

A Bayesian Approach to Audio-Visual Speaker Identification

Ara V Nefian¹, Lu Hong Liang¹, Tieyan Fu², and Xiao Xing Liu¹

¹ Microprocessor Research Labs, Intel Corporation,
{ara.nefian, lu.hong.liang, xiao.xing.liu}@intel.com,

² Computer Science and Technology Department, Tsinghua University,
futieyan00@mails.tsinghua.edu.cn

Abstract. In this paper we describe a text dependent audio-visual speaker identification approach that combines face recognition and audio-visual speech-based identification systems. The temporal sequence of audio and visual observations obtained from the acoustic speech and the shape of the mouth are modeled using a set of coupled hidden Markov models (CHMM), one for each phoneme-viseme pair and for each person in the database. The use of CHMM in our system is justified by the capability of this model to describe the natural audio and visual state asynchrony as well as their conditional dependence over time. Next, the likelihood obtained for each person in the database is combined with the face recognition likelihood obtained using an embedded hidden Markov model (EHMM). Experimental results on XM2VTS database show that our system improves the accuracy of the audio-only or video-only speaker identification at all levels of acoustic signal-to-noise ratio (SNR) from 5 to 30db.

1 Introduction

Increased interest in robust person identification systems leads to complex systems that rely often on the fusion of several type of sensors. Audio-visual speaker identification (AVSI) systems are particularly interesting due to their increased robustness to acoustic noise. These systems combine acoustic speech features with facial or visual speech features to reveal the identity of the user. As in audio-visual speech recognition the key issues for robust AVSI systems are the visual feature extraction and the audio-visual decision strategy.

Audio-visual fusion methods [22, 3] can be broadly grouped into two categories: feature fusion and decision systems. In feature fusion systems the observation vectors, obtained through the concatenation of acoustic and visual speech feature vectors, are described using a hidden Markov model (HMM). However, the audio and visual state synchrony assumed by these systems may not describe accurately the audio-visual speech generation. In comparison, in decision level systems the class conditional likelihood of each modality is combined at phone or word levels. Some of the most successful decision fusion models include the multi-stream HMM [18], or the product HMM [21, 4].

The Bayesian models also revealed their high modeling accuracy for face recognition. Recent face recognition systems using embedded Bayesian networks [14]

showed their improved performance over some template-based approaches [23, 1, 6, 17].

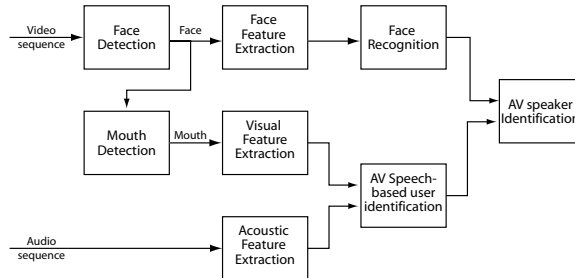


Fig. 1. The audio-visual speaker identification system.

The Bayesian approach to audio-visual speaker identification described in this paper (Figure 1) starts with the detection of the face and mouth in a video sequence. The facial features are used in the computation of the face likelihood (Section 2) while the visual features of the mouth region together with the acoustic features determine the likelihood of the audio-visual speech (Sections 3 and 4). Finally the face and the audio-visual speech likelihood are combined in a late integration scheme to reveal the identity of the user (Section 5).

2 The Face Model

While HMM are very successful in speech or gesture recognition, an equivalent two-dimensional HMM for images has been shown to be impractical due to its complexity [5]. Figure 2a shows a graph representation of the 2D HMM with the square nodes representing the discrete hidden nodes and the circles describing the continuous observation nodes. In recent years, several approaches to approx-

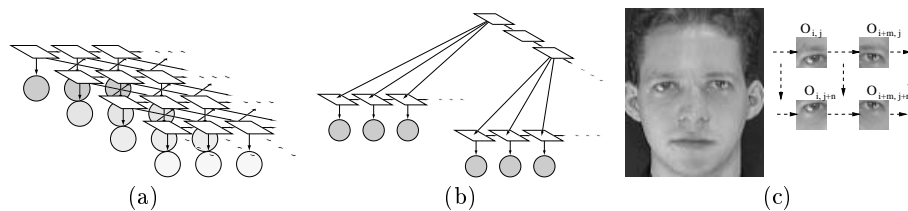


Fig. 2. A 2D HMM for face recognition (a), an embedded HMM (b) and the facial feature block extraction for face recognition (c).

