Realistic Mouth-Synching for Speech-Driven Talking Face Using Articulatory Modelling

Lei Xie and Zhi-Qiang Liu, *Senior Member, IEEE*

***Abstract*—This paper presents an articulatory modelling approach to convert acoustic speech into realistic mouth animation. We directly model the movements of articulators, such as lips, tongue, and teeth, using a dynamic Bayesian network (DBN)-based audio-visual articulatory model (AVAM). A multiple-stream structure with a shared articulator layer is adopted in the model to synchronously associate the two building blocks of speech, i.e., audio and video. This model not only describes the synchronization between visual articulatory movements and audio speech, but also reflects the linguistic fact that different articulators evolve asynchronously. We also present a Baum–Welch DBN inversion (DBNI) algorithm to generate optimal facial parameters from audio given the trained AVAM under maximum likelihood (ML) criterion. Extensive objective and subjective evaluations on the JEWEL audio-visual dataset demonstrate that compared with phonemic HMM approaches, facial parameters estimated by our approach follow the true parameters more accurately, and the synthesized facial animation sequences are so lively that 38% of them are undistinguishable.**

***Index Terms*—Articulatory model, Baum–Welch DBN inversion (DBNI), dynamic Bayesian networks (DBNs), facial animation, mouth-synching, talking face.**

I. Introduction

COMPUTER-ANIMATED talking faces have become more popular in multimedia applications, such as newsreaders, online virtual avatars, video games, and videophones. Experiments show that the trust and attention of humans towards machines are able to increase by 30% if humans are communicating with talking faces instead of text-only [1]. However, realistic facial animation still remains to be one of the most challenging tasks despite decades of extensive research. This is mainly due to the fact that the mechanisms of human facial expressions are not yet well understood.

According to the underlying face/head model, talking faces can be 3-D-model-based or image-based [2]. The former approaches usually start with a mesh of 3-D polygons that define the head shape, which can be deformed parametrically to perform facial actions. A texture-image is mapped over the mesh to render the skin and facial parts [3]. As another kind of 3-D approach, the physics-based animation makes use of laws of physics or muscle forces based on the anatomical structure of a human face [4]. Targeting to video-realistic performance, the image-based approaches [2], [5]–[8] use recorded image sequences of faces and render the facial movements directly at image level. Ezzat *et al.* [6] proposed a multidimensional morphable model (MMM), which is capable of morphing between 46 prototype mouth images statistically collected from a small sample set. Cosatto *et al.* [2], [7], [8] described another image-based approach with higher realism and flexibility, which searched within a large database of recorded motion samples for the closest matches. According to the input, talking faces can be text-driven or speech-driven [2]. Although most state-of-the-art text-driven talking faces employ concatenative speech synthesizers [9], they still lack natural speech prosody and emotions. Therefore, many researchers investigate how to drive a talking face from real human speech, i.e., speech-driven. These approaches use speech signals to generate more natural facial animations with high fidelities of both audio and video.

The essential problem of speech-driven talking face is *mouth-synching*: the synthesis of mouth movements matching an input audio naturally. Actually, the mouth-synching problem can be considered theoretically as an audio-to-visual conversion (or mapping) problem, which is rather complicated due to the *co-articulation* phenomena [10]. The 3-D-model-based approaches usually use articulation rules [10] to capture the speech dynamics, while the image-based approaches implicitly model co-articulation by capturing and rewriting various video segments of articulation. During the last two decades, machine learning methods have been used extensively in the mapping problem, such as vector quantization (VQ), neural networks (NNs) [11], time-delay neural networks (TDNN) [12], and hidden Markov models (HMMs) [5], [13]–[19], [21].

HMMs [13] have recently been used in mouth-synching. One of the earliest HMM approaches was proposed by Yamamoto *et al.* [14]. They trained HMMs from audio data, and aligned the corresponding visual parameters to the HMM states. During synthesis, they used the Viterbi algorithm [13] to collect a time-labelled HMM state sequence, and then the visual parameters associated with each state were retrieved by mapping. This simple approach achieved jerky animations since the visual parameters were just averages of the Gaussian mixture components associated to the current state, and these visual parameters were only indirectly related to the current audio input. Interpolating, splining and morphing might eliminate the jerkiness, but these *ad hoc* solutions ignore the natural speech dynamics, resulting in limited success [15].

To avoid jerkiness and preserve dynamics, Chen *et al.* [16] proposed a least-mean squared HMM (LMS-HMM) method

Manuscript received September 27, 2005; revised July 18, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jie Yang.

The authors are with the School of Creative Media, City University of Hong Kong, Hong Kong, China (e-mail: xielei21st@gmail.com; zq.liu@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2006.888009

using joint audio-visual HMMs, where the visual output was made dependent not only on the current state, but also on the current audio input. They trained AV-HMMs using joint audio-visual observations, and extracted the distributions of audio HMMs from the AV-HMMs. The synthesis involved two steps. First, an optimal state sequence was found using the Viterbi algorithm. Second, the audio input and the Gaussian mixtures associated to each state were used to derive visual parameters by a least-mean-square regression.

Bregler *et al.* [5] described an image-based talking face called *Video Rewrite* that used triphones (three consecutive phonemes) to model co-articulations. Jerkiness was avoided by using more speech segments and smooth concatenation of real video clips. As a similar approach, Cao *et al.* [17] organized the phonemic video clips by an *Anime Graph*, and proposed a greedy graph search algorithm to improve the efficiency of the clip selection process.

Since the above approaches solely rely on the phonemic sequences derived from speech recognition, the synthesis performance heavily depends on the Viterbi search. If the speech is contaminated by noise, the incorrect HMM state assignments achieved by the Viterbi search may result in incorrect mouth animation. Moreover, the Viterbi sequence may represent only a small fraction of the total probability mass, and many other slightly different state sequences are nearly as likely [15]. Brand [15] proposed a remapping HMM method and an entropy minimization training scheme in the *Voice Puppetry*, which enabled the Viterbi sequence to capture a large proportion of the total probability mass.

Choi *et al.* [18] presented a Baum–Welch HMM inversion (HMMI) approach, thus avoiding the Viterbi search. As the dual procedure to the Baum–Welch HMM re-estimation [13], HMMI was first proposed for robust speech recognition [19]. Choi *et al.* [18] extended this algorithm to audio-visual data space. After training of audio-visual phoneme HMMs (we call this model AVPM) using joint audio-visual observations, optimal visual parameters were generated directly by Baum–Welch iterations under maximum likelihood (ML) sense. Later they implemented the HMMI in their *Virtual-Face* [20] in a netmeeting application. More recently, Fu *et al.* [21] demonstrated that the HMMI method outperformed the remapping HMM and the LMS-HMM on a common test bed.

Even though these HMM-based approaches can provide reasonable lip movements, they are still far from natural compared with real recordings. This is probably because these approaches adopt *phoneme-based* or *word-based* HMM modelling, which do not incorporate any knowledge of the source that produced the speech.

In this paper, we propose an audio-visual articulatory model (AVAM) for mouth-synching in speech-driven talking face. In contrast to phoneme-based HMM techniques, to reflect co-articulation explicitly and concisely, we use dynamic Bayesian networks (DBNs) to model the articulator actions, simulating the speech production process. We propose the Baum–Welch DBN inversion (DBNI) algorithm, which directly converts audio to visual parameters under ML criterion given the trained AVAMs, preserving the speech dynamics. We built up an image-based mouth animation framework to test the proposed approach. Experimental results show that compared with the AVPM [20] and its triphone variant [5], the proposed AVAM can generate more realistic mouth animations.

