

AUDIO-VISUAL SPEAKER IDENTIFICATION USING COUPLED HIDDEN MARKOV MODELS

Tieyan Fu¹, Xiao Xing Liu², Lu Hong Liang², Xiaobo Pi² and Ara V. Nefian²

¹Department of Computer Science and Technology, Tsinghua University
futieyan00@mails.tsinghua.edu.cn

²Microprocessor Research Labs, Intel Corporation
{xiao.xing.liu, lu.hong.liang, xiaobo.pi, ara.nefian}@intel.com

ABSTRACT

In this paper, we investigate the use of the coupled hidden Markov models (CHMM) for the task of audio-visual text dependent speaker identification. Our system determines the identity of the user from a temporal sequence of audio and visual observations obtained from the acoustic speech and the shape of the mouth, respectively. The multi modal observation sequences are then modeled using a set of CHMMs, one for each phoneme-viseme pair and for each person in the database. The use of CHMMs in our system is justified by the capacity of this model to describe the natural audio and visual state asynchrony as well as their conditional dependency over time. To train a CHMM we first train a speaker independent model using expectation-maximization (EM), and then we build a speaker dependent model using maximum a posteriori (MAP) training. Experimental results on XM2VTS database show that our system improves the accuracy of audio-only or video-only speaker identification at all levels of acoustic signal-to-noise ratio (SNR) from 0 to 30db.

1. INTRODUCTION

The increased interest for robust person identification systems led to complex system that rely often on the fusion of several type of sensors. Audio-visual speaker identification systems are particularly interesting due to their increased robustness to acoustic and visual noise. In our work the sequence of visual features extracted from the mouth shape over time is combined with the features of acoustic speech to obtain a reliable speaker identification system. The motivation of our approach is supported by recent audio-visual speech and speaker recognition systems that empirically demonstrate the strong correlation between acoustic and visual speech [7, 1, 17, 13]. udio-visual fusion methods [17, 4] can be broadly grouped into two categories: feature fusion and decision systems. In feature fusion systems the observation vectors, obtained through the concate-

nation 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 [14], or the product HMM [16, 5].

The coupled hidden Markov model (CHMM) based audio-visual speaker identification system presented in this paper can be seen as an extension of the decision fusion system at phoneme-viseme level. A CHMM allows for audio-visual state asynchrony as well as captures the natural conditional dependencies between the two modalities at the state level. Recently it has been shown that CHMMs outperform both the product and the multi-stream HMM for the task isolated word audio-visual speech recognition [12].

The outline of this paper is as follows. In Section 2 we give a formal definition of the CHMM used in our system. The training and decoding of the CHMM for the speaker identification task are detailed in Sections 3 and 4 respectively. The experimental results are discussed in Section 5, while Section 6 presents the conclusions of our work and proposes directions for future research.

2. THE AUDIO-VISUAL CHMM

A CHMM [2] can be seen as a collection of hidden Markov models, 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 the other for visual observations. The parameters of a CHMM with two channels are defined below:

$$\begin{aligned}\pi_0^c(i) &= P(q_1^c = i) \\ b_t^c(i) &= P(\mathbf{O}_t^c | q_t^c = i)\end{aligned}$$

$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^a = j, q_{t-1}^v = k)$$

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 . In 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, m th mixture and the c th channel. M_i^c represents 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 phoneme-viseme pairs, defined in [13], for each person in the database.

3. TRAINING

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 phoneme-viseme 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.

3.1. 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 [11] followed by the estimation-maximization (EM) algorithm [9]. Each of the models obtained in the first stage are extended with one entry and one exit non-emitting states. The use of the non-emitting states also enforces the phoneme-viseme synchrony at the model boundaries (Figure 1). Next,

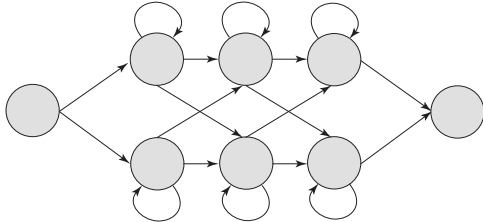


Fig. 1. The state diagram of the coupled HMM used in our audio-visual speaker identification system.

the parameters of the CHMMs are refined through the embedded training of all CHMM from continuous audio-visual speech [9]. In this stage, the labels of the training sequences consist of the sequence of phoneme-viseme 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.

3.2. 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 [15]:

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

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

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

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:

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

where

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

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 [9] 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 obtained as

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

where δ is the relevance factor which is set $\delta = 16$ in our experiments. Note that as more speaker dependent data for 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 side, when less speaker specific data is available, the MAP parameters are very close to the parameters of the background model.

4. RECOGNITION

The decoding of a test audio-visual sequence is carried out via a graph decoder [6, 9]. The speaker model with the highest likelihood reveals the identity of the user. Note that our text dependent speaker identification does not make use of the phoneme-viseme forced alignment. To deal with the variation in the relative reliability of the audio and visual features at different levels of acoustic noise, we modified the observation probabilities used in decoding such that $\tilde{b}_t^c(i) = [b_t^c]^{\lambda_c}$ where the audio and video stream exponents λ_a and λ_v satisfy $\lambda_a, \lambda_v \geq 0$ and $\lambda_a + \lambda_v = 1$. The values λ_a, λ_v corresponding to a specific acoustic SNR level are obtained experimentally to minimize the average word error rate.

5. EXPERIMENTAL RESULTS

In our experiments the acoustic observation vectors consist of 13 Mel frequency cepstral coefficients (MFCC) with their first and second order time derivatives extracted from a window of 25.6 ms, with an overlap of 15.6 ms. The visual features are obtained from the mouth region through a cascade algorithm described in [8]. The extraction of the visual features starts with the neural network based face detection system followed by the detection and tracking of the mouth region using a set of support vector machine classifiers (Figure 2).

