EURASIP Journal on Applied Signal Processing 2002:11, 1274–1288


# Dynamic Bayesian Networks for Audio-Visual Speech Recognition

**Ara V. Nefian**
*Intel Corporation, Microprocessor Research Labs, 2200 Mission College Blvd., Santa Clara, CA 95052-8119, USA*
*Email: ara.nefian@intel.com*

**Luhong Liang**
*Intel Corporation, Microcomputer Research Labs, Guanghua Road, 100020 Chaoyang District, Beijing, China*
*Email: luhong.liang@intel.com*

**Xiaobo Pi**
*Intel Corporation, Microcomputer Research Labs, Guanghua Road, 100020 Chaoyang District, Beijing, China*
*Email: xiaobo.pi@intel.com*

**Xiaoxing Liu**
*Intel Corporation, Microcomputer Research Labs, Guanghua Road, 100020 Chaoyang District, Beijing, China*
*Email: xiaoxing.liu@intel.com*

**Kevin Murphy**
*Computer Science Division, University of California, Berkeley, Berkeley, CA 94720-1776, USA*
*Email: murphyk@cs.berkeley.edu*

*Received 30 November 2001 and in revised form 6 August 2002*

The use of visual features in audio-visual speech recognition (AVSR) is justified by both the speech generation mechanism, which is essentially bimodal in audio and visual representation, and by the need for features that are invariant to acoustic noise perturbation. As a result, current AVSR systems demonstrate significant accuracy improvements in environments affected by acoustic noise. In this paper, we describe the use of two statistical models for audio-visual integration, the coupled HMM (CHMM) and the factorial HMM (FHMM), and compare the performance of these models with the existing models used in speaker dependent audio-visual isolated word recognition. The statistical properties of both the CHMM and FHMM allow to model the state asynchrony of the audio and visual observation sequences while preserving their natural correlation over time. In our experiments, the CHMM performs best overall, outperforming all the existing models and the FHMM.

**Keywords and phrases:** audio-visual speech recognition, hidden Markov models, coupled hidden Markov models, factorial hidden Markov models, dynamic Bayesian networks.

## 1. INTRODUCTION

The variety of applications of automatic speech recognition (ASR) systems for human computer interfaces, telephony, and robotics has driven the research of a large scientific community in recent decades. However, the success of the currently available ASR systems is restricted to relatively controlled environments and well-defined applications such as dictation or small to medium vocabulary voice-based control commands (e.g., hand-free dialing). Often, robust ASR systems require special positioning of the microphone with respect to the speaker resulting in a rather unnatural human-machine interface. In recent years, together with the investigation of several acoustic noise reduction techniques, the study of visual features has emerged as attractive solution to speech recognition under less constrained environments. The use of visual features in audio-visual speech recognition (AVSR) is motivated by the speech formation mechanism and the natural ability of humans to reduce audio ambiguity using visual cues [1]. In addition, the visual information provides complementary features that cannot be corrupted by the acoustic noise of the environment. The importance of

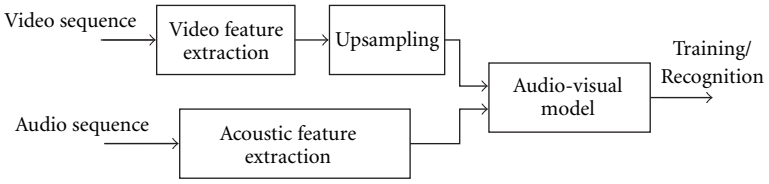


Figure 1: The audio-visual speech recognition system.

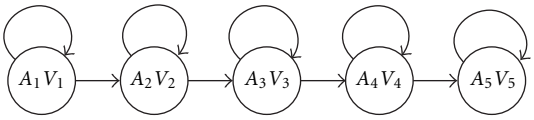


Figure 2: The state transition diagram of a left-to-right HMM.

visual features for speech recognition, especially under noisy environments, has been demonstrated by the success of recent AVSR systems [2]. However, problems such as the selection of the optimal set of visual features, or the optimal models for audio-visual integration remain challenging research topics. In this paper, we describe a set of improvements to the existing methods for visual feature selection and we focus on two models for isolated word audio-visual speech recognition: the coupled hidden Markov model (CHMM) [3] and the factorial hidden Markov model (FHMM) [4], which are special cases of the dynamic Bayesian networks [5]. The structure of both models investigated in this paper describes the state synchrony of the audio and visual components of speech while maintaining their natural correlation over time. The isolated word AVSR system illustrated in Figure 1 is used to analyze the performance of the audio-visual models introduced in this paper. First, the audio and visual features (Section 3) are extracted from each frame of the audio-visual sequence. The sequence of visual features, which describe the mouth deformation over consecutive frames, is upsampled to match the frequency of the audio observation vectors. Finally, both the factorial and the coupled HMM (Section 4) are used for audio-visual integration, and their performance for AVSR in terms of parameter complexity, computational efficiency (Section 5), and recognition accuracy (Section 6) is compared to existing models used in current AVSR systems.

## 2. RELATED WORK

Audio-visual speech recognition has emerged in recent years as an active field, gathering researchers in computer vision, signal and speech processing, and pattern recognition [2]. With the selection of acoustic features for speech recognition well understood [6], robust visual feature extraction and selection of the audio-visual integration model are the leading research areas in audio-visual speech recognition.

Visual features are often derived from the shape of the mouth [7, 8, 9, 10]. Although very popular, these methods rely exclusively on the accurate detection of the lip contours which is often a challenging task under varying illumination conditions and rotations of the face. An alternative approach is to obtain visual features from the transformed gray scale intensity image of the lip region. Several intensity or appearance modeling techniques have been studied, including principal component analysis [9], linear discriminant analysis (LDA), discrete cosine transform (DCT), and maximum likelihood linear transform [2]. Methods that combine shape and appearance modeling were presented in [2, 11].

Existing techniques for audio-visual (AV) integration [2, 10, 12], consist of feature fusion and decision fusion methods. In feature fusion method, the observation vectors are obtained by the concatenation of the audio and visual features, that can be followed by a dimensionality reduction transform [13]. The resulting observation sequences are modeled using a left-to-right hidden Markov model (HMM) [6] as described in Figure 2. In decision fusion systems the class conditional likelihood of each modality is combined at different levels (state, phone, or word) to generate an overall conditional likelihood used in recognition. Some of the most successful decision fusion models include the multistream HMM, the product HMM, or the independent HMM. The multistream HMM [14] assumes that the audio and video sequences are state synchronous but, unlike the HMM for feature fusion, allows the likelihood of the audio and visual observation sequences to be computed independently. This allows to weigh the relative contribution of the audio and visual likelihood to the overall likelihood based on the reliability of the corresponding stream at different levels of acoustic noise. Although more flexible than the HMM, the multistream HMM cannot accurately describe the natural state asynchrony of the audio-visual speech. The audio-visual multistream product HMM [11, 14, 15, 16, 17] illustrated in Figure 3, can be seen as an extension of the previous model by representing each hidden state of the multistream HMM as a pair of one audio and one visual state. Due to its structure, the multistream product HMM allows for audio-video state asynchrony, controlled through the state transition matrix of the model, and forces the audio and video streams to be in synchrony at the model boundaries (phone level in continuous speech recognition systems or word level in isolated word recognition systems). The audio-visual sequences can also be modeled using two independent HMMs [2], one for audio and one for visual features. This model extends the level of asynchrony between the audio and visual states of the previous models, but fails to preserve the natural dependency over time of the acoustic and visual features of speech.

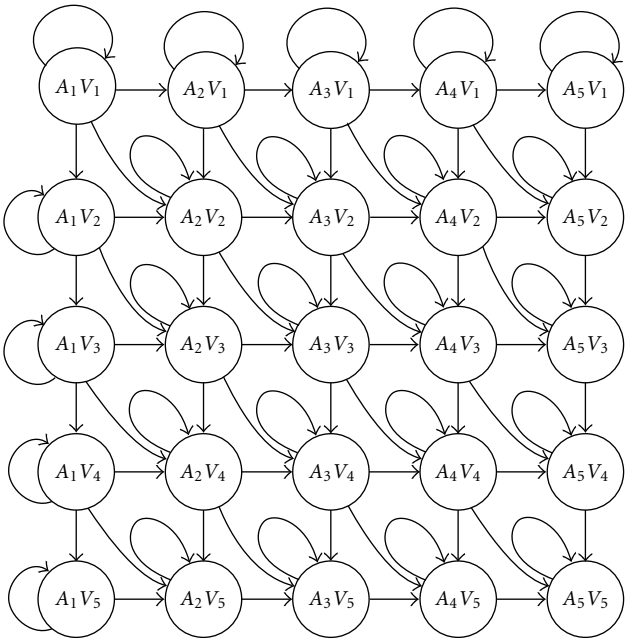


FIGURE 3: The state transition diagram of a product HMM.

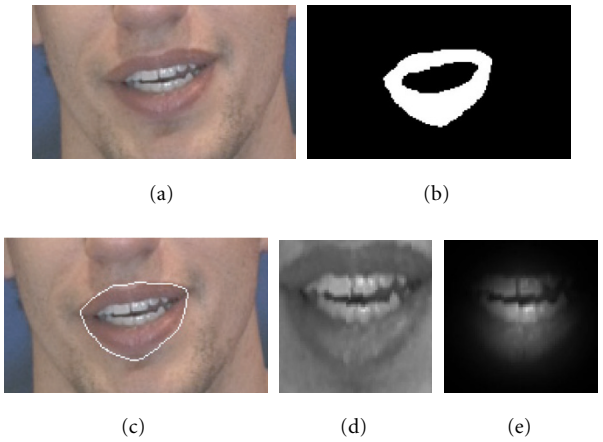


FIGURE 4: (a) The lower region of the face used as an initial estimate for the mouth location, (b) binary image representing the mouth segmentation results, (c) the result of the lip contour extraction, (d) the scale and rotation normalized mouth region, (e) the result of the normalized mouth region windowing.