The following section describes our DBN-based articulatory modelling technique. In Section III, the Baum–Welch DBN inversion algorithm for audio-to-visual conversion is proposed in detail. Section IV provides the objective and subjective evaluations. Conclusions are drawn in Section V.

II. Articulatory Modelling

A. Motivations

The autosegmental phonology [22] holds the view that speech is produced not from a single stream of phonemes, but from multiple streams of linguistic features. These features can evolve asynchronously and do not necessarily form phonetic segments [22]. Generally speaking, linguistic features include tone, duration, and articulators. We know that speech is formed by the glottal excitement of the vocal tract comprised of articulators which shape the sound in complex ways. Therefore, a model that directly simulates the articulator configurations could improve the mouth-synching performance.

Our previous experimental results [23] on audio-visual speech recognition have shown that articulatory models outperform the phoneme-based models and provide abundant orofacial motion information which can improve the speech intelligibility. Articulator modelling has many advantages such as being better able to predict co-articulation effects. Furthermore, by modelling articulators, we allow asynchrony between their configurations, which may more accurately model speech production as autosegmental phonology has indicated. We have also seen mounting evidences that a phoneme-based model of speech is inadequate for speech modelling, especially for spontaneous, conversational speech. Phoneme-based models for speech recognition do not explicitly incorporate any knowledge of the speech production source. Moreover, speech recognition systems based on articulator features (AFs) are more robust to noise and reverberation [24].

Finally in view of visual articulation, since our goal is to derive orofacial motions from speech signals, a model directly analogous to the human articulatory system should better reflect the co-articulation phenomenon, leading to more realistic facial animations. Current co-articulation engines in facial animation are solely derived from physical gestural theories of speech production, such as articulation rules of facial muscles, while they ignore the consanguinity between visual speech and audio speech. Facial animation may benefit from audio-visual associations by directly describing the actions of lips, tongue and teeth in the two building blocks of speech.

B. Dynamic Bayesian Networks

During the last decade, Bayesian networks (BNs) [25] have become popular in many fields due to their great expressive power and capability in inference and learning. Recently, the use of BNs in speech modelling has gained much attention [26]. Although HMMs are widely used in speech modelling, they have inherent drawbacks in describing real-world speech

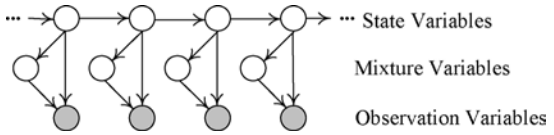


Fig. 1. DBN representation of HMM.

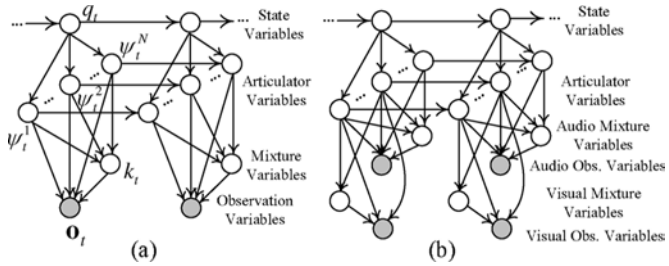


Fig. 2. (a) Audio (or visual) articulatory model and (b) audio-visual articulatory model. For clarity, only two frames are drawn.

phenomena, such as gender and age differences, pronunciation variability, and channel variability, due to their limited expressive power. In speech modelling, the widely used left-to-right HMMs allow only one hidden state in each time frame. In contrast, BNs have the ability to express arbitrary sets of variables in each time frame and interpret causality between variables, and offer a highly systematic and unified means to model the details of speech phenomena.

Dynamic Bayesian networks (DBNs) [26] extend the BN framework by representing multiple collections of random variables as they evolve over time. Fig. 1 illustrates a simple left-to-right HMM represented by DBN. Each time frame possesses three variables: the hidden state variable, the mixture variable and the observation variable. Compared with the conventional HMM, time is explicitly depicted in the unrolling DBN diagram, i.e., each time frame gets its own separate segment in the model. Fig. 1 represents exactly four time frames; to represent longer time series requires more segments.

C. Articulatory Modelling Via DBNs

Although HMMs can model articulators by encoding every possible combination of articulator values as a separate state [27], this indirect modelling is cumbersome. DBNs provide a fairly direct and natural way to mimic articulators, since they allow for an arbitrary number of variables and flexible model structures. Bilmes *et al.* [28] proposed a prototype of DBN-based articulatory model for speech recognition with multiple layers of variables explicitly representing words, phonemes and sub-phonemes. We extend this model to solving mouth-synching problem for speech-driven talking face.

Our articulatory model for a single observation stream (audio or video) is shown in Fig. 2(a), which is called the *Audio (or Visual) Articulatory Model* (AAM and VAM). Similar to that in [28], a layer representing various articulators is incorporated between state and observation variables. The state variable usually describes phonemes or sub-phonemes. Each articulator variable represents a particular kind of articulator feature, such as *voicing*, *velum*, *lipRounding*, and *tongueShow*. The articulator feature set and their values used in this paper are defined in Section II-D. As indicated in [28], the value of an articulator variable depends on its own value in the previous frame as well as the current state. The dependency on its previous frame is to model the natural continuity constraints on feature values, since articulators cannot change from one value to another totally different one without going through intermediate value(s). For example, lips cannot change from "round" to "wide" without passing through "mid". This model structure also allows articulators to change their values independently and asynchronously, reflecting the linguistic fact indicated in Section II-A [22]. The local conditional probability distribution (CPD) of $P(\mathbf{o}_t|\Psi_t, k_t)$ is defined as

$$P(\mathbf{o}_t|\Psi_t, k_t) = b_i(\mathbf{o}_t) = \sum_{k=1}^{K} \omega_{ik}\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}) \quad (1)$$

where $\mathbf{o}_t$ and k_t denote the observation and the mixture component at time frame t, respectively. $\Psi_t = \{\psi_t^1, \psi_t^2, \ldots, \psi_t^N\}$ denotes the value set of the N articulators at t, and i denotes a possible value set of articulators. K is the total number of mixture components and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, that is

$$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^P|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{o}-\boldsymbol{\mu})\right) \quad (2)$$

where P is the dimensionality of observation $\mathbf{o}$.

Compared with the articulatory model for speech recognition in [28], our model removes the complicated syntax layers used in word decoding, since we are not interested in the syntax the utterance conveys, but how the visual observations match the acoustic speech. Another difference in the model structure is that we include a layer of mixture variables to describe the causal probability between articulators and their outcomes–observed audio or visual data.

To describe audio-visual speech in an articulatory way, we further extend the model in Fig. 2(a) to multiple observation streams as shown in Fig. 2(b). Correspondingly the *Audio-Visual Articulatory Model* (AVAM) is composed of two observation streams each of which describes one modality of speech—audio or visual. Since the audio and visual observations are generated from the same articulation source, only a shared articulator layer is incorporated. As several articulators such as *velum* cannot be observed visually and only visible articulators contribute to the visual building block of speech, visual observations are up-linked only to those visible articulators. Not only does the subtle structural design mimic the true human articulatory system, but also reduces the number of variables. This multistream structure also encapsulates the synchronization between the audio and video, and may lead to a better mouth-synching performance. Since not all articulators contribute to the visual modality, we use separate mixture variables for audio and visual streams.

