beta_{t-1}(\mathbf{q}_{t-1}).\end{aligned} \tag{A.10}$$

The probability of the lth observation sequence $\mathbf{O}_l^{av}$ with length T_l is computed as

$$P_l = \alpha_{l,T_l}(\mathbf{q}_{l,T_l}) = \beta_{l,1}(\mathbf{q}_{l,1}). \tag{A.11}$$

M Step: The forward and backward probabilities obtained in the E step are used to reestimate the following parameters:

$$\mu_{q_t^s k} = \frac{\sum_l (1/P_l) \sum_t \gamma_{l,t}(q_t^s, k) \mathbf{o}_{l,t}^s}{\sum_l (1/P_l) \sum_t \gamma_{l,t}(q_t^s, k)}, \tag{A.12}$$

$$\Sigma_{q_t^s k} = \frac{\sum_l (1/P_l) \sum_t \gamma_{l,t}(q_t^s, k)(\mathbf{o}_{l,t}^s - \mu_{q_t^s k})(\mathbf{o}_{l,t}^s - \mu_{q_t^s k})^{\mathrm{T}}}{\sum_l (1/P_l) \sum_t \gamma_{l,t}(q_t^s, k)}, \tag{A.13}$$

$$w_{q_t^s k} = \frac{\sum_l (1/P_l) \sum_t \gamma_{l,t}(q_t^s, k)}{\sum_l (1/P_l) \sum_t \sum_{k'} \gamma_{l,t}(q_t^s, k')}, \tag{A.14}$$

where

$$\gamma_{l,t}(q_t^s, k) = \frac{\sum_{\mathbf{q}_t'} \alpha_{l,t}(\mathbf{q}_t') \beta_{l,t}(\mathbf{q}_t')}{\sum_{\mathbf{q}_t} \alpha_{l,t}(\mathbf{q}_t) \beta_{l,t}(\mathbf{q}_t)} \times \frac{w_{q_t^s k} \mathscr{N}(\mathbf{o}_{l,t}^s, \mu_{q_t^s k}, \Sigma_{q_t^s k})}{\sum_{k'} w_{q_t^s k'} \mathscr{N}(\mathbf{o}_{l,t}^s, \mu_{q_t^s k'}, \Sigma_{q_t^s k'})} \tag{A.15}$$

is called the state occupation probability; and $\mathbf{q}_t'$ can be any state vector such that $q_t^s \in \mathbf{q}_t'$.

The state transition probabilities $P(q_t^s = i^s|\mathbf{q}_{t-1} = \mathbf{j})$ can be estimated by

$$P(i^s|\mathbf{j}) = \frac{\sum_l (1/P_l) \sum_{\mathbf{i}} \sum_t \alpha_{l,t}(\mathbf{j}) P(\mathbf{i}|\mathbf{j}) P(\mathbf{o}_{l,t+1}^{av}|\mathbf{i}) \beta_{l,t+1}(\mathbf{i})}{\sum_l (1/P_l) \sum_t \alpha_{l,t}(\mathbf{j}) \beta_{l,t}(\mathbf{j})}, \tag{A.16}$$

where $P(\mathbf{i}|\mathbf{j}) = \prod_s P(i^s|\mathbf{j})$, and vectors $\mathbf{i}$ and $\mathbf{j}$ can be any state vectors such that $i^s \in \mathbf{i}$ and $j^s \in \mathbf{j}$.

References

[1] R. Lippman, Speech recognition by machines and humans, Speech Commun. 22 (1) (1997) 1–15.

[2] J. Ostermann, A. Weissenfeld, Talking faces—technologies and applications, in: Proceedings of ICPR'04, vol. 3, 2004, pp. 826–833.

[3] M.M. Cohen, D.W. Massaro, Modeling coarticulation in synthetic visual speech, in: M. Magnenat-Thalmann, D. Thalmann (Eds.), Models and Techniques in Computer Animation, Springer, Tokyo, 1993, pp. 139–156.

[4] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D.H. Salesin, Synthesizing realistic facial expressions from photographs, in: Proceedings of ACM SIGGRAPH'98, vol. 3, 1998, pp. 75–84.