## 3. VISUAL FEATURE EXTRACTION

Robust location of the facial features, specially the mouth region, and the extraction of a discriminant set of visual observation vectors are the two key elements of the AVSR system. The cascade algorithm for visual feature extraction used in our AVSR system consists of the following steps: face detection, mouth region detection, lip contour extraction, mouth region normalization and windowing, 2D-DCT and LDA coefficient extraction. Next, we will describe the steps of the cascade algorithm in more detail.

The extraction of the visual features starts with the detection of the speaker's face in the video sequence. The face detector used in our system is described in [18]. The lower half of the detected face (Figure 4a) is a natural choice for the initial estimate of the mouth region.

Next, LDA is used to assign the pixels in the mouth region to the lip and face classes. LDA transforms the pixel values from the RGB chromatic space into a one-dimensional space that best separates the two classes. The optimal linear discriminant space [19] is computed off-line using a set of manually segmented images of the lip and face regions. Figure 4b shows a binary image of the lip segmentation from the lower region of the face in Figure 4a.

The contour of the lips (Figure 4c) is obtained through the binary chain encoding method [20] followed by a smoothing operation. Figures 5a, 5b, 5c, 5d, 5e, 5f, and 5h show several successful results of the lip contour extraction. Due to the wide variety of skin and lip tones, the mouth segmentation and therefore the lip contour extraction may result in inaccurate results (Figures 5i and 5j).

The lip contour is used to estimate the size and the rotation of the mouth in the image plane. Using an affine transform a rotation and size normalized grayscale region of the mouth ($64 \times 64$ pixels) is obtained from each frame of the video sequence (Figure 4d). However, not all the pixels in the mouth region have the same relevance for visual speech recognition. In our experiments we found that, as expected, the most significant information for speech recognition is contained in the pixels inside the lip contour. Therefore, we use an exponential window $w[x, y] = \exp(-((x - x_0)^2 + (y - y_0)^2)/\sigma^2)$, $\sigma = 12$, to multiply the pixels values in the grayscale normalized mouth region. The window of size $64 \times 64$ is centered in the center of the mouth region $(x_0, y_0)$. Figure 4e illustrates the result of the mouth region windowing.

Next, the normalized and windowed mouth region is decomposed into eight blocks of height 32 and width 16, and the 2D-DCT transform is applied to each of these blocks. A set of four 2D-DCT coefficients from a window of size $2 \times 2$ in the lowest frequency in the 2D-DCT domain are extracted from each block. The resulting coefficients extracted are arranged in a vector of size 32.

In the final stage of the video feature extraction cascade, the multiclass LDA [19] is applied to the vectors of 2D-DCT coefficients. For our isolated word speech recognition system, the classes of the LDA are associated to the words available in the database. A set of 15 coefficients, corresponding to the most significant generalized eigenvalues of the LDA decomposition are used as visual observation vectors.

## 4. THE AUDIO-VISUAL MODEL

The audio-visual models used in existing AVSR systems, as well as the audio-visual models discussed in this paper, are special cases of dynamic Bayesian networks (DBN) [5, 21, 22]. DBNs are directed graphical models of stochastic processes in which the hidden states are represented in

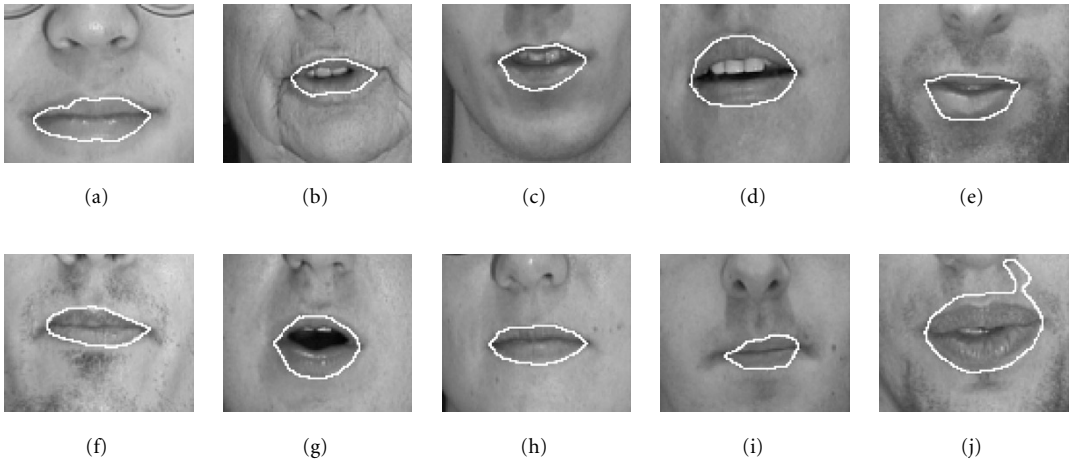


Figure 5: Examples of the mouth contour extraction.

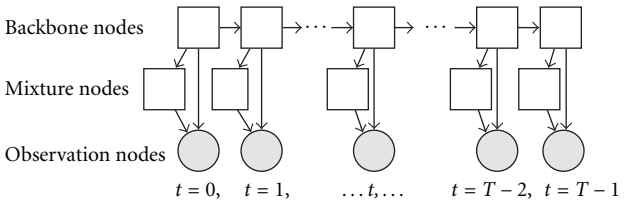


Figure 6: The audio-visual HMM.

terms of individual variables or factors. A DBN is specified by a directed acyclic graph, which represents the conditional independence assumptions and the conditional probability distributions of each node [23, 24]. With the DBN representation, the classification of the decision fusion models can be seen in terms of independence assumptions of the transition probabilities and of the conditional likelihood of the observed and hidden nodes. Figure 6 represents an HMM as a DBN. The transparent squares represent the hidden discrete nodes (variables), while the shaded circles represent the observed continuous nodes. Throughout this paper, we will refer to the hidden nodes conditioned over time as coupled or backbone nodes and to the remaining hidden nodes as mixture nodes. The variables associated with the backbone nodes represent the states of the HMM, while the values of the mixture nodes represent the mixture component associated with each of the state of the backbone nodes. The parameters of the HMM [6] are

$$\begin{aligned} \pi(i) &= P(q_1 = i), \\ b_t(i) &= P(\mathbf{O}_t | q_t = i), \\ a(i|j) &= P(q_t = i | q_{t-1} = j), \end{aligned} \tag{1}$$

where $q_t$ is the state of the backbone node at time $t$, $\pi(i)$ is the initial state distribution for state $i$, $a(i|j)$ is the state transition probability from state $j$ to state $i$, and $b_t(i)$ represents the probability of the observation $\mathbf{O}_t$ given the $i$th state of the backbone nodes. The observation probability is generally modeled using a mixture of Gaussian components.

Introduced for audio-only speech recognition, the multistream HMM a (MSHMM) became a popular model for multimodal sequences such as the audio-visual speech. In

$$b_t(i) = \prod_{s=1}^{S} \left[ \sum_{m=1}^{M_i^s} w_{i,m}^s N\left(\mathbf{O}_t^s, \boldsymbol{\mu}_{i,m}^s, \mathbf{U}_{i,m}^s\right) \right]^{\lambda_s}, \tag{2}$$

where $S$ represents the total number of streams, $\lambda_s$ ($\sum_s \lambda_s = 1$, $\lambda_s \geq 0$) are the stream exponents, $\mathbf{O}_t^s$ is the observation vector of the $s$th stream at time $t$, $M_i^s$ is the number of mixture components in stream $s$ and state $i$, and $\boldsymbol{\mu}_{i,m}^s$, $\mathbf{U}_{i,m}^s$, $w_{i,m}^s$ are the mean, covariance matrix, and mixture weight for the $s$th stream, $i$th state, and $m$th Gaussian mixture component, respectively. The two streams ($S = 2$) of the audio-visual MSHMM (AV MSHMM) model the audio and the video sequence. For the AV MSHMM, as well as for the HMM used in video-only or audio-only speech recognition, all covariance matrices are assumed diagonal, and the transition probability matrix reflects the left-to-right state evolution

$$a(i|j) = 0, \quad \text{if } i \notin \{j, j+1\}. \tag{3}$$

The audio and visual state synchrony imposed by the AV MSHMM can be relaxed using models that allow one hidden backbone node per stream at each time $t$. Figure 7 illustrates a two-stream independent HMM (IHMM) represented as a DBN. Let $\mathbf{i} = \{i_1, \ldots, i_S\}$ be some set of states of the backbone nodes, $N_s$ the number of states of the backbone nodes in stream $s$, $q_t^s$ the state of the backbone node in stream $s$ at time $t$ and $\mathbf{q}_t = \{q_t^1, \ldots, q_t^S\}$. Formally, the parameters of an

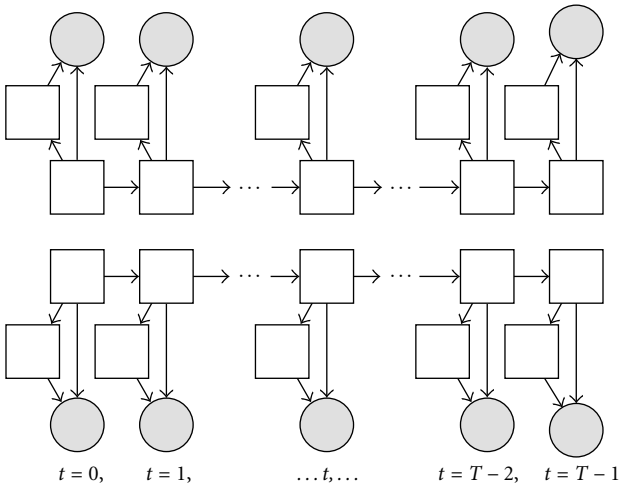


Figure 7: A two-stream independent HMM.

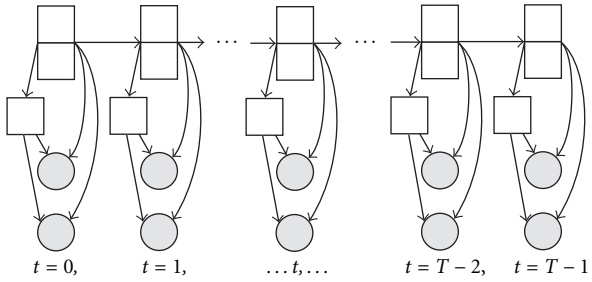


Figure 8: The audio-visual product HMM.