D. Articulatory Features

Articulator features specify the states of vocal tract directly or implicitly over time. For acoustic speech production, glottis, velum, tongue and lips are the most important articulators [28]. Visual capture system records only the visible articulators such as lips, frontal tongue, teeth, and jaw. These visible articulators

TABLE I
ARTICULATOR FEATURE SET FOR ARTICULATORY MODELLING

Label	Feature	Values
ψ^1	voicing	on, off
ψ^2	velum	open, closed
ψ^3	manner	closure, sonorant, fricative, burst
ψ^4	lipRounding	rounded, slightly rounded, mid, wide
ψ^5	tongueShow	touching top teeth, near alveolar ridge, touching alveolar, others
ψ^6	teethShow	on, off

affect the post formulation of uttering. Lipreading [23] experiments tell us that these articulators do provide discriminative information for linguistic classification.

Considering the variable space, computing complexity, and coverage of phonemes, we manually define an articulator feature set (shown in Table I) for our articulatory modelling according to the evolution of vocal tract. Note that *lipRongding*, *tongueShow* and *teethShow* are visible articulator features directly corresponding to the actions of lips, tongue and teeth. The features are subjectively quantized to discrete values based on the physical actions or articulator manners. For example, the *velum* usually has two actions, i.e., open and closed.

III. BAUM–WELCH DYNAMIC BAYESIAN NETWORK INVERSION

Since our DBN-structured articulatory model is designed to output appropriate mouth shapes synchronized with audio given the audio signal, we need an audio-to-visual conversion algorithm. As described in Section I, Choi *et al.* [20] proposed the HMMI algorithm, which directly generated visual parameters under ML criterion from acoustic speech, eliminating the Viterbi search and preserving the speech dynamics. Based on their approach, we propose the inversion algorithm for DBNs, namely *Baum–Welch DBN inversion* (DBNI) algorithm. We use ML as the criterion to find the optimal visual parameters, i.e., mouth shapes, maximizing the likelihood of visual parameters given the audio data and the articulatory model.

In the following, we first derive the DBNI for the model topology with a single observation stream shown in Fig. 3(a). The AAM/VAM illustrated in Fig. 2(a) fall into this case. Then we present the DBNI for the model topology with multiple observation streams shown in Fig. 3(b), which is finally applied to the AVAM in Fig. 2(b) for audio-to-visual conversion.

According to the Baum–Welch algorithm [13], optimal observations $\mathbf{O}^*$ can be found by iteratively maximizing the Baum's auxiliary function [29] $\mathcal{Q}(\lambda, \lambda; \mathbf{O}, \mathbf{O}')$, i.e.,

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

where $\mathbf{O}$ and $\mathbf{O}'$ denote the old and new observation sequences in the observation space $\mathcal{O}$ respectively.

In the DBN topology as illustrated in Fig. 3(a), we consider $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ to be the only observed data and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T\}$ their underlying hidden parents. Mixture variables which are also parents of $\mathbf{O}$ are denoted by $\kappa = \{\kappa_1, \kappa_2, \ldots, \kappa_T\}$, and other hidden variables are denoted by $q = \{q_1, q_2, \ldots, q_T\}$. Thus, the variable set in time frame t is composed of $\{q_t, \theta_t, \kappa_t, \mathbf{o}_t\}$. As described in the expectation maximization (EM) algorithm [30], the incomplete-data likelihood function is given by $P(\mathbf{O}|\lambda)$, whereas the complete-data likelihood function is $P(\mathbf{O}, q, \theta, \kappa|\lambda)$. Given q, θ, κ, and a well-trained model parameter set λ, according to the Markov property of independent relationships between variables, the likelihood of the complete-data can be formed as follows:

$$P(\mathbf{O}, q, \theta, \kappa|\lambda) = \prod_{t=1}^{T} \left[\prod_{i=1}^{D} P(X_t^i | pa(X_t^i)) P(\mathbf{o}_t | \theta_t, k_t) \right] \tag{4}$$

where X_t^i denotes a particular value of the hidden variable x^i in frame t. The immediate predecessors of x^i are referred to as its parents, with values $pa(X_t^i)$ in frame t, and D is the number of hidden variables of a time frame. The auxiliary function can be explicitly expressed as shown in (5) at the bottom of the page, where θ, q, and κ denote the possible sequences of $\mathbf{O}$'s hidden parents, possible sequences of other hidden variables and possible mixture segmentation sequences, respectively. By taking the derivative of the auxiliary function $\mathcal{Q}(\lambda, \lambda; \mathbf{O}, \mathbf{O}')$ with respect to $\mathbf{o}_t'$ to zero, we get

$$\begin{aligned}
&\frac{\partial \mathcal{Q}(\lambda, \lambda; \mathbf{O}, \mathbf{O}')}{\partial \mathbf{o}_t'} \\
&= \frac{\partial}{\partial \mathbf{o}_t'} \left[\sum_q \sum_\theta \sum_\kappa P(\mathbf{O}, q, \theta, \kappa|\lambda) \cdot \sum_{t=1}^{T} \log P(\mathbf{o}_t'|\theta_t, k_t) \right] \\
&= \sum_q \sum_\theta \sum_\kappa P(\mathbf{O}, q, \theta, \kappa|\lambda) \frac{\partial}{\partial \mathbf{o}_t'} [\log P(\mathbf{o}_t'|\theta_t, k_t)] \\
&= \sum_q \sum_\theta \sum_\kappa P(\mathbf{O}, q, \theta, \kappa|\lambda) \frac{1}{b_{\theta_t k_t}(\mathbf{o}_t')} \cdot \frac{\partial b_{\theta_t k_t}(\mathbf{o}_t')}{\partial \mathbf{o}_t'} \\
&= \sum_{l=1}^{L} \sum_{j=1}^{N} \sum_{k=1}^{K} P(\mathbf{O}, q_t = l, \theta_t = j, \kappa_t = k|\lambda) \Sigma_{jk}^{-1} (\mathbf{o}_t' - \boldsymbol{\mu}_{jk}) \\
&= 0
\end{aligned} \tag{6}$$

where L denotes the number of possible value sets of other hidden variables, and N denotes the number of possible value

$$\begin{aligned}
&\mathcal{Q}(\lambda, \lambda; \mathbf{O}, \mathbf{O}') \\
&= \sum_q \sum_\theta \sum_\kappa P(\mathbf{O}, q, \theta, \kappa|\lambda) \log P(\mathbf{O}', q, \theta, \kappa|\lambda) \\
&= \sum_q \sum_\theta \sum_\kappa P(\mathbf{O}, q, \theta, \kappa|\lambda) \left\{ \sum_{t=1}^{T} \sum_{i=1}^{D} \log P(X_t^i | pa(X_t^i)) + \sum_{t=1}^{T} \log P(\mathbf{o}_t'|\theta_t, k_t) \right\}
\end{aligned} \tag{5}$$