[5] L. Yin, A. Basu, S. Bernogger, A. Pinz, Synthesizing realistic facial animations using energy minimization for model-based coding, Pattern Recognition 34 (11) (2001) 2201–2213.

[6] E. Cosatto, J. Ostermann, H.P. Graf, J. Schroeter, Lifelike talking faces for interactive services, Proc. IEEE 91 (9) (2003) 1406–1428.

[7] C. Bregler, M. Covell, M. Slaney, Video rewrite: driving visual speech with audio, in: Proceedings of ACM SIGGRAPH'97, 1997.

[8] T. Ezzat, G. Geiger, T. Poggio, Trainable videorealistic speech animation, in: Proceedings of ACM SIGGRAPH, 2002, pp. 388–397.

[9] E. Cosatto, H. Graf, Sample-based synthesis of photo-realistic talking heads, in: Proceedings of IEEE Computer Animation, 1998, pp. 103–110.

[10] E. Cosatto, H. Graf, Photo-realistic talking heads from image samples, IEEE Trans. Multimedia 2 (3) (2000) 152–163.

[11] P. Hong, Z. Wen, T.S. Huang, Real-time speech-driven face animation with expressions using neural networks, IEEE Trans. Neural Networks 13 (4) (2002) 916–927.

[12] F.J. Huang, T. Chen, Real-time lip-synch face animation driven by human voice, in: IEEE Second Workshop on Multimedia Signal Processing, 1998, pp. 352–357.

[13] R.R. Rao, T. Chen, R.M. Mersereau, Audio-to-visual conversion for multimedia communication, IEEE Trans. Ind. Electron. 45 (1) (1998) 15–22.

[14] E. Yamamoto, S. Nakamura, K. Shikano, Lip movement synthesis from speech based on Hidden Markov Models, Speech Commun. 26 (1–2) (1998) 105–115.

[15] M. Brand, Voice puppetry, in: SIGGRAPH'99, Los Angeles, 1999, pp. 21–28.

[16] K. Choi, J. N. Hwang, Baum–Welch hidden Markov model inversion for reliable audio-to-visual conversion, in: Proceedings of the IEEE 3rd Workshop Multimedia Signal Processing, 1999, pp. 175–180.

[17] K. Choi, Y. Luo, J.N. Hwang, Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system, J. VLSI Signal Process. 29 (1–2) (2001) 51–61.

[18] S. Lee, D. Yook, Audio-to-visual conversion using hidden Markov models, in: M. Ishizuka, S. A. (Eds.), Proceedings of PRICAI2002, Lecture Notes in Artificial Intelligence, Springer, Berlin, 2002, pp. 563–570.

[19] L. Xie, D.-M. Jiang, I. Ravyse, W. Verhelst, H. Sahli, V. Slavova, R.-C. Zhao, Context dependent viseme models for voice driven animation, in: The 4th EURASIP Conference on Video/Image Processing and Multimedia Communications, vol. 2, 2003, pp. 649–654.

[20] P.S. Aleksic, A.K. Katsaggelos, Speech-to-video synthesis using MPEG-4 compliant visual features, IEEE Trans. Circuits Systems Video Technol. 14 (5) (2004) 682–692.

[21] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech animation, Proc. IEEE 77 (2) (1989) 257–286.

[22] S. Fu, R. Gutierrez-Osuna, A. Esposito, K.P. Kakumanu, O.N. Garcia, Audio/visual mapping with cross-modal hidden Markov models, IEEE Trans. Multimedia 7 (2) (2005) 243–251.

[23] S.Y. Moon, J.N. Hwang, Noisy speech recognition using robust inversion of hidden Markov models, in: Proceedings of ICASSP'95, 1995, pp. 145–148.

[24] T. Ezzat, T. Poggio, Miketalk: A talking facial display based on morphing visemes, in: Proceedings of the Computer Animation Conference, 1998, pp. 96–102.