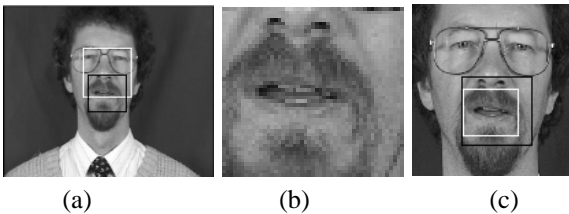


Fig. 2. (a) An example of the face detection (white rectangle), and the estimated region of search for the mouth (black rectangle). (b) The estimated region of search for the mouth, enlarged. (c) The mouth detection result (white rectangle) from the initial region of search for the mouth (black rectangle).

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 of size 13, 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 1). Each state has 32 mixture components with diagonal covariance matrices. The audio-visual speaker identification system presented in this paper was tested on digit enumeration sequences from the XM2VTS database [10]. 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. 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. Table 1 describes the error rate in speaker identification at different levels of acoustic SNR, and for different values of the visual stream exponent.

SNR(db)	30	20	10	0
$\lambda_v=0.0$	0.6	80.3	95.9	99.7
$\lambda_v=0.2$	0.0	34.7	80.9	95.3
$\lambda_v=0.3$	0.0	13.8	55.0	81.6
$\lambda_v=0.5$	0.3	3.4	14.1	25.5
$\lambda_v=0.7$	1.2	2.5	4.4	5.9
$\lambda_v=0.8$	1.3	2.5	3.4	4.4
$\lambda_v=1.0$	5.3	5.3	5.3	5.3

Table 1. The error rate of the audio-visual speaker identification system for several SNR levels and visual stream exponents λ_v .

It can be seen that the improvement in speaker identification at lower SNR is obtained for higher values of λ_v , while at higher SNR a smaller error rate is obtained for smaller λ_v . When the audio features are ignored ($\lambda_v = 1.0$), the error rate of the visual only speaker identification is, as expected, independent of the acoustic SNR. However, at low values of SNR, the use of visual features decreases significantly the error rate of the audio-only system from 99.7% to 4.4% for the best $\lambda_v=0.8$. Also note that, at all SNR values the best performance is obtained for a combination of audio and video features, no single modality system outperforming the audio-visual speaker identification. Figure 3 displays the error rate of the audio-only video-only and audio-visual speaker identification system for best values of λ_v at different SNR levels.

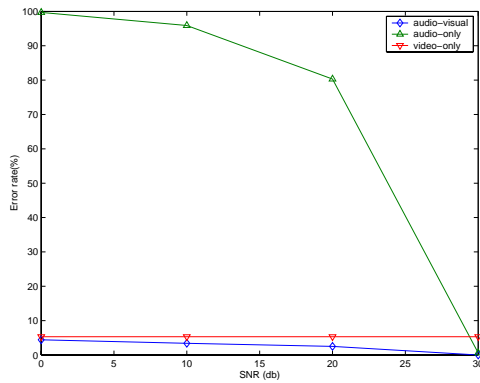


Fig. 3. The error rate of the audio-only, video-only and audio-visual speaker identification system.

6. CONCLUSIONS

In this paper, we introduce a novel text dependent audio-visual speaker identification system that models jointly the acoustic and visual features of speech using a CHMM. The use of strongly correlated acoustic and visual temporal features makes the current system more difficult to break, and more accurate than acoustic only speaker identification systems.

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. In fact the error rate of the audio-only system at SNR=0db is reduced by over 95%.

Future research will be directed toward the integration of the current audio-visual speaker identification with a face recognition system. It is also interesting to study the performance of the current system for different types of acoustic noise as well as visual noise. The similarity between the visual features used in this work and the visual features used in audio visual speech recognition allows for the integration of the two systems in a text independent audio-visual speaker identification.

7. REFERENCES

- [1] A.Senior, C.Neti, and B.Maison. On the use of visual information for improving audio-based speaker recognition. In *In Proc. of Audio Visual Speech Processing Conference*, 1999.
- [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] T. Chen. Audiovisual speech processing. *Signal Processing Magazine*, 18:9–21, January 2001.
- [4] 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.
- [5] S. Dupont and J. Luetin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2:141–151, September 2000.
- [6] S. Young et. al. *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, UK, 1995.
- [7] J.Luetin, N.Thacker, and S.Beer. Speaker identification by lipreading. In *In Proc. of International Conference on Spoken Language Processing*, pages 62–64, 1996.
- [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] 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.
- [10] J. Luetin and G. Maitre. Evaluation protocol for the XM2FDB database. In *IDIAP-COM 98-05*, 1998.
- [11] 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.
- [12] A.V. Nefian, L. Liang, X. Liu, X. Pi, C. Mao, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Applied Signal Processing, special issue on Audio Visual Signal Processing*, 2002, to appear.
- [13] C. Neti, G. Potamianos, J. Luetin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio visual speech recognition. In *Final Workshop 2000 Report*, 2000.
- [14] G. Potamianos, J. Luetin, 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.
- [15] 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.
- [16] 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.
- [17] S.B.Yacouband, S.Luetin, 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.
- [18] P. Teissier, J. Rober-Ribes, J.-L. Schwartz, and A. Guerin-Dugue. Comparing models for audio-visual fusion in a noisy vowel recognition task. *IEEE Transactions on Speech and Audio and Signal Processing*, 7:629–642, 1999.