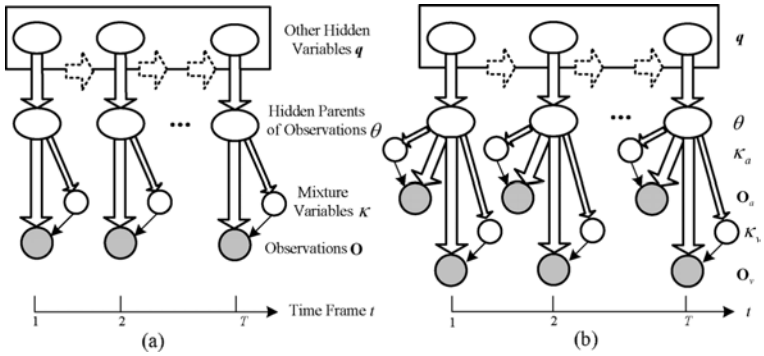


Fig. 3. (a) DBN topology with a single observation stream and (b) DBN topology with multiple observation streams. An elliptical node denotes a group of variables, while a round node denotes a single variable. A wide arc represents multiple dependencies between nodes, and a slim arc represents a single dependency between nodes. The dotted wide arc sketches the time dependencies among nodes between two consecutive frames.

sets of hidden parents of observation $\mathbf{o}_t$. For clarity, we define a new quantity $\gamma_t(l,j,k) = P(\mathbf{O}, q_t = l, \theta_t = j, \kappa_t = k|\lambda)$.

Thus, we can find the re-estimated inputs $\mathbf{o}'_t$ by

$$\mathbf{o}'_t = \frac{\sum_{l=1}^{L}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(l,j,k)\Sigma_{jk}^{-1}\boldsymbol{\mu}_{jk}}{\sum_{l=1}^{L}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(l,j,k)\Sigma_{jk}^{-1}} \tag{7}$$

where $\gamma_t(l,j,k)$ can be computed using the frontier algorithm [31].

For audio-to-visual conversion, the problem is to estimate missing visual parameters based on well-trained AVAMs and the audio input given. Thus, we extend the DBNI to multiple observation streams [as shown in Fig. 3(b)] to convert audio input to optimal visual parameters under the ML sense. Similar to the DBNI for a single observation stream, we can easily get the re-estimated visual parameters $\mathbf{o}'_{vt}$ by

$$\mathbf{o}'_{vt} = \frac{\sum_{l=1}^{L}\sum_{j=1}^{N}\sum_{k_a=1}^{K_a}\sum_{k_v=1}^{K_v}\gamma_t(l,j,k_a,k_v)\Sigma_{jk_v}^{-1}\boldsymbol{\mu}_{jk_v}}{\sum_{l=1}^{L}\sum_{j=1}^{N}\sum_{k_a=1}^{K_a}\sum_{k_v=1}^{K_v}\gamma_t(l,j,k_a,k_v)\Sigma_{jk_v}^{-1}} \tag{8}$$

where a and v denote the audio and visual streams respectively, and $\gamma_t(l,j,k_a,k_v)$ can be computed through the frontier algorithm [31].

Equation (8) shows that the DBNI is able to move the visual observations $\mathbf{o}_{vt}$ closer to the mean μ_{jk_v} of a visual Gaussian mixture by fixing the mean location of each mixture, while still retaining the original distributions of visual data. The calculation of γ_t in (8) involves observations of both audio and visual data, and the shared layer of articulators. Thus, this calculation process incorporates both modalities of speech. Moreover, since the estimation of $\mathbf{o}'_{vt}$ involves all possible values of hidden variables, the estimated visual parameters are more likely to approach the global shape of the actual parameters.

IV. Experiments

To evaluate the performance of the proposed mouth-synching technique, we have carried out extensive experiments on an image-based mouth animation approach. We have compared our method with the AVPM [18] (which uses the HMMI [20] algorithm as its audio-to-visual conversion method) since they both use integrated audio-visual models with similar conversion mechanisms. Also, the AVPM is a typical phoneme-based model which can generate reasonable lip movements. Therefore, it is appropriate to benchmark our AVAM which makes use of a novel articulatory modelling technique. Since triphones have been used recently to capture co-articulation effect for a considerable success [5], we also have compared our method with a triphone approach which can be considered as an extension of AVPM. This approach also uses the HMMI algorithm as its conversion method, and is named as AVTM (audio-visual triphone model).

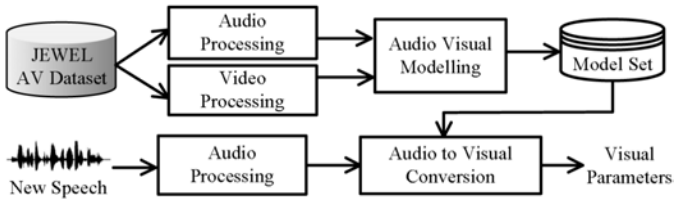


Fig. 4. Block diagram of the experiment setup.

A. Experiment Setup

Fig. 4 shows the block diagram of the experiment setup, which involves audio and video processing, audio-visual modelling and audio-to-visual conversion.

1) Audio-Visual Dataset: We used the JEWEL audio-visual dataset [32] in the experiments, which contains 524 recordings of one female speaker uttering sentences from the TIMIT corpus. The training set is composed of two SA sentences and 450 SX sentences, and the testing set contains 72 SI sentences. In the dataset, the speaker's head-and-shoulder front view against a white background is shot by a digital video camera in a studio environment, where synchronized audio and video are recorded. The audio is acquired at a rate of 16 Hz with a 30 dB SNR. The video is of 720×576 pixel in dimension, interlaced, captured in RGB color at 25 frames/s. In total, the audio-visual recordings are about 2500 s in duration. The speaker's mouth region is extracted and normalized [23].

2) Audio and Video Processing: Mel-frequency cepstral coefficients (MFCCs) as well as their velocity and acceleration

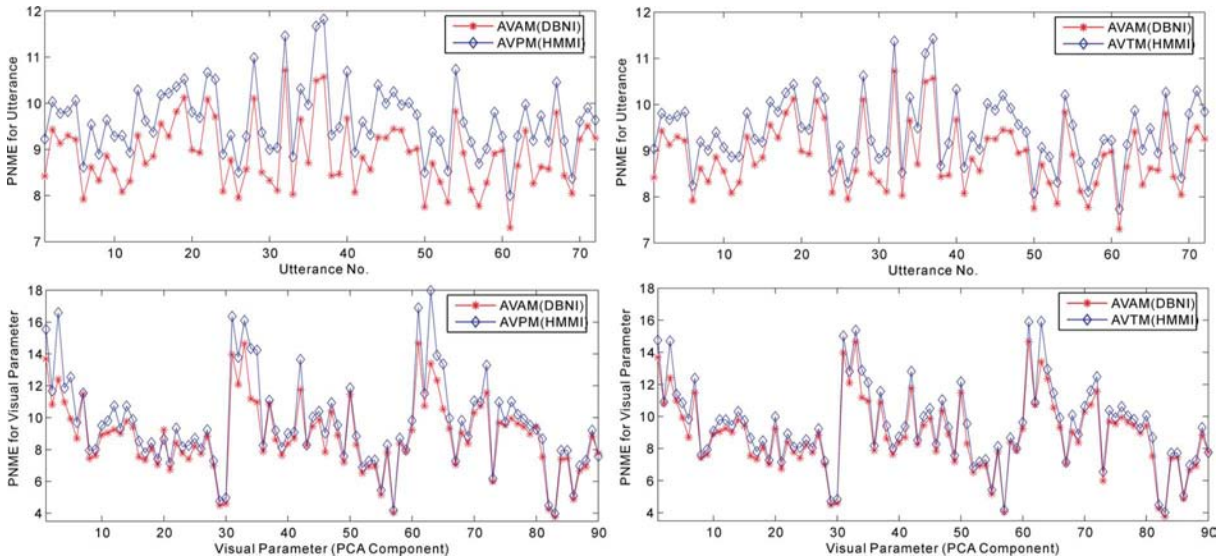


Fig. 5. PNMEs for 72 JEWEL testing utterances (top) and 90 visual parameters (bottom).