IHMM are

$$\pi(\mathbf{i}) = \prod_s \pi^s(i_s) = \prod_s P(q_1^s = i_s), \tag{4}$$

$$b_t(\mathbf{i}) = \prod_s b_t^s(i_s) = \prod_s P(\mathbf{O}_t^s | q_t^s = i_s), \tag{5}$$

$$a(\mathbf{i}|\mathbf{j}) = \prod_s a^s(i_s|j_s) = \prod_s P(q_t^s = i_s | q_{t-1}^s = j_s), \tag{6}$$

where $\pi^s(i_s)$ and $b_t^s(i_s)$ are the initial state distribution and the observation probability of state $i_s$ in stream $s$, respectively, and $a^s(i_s|j_s)$ is the state transition from state $j_s$ to state $i_s$ in stream $s$. For the audio-visual IHMM (AV IHMM) each of the two HMMs, describing the audio or video sequence, is constrained to a left-to-right structure, and the observation likelihood $b_t^s(i)$ is computed using a mixture of Gaussian density functions, with diagonal covariance matrices. The AV IHMM allows for more flexibility than the AV MSHMM in modeling the state asynchrony but fails to model the natural correlation in time between the audio and visual components of speech. This is a result of the independent modeling of the transition probabilities (see (6)) and of the observation likelihood (see (5)).

A product HMM (PHMM) can be seen as a standard HMM, where each backbone state is represented by a set of states, one for each stream [17]. The parameters of a PHMM are

$$\pi(\mathbf{i}) = P(\mathbf{q}_1 = \mathbf{i}), \tag{7}$$

$$b_t(\mathbf{i}) = P(\mathbf{O}_t | \mathbf{q}_t = \mathbf{i}), \tag{8}$$

$$a(\mathbf{i}|\mathbf{j}) = P(\mathbf{q}_t = \mathbf{i} | \mathbf{q}_{t-1} = \mathbf{j}), \tag{9}$$

where $\mathbf{O}_t$ can be obtained through the concatenation of the observation vectors in each stream

$$\mathbf{O}_t = \left[ (\mathbf{O}_t^1)^T, \ldots, (\mathbf{O}_t^S)^T \right]^T. \tag{10}$$

The observation likelihood can be computed using a Gaussian density or a mixture with Gaussian components. The use of PHMM in AVSR is justified primarily because it allows for state asynchrony, since each of the coupled nodes can be in any combination of audio and visual states. In addition, unlike the IHMM, the PHMM preserves the natural correlation of the audio and visual features due the joint probability modeling of both the observation likelihood (see (8)) and transition probabilities (see (9)). For the PHMM used in AVSR, denoted in this paper as the audio-visual PHMM (AV PHMM), the audio and visual state asynchrony is limited to a maximum of one state. Formally, the transition probability matrix from state $\mathbf{j} = [j_a, j_v]$ to state $\mathbf{i} = [i_a, i_v]$ is given by

$$a(\mathbf{i}|\mathbf{j}) = 0 \quad \text{if} \begin{cases} i_s \notin \{j_s, j_s + 1\}, & s \in \{a, v\}, \\ |i_a - i_v| \geq 2, & \mathbf{i} \neq \mathbf{j}, \end{cases} \tag{11}$$

where indices $a$ and $v$ denote the audio and video stream, respectively. In the AV PHMM described in this paper (Figure 8) the observation likelihood is computed using

$$b_t(\mathbf{i}) = \sum_{m=1}^{M_\mathbf{i}} w_{\mathbf{i},m} \prod_s \left[ N(\mathbf{O}_t^s, \boldsymbol{\mu}_{\mathbf{i},m}^s, \mathbf{U}_{\mathbf{i},m}^s) \right]^{\lambda_s}, \tag{12}$$

where $M_\mathbf{i}$ represents the number of mixture components associated with state $\mathbf{i}$, $\boldsymbol{\mu}_{\mathbf{i},m}^s$ and $\mathbf{U}_{\mathbf{i},m}^s$ are the mean and the diagonal covariance matrices corresponding to stream $s$ given the state $\mathbf{i}$ and mixture component $m$, and $w_{\mathbf{i},m}$ are the mixture weights corresponding to the state $\mathbf{i}$. Unlike the MSHMM (see (2)) the likelihood representation for the PHMM used in (12) models the stream observations jointly through the dependency of the same mixture node. In this paper, the model parameters are trained for fixed values of the stream exponents $\lambda_s = 1$. For testing, the stream exponents are chosen to maximize the average recognition rate at different acoustic signal-to-noise ratio (SNR) levels. Since in the PHMM both the transition and observation likelihood are jointly computed, and in the IHMM both transition and observation likelihood in each stream are independent, these models can be considered extreme cases of a range of models that combine the joint and independent modeling of the transition probabilities and observation likelihoods. Two of these models, namely the factorial HMM and the coupled HMM, and their application in audio-visual integration will be discussed next.

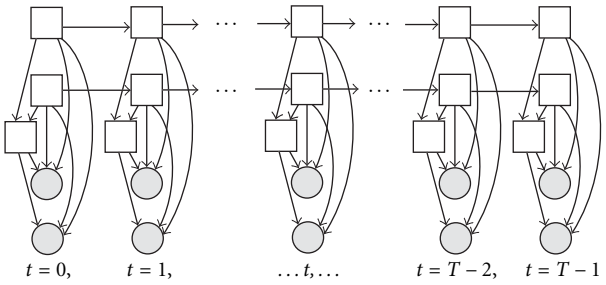


Figure 9: The audio-visual factorial HMM.

### 4.1. The audio-visual factorial hidden Markov model

The factorial HMM (FHMM) [4] is a generalization of the HMM suitable for a large range of multimedia applications that integrate two or more streams of data. The FHMM generalizes an HMM by representing the hidden state by a set of variables or factors. In other words, it uses a distributed representation of the hidden state. These factors are assumed to be independent of each other, but they all contribute to the observations, and hence become coupled indirectly due to the "explaining away" effect [23]. The elements of a factorial HMM are described as

$$\pi(\mathbf{i}) = P(\mathbf{q}_1 = \mathbf{i}), \tag{13}$$

$$b_t(\mathbf{i}) = P(\mathbf{O}_t | \mathbf{q}_t = \mathbf{i}), \tag{14}$$

$$a(\mathbf{i}|\mathbf{j}) = \prod_s a^s(i_s|j_s) = \prod_s P(q_t^s = i_s | q_{t-1}^s = j_s). \tag{15}$$

It can be seen that as with the IHMM, the transition probabilities of the FHMM are computed using the independence assumption between the hidden states or factors in each of the HMMs (see (15)). However, as with the PHMM, the observation likelihood is jointly computed from all the hidden states (see (14)). The observation likelihood can be computed using a continuous mixture with Gaussian components. The FHMM used in AVSR, denoted in this paper as the audio-visual FHMM (AV FHMM), has a set of modifications from the general model. In the AV FHMM used in this paper (Figure 9), the observation likelihoods are obtained from the multistream representation as described in (12). To model the causality in speech generation, the following constraint on the transition probability matrices of the AV FHMM is imposed:

$$a^s(i_s|j_s) = 0, \quad \text{if } i_s \notin \{j_s, j_s + 1\}, \tag{16}$$

where $s \in \{a, v\}$.

#### 4.1.1 Training factorial HMMs

As is well known, DBNs can be trained using the expectation-maximization (EM) algorithm (see, e.g., [22]). The EM algorithm for the FHMM is described in Appendix A. However, this only converges to a local optimum, making the choice of the initial parameters of the model a critical issue. In this paper, we present an efficient method for initialization using a Viterbi algorithm derived for the FHMM. The Viterbi algorithm for FHMMs is described below for an utterance $\mathbf{O}_1, \ldots, \mathbf{O}_T$ of length $T$.

(i) Initialization

$$\delta_1(\mathbf{i}) = \pi(\mathbf{i}) b_1(\mathbf{i}), \tag{17}$$

$$\psi_1(\mathbf{i}) = 0; \tag{18}$$

(ii) Recursion

$$\begin{aligned} \delta_t(\mathbf{i}) &= \max_{\mathbf{j}} \{\delta_{t-1}(\mathbf{j}) a(\mathbf{i}|\mathbf{j})\} b_t(\mathbf{i}), \\ \psi_t(\mathbf{i}) &= \arg\max_{\mathbf{j}} \{\delta_{t-1}(\mathbf{j}) a(\mathbf{i}|\mathbf{j})\}; \end{aligned} \tag{19}$$

(iii) Termination

$$\begin{aligned} P^* &= \max_{\mathbf{i}} \{\delta_T(\mathbf{i})\}, \\ \mathbf{q}_T &= \arg\max_{\mathbf{i}} \{\delta_T(\mathbf{i})\}; \end{aligned} \tag{20}$$

(iv) Backtracking

$$\mathbf{q}_t = \psi_{t+1}(\mathbf{q}_{t+1}), \tag{21}$$

where $P^* = \max_{\mathbf{q}_1, \ldots, \mathbf{q}_T} P(\mathbf{O}_1, \ldots, \mathbf{O}_T, \mathbf{q}_1, \ldots, \mathbf{q}_T)$, and $a(\mathbf{i}|\mathbf{j})$ is obtained using (15). Note that, as with the HMM, the Viterbi algorithm can be computed using the logarithms of the model parameters, and additions instead of multiplication.

The initialization of the training algorithm iteratively updates the initial parameters of the model from the optimal segmentation of the hidden states. The state segmentation algorithm described in this paper reduces the complexity of the search for the optimal sequence of backbone and mixture nodes using the following steps. First, we use the Viterbi algorithm, as described above, to determine the optimal sequence of states for the backbone nodes. Second, we obtain the most likely assignment to the mixture nodes. Given these optimal assignments to the hidden nodes, the appropriate sets of parameters are updated. For the FHMM with $\lambda_s = 1$ and general covariance matrices the initialization of the training algorithm is described below.