imate a 2D HMM with computationally practical models have been investigated

[20, 14, 7]. In this paper, the face images are modeled using an embedded HMM (EHMM) [15]. The EHMM used for face recognition is a hierarchical statistical model with two layers of discrete hidden nodes (one layer for each data dimension) and a layer of observation nodes. In an EHMM both the “parent” and “child” layer of hidden nodes are described by a set of HMMs (Figure 2b). The states of the HMM in the “parent” and “child” layers are referred to as the *super states* and the states of the model respectively. The hierarchical structure of the EHMM or the embedded Bayesian networks [14] in general reduces significantly the complexity of these models compared to the 2D HMM. The sequence of observation vectors for an EHMM are obtained from a window that scans the image from left to right and top to bottom as shown in Figure 2c. Using the images in the training set, an EHMM is trained for each person in the database by means of the EM algorithm described in [15]. Recognition is carried out via the Viterbi decoding algorithm [12].

3 The Audio-Visual Speech Model

A coupled HMM (CHMM) [2] can be seen as a collection of HMMs, one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t - 1$ for all the related HMMs. Throughout this paper we will use CHMM with two channels, one for audio and

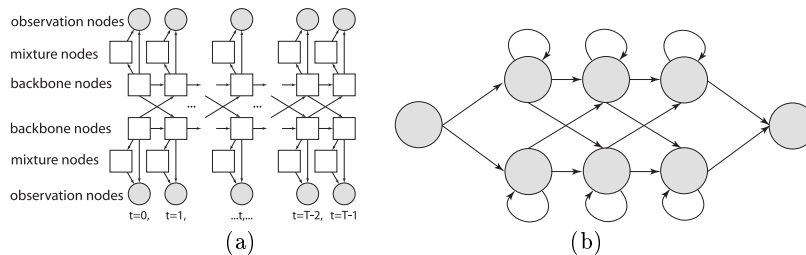


Fig. 3. A directed graph representation of a two channel CHMM with mixture components (a) and the state diagram representation of the CHMM used in our audio-visual speaker identification system (b).

the other for visual observations (Figure 3a). The parameters of a CHMM with two channels are defined below:

$$\begin{aligned}\pi_0^c(i) &= P(q_1^c = i) \\ b_t^c(i) &= P(\mathbf{O}_t^c | q_t^c = i) \\ a_{i|j,k}^c &= P(q_t^c = i | q_{t-1}^a = j, q_{t-1}^v = k)\end{aligned}$$

where $c \in \{a, v\}$ denotes the audio and visual channels respectively, and q_t^c is the state of the backbone node in the c th channel at time t . For a continuous

mixture with Gaussian components, the probabilities of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N(\mathbf{O}_t^c, \mu_{i,m}^c, \mathbf{U}_{i,m}^c)$$

where \mathbf{O}_t^c is the observation vector at time t corresponding to channel c , and $\mu_{i,m}^c$ and $\mathbf{U}_{i,m}^c$ and $w_{i,m}^c$ are the mean, covariance matrix and mixture weight corresponding to the i th state, the m th mixture and the c th channel. M_i^c is the number of mixtures corresponding to the i th state in the c th channel. In our audio-visual speaker identification system, each CHMM describes one of the possible phoneme-viseme pairs as defined in [16], for each person in the database.

4 Training the CHMM

The training of the CHMM parameters for the task of audio-visual speaker identification is performed in two stages. First, a speaker-independent background model (BM) is obtained for each CHMM corresponding to a viseme-phoneme pair. Next, the parameters of the CHMMs are adapted to a speaker specific model using a maximum a posteriori (MAP) method. To deal with the requirements of a continuous speech recognition systems, two additional CHMMs are trained to model the silence between consecutive words and sentences.