derivatives were adopted as audio features, leading to a set of 39 parameters for each frame. Principal component analysis (PCA) was implemented to the red, green and blue channels of 1500 representative mouth images chosen from the training set, generating a set of *eigenlips*. A mouth image can be approximately represented by a linear combination of these eigenlips. The combination weights (i.e., PCA coefficients) were used as visual parameters. In total, a set of 90 visual features (30 for each channel) was collected for each video frame. Visual features were up-sampled to 100 frames/s to meet the audio feature sampling rate (100 Hz).

3) Audio-Visual Modelling: In the AVPM system, a set of 47 three-state, left-to-right phoneme HMMs was trained using the Baum–Welch algorithm [13]. We performed an iterative mixture splitting approach, and achieved five continuous Gaussian mixtures for each HMM state describing the distributions of audio-visual signals. It should be noted that silence ("sil") and short pause ("sp") were included in the model set. We trained 327 context-dependent HMMs including triphones, biphones, and phonemes for the AVTM system according to the availability of the corresponding training data. They had the same model topology with the phoneme models in the AVPM system.

In the AVAM system, we trained the AVAMs whose structure is shown in Fig. 2(b). The state variables represent the English sub-phonemes, and a set of 6 articulator variables describes the articulator features defined in Table I. To make a comparative study, we designated a set of 141 values (47×3) to the state variables, corresponding to the 47 three-state phoneme HMMs in the AVPM system. The CPDs $P(\mathbf{o}_{st}|\theta_t, k_{st}),\ s \in \{a, v\}$ were trained using the standard EM algorithm [30]. According to the model structure illustrated in Fig. 2(b), a set of 5 continuous Gaussian mixtures for each possible combination of articulatory feature values was trained for the audio and visual streams, respectively. We designed a phoneme-state-to-articulators mapping table according to the English phonetics, and other CPDs in Fig. 2(b) were trained using a supervised learning method [23]. Considering the physical constraints among the articulators [23] and the dataset context, we trained totally 472 and 236 sets of Gaussian mixtures for the audio and visual streams, respectively. In all the systems, we used manually labelled phoneme level transcriptions as well as frame-synchronized audio-visual features during the training process.

4) Audio-to-Visual Conversion: During audio-to-visual conversion, the AVPM and AVTM systems used the HMMI algorithm [20], and the AVAM system used the proposed DBNI algorithm. The global mean of all visual Gaussian mixtures was used as the initial values of visual parameters. According to the convergence property of EM, the estimated optimal visual parameters were collected framewise for a dataset utterance within a very few iterations.

B. Objective Evaluation: Estimation Error

We have carried out objective evaluations by directly comparing the visual parameters estimated by the 3 testing systems with the real parameters extracted from the original video. To make a quantitative evaluation, we have calculated the *percentage normalized mean error* (PNME) for each testing utterance (72 in total) of the JEWEL dataset and for each visual parameter (90 in total). The PNMEs for utterance and visual parameter are defined, respectively, as

$$PNME_k = \frac{\sum_{i=1}^{I_k}\sum_{j=1}^{90}\left|\hat{C}_{ijk} - C_{ijk}\right|}{I_k \times 90} \times 100\%, \tag{9}$$

$$PNME_j = \frac{\sum_{k=1}^{72}\sum_{i=1}^{I_k}\left|\hat{C}_{ijk} - C_{ijk}\right|}{72 \times I_k} \times 100\% \tag{10}$$

where I_k denotes the frame number of utterance k ($k = 1, 2, \ldots, 72$), and C_{ijk} and $\hat{C}_{ijk}$ denote the actual and estimated visual parameter j ($j = 1, 2, \ldots, 90$) after normalization for frame i in utterance k respectively.

The PNME curves in Fig. 5 reveal that the proposed AVAM system can effectively reduce the estimation errors in terms of

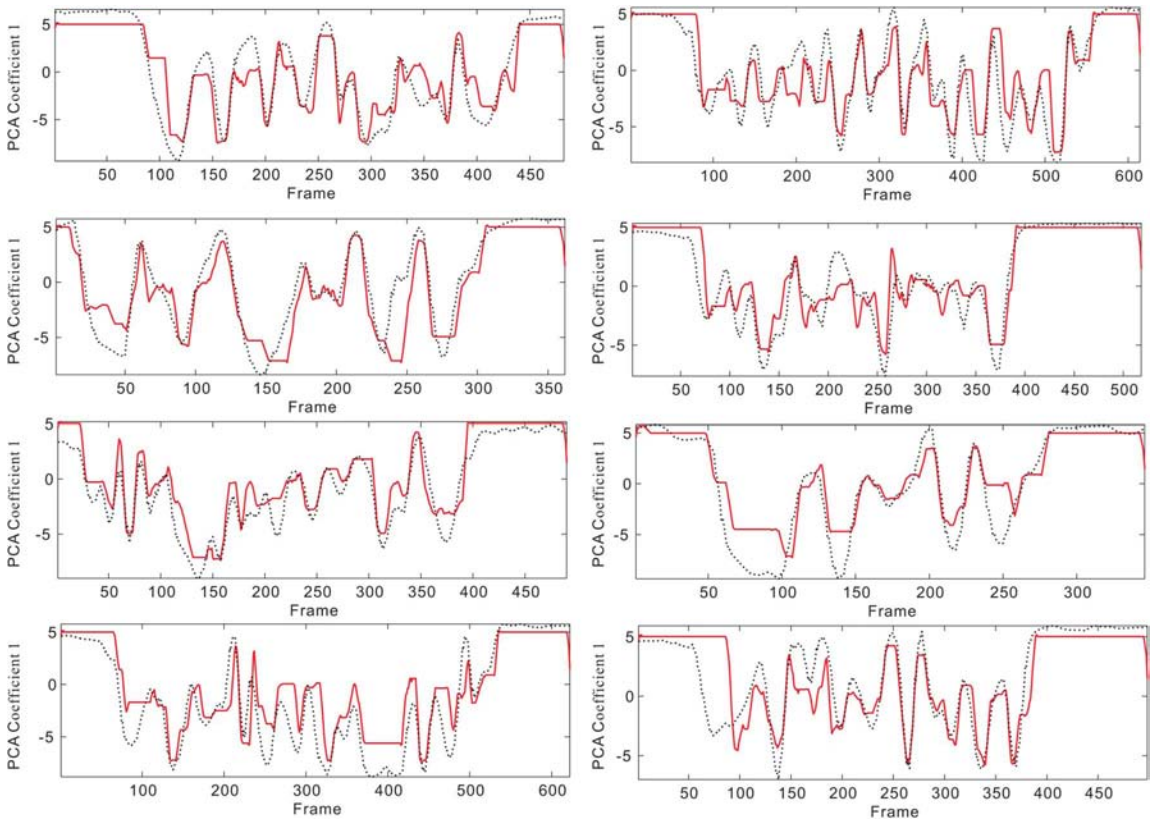


Fig. 6. Estimated visual parameter (red solid line) by the AVAM system versus actual parameter (black dotted line) trajectories for eight testing utterances from the JEWEL dataset.