*Step* 1. Let $R$ be the number of training examples and let $\mathbf{O}_{r,1}^s, \ldots, \mathbf{O}_{r,T_r}^s$ be the observation sequence of length $T_r$ corresponding to the $s$th stream of the $r$th ($1 \le r \le R$) training example. First, the observation sequences $\mathbf{O}_{r,1}^s, \ldots, \mathbf{O}_{r,T_r}^s$ are uniformly segmented according to the number of states of the backbone nodes $N_s$. Then, a new sequence of observation vectors is obtained by concatenating the observation vectors assigned to each state $i_s$, $s = 1, \ldots, S$. For each state set $\mathbf{i}$ of the backbone nodes, the mixture parameters are initialized using the K-means algorithm [19] with $M_{\mathbf{i}}$ clusters.

*Step* 2. The new parameters of the model are estimated from the segmented data

$$\begin{aligned}
\boldsymbol{\mu}^s_{\mathbf{i},m} &= \frac{\sum_{r,t}\gamma_{r,t}(\mathbf{i},m)\mathbf{O}^s_{r,t}}{\sum_{r,t}\gamma_{r,t}(\mathbf{i},m)},\\
\mathbf{U}^s_{\mathbf{i},m} &= \frac{\sum_{r,t}\gamma_{r,t}(\mathbf{i},m)\left(\mathbf{O}^s_{r,t}-\boldsymbol{\mu}^s_{\mathbf{i},m}\right)\left(\mathbf{O}^s_{r,t}-\boldsymbol{\mu}^s_{\mathbf{i},m}\right)^T}{\sum_{r,t}\gamma_{r,t}(\mathbf{i},m)},\\
w_{\mathbf{i},m} &= \frac{\sum_{r,t}\gamma_{r,t}(\mathbf{i},m)}{\sum_{r,t}\sum_{m'}\gamma_{r,t}(\mathbf{i},m')},\\
a^s(i|j) &= \frac{\sum_{r,t}\epsilon^s_{r,t}(i,j)}{\sum_{r,t}\sum_{l}\epsilon^s_{r,t}(i,l)},
\end{aligned} \tag{22}$$

where

$$\begin{aligned}
\gamma_{r,t}(\mathbf{i},m) &= \begin{cases} 1, & \text{if } \mathbf{q}_{r,t}=\mathbf{i},\ c_{r,t}=m,\\ 0, & \text{otherwise,}\end{cases}\\
\epsilon^s_{r,t}(i,j) &= \begin{cases} 1, & \text{if } q^s_{r,t}=i,\ q^s_{r,t-1}=j,\\ 0, & \text{otherwise,}\end{cases}
\end{aligned} \tag{23}$$

where $q^s_{r,t}$ represents the state of the $t$th backbone node in the $s$th stream of the $r$th observation sequence, and $c_{r,t}$ is the mixture component of the $r$th observation sequence at time $t$.

*Step* 3. An optimal state sequence $\mathbf{q}_{r,1},\ldots,\mathbf{q}_{r,T_r}$ of the backbone nodes is obtained for the $r$th observation sequence using the Viterbi algorithm (see below). The mixture component $c_{r,t}$ is obtained as

$$c_{r,t} = \max_{m=1,\ldots,M_{\mathbf{i}}} P(\mathbf{O}_{r,t}|\mathbf{q}_{r,t}=\mathbf{i},\ c_{r,t}=m). \tag{24}$$

*Step* 4. The iterations in Steps 2, 3, and 4 are repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below a convergence threshold.

#### 4.1.2 Recognition using the factorial HMM

To classify a word, the log likelihood of each model is computed using the Viterbi algorithm described in the previous section. The parameters of the FHMM corresponding to each word in the database are obtained in the training stage using clean audio signals (SNR = 30 dB). In the recognition stage, the audio tracks of the testing sequences are altered by white noise with different SNR levels. The influence of the audio and visual observation streams is weighted based on the relative reliability of the audio and visual features for different levels of the acoustic SNR. Formally, the observation likelihoods are computed using the multistream representation in (12). The values of the audio and visual exponents $\lambda_s$, $s\in\{a,v\}$, corresponding to a specific acoustic SNR level are obtained experimentally to maximize the average recognition rate. Figure 10 illustrates the variation of the audio-visual speech recognition rate for different values of the audio exponent $\lambda_a$ and different values of SNR. Note that each of the AVSR curves at all SNR levels reaches smooth maximum levels. This is particularly important in designing robust AVSR systems and allows for the exponents to be chosen in a relatively large range of values. Table 1 describes the audio exponents $\lambda_a$ used in our system which were derived from Figure 10. As expected, the value of the optimal audio exponents decays with the decay of the SNR levels, showing the increased reliability of the video at low acoustic SNR.

Table 1: The optimal set of exponents for the audio stream $\lambda_a$ for the FHMM at different SNR values of the acoustic speech.

| SNR (dB) | 30 | 28 | 26 | 24 | 22 | 20 | 18 | 16 | 14 | 12 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_a$ | 0.8 | 0.8 | 0.7 | 0.6 | 0.6 | 0.3 | 0.2 | 0.1 | 0.1 | 0.1 | 0.0 |

### 4.2. The audio-visual coupled hidden Markov model

The coupled HMM (CHMM) [3] is a DBN that allows the backbone nodes to interact, and at the same time to have their own observations. In the past, CHMM have been used to model hand gestures [3], the interaction between speech and hand gestures [25], or audio-visual speech [26, 27]. Figure 11 illustrates a continuous mixture two-stream CHMM used in our audio-visual speech recognition system. The elements of the coupled HMM are described as

$$\pi(\mathbf{i}) = \prod_s \pi^s(i_s) = \prod_s P(q^s_1=i_s), \tag{25}$$

$$b_t(\mathbf{i}) = \prod_s b^s_t(i_s) = \prod_s P(\mathbf{O}^s_t|q^s_t=i_s), \tag{26}$$

$$a(\mathbf{i}|\mathbf{j}) = \prod_s a^s(i_s|\mathbf{j}) = \prod_s P(q^s_t=i_s|\mathbf{q}_{t-1}=\mathbf{j}). \tag{27}$$

Note that in general, to decrease the complexity of the model, the dependency of a backbone node at time $t$ is restricted to its neighbor backbone nodes at time $t-1$. As with the IHMM, in the CHMM the computation of the observation likelihood assumes the independence of the observation likelihoods in each stream. However, the transition probability of each coupled node is computed as joint probability of the set of states at previous time. With the constraint $a^s(i_s|\mathbf{j}) = a^s(i_s|j_s)$ a CHMM is reduced to an IHMM.

For the audio-visual CHMM (AV CHMM) the observation likelihoods of the audio and video streams are computed using a mixture of Gaussians with diagonal covariance matrices, and the transition probability matrix is constrained to reflect the natural audio-visual speech dependencies

$$a^s(i_s|\mathbf{j}) = 0 \quad \text{if} \begin{cases} i_s \notin \{j_s, j_s+1\},\\ |\,i_s - j_{s'}\,| \geq 2, \quad s'\neq s,\end{cases} \tag{28}$$

where $s,s'\in\{a,v\}$. The CHMM relates also to the Boltzmann zipper [28] used in audio-visual speech recognition. The Boltzmann zipper consists of two linear Boltzmann networks connected such that they can influence each other. Figure 12 illustrates a Boltzmann zipper where each of the Boltzmann chains is represented as an HMM. Note that although the connections between nodes within the same Boltzmann chain can be seen as transition probabilities of an HMM, the connections between nodes of different chains do not have the same significance [19]. Due to its structure, the

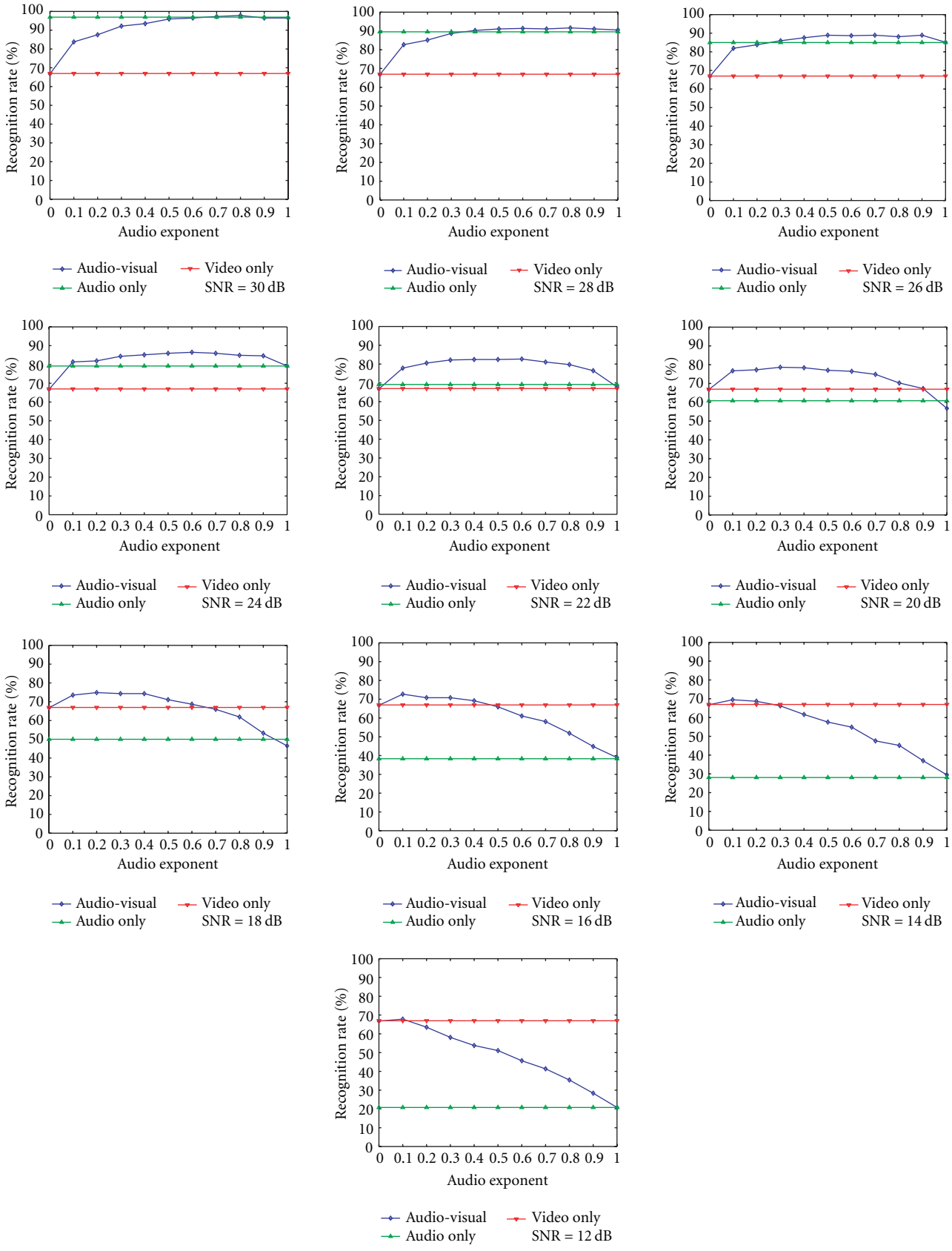


Figure 10: The FHMM recognition rate against SNR for different audio exponents.