4.1 Maximum Likelihood Training of the Background Model

In the first stage, the CHMMs for isolated phoneme-viseme pairs are initialized using the Viterbi-based method described in [13] followed by the estimation-maximization (EM) algorithm [10]. Each of the models obtained in the first stage is extended with one entry and one exit non-emitting state (Figure 3 b). The use of the non-emitting states also enforces the phoneme-viseme synchrony at the model boundaries. Next, the parameters of the CHMMs are refined through the embedded training of all CHMM from continuous audio-visual speech [10]. In this stage, the labels of the training sequences consist only of the sequence of phoneme-visemes with all boundary information being ignored. We will denote the mean, covariance matrices and mixture weights for mixture m , state i , and channel c of the trained CHMM corresponding to the background model as $(\mu_{i,m}^c)_{BM}$, $(\mathbf{U}_{i,m}^c)_{BM}$, and $(w_{i,m}^c)_{BM}$ respectively.

4.2 Maximum A Posteriori Adaptation

In this stage of the training, the state parameters of the background model are adapted to the characteristics of each speaker in the database. The new state parameters for all CHMMs $\hat{\mu}_{i,m}^c$, $\hat{\mathbf{U}}_{i,m}^c$ and $\hat{w}_{i,m}^c$ are obtained through Bayesian

adaptation [19]:

$$\hat{\mu}_{i,m}^c = \theta_{i,m}^c \mu_{i,m}^c + (1 - \theta_{i,m}^c) (\mu_{i,m}^c)_{BM} \quad (1)$$

$$\hat{\mathbf{U}}_{i,m}^c = \theta_{i,m}^c \mathbf{U}_{i,m}^c - (\mu_{i,m}^c)^2 + (\mu_{i,m}^c)_{BM}^2 + (1 - \theta_{i,m}^c) (\mathbf{U}_{i,m}^c)_{BM} \quad (2)$$

$$\hat{w}_{i,m}^c = \theta_{i,m}^c w_{i,m}^c + (1 - \theta_{i,m}^c) (w_{i,m}^c)_{BM}, \quad (3)$$

where $\theta_{i,m}^c$ is a parameter that controls the MAP adaptation for mixture component m in channel c and state i . The sufficient statistics of the CHMM states corresponding to a specific user, $\mu_{i,m}^c$, $\mathbf{U}_{i,m}^c$ and $w_{i,m}^c$ are obtained using the EM algorithm from the available speaker dependent data as follows:

$$\begin{aligned} \mu_{i,m}^c &= \frac{\sum_{r,t} \gamma_{r,t}^c(i, m) \mathbf{O}_{r,t}}{\sum_{r,t} \gamma_{r,t}^c(i, m)} \\ \mathbf{U}_{i,m}^c &= \frac{\sum_{r,t} \gamma_{r,t}^c(i, m) (\mathbf{O}_{r,t}^c - \mu_{i,m}^c) (\mathbf{O}_{r,t}^c - \mu_{i,m}^c)^T}{\sum_{r,t} \gamma_{r,t}^c(i, m)} \\ w_{i,m}^c &= \frac{\sum_{r,t} \gamma_{r,t}^c(i, m)}{\sum_{r,t} \sum_k \gamma_{r,t}^c(i, k)}, \end{aligned}$$

where

$$\gamma_{r,t}^c(i, m) = \frac{\sum_j \frac{1}{P_r} \alpha_{r,t}(i, j) \beta_{r,t}(i, j)}{\sum_{i,j} \frac{1}{P_r} \alpha_{r,t}(i, j) \beta_{r,t}(i, j)} \frac{w_{i,m}^c N(\mathbf{O}_{r,t}^c | \mu_{i,m}^c, \mathbf{U}_{i,m}^c)}{\sum_k w_{i,k}^c N(\mathbf{O}_{r,t}^c | \mu_{i,k}^c, \mathbf{U}_{i,k}^c)},$$

and $\alpha_{r,t}(i, j) = P(\mathbf{O}_{r,1}, \dots, \mathbf{O}_{r,t} | q_{r,t}^a = i, q_{r,t}^v = j)$ and $\beta_{r,t}(i, j) = P(\mathbf{O}_{r,t+1}, \dots, \mathbf{O}_{r,T_r} | q_{r,t}^a = i, q_{r,t}^v = j)$ are the forward and backward variables respectively [10] computed for the r th observation sequences $\mathbf{O}_{r,t} = [(\mathbf{O}_{r,t}^a)^T, (\mathbf{O}_{r,t}^v)^T]^T$. The adaptation coefficient is

$$\theta_{i,m}^c = \frac{\sum_{r,t} \gamma_{r,t}^c(i, m)}{\sum_{r,t} \gamma_{r,t}^c(i, m) + \delta},$$

where δ is the relevance factor, which is set $\delta = 16$ in our experiments. Note that as more speaker dependent data for a mixture m of state i and channel c becomes available, the contribution of the speaker specific statistics to the MAP state parameters increases (Equations 1- 3). On the other hand, when less speaker specific data is available, the MAP parameters are very close to the parameters of the background model.