[25] D.G. Stork, M.E. Hennecke (Eds.), Speechreading by Humans and Machines, Springer, Berlin, 1996.
[26] L. Xie, Research on key issues of audio visual speech recognition, Ph.D. Thesis, Northwestern Polytechnical University, September 2004.
[27] K.W. Grant, S. Greenberg, Speech intelligibility derived from asynchronous processing of auditory-visual information, in: Proceedings of the International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark, 2001, pp. 132–37.
[28] F. Lavagetto, Converting speech into lip movements: a multimedia telephone for hard hearing people, IEEE Trans. Rehabil. Eng. 3 (1995) 90–102.
[29] M. McGrath, Q. SummerLeld, Intermodal timing relations and audio–visual speech recognition, J. Acoust. Soc. Am. 77 (1985) 678–685.
[30] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audio–visual speech, Proc. IEEE 91 (9) (2003) 1306–1326.
[31] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, Berlin, 2001.
[32] K. Murphy, Dynamic Bayesian networks: representation, inference and learning, Ph.D. Thesis, University of California, Berkeley, 2002.
[33] H. Bourlard, S. Dupont, A new ASR approach based on independent processing and recombination of partial frequency bands, in: Proceedings of the International Conference on Spoken Language Processing, Philadelphia, 1996, pp. 426–429.
[34] B. Logan, P.J. Moreno, Factorial hidden Markov models for speech recognition: preliminary experiments, Technical Reports of Cambridge Research Lab (CRL-97-7).
[35] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, in: IEEE International Conference on Computer Vision and Pattern Recognition, 1997, pp. 994–999.
[36] A.V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, K. Murphy, A coupled HMM for audio–visual speech recognition, in: Proceedings of ICASSP'02, 2002.
[37] F. Pernkopf, 3D surface inspection using coupled HMMs, in: Proceedings of 17th ICPR'04, 2004.
[38] S. Ananthakrishnan, S.S. Narayanan, An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model, in: Proceedings of ICASSP'05, 2005.
[39] L. Xie, Z. Ye, The JEWEL audio visual dataset for facial animation, URL ⟨http://www.cityu.edu.hk/rcmt/mouth-synching/jewel.htm⟩.
[40] L. Xie, X.-L. Cai, R.-C. Zhao, A robust hierarchical lip tracking approach for lipreading and audio visual speech recognition, in: The 3rd IEEE International Conference on Machine Learning and Cybernetics, vol. 6, Shanghai, China, 2004, pp. 3620–3624.
[41] A. Dempster, A.N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Statist. Soc. (Ser. B) 39 (1977) 89–111.
[42] S. Young, G. Evermann, D. Kershaw, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (Version 3.2), Cambirdge University Engineering Department, Cambridge, 2002, URL ⟨http://htk.eng.cam.ac.uk/⟩.
[43] S. Dupont, J. Luettin, Audio–visual speech modelling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[44] P. Pèrez, M. Gangnet, A. Blake, Poisson image editing, ACM Trans. Graphics (SIGGRAPH) 22 (3) (2003) 313–318.

About the Author—LEI XIE received the B.Eng., M.Eng. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1999, 2001 and 2004, respectively, all in computer science. He was granted IBM Excellent Chinese Student Awards twice in 1999 and 2002. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium as a Visiting Scientist. From 2004 to 2006, he was a Senior Research Associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Hong Kong SAR, China. Dr. Xie is currently a Postdoctoral Fellow in the Human–Computer Communications Laboratory (HCCL), Department of Systems Engineering & Engineering Management, the Chinese University of Hong Kong. His current research interest includes talking face, multimedia retrieval, speech recognition, multimedia signal processing and pattern recognition.

About the Author—ZHI-QIANG LIU received the M.A.Sc. degree in Aerospace Engineering from the Institute for Aerospace Studies, The University of Toronto, and the Ph.D. degree in Electrical Engineering from The University of Alberta, Canada. He is currently with School of Creative Media, City University of Hong Kong. He has taught computer architecture, computer networks, artificial intelligence, programming languages, machine learning, pattern recognition, computer graphics, and art & technology.
His interests are scuba diving, neural-fuzzy systems, painting, gardening, machine learning, mountain/beach trekking, human–media systems, horse riding, computer vision, serving the community, mobile computing, computer networks, and fishing.