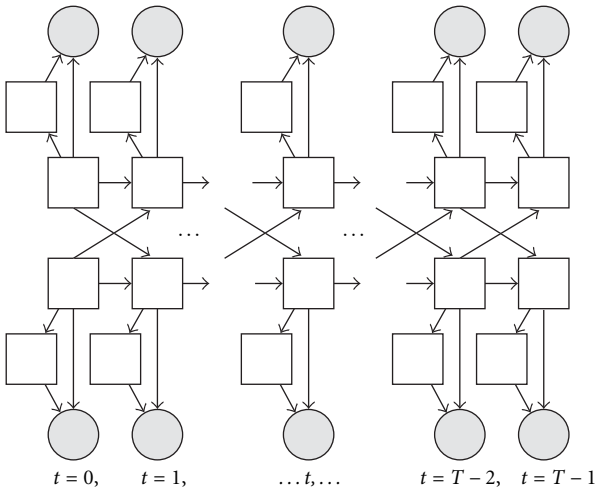


Figure 11: The audio-visual coupled HMM.

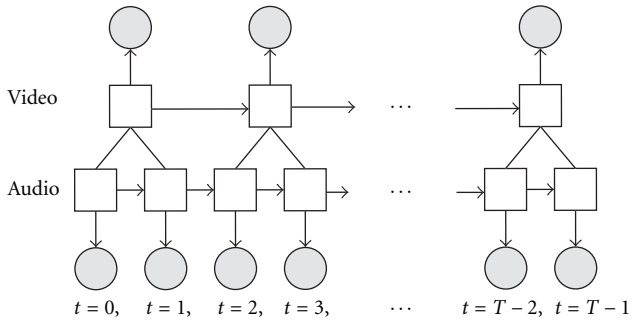


Figure 12: The Boltzmann zipper used in audio-visual integration.

Boltzmann zipper can address the problem of "fast" (audio) and "slow" (visual) observation vector integration.

#### 4.2.1 Training the coupled HMM

In the past, several training techniques for the CHMM were proposed including the Monte Carlo sampling method and the $N$-head dynamic programming method [3, 26]. The CHMM in this paper is trained using EM (Appendix B) which makes the choice of robust initial parameters very important. In this section we describe an efficient initialization method of the CHMM parameters, which is similar to the initialization of the FHMM parameters described previously. The initialization of the training algorithm for the CHMM is described by the following steps:

*Step* 1. Given $R$ training examples, the observation sequence of length $T_r$ corresponding to the $r$th example ($1 \le r \le R$) and $s$th stream, $\mathbf{O}^s_{r,1}, \ldots, \mathbf{O}^s_{r,T_r}$, is uniformly segmented according to the number of states of the backbone nodes $N_s$. Hence an initial state sequence for the backbone nodes $q^s_{r,1}, \ldots, q^s_{r,t}, \ldots, q^s_{r,T_r}$ is obtained for each data stream $s$. For each state $i$ in stream $s$ the mixture segmentation of the data assigned to it is obtained using the K-means algorithm [19] with $M^s_i$ clusters. Consequently the mixture components $c^s_{r,t}$ for the $r$th observation sequence at time $t$ and stream $s$ is obtained.

*Step* 2. The new parameters of the model are estimated from the segmented data

$$\begin{aligned}
\boldsymbol{\mu}^s_{i,m} &= \frac{\sum_{r,t} \gamma^s_{r,t}(i,m)\mathbf{O}^s_{r,t}}{\sum_{r,t} \gamma^s_{r,t}(i,m)},\\
\mathbf{U}^s_{i,m} &= \frac{\sum_{r,t} \gamma^s_{r,t}(i,m)\left(\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{i,m}\right)\left(\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{i,m}\right)^T}{\sum_{r,t} \gamma^s_{r,t}(i,m)},\\
w^s_{i,m} &= \frac{\sum_{r,t} \gamma^s_{r,t}(i,m)}{\sum_{r,t}\sum_{m'} \gamma^s_{r,t}(i,m')},\\
a^s(i|\mathbf{j}) &= \frac{\sum_{r,t} \epsilon^s_{r,t}(i,\mathbf{j})}{\sum_{r,t}\sum_{\mathbf{j}} \epsilon^s_{r,t}(i,\mathbf{j})},
\end{aligned} \tag{29}$$

where

$$\begin{aligned}
\gamma^s_{r,t}(i,m) &= \begin{cases} 1, & \text{if } q^s_{r,t} = i,\ c^s_{r,t} = m,\\ 0, & \text{otherwise,}\end{cases}\\
\epsilon^s_{r,t}(i,\mathbf{j}) &= \begin{cases} 1, & \text{if } q^s_{r,t} = i,\ \mathbf{q}_{r,t-1} = \mathbf{j},\\ 0, & \text{otherwise,}\end{cases}
\end{aligned} \tag{30}$$

where $c^s_{r,t}$ is the mixture component for the $s$th stream of the $r$th observation sequence at time $t$.

*Step* 3. An optimal state sequence of the backbone nodes $\mathbf{q}_{r,1}, \ldots, \mathbf{q}_{r,T_r}$ is obtained using the Viterbi algorithm for the CHMM [29]. The steps of the Viterbi segmentation for the CHMM are described by (17), (18), (19), (20), and (21), where the initial state probability $\pi(\mathbf{i})$, the observation likelihood $b(\mathbf{i})$ and transition probabilities $a(\mathbf{i}|\mathbf{j})$ are computed using (25), (26), and (27), respectively. The mixture components $c^s_{r,t}$ are obtained using

$$c^s_{r,t} = \max_{m=1,\ldots,M^s_i} P\left(\mathbf{O}^s_{r,t} | q^s_{r,t} = i,\ c^s_{r,t} = m\right). \tag{31}$$

*Step* 4. The iterations in Steps 2, 3, and 4 are repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below a convergence threshold.

#### 4.2.2 Recognition using the coupled HMM

The isolated word recognition is carried out via the Viterbi algorithm for CHMM, where the observation probability for each observation conditional likelihood is modified to handle different levels of noise

$$\tilde{b}^s_t(i_s) = b_t\left(\mathbf{O}^s_t | q^s_t = i_s\right)^{\lambda_s}, \tag{32}$$

where $\lambda_s, s \in \{a, v\}$, are the exponents of the audio and video streams respectively obtained experimentally to maximize the average recognition rate for a specific acoustic SNR level. Table 2 describes the audio exponents $\lambda_a$ used in our system, and Figure 13 shows the variation of CHMM-based audio-visual recognition rate for different values of the audio exponent $\lambda_a$ and different values of SNR. In all our experiments

Table 2: The optimal set of exponents for the audio stream $\lambda_a$ at different SNR values of the acoustic speech for the CHMM.

| SNR (dB) | 30 | 28 | 26 | 24 | 22 | 20 | 18 | 16 | 14 | 12 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_a$ | 0.9 | 0.7 | 0.8 | 0.7 | 0.5 | 0.4 | 0.3 | 0.2 | 0.2 | 0.2 | 0.1 |

the audio sequences were perturbed by white noise. The average audio-only, video-only, and CHMM-based audio-visual recognition rates for different levels of SNR are shown in Figure 14.

## 5. MODEL COMPLEXITY ANALYSIS

Together with the recognition accuracy, the number of parameters of the model and the computational complexity required by the recognition process are very important in the analysis of a model. Models with a small number of parameters produce better estimates for the same amount of training data. Tables 3, 4, 5, and 6 describe the size of the parameter space and the computational complexity required for recognition using the PHMM, IHMM, FHMM, and CHMM. We consider both the general case as well as the specific models used in AVSR (i.e., AV PHMM, AV IHMM, AV FHMM, and AV CHMM) which include the use of diagonal covariance matrices, and sparse transition probability matrices as described in Section 4. In addition to the notations introduced in the previous section, the size of the observation vector in modality $s$ is denoted by $V_s$. For simplification, for the IHMM and CHMM, we consider that all the mixture nodes in stream $s$ have the same number of components $M_s$, independent of the state of the parent backbone nodes. For the PHMM and FHMM we assume that all state sets have the same number of mixture components $M$.