5 Recognition

Given an audio-visual test sequence, the recognition is performed in two stages. First the face likelihood and the audio-visual speech likelihood are computed separately. To deal with the variation in the relative reliability of the audio and visual features of speech at different levels of acoustic noise, we modified the

observation probabilities used in decoding such that $\tilde{b}_t^c(i) = [b_t^c]^{\lambda_c}$, $c \in \{a, v\}$ where the audio and video stream exponents λ_a and λ_v satisfy $\lambda_a, \lambda_v \geq 0$ and $\lambda_a + \lambda_v = 1$. Then the overall matching score of the audio-visual speech and face model is computed as

$$L(\mathbf{O}^f, \mathbf{O}^a, \mathbf{O}^v | k) = \lambda_f L(\mathbf{O}^f | k) + \lambda_{av} L(\mathbf{O}^a, \mathbf{O}^v | k) \quad (4)$$

where \mathbf{O}^a , \mathbf{O}^v and \mathbf{O}^f are the acoustic speech, visual speech and facial sequence of observations, $L(*|k)$ denotes the observation likelihood for the k th person in the database and $\lambda_f, \lambda_{av} \geq 0, \lambda_f + \lambda_{av} = 1$ are some weighting coefficients for the face and audio-visual speech likelihoods.

6 Experimental Results

The audio-visual speaker identification system presented in this paper was tested on digit enumeration sequences from the XM2VTS database [11]. For parameter adaptation we used four training sequences from each of the 87 speakers in our training set while for testing we used 320 sequences.

In our experiments the acoustic observation vectors consist of 13 Mel frequency cepstral (MFC) coefficients with their first and second order time derivatives, extracted from windows of 25.6ms, with an overlap of 15.6ms. The extraction of the visual features starts with the face detection system described in [9] followed by the detection and tracking of the mouth region using a set of support vector machine classifiers. The features of the visual speech are obtained from the mouth region through a cascade algorithm described in [8]. The pixels in the mouth region are mapped to a 32-dimensional feature space using the principal component analysis. Then, blocks of 15 consecutive visual observation vectors are concatenated and projected on a 13 class, linear discriminant space. Finally, the resulting vectors, with their first and second order time derivatives are used as the visual observation sequences. The audio and visual features are integrated using a CHMM with three states in both the audio and video chains with no back transitions (Figure 3b). Each state has 32 mixture components with diagonal covariance matrices.

In our system, the facial features are obtained using a sampling window of size 8×8 with 75% overlap between consecutive windows. The observation vectors corresponding to each position of the sampling window consist of a set of 2D discrete Cosine transform (2D DCT) coefficients. Specifically, we used nine 2D DCT coefficients obtained from a 3×3 region around the lowest frequency in the 2D DCT domain. The faces of all people in the database are modeled using EHMM with five super states and 3,6,6,6,3 states per super state respectively. Each state of the hidden nodes in the ‘‘child’’ layer of the EHMM is described by a mixture of three Gaussian density functions with diagonal covariance matrices.

To evaluate the behavior of our speaker identification system in environments affected by acoustic noise, we corrupted the testing sequences with white Gaussian noise at different SNR levels, while we trained on the original clean acoustic sequences. Figure 4 shows the error rate of the audio-only $(\lambda_v, \lambda_f) = (0.0, 0.0)$,

video-only $(\lambda_v, \lambda_f) = (1.0, 0.0)$, audio-visual $(\lambda_f = 0.0)$ and face-audio-visual $(\lambda_v, \lambda_f) = (0.5, 0.3)$ speaker identification system at different SNR levels.