PNME as compared with the AVPM system. With the introduction of context dependent HMMs (triphones and biphones), the AVTM system can also reduce the PNMEs as compared with the AVPM system. However, the proposed AVAM system performed the best among the three testing systems. On average the AVAM system reduced the PNME by as much as 8.0% as compared with the AVPM system; and 5.6% as compared with the AVTM system.

Fig. 6 depicts some examples of the trajectories for the first visual parameter for eight testing utterances in the JEWEL dataset. These trajectories were generated by the AVAM system using the DBNI conversion algorithm, and further smoothed by a mean filter with a 3-frame width to remove jitter. We can clearly observe that the estimated visual parameters match the original ones very well.

C. Objective Evaluation: Audio-Visual Speech Recognition

Since the use of visual speech information in addition to audio is able to effectively improve speech understanding in noisy conditions [23], we have conducted audio-visual speech recognition (AVSR) experiments for a more perceptual evaluation of the estimated visual parameters. Ninety visual parameters (estimated or the ground truth) were combined framewise with 39 MFCC coefficients to improve the speech recognition rate under noisy acoustic conditions. Different from direct comparison in terms of estimation error (Section IV-B), this evaluation provides a way to quantify the amount of lipreading information contained in the estimated visual speech. We have used the multistream HMM (MSHMM) [23] as the audio-visual fusion scheme, which is a popular model in the AVSR literature due to its ability to model the reliability of audio and visual streams easily and effectively via the following weighted fusion:

$$b_i(\mathbf{o}_t^a, \mathbf{o}_t^v) = \prod_{s\in\{a,v\}} \left[\sum_{k=1}^{K_s} \omega_{iks}\mathcal{N}(\mathbf{o}_t^s; \mu_{iks}, \Sigma_{iks})\right]^{\delta_s} \quad (11)$$

where the stream reliability is described by stream exponents δ_s and $\delta_a + \delta_v = 1$.

We manually corrupted the audio data in the JEWEL dataset with additive speech babble noise at various SNRs (28 dB, 25 dB, 23 dB, 20 dB, 15 dB, and 10 dB). Matched training-testing conditions were considered in the experiments to accurately measure the influence of the visual parameters on recognition performance. At each testing SNR, audio from the JEWEL training set was used to train the 47 three-state, left-to-right, state-synchronous, five-continuous-Gaussian-mixture, phoneme MSHMMs; and audio from the JEWEL testing set was used for a recognition test. Note that the visual parameters used for MSHMM training and recognition test were estimated from the the noise-free ($\mathrm{SNR} = 30$ dB) audio or extracted from the original data. Stream exponents δ_s were empirically selected *a priori* on a small development set for each testing SNR by minimizing the word error rate (WER). According to the different sets of visual parameters, we built four AVSR systems, namely MSHMM-AVPM, MSHMM-AVTM, MSHMM-AVAM and MSHMM-Act (actual PCA parameters). An audio-only (AO) system with 47 conventional

TABLE II
AUDIO-VISUAL SPEECH RECOGNITION RESULTS IN TERMS OF WER

System	30dB	28dB	25dB	23dB	20dB	15dB	10dB
AO	17.0	18.3	20.8	24.3	30.6	39.7	61.7
MSHMM-AVPM	16.9	17.2	19.0	22.9	27.0	33.3	50.2
MSHMM-AVTM	16.7	16.9	18.2	21.3	24.6	31.8	46.5
MSHMM-AVAM	16.3	16.9	17.5	20.6	24.0	30.1	44.5
MSHMM-Act	16.1	16.4	17.4	19.4	23.9	29.9	43.0

left-to-right, three-state, five-continuous-Gaussian-mixture, phoneme HMMs was also built to benchmark the experiments. All the systems were developed using the HTK Toolkit 3.2 [33]. The experimental results are summarized in Table II.

As can be seen clearly, the performance of the AO system is heavily affected by acoustic noise. Analysis unveils that lots of insertion errors occur when speech is contaminated by the babble noise, which contributes significantly to the WER. This indicates that even matched training-testing condition is considered, the additive noise still severely affects the recognition performance. Not surprisingly, with the help of the visual information, all the testing AVSR systems outperform the AO system under noisy conditions. The MSHMM-Act system, which uses the actual visual parameters extracted from original videos, performs the best. The MSHMM-AVAM system, which uses the visual parameters estimated by AVAM and DBNI, performs the best among the three systems using estimated visual parameters, and observed the closest WERs with the MSHMM-Act system. Its relative reduction in WER compared to the AO system ranges from 7.7% for a noisy audio with a 28 dB SNR to 27.9% for a noisy audio with a 10 dB SNR. Indeed, the overall results clearly demonstrate that the lipreading information provided in our estimated visual parameters is capable of significantly reducing speech recognition error under noisy conditions.

D. Subjective Evaluation

1) Performance Refinement: We have carried out subjective evaluation which compares the synthesized image sequences with the original recordings. The synthesized mouth sequences were realized from the estimated visual parameters through PCA expansion. 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 dimensions of the visual parameters. Introducing more visual parameters will result in more prediction errors due to the problem of data sparseness. Therefore, we used a performance refinement process to improve the realism of the animation. We selected a set of 500 typical mouth images (normalized) from the JEWEL dataset, and saved their full-dimension PCA components (1500 here) to a candidate set $\mathbf{G}$ (green channel for example)

$$\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{500}\},\ \mathbf{g} = [\mathbf{g}^r, \mathbf{g}^d] \tag{12}$$

where $r = 1:30$, $d = 31:1500$, and $\mathbf{g}$ denotes the PCA coefficient vector for the green channel. We chose the visual parameters $\hat{\mathbf{g}}_t^d$ as the fine detail augments of green channel of frame t using the following criterion:

$$\hat{\mathbf{g}}_t^d = \arg\min_{\mathbf{g}_j^d} |\mathbf{o}_t^v - \mathbf{g}_j^r|,\ j = 1, 2, \ldots, 500 \tag{13}$$

where $\mathbf{o}_t^v$ denotes the estimated visual parameters. Therefore, the visual parameter vector will be $[\mathbf{o}_t^v, \hat{\mathbf{g}}_t^d]$. The mouth images were resembled by PCA expansion using the full visual parameter vector. Fig. 7 demonstrates some snapshots from a mouth animation sequence synthesized by the AVAM system for an utterance from the JEWEL dataset.

2) Driving a Talking Face: We have built up a prototype of taking face system. To avoid the "zombie"-like effect and make the talking face more natural, a *4-layer* overlaying process was adopted, in which the synthesized mouth frames, the corresponding jaws, eye blinks, and the base faces are sewed up to generate a facial animation sequence.