In terms of the space required by the parameters of the models, we count the elements of the transition probability matrices ($\mathbf{A}$), the mean vectors ($\boldsymbol{\mu}$), covariance matrices ($\mathbf{U}$), and the weighting coefficients ($w$) per HMM word. From Tables 3 and 6 we see that the IHMM and CHMM as well as the AV IHMM and AV CHMM, require the same number of parameters for $\boldsymbol{\mu}$, $\mathbf{U}$, and $w$. This is due to the fact that in these models, the probability of the observation vector $\mathbf{O}_t^s$ in stream $s$ at time $t$ depends only on its private mixture and backbone node. Due to the coupling of the backbone nodes, the space required by the $\mathbf{A}$ parameters of the CHMM and AV CHMM is larger than that for the IHMM and AV IHMM, respectively. However, if $V_s \gg N_s$, the space required by the $\mathbf{A}$ parameters is negligible compared to $\boldsymbol{\mu}$, $\mathbf{U}$ and $w$, making the AV CHMM and AV IHMM very similar from the point of view of the parameter space requirements. The joint dependency of the observation vector $\mathbf{O}_t$ on all backbone nodes at time $t$ for the PHMM, FHMM increases significantly the number of parameters of these models compared to the IHMM and CHMM (Tables 4 and 5). Note that the left-to-right topology of each HMM in an AV FHMM (see (16)) does not reduce the number of audio-visual state combinations which remains of the order of $\prod_{s=1}^{2} N_s$. On the other hand, the sparse transition probability matrix of the AV PHMM (see (11)) only allows a number of audio-visual states of the order of $\sqrt{\prod_{s=1}^{2} N_s}$. This reduces the parameter space of the AV PHMM compared to the AV FHMM while still remaining more complex than the AV IHMM or AV CHMM.

In terms of time required by the recognition process, we count the number of log-likelihood additions required by the Viterbi algorithm per time instant (in the row labeled "Viterbi") and the number of operations required for the evaluation of the observation likelihoods. The Viterbi algorithm with the lowest complexity is obtained with the AV IHMM where the transition and observation probabilities in one stream are independent from the transition and observation probabilities in the other stream. On the other hand, the

Table 3: The number of parameters and running time needed for independent HMMs, and for the specific model used in AVSR.

| Space | IHMM | AV IHMM |
|---|---|---|
| $\mathbf{A}$ | $O(\sum_s N_s^2)$ | $O(\sum_{s=1}^{2} N_s)$ |
| $\boldsymbol{\mu}$ | $\sum_s N_s M_s V_s$ | $\sum_{s=1}^{2} N_s M_s V_s$ |
| $\mathbf{U}$ | $\sum_s N_s M_s V_s^2$ | $\sum_{s=1}^{2} N_s M_s V_s$ |
| $w$ | $\sum_s N_s M_s$ | $\sum_{s=1}^{2} N_s M_s$ |
| Time | IHMM | AV IHMM |
| Viterbi | $O(\sum_s N_s^2)$ | $O(\sum_{s=1}^{2} N_s)$ |

Table 4: The number of parameters and running time needed for the product HMM, and for the specific model used in AVSR.

| Space | PHMM | AV PHMM |
|---|---|---|
| $\mathbf{A}$ | $O(\prod_s N_s^2)$ | $O(\sqrt{\prod_{s=1}^{2} N_s})$ |
| $\boldsymbol{\mu}$ | $(\prod_s N_s) M (\sum_s V_s)$ | $O(\sqrt{\prod_{s=1}^{2} N_s} M \sum_{s=1}^{2} V_s)$ |
| $\mathbf{U}$ | $(\prod_s N_s) M (\sum_s V_s)^2$ | $O(\sqrt{\prod_{s=1}^{2} N_s} M \sum_{s=1}^{2} V_s)$ |
| $w$ | $(\prod_s N_s) M$ | $O(\sqrt{\prod_{s=1}^{2} N_s} M)$ |
| Time | PHMM | AV PHMM |
| Viterbi | $O(\prod_s N_s^2)$ | $O(\sqrt{\prod_{s=1}^{2} N_s})$ |

Table 5: The number of parameters and running time needed for the factorial HMM, and for the specific model used in AVSR.

| Space | FHMM | AV FHMM |
|---|---|---|
| $\mathbf{A}$ | $O(\sum_s N_s^2)$ | $O(\sum_{s=1}^{2} N_s)$ |
| $\boldsymbol{\mu}$ | $(\prod_s N_s) M (\sum_s V_s)$ | $\sum_{s=1}^{2} (\prod_{s=1}^{2} N_s) M V_s$ |
| $\mathbf{U}$ | $(\prod_s N_s) M (\sum_s V_s)^2$ | $\sum_{s=1}^{2} (\prod_{s=1}^{2} N_s) M V_s$ |
| $w$ | $(\prod_s N_s) M$ | $(\prod_{s=1}^{2} N_s) M$ |
| Time | FHMM | AV FHMM |
| Viterbi | $O(\prod_s N_s^2)$ | $O(\prod_{s=1}^{2} N_s^2)$ |

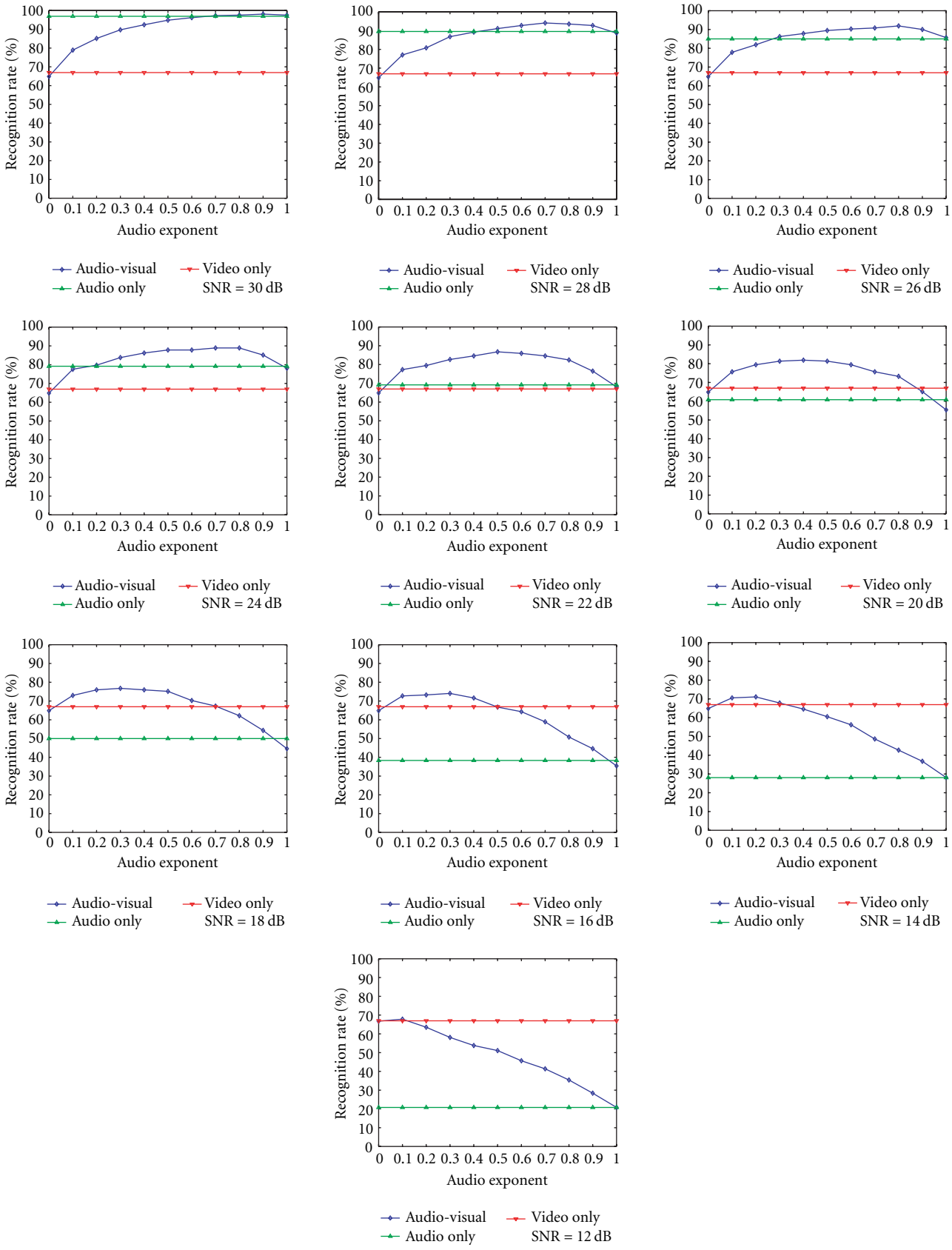


Figure 13: The coupled HMM-based audio-visual speech recognition rate dependency on the audio exponent for different values of the SNR.

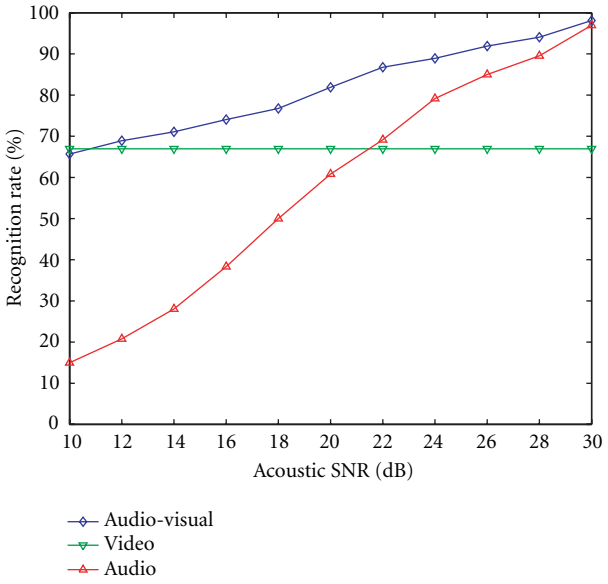


FIGURE 14: Comparison of the recognition rate of the audio-only, video-only and CHMM-based audio-visual speech recognition.