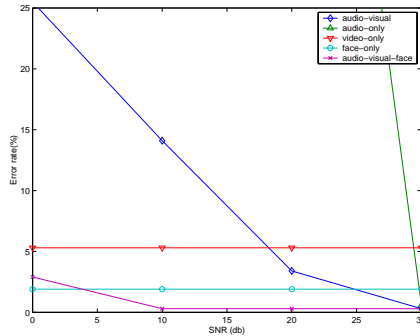


Fig. 4. The error rate of the audio-only $(\lambda_v, \lambda_f) = (0.0, 0.0)$, video-only $(\lambda_v, \lambda_f) = (1.0, 0.0)$, audio-visual $(\lambda_f = 0.0)$ and face+audio-visual $(\lambda_v, \lambda_f) = (0.5, 0.3)$ speaker identification system.

7 Conclusions

In this paper, we described a Bayesian approach for text dependent audio-visual speaker identification. Our system uses a hierarchical decision fusion approach. In the lower level the acoustic and visual features of speech are integrated using a CHMM and the face likelihood is computed using an EHMM. In the upper level, a late integration scheme combines the likelihood of face and audio-visual speech to reveal the identity of the speaker.

The use of strongly correlated acoustic and visual temporal features of speech together with overall facial characteristics makes the current system very difficult to break, and increases its accuracy to acoustic noise. The study of the recognition performance in environments corrupted by acoustic noise shows that our system outperforms the audio-only baseline system by a wide margin.

References

1. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs Fisherfaces: Recognition using class specific linear projection. In *Proceedings of Fourth European Conference on Computer Vision, ECCV'96*, pages 45–58, April 1996.
2. M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997.
3. C. Chibelushi, F. Deravi, and S.D. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, March 2002.
4. S. Dupont and J. Luetttin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2:141–151, September 2000.

5. S. Kuo and O.E. Agazzi. Keyword spotting in poorly printed documents using pseudo 2-D Hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):842–848, August 1994.
6. A. Lawrence, C.L. Giles, A.C Tsoi, and A.D. Back. Face recognition : A convolutional neural network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.
7. Jia Lia, A. Najmi, and R.M. Gray. Image classification by a two-dimensional hidden markov model. *IEEE Transactions on Signal Processing*, 48(2):517–533, February 2000.
8. L. Liang, X. Liu, X. Pi, Y. Zhao, and A. V. Nefian. Speaker independent audio-visual continuous speech recognition. In *International Conference on Multimedia and Expo*, volume 2, pages 25–28, 2002.
9. R. Lienhart and J. Maydt. An extened set of Haar-like features for rapid objection detection. In *IEEE International Conference on Image Processing*, volume 1, pages 900–903, 2002.
10. X. Liu, L. Liang, Y. Zhao, X. Pi, and A. V. Nefian. Audio-visual continuous speech recognition using a coupled hidden Markov model. In *International Conference on Spoken Language Processing*, 2002.
11. J. Luetttin and G. Maitre. Evaluation protocol for the XM2FDB database. In *IDIAP-COM 98-05*, 1998.
12. A. V. Nefian and M. H. Hayes. Face recognition using an embedded HMM. In *Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication*, pages 19–24, March 1999.
13. A. V. Nefian, L. Liang, X. Pi, X. Liu, and C. Mao. A coupled hidden Markov model for audio-visual speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 2013–2016, 2002.
14. Ara V. Nefian. Embedded Bayesian networks for face recognition. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 25–28, 2002.
15. Ara V. Nefian and Monson H. Hayes III. Maximum likelihood training of the embedded HMM for face detection and recognition. In *IEEE International Conference on Image Processing*, volume 1, pages 33–36, 2000.
16. C. Neti, G. Potamianos, J. Luetttin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio visual speech recognition. In *Final Workshop 2000 Report*, 2000.
17. Jonathon Phillips. Matching pursuit filters applied to face identification. *IEEE Transactions on Image Processing*, 7(8):1150–1164, August 1998.
18. G. Potamianos, J. Luetttin, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 169–172, 2001.
19. D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using an adapted gaussian mixture model. *Digital Signal Processing*, 10:19–41, 2000.
20. F. Samaria. *Face Recognition Using Hidden Markov Models*. PhD thesis, University of Cambridge, 1994.
21. L.K. Saul and M.L. Jordan. Boltzmann chains and hidden Markov models. In G. Tesauro, David S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. The MIT Press, 1995.
22. S.B.Yacouband, S.Luetttin, J.Jonsson, K.Matas, and J.Kittler. Audio-visual person verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 580–585, 1999.
23. M. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings of International Conference on Pattern Recognition*, pages 586 – 591, 1991.