As head movement is important for facial animation to appear natural, five recorded snippets with neutral expressions and tiny head movements were selected as the base face sequences. This also ensured that the viewers mainly focused the evaluations on the mouth articulation. The head pose, eye and jaw positions were manually annotated for exactly stitching the facial parts with the base face in the overlaying process. An eye blink process were also extracted and saved. A set of 28 typical jaw image masks with different downward actions in articulation was collected.

In the overlaying process, a base face sequence was randomly selected from the snippet set. The eye-blink sequence was inserted once in each N-s period ($2 < N < 4$). We associated an appropriate jaw from the jaw candidate set to each synthesized mouth according to the mouth opening scale and the waveform energy. To avoid any boundary artifacts, we used the Poisson cloning technique [34] to merge together the four layers according to the annotated tilting angles of head pose, eye, and jaw positions. Fig. 8 shows some snapshots from our talking face.

3) Subjective Evaluation: Two type of subjective assessments were performed: *scoring test* and *Turing test*; and a group of 20 viewers with no prior experience were involved. To get a fair evaluation focused on the mouth animation and to separate the different factors influencing the speech perception [2], mouth region was cropped from the 72 original full-face video, and overlaid to the base face using the same 4-layer overlaying process, named original video (Ori). Prior to testing, we randomly named and mixed together the synthesized videos (by AVPM, AVTM, and AVAM) and the Ori videos in AVI format from the JEWEL testing set, achieving a set of 288 (72×4) videos with accompanied real audio. The video set was separated to five sessions in order to avoid any unfair judgements from viewers due to fatigue and boredom. In each test session, we randomly presented the videos to the viewer. The viewer was first asked to score each video in terms of naturalness of the uttering mouth matching the audio, namely scoring test. A five-point assessment was adopted (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). Then the viewer was asked to judge whether the video was synthesized artificially or real recordings, i.e., the Turing test. We did not perform simultaneous side-by-side presentations of the synthesized

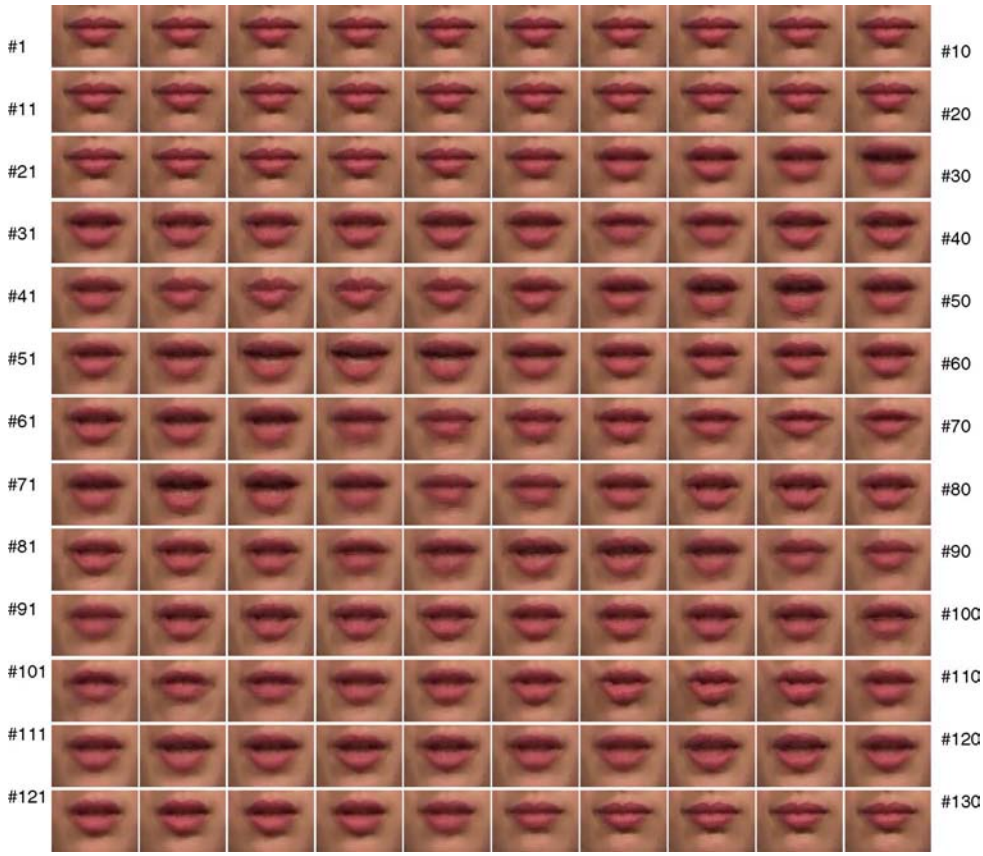


Fig. 7. Snapshots from the synthesized mouth sequence (25 frames/s) for the utterance "*Scientific progress comes from the development of new techniques.*"

Fig. 8. Some snapshots from the speech-driven talking face.

and real videos, because the viewers would have shifted their gaze from one to another while the utterance was played. As a negative effect, the viewers could have compared local features but not the impression the moving mouth would have given throughout the video. The results of scoring and Turing tests are summarized in Table III and Fig. 9, respectively.

TABLE III
SCORING TEST RESULTS

System	Score (5=Excellent, 1=Bad)					
	5	4	3	2	1	MOS
AVPM	14	15	23	12	7	3.2
AVTM	20	18	18	14	1	3.6
AVAM	23	20	16	13	0	3.7
Ori	28	27	17	0	0	4.2

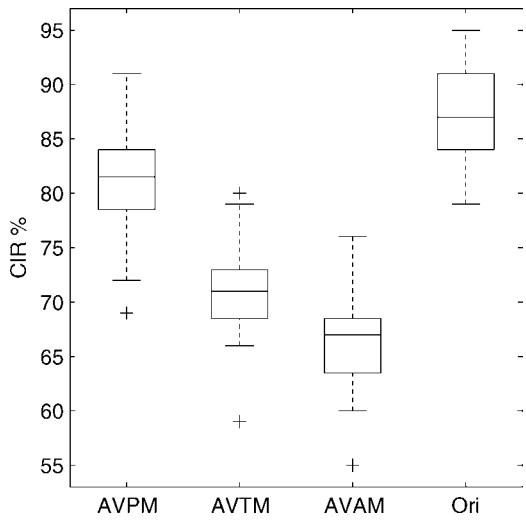


Fig. 9. Turing test results.

Table III shows that the viewers give relatively high scores to the synthesized videos generated by the AVAM and AVTM compared with those of AVPM. On average, 23 synthesized videos by AVAM (in total 72) are given the highest score, which is close to that of Ori videos (28 out of 72). Among the three testing systems which synthesize mouth animations from speech, the proposed AVAM system performs the best with a mean opinion score (MOS) of 3.7. It is interesting that the viewers did not unanimously give the original videos the excellent score, this is mainly because prosodic movements of mouth exists in the small mouth region cropped from the original recorded videos, therefore acentric movements are considered as weird and unnatural.