TABLE 6: The number of parameters and running time needed for the coupled HMM, and for the specific model used in AVSR.

| Space | CHMM | AV CHMM |
|---|---|---|
| $\mathbf{A}$ | $O((\sum_s N_s)(\prod_s N_s))$ | $O(\sqrt{\prod_{s=1}^{2} N_s})$ |
| $\boldsymbol{\mu}$ | $\sum_s N_s M_s V_s$ | $\sum_{s=1}^{2} N_s M_s V_s$ |
| $\mathbf{U}$ | $\sum_s N_s M_s V_s^2$ | $\sum_{s=1}^{2} N_s M_s V_s$ |
| $w$ | $\sum_s N_s M_s$ | $\sum_{s=1}^{2} N_s M_s$ |
| Time | CHMM | AV CHMM |
| Viterbi | $O(\prod_s N_s^2)$ | $O(\sqrt{\prod_{s=1}^{2} N_s})$ |

joint dependency of the observation node from all the backbone nodes in the same time slice $t$ for the AV FHMM and AV PHMM, or the joint state transition probabilities from all the backbone nodes at time $t-1$ for the AV PHMM and AV CHMM increases significantly the complexity of the Viterbi algorithm for these models. Unlike the AV FHMM, in the AV PHMM and AV CHMM the total number of possible audio-visual states is restricted by the sparse transition probability matrix (see (11) and (28)) reducing the stated decoding complexity of these models. Note that with $M_s \approx N_s$, which is the case for our experiments, or $M_s \gg N_s$, which is the case in large vocabulary applications, the dominant role in the complexity required by the recognition process is played by the number of calls to the exponential function needed per time step to evaluate the observation likelihoods. This number equals the total number of elements of mixture weights shown in the row labeled $w$. Therefore, we can conclude that in terms of both the size of the parameter space and the recognition complexity, the AV CHMM and AV IHMM compare closely and outperform the AV PHMM, especially the AV FHMM. However, unlike the AV IHMM, the coupling of the backbone nodes in the AV CHMM can model the correlation of the audio-visual components of speech. In the next section, we will complete the analysis of the above models with the experimental results in audio-visual speech recognition.

TABLE 7: A comparison of the video-only speech recognition rates for different video feature extraction techniques.

| Video features | Recognition rate |
|---|---|
| 1D DCT, LDA | 43.06% |
| Window, 1D DCT, LDA | 52.50% |
| 2D DCT blocks, LDA | 64.17% |
| Window, 2D DCT blocks, LDA | 66.94% |

## 6. EXPERIMENTAL RESULTS

We tested the speaker dependent isolated word audio visual recognition system on the CMU database [18]. Each word in the database is repeated ten times by each of the ten speakers in the database. For each speaker, nine examples of each word were used for training and the remaining example was used for testing. In our experiments we compared the accuracy of the audio-only, video-only and audio-visual speech recognition systems using the AV MSHMM, AV CHMM, AV FHMM, AV PHMM, and AV IHMM described in Section 4. For each of the audio-only and video-only recognition tasks, we model the observation sequences using a left-to-right HMM with five states, three Gaussian mixtures per state and diagonal covariance matrices. In the audio-only and all audio-visual speech recognition experiments, the audio sequences used in training are captured in clean acoustic conditions and the audio track of the testing sequences was altered by white noise at various SNR levels from 30 dB (clean) to 12 dB. The audio observation vectors consist of 13 MFC coefficients [6], extracted from overlapping frames of 20 ms. The visual observations are obtained using the cascade algorithm described in Section 3.

Table 7 shows the effect of the mouth region windowing and 2D-DCT coefficients extraction (window, 2D-DCT, LDA) on the visual-only recognition rate. It can be seen that the cascade algorithm that uses 2D-DCT coefficients extracted from eight non overlapping blocks of the mouth region followed by LDA (2D-DCT, LDA) outperforms the system that uses 32 1D-DCT coefficients extracted from the mouth region followed by LDA with the same number of coefficients (1D-DCT, LDA). In addition the use of mouth region windowing in the cascade algorithm (window, 1D-DCT, LDA or window, 2D-DCT, LDA) increases the recognition rate of the system without data windowing (1D-DCT, LDA and 2D-DCT, LDA, respectively).

In all audio-visual models the backbone nodes have five states and all mixture nodes are modeled using a mixture of

Table 8: A comparison of the speech recognition rate at different levels of acoustic SNR using an HMM for video-only features (V HMM), an HMM for audio-only features (A HMM), an MSHMM for audio-visual features (AV MSHMM), the independent audio-visual HMM (AV IHMM), the product audio-visual HMM (AV PHMM), the factorial audio-visual HMM (AV FHMM) and the coupled audio-visual HMM (AV CHMM).

| SNR (dB) | 30 | 28 | 26 | 24 | 22 | 20 | 18 | 16 | 14 | 12 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| V HMM (%) | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 | 66.9 |
| A HMM (%) | 96.9 | 89.5 | 85.0 | 79.2 | 69.2 | 60.8 | 50.0 | 38.3 | 28.0 | 20.8 | 15.0 |
| AV MSHMM (%) | 98.6 | 93.5 | 90.5 | 87.0 | 84.3 | 79.2 | 74.6 | 72.7 | 70.3 | 68.1 | 67.8 |
| AV IHMM (%) | 97.6 | 93.0 | 90.5 | 87.8 | 84.0 | 78.9 | 76.2 | 71.6 | 69.2 | 67.6 | 67.6 |
| AV PHMM (%) | 97.8 | 91.6 | 89.2 | 86.8 | 83.5 | 78.9 | 74.9 | 73.0 | 71.1 | 68.6 | 67.3 |
| AV FHMM (%) | 97.8 | 91.6 | 88.9 | 86.5 | 82.7 | 78.6 | 74.9 | 72.7 | 69.5 | 67.8 | 66.8 |
| AV CHMM (%) | 98.1 | 94.1 | 91.9 | 88.9 | 86.8 | 81.9 | 76.8 | 74.1 | 71.1 | 68.9 | 65.7 |

three Gaussian density functions, with diagonal covariance matrices. We trained all AV models using equal stream exponents ($\lambda_a = \lambda_v = 1$). In testing, the value of the stream exponents were chosen to maximize the average recognition rate for each value of the acoustic SNR. Our experimental results shown in Table 8 indicate that the CHMM-based audio-visual speech recognition system performs best overall, achieving the highest recognition rates in a wide range of SNR from 12 dB to 30 dB. As expected, all the audio-visual systems outperform significantly the audio-only recognition rate in noisy conditions, reaching about 50% reduction in the word error rate at SNR = 10 dB. Note that at SNR = 10 dB the AVSR recognition rate is practically bounded by the video-only recognition.

## 7. CONCLUSIONS

This paper studies the use of two types of dynamic Bayesian networks, the factorial and the coupled HMM, and compares their performances with existing models for audio-visual speech recognition. Both the FHMM and CHMM are generalizations of the HMM suitable for a large variety of multimedia applications that involve two or more streams of data. The parameters of the CHMM and FHMM, as special cases of DBN, can be trained using EM. However, EM is a local optimization algorithm that makes the choice of the initial parameters a critical issue. In this paper, we present an efficient method for the parameter initialization, using a Viterbi algorithm derived for each of the two models. For AVSR, the CHMM and the FHMM with two streams, one for audio and one for visual observation sequences, are particularly interesting. Both models allow for audio and visual state asynchrony, while still preserving the natural correlation of the audio and visual observations over time. With the FHMM, the audio and visual states are independent of each other, but they jointly model the likelihood of the audio-visual observation vector, and hence become correlated indirectly. On the other hand, with the CHMM, the likelihoods of the audio and visual observation vectors are modeled independently of each other, but each of the audio and visual states are conditioned jointly by the previous set of audio and visual states. The performance of the FHMM and the CHMM for speaker dependent isolated word AVSR was compared with existing models such as the multistream HMM, the independent HMM and the product HMM. The coupled HMM-based system outperforms all the other models at all SNR levels from 12 dB to 30 dB. The lower performance of the FHMM can be an effect of the large number of parameters required by this model, and the relatively limited amount of data in our experiments. In contrast, the efficient structure of the CHMM requires a small number of parameters, comparable to the independent HMM, without reducing the flexibility of the model. The best recognition accuracy in our experiments, the low parameter space, and the ability to exploit parallel computation make the CHMM a very attractive choice for audio visual integration. Our preliminary experimental results [30] show that the CHMM is a viable tool for speaker independent audio-visual continuous speech recognition.

## APPENDICES

### A. THE EM ALGORITHM FOR FHMM

The EM algorithm for the multistream FHMM (see (12)) with $\lambda_s = 1$ and general covariance matrices is described by the following steps.

*E Step*. The forward probability, defined as $\alpha_t(\mathbf{i}) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, \mathbf{q}_t = \mathbf{i})$, and the backward probability $\beta_t(\mathbf{i}) = P(\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T | \mathbf{q}_t = \mathbf{i})$ are computed as follows. Starting with the initial conditions

$$\alpha_1(\mathbf{i}) = \pi(\mathbf{i}) b_1(\mathbf{i}), \tag{A.1}$$

the forward probabilities are computed recursively from

$$\alpha_t(\mathbf{i}) = b_t(\mathbf{i}) \sum_{\mathbf{j}} a(\mathbf{i}|\mathbf{j}) \alpha_{t-1}(\mathbf{j}) \tag{A.2}$$

for $t = 2, 3, \ldots, T$. Similarly, from the initial conditions

$$\beta_T(\mathbf{i}) = 1, \tag{A.3}$$

the backward probabilities are computed recursively from

$$\beta_t(\mathbf{j}) = \sum_{\mathbf{i}} b_{t+1}(\mathbf{i}) a(\mathbf{i}|\mathbf{j}) \beta_{t+1}(\mathbf{i}) \tag{A.4}$$

for $t = T-1, T-2, \ldots, 1$. The transition probability $a(\mathbf{i}|\mathbf{j})$ is computed according to (15). The probability of the $r$th observation sequence $\mathbf{O}_r$ of length $T_r$, is computed as $P_r = \alpha_{r,T_r}(N_1, \ldots, N_S) = \beta_{r,1}(1, \ldots, 1)$ where $\alpha_{r,t}$, and $\beta_{r,t}$ are the forward and backward variables for the $r$th observation sequence.

*M Step*. The forward and backward probabilities obtained in the E step are then used to re-estimate the state parameters using

$$\begin{aligned}
\boldsymbol{\mu}^s_{\mathbf{i},m} &= \frac{\sum_r (1/P_r) \sum_t \gamma_{r,t}(\mathbf{i}, m) \mathbf{O}^s_{r,t}}{\sum_r (1/P_r) \sum_t \gamma_{r,t}(\mathbf{i}, m)}, \\
\mathbf{U}^s_{\mathbf{i},m} &= \frac{\sum_r (1/P_r) \sum_t \gamma_{r,t}(\mathbf{i}, m) (\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{\mathbf{i},m}) (\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{\mathbf{i},m})^T}{\sum_r (1/P_r) \sum_t \gamma_{r,t}(\mathbf{i}, m)}, \\
w_{\mathbf{i},m} &= \frac{\sum_r (1/P_r) \sum_t \gamma_{r,t}(\mathbf{i}, m)}{\sum_r (1/P_r) \sum_t \sum_{m'} \gamma_{r,t}(\mathbf{i}, m')},
\end{aligned} \tag{A.5}$$

where

$$\begin{aligned}
\gamma_{r,t}(\mathbf{i}, m) = {} & \frac{\alpha_{r,t}(\mathbf{i}) \beta_{r,t}(\mathbf{i})}{\sum_{\mathbf{i}} \alpha_{r,t}(\mathbf{i}) \beta_{r,t}(\mathbf{i})} \\
& \times \frac{w_{\mathbf{i},m} \prod_s N(\mathbf{O}^s_{r,t}, \boldsymbol{\mu}^s_{\mathbf{i},m}, \mathbf{U}^s_{\mathbf{i},m})}{\sum_{m'} w_{\mathbf{i},m'} \prod_s N(\mathbf{O}^s_{r,t}, \boldsymbol{\mu}^s_{\mathbf{i},m'}, \mathbf{U}^s_{\mathbf{i},m'})}.
\end{aligned} \tag{A.6}$$

The state transition probabilities can be estimated using

$$\tilde{a}^s(i|j) = \frac{\sum_r (1/P_r) \sum_{\mathbf{i},\mathbf{j}} \sum_t \alpha_{r,t}(\mathbf{j}) a(\mathbf{i}|\mathbf{j}) b_{r,t+1}(\mathbf{i}) \beta_{r,t+1}(\mathbf{i})}{\sum_r (1/P_r) \sum_t \sum_{\mathbf{j}} \alpha_{r,t}(\mathbf{j}) \beta_{r,t}(\mathbf{j})}, \tag{A.7}$$

where vectors $\mathbf{i}$ and $\mathbf{j}$ in (A.7) can be any state vectors such that $i_s = i$ and $j_s = j$, respectively.

## B. THE EM ALGORITHM FOR CHMM

The EM algorithm for the CHMM is described by the following steps.

*E Step*. The forward probability and backward probability and the observation probability $P_r$ are computed as in Appendix A, where the initial state probability $\pi(\mathbf{i})$, the observation probability $b_t(\mathbf{i})$, and the transition probability $a(\mathbf{i}|\mathbf{j})$ are computed as in (25), (26), and (27).

*M Step*. The forward and backward probabilities obtained in the E step are used to re-estimate the state parameters as follows:

$$\begin{aligned}
\boldsymbol{\mu}^s_{i,m} &= \frac{\sum_r (1/P_r) \sum_t \gamma^s_{r,t}(i, m) \mathbf{O}^s_{r,t}}{\sum_r (1/P_r) \sum_t \gamma^s_{r,t}(i, m)}, \\
\mathbf{U}^s_{i,m} &= \frac{\sum_r (1/P_r) \sum_t \gamma^s_{r,t}(i, m) (\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{i,m}) (\mathbf{O}^s_{r,t} - \boldsymbol{\mu}^s_{i,m})^T}{\sum_r (1/P_r) \sum_t \gamma^s_{r,t}(i, m)}, \\
w^s_{i,m} &= \frac{\sum_r (1/P_r) \sum_t \gamma^s_{r,t}(i, m)}{\sum_r (1/P_r) \sum_t \sum_{m'} \gamma^s_{r,t}(i, m')},
\end{aligned} \tag{B.1}$$

where

$$\begin{aligned}
\gamma^s_{r,t}(i, m) = {} & \frac{\sum_{\mathbf{i} \text{ s.t. } i_s = i} \alpha_{r,t}(\mathbf{i}) \beta_{r,t}(\mathbf{i})}{\sum_{\mathbf{i}} \alpha_{r,t}(\mathbf{i}) \beta_{r,t}(\mathbf{i})} \\
& \times \frac{w^s_{i,m} N(\mathbf{O}^s_{r,t}, \boldsymbol{\mu}^s_{i,m}, \mathbf{U}^s_{i,m})}{\sum_{m'} w^s_{i,m'} N(\mathbf{O}^s_{r,t}, \boldsymbol{\mu}^s_{i,m'}, \mathbf{U}^s_{i,m'})}.
\end{aligned} \tag{B.2}$$

The state transition probabilities can be estimated using

$$\tilde{a}^s(i|\mathbf{j}) = \frac{\sum_r (1/P_r) \sum_{\mathbf{i}} \sum_t \alpha_{r,t}(\mathbf{j}) a(\mathbf{i}|\mathbf{j}) b_{r,t+1}(\mathbf{i}) \beta_{r,t+1}(\mathbf{i})}{\sum_r (1/P_r) \sum_t \alpha_{r,t}(\mathbf{j}) \beta_{r,t}(\mathbf{j})}, \tag{B.3}$$

where $\mathbf{i}$ in (B.3) can be any state vector such that $i_s = i$.

## REFERENCES


[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] C. Neti, G. Potamianos, J. Luettin, et al., "Audio visual speech recognition, Final workshop 2000 report," Tech. Rep., Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA, 2000.

[3] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 994–999, San Juan, Puerto Rico, June 1997.

[4] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," in *Proc. Conf. Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., vol. 8, pp. 472–478, MIT Press, Cambridge, Mass, USA, 1995.

[5] T. Dean and K. Kanazawa, "A model for reasoning about persistence and causation," *Artificial Intelligence*, vol. 93, no. 1-2, pp. 1–27, 1989.

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

[7] J. Luettin, N. Thacker, and S. Beet, "Speechreading using shape and intensity information," in *Proc. the 4th IEEE International Conf. on Spoken Language Processing*, vol. 1, pp. 58–61, Philadelphia, Pa, USA, 1996.

[8] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 9–21, January 2001.

[9] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *Proc. IEEE International Conf. on Computer Vision*, pp. 494–499, Boston, Mass, USA, 1995.

[10] R. Kober, U. Harz, and J. Schiffers, "Fusion of visual and acoustic signals for command-word recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1495–1497, Munich, Germany, April 1997.

[11] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[12] A. Adjoudani and C. Benoît, "Audio-visual speech recognition compared across two architectures," in *European Conference on Speech Communication and Technology*, pp. 1563–1566, Madrid, Spain, 1995.
[13] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 165–168, Salt Lake City, Utah, USA, May 2001.
[14] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 169–172, Salt Lake City, Utah, USA, 2001.
[15] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 821–824, Atlanta, Ga, USA, May 1996.
[16] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition," in *IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 1073–1076, New York, NY, USA, 2000.
[17] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Human Language Technology Conference*, San Diego, Calif, USA, March 2002.
[18] Advanced Multimedia Processing Lab, http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/, Carnegie Mellon University, Pittsburgh, Pa, USA.
[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2000.
[20] K. R. Castleman, *Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1996.
[21] U. Kjaerulff, "A computational scheme for reasoning in dynamic probabilistic networks," in *Proc. the 8th International Conference on Uncertainty in Artificial Intelligence*, pp. 121–129, Stanford, Calif, USA, 1992.
[22] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive Processing of Sequences and Data Structures*, C. Giles and M. Gori, Eds., Lecture Notes in Artificial Intelligence, pp. 168–197, Springer-Verlag, Berlin, Germany, 1998.
[23] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Mateo, Calif, USA, 1988.
[24] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, USA, 2001.
[25] V. Pavlovic, *Dynamic Bayesian networks for information fusion with applications to human-computer interfaces*, Ph.D. thesis, University of Illinois, Urbana-Champaign, Ill, USA, 1999.
[26] S. Chu and T. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Proc. IEEE International Conf. on Spoken Language Processing*, vol. 2, pp. 747–750, Beijing, China, 2000.
[27] S. Chu and T. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2009–2012, Orlando, Fla, USA, May 2002.
[28] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, pp. 331–349, Springer-Verlag, Berlin, Germany, 1996.
[29] A. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2013–2016, Orlando, Fla, USA, May 2002.
[30] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. Nefian, "Speaker independent audio-visual continuous speech recognition," in *IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, August 2002.

**Ara V. Nefian** is a Senior Researcher at Intel Corporation, Microprocessor Research Labs in Santa Clara, California, USA. Ara received the engineering Diploma degree in electrical engineering in 1994 from the "Politehnica" University of Bucharest, Romania. In 1995, he received the MSEE degree and in 1999, the Ph.D. degree, all in electrical engineering from Georgia Tech, Atlanta. Current research interests include the study of graphical models for face and gesture recognition and audio-visual signal processing.

**Luhong Liang** is a Researcher at Intel China Research Center in Beijing. He received the Bachelor degree in 1997 and the Ph.D. degree in 2001, all in computer science from Tsinghua University in Beijing. His research interests include face detection and recognition, audio-visual signal process and biometrics.

**Xiaobo Pi** is a Researcher at Intel China Research Center in Beijing. Xiaobo received the Bachelor degree in electronic engineering in 1991 from the University of Electronic Science and Technology of China. In 1994, he received the Master degree in electronic engineering from Beijing Institute of Technology and in 1997, the Ph.D. degree in acoustics from Institute of Acoustics, Chinese Academy of Sciences. Current research interests include speech recognition and audio-visual signal processing.

**Xiaoxing Liu** received his M.S. in computer science from Fudan University. He joined Intel China Research Center in 1999. His research interests include speech recognition, speaker identification, speech analysis, and stochastic modeling.

**Kevin Murphy** received his Ph.D. in computer science from University of California, Berkeley in July 2002. His thesis was on "Dynamic Bayesian Networks: Representation, Inference, and Learning." He is about to start a postdoctorate at the MIT AI Lab, where he is planning to apply graphical models to computer vision. He will continue to consult for Intel.