For image-based talking faces, the ultimate goal is to produce animations that pass the Turing test, that is, that viewers cannot distinguish between animations and real recordings. The Turing test can be quantified in terms of correct identifying rate (*CIR*), which is defined as

$$CIR = \frac{\text{Number of correctly identified videos}}{\text{Number of total testing videos}} \times 100\% \quad (14)$$

Fig. 9 illustrates the boxplot of CIR calculated on the 20 viewers. The inter-quartile ranges (IQRs) of AVPM, AVTM, AVAM and Ori are 78–84%, 67–73%, 63–67%, and 83–92%. On average, 18%, 29%, and 38% of the synthesized videos by AVPM, AVTM, and AVAM are detected as real recordings, while 13% of the real recorded videos are mistakenly considered as synthesized video. Although the viewers had quite different judgements on whether each video was real or artificially synthesized (a wide CIR range), the majority opinions indicate that for 38% of the videos synthesized by our AVAM system, the viewers were unable to tell whether the presented video was a synthetic one or a real one.

V. Conclusions

This paper presents an articulatory modelling technique for realistic mouth-synching in speech-driven talking face. Motivated by the fact that mouth movement is originated by articulation, we directly model the configurations of articulators, such as lips, tongue and teeth, using a DBN-based audio-visual articulatory model (AVAM). Audio speech and visual speech are synchronously associated by two streams, and a shared articulator layer is incorporated for both streams. This structure not only reflects the consanguinity between facial expressions and audio speech but also depicts the linguistic fact that different articulators evolve asynchronously. To output appropriate mouth sequences with natural speech dynamics, we present a Baum–Welch DBN inversion (DBNI) algorithm, which converts audio to optimal visual parameters by maximizing the likelihood of the visual parameters given the audio data and the AVAM.

We compared the proposed AVAM with the audio-visual phoneme model (AVPM) and its triphone variant (AVTM) by both objective and subjective evaluations on the JEWEL audio-visual dataset. Objective evaluations show that compared with AVPM and AVTM, the proposed AVAM can effectively reduce the estimation errors on the visual parameters, and resultant parameters match the true parameters more accurately. AVSR experiments also show that with the help of the estimated visual parameters, recognition error rates are effectively reduced under noisy acoustic conditions. We have built up a prototype of talking face to make subjective evaluations. Results show that synthesized animation generated by the proposed AVAM matches the corresponding audio naturally. More encouragingly, 38% of the synthesized animation sequences generated by the AVAM are so lively that the viewers could not even distinguish them from the real recordings.

Since articulatory modelling has been proven more robust to ambient noise [24], we are currently trying to realize natural mouth-synching under adverse acoustic conditions. Finally, as DBNs have great expressive power, emotions that speech convey may be encapsulated and converted to visual parameters, achieving an expressive talking face.

References

[1] J. Ostermann and A. Weissenfeld, "Talking faces-technologies and applications," in *Proc. of ICPR'04*, Aug. 2004, vol. 3, pp. 826–833.

[2] E. Cosatto, J. Ostermann, H. P. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *Proc. IEEE*, vol. 91, no. 9, pp. 1406–1428, Sep. 2003.

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

[4] K. Kaehler, J. Haber, H. Yamauchi, and HP Seidel, "Head shop: Generating animated head models with anatomical structure," in *Proc. ACM SIGGRAPH'02*, 2002, pp. 55–63.

[5] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. ACM SIGGRAPH'97*, 1997.

[6] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proc. ACM SIGGRAPH*, 2002, pp. 388–397.

[7] E. Cosatto and H. Graf, "Sample-based synthesis of photo-realistic talking heads," in *Proc. IEEE Computer Animation*, 1998, pp. 103–110.

[8] ——, "Photo-realistic talking heads from image samples," *IEEE Trans. Multimedia*, vol. 2, pp. 152–163, 2000.

[9] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan, "Recent improvements to the IBM trainable speech synthesis system," in *Proc. ICASSP'03*, 2003, vol. 1, pp. 708–711.

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

[11] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech-driven face animation with expressions using neural networks," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 916–927, 2002.

[12] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proc. AVSP'99*, 1999, pp. 133–138.

[13] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[14] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Commun.*, vol. 26, no. 1–2, pp. 105–115, 1998.

[15] M. Brand, "Voice puppetry," in *Proc. ACM SIGGRAPH'99*, 1999, pp. 21–28.

[16] T. Chen, "Audiovisual speech processing: Lip reading and lip synchronization," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, 2001.

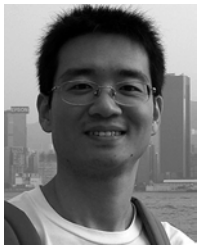
[17] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, "Real-time speech motion synthesis from recorded motions," in *Eurographics/ACM SIGGRAPH Symp. Computer Animation*, 2004, pp. 347–355.

[18] K. Choi and J. N. Hwang, "Baum–Welch hidden Markov model inversion for reliable audio-to-visual conversion," in *Proc. IEEE 3rd Workshop Multimedia Signal Processing*, 1999, pp. 175–180.

[19] S. Y. Moon and J. N. Hwang, "Noisy speech recognition using robust inversion of hidden Markov models," in *Proc. ICASSP'95*, 1995, pp. 145–148.

[20] K. Choi, Y. Luo, and J. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *J. VLSI Signal Process.*, no. 29, pp. 51–61, 2001.

[21] S. Fu, R. Gutierrez-Osuna, A. Esposito, K. P. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden Markov models," *IEEE Trans. Multimedia*, vol. 7, pp. 243–251, 2005.
[22] J. Goldsmith, *Autosegmental and Metrical Phonology*, W. Hardcastle and N. Hewlett, Eds. New York: Basil Blackwell, 1990.
[23] L. Xie, "Research on Key Issues of Audio Visual Speech Recognition," Ph.D. dissertation, Northwestern Polytechnical Univ., Xian, China, 2004.
[24] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP'98*, 1998, pp. 891–894.
[25] F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2001.
[26] G. G. Zweig, "Bayesian network structures and inference techniques for automatic speech recognition," *Comput. Speech Lang.*, vol. 17, pp. 173–193, 2003.
[27] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Speech Commun.*, vol. 41, pp. 511–529, 2003.
[28] J. A. Bilmes, G. Zweig, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne, "Discriminatively structured graphical models for speech recognition," in *Tech. Rep. JHU 2001 Summer Workshop*, 2001.
[29] L. E. Baum and G. R. Sell, "Growth functions for transformations on manifolds," *Pacific J. Math.*, vol. 27, no. 2, pp. 211–227, 1968.
[30] A. Dempster, A. N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 89–111, 1977.
[31] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, Univ. California, Berkeley, 2002.
[32] L. Xie and Z. Ye, The JEWEL Audio-Visual Dataset for Facial Animation 2005, Tech. Rep. RCMT 05–11.
[33] S. Young, G. Evermann, D. Kershaw, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book Eng. Dept., Cambridge Univ., Cambridge, U.K., 2002 [Online]. Available: http://htk.eng.cam.ac.uk/, 3.2
[34] P. Pèrez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graphics (SIGGRAPH'03)*, vol. 22, no. 3, pp. 313–318, 2003.



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.

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, China. He is currently a Postdoctoral Fellow in the Human-Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong. His current research interest includes talking face, multimedia retrieval, speech recognition, multimedia signal processing, and pattern recognition.

Zhi-Qiang Liu (S'82–M'86–SM'91) received the M.A.Sc. degree in aerospace engineering from the Institute for Aerospace Studies, University of Toronto, Toronto, ON, Canada, 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 and technology. His interests are neural-fuzzy systems, machine learning, human-media systems, computer vision, mobile computing, and computer networks.