A COUPLED HMM FOR AUDIO-VISUAL SPEECH RECOGNITION

Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao and Kevin Murphy

Microcomputer Research Labs, Intel Corporation
Santa Clara, CA, 95052
{ara.nefian, luhong.laing, xioabo.pi, xioaxing.liu, crusoe.mao, kevin.murphy}@intel.com

ABSTRACT

In recent years several speech recognition systems that use visual together with audio information showed significant increase in performance over the standard speech recognition systems. The use of visual features is justified by both the bimodality of the speech generation and by the need of features that are invariant to acoustic noise perturbation. The audio-visual speech recognition system presented in this paper introduces a novel audio-visual fusion technique that uses a coupled hidden Markov model (HMM). The statistical properties of the coupled-HMM allow us to model the state asynchrony of the audio and visual observations sequences while still preserving their natural correlation over time. The experimental results show that the coupled HMM outperforms the multistream HMM in audio visual speech recognition.

1. INTRODUCTION

The variety of applications of automatic speech recognition (ASR) systems, for human computer interfaces, telephony, or robotics has driven the research of a large scientific community over last decades. The success of currently available ASR systems is however restricted to relatively controlled environments and well defined applications such as dictation or small to medium vocabulary voice-based control commands (hand free dialing, etc). Often, robust ASR systems require special positioning of the microphone with respect to the speaker resulting in a rather unnatural human-machine interface. Together with the investigation of several acoustic noise reduction techniques, in recent years the study of visual features has emerged as an attractive solution to speech recognition under less constrained environments. The use of visual features in audio-visual speech recognition (AVSR) is motivated by the bimodality of the speech formation and the ability of humans to better distinguish spoken sounds when both audio and video are available [10]. In addition, the visual information provides the system with complementary features that can not be corrupted by the acoustic noise of the environment. Although the use of visual features for a robust speech recognition system appears natural, there are several questions to be answered such as what is a robust set of visual features, what is the best means of audio and visual feature integration, and what represents the best model for audio-visual data. In this paper, we will describe an AVSR system that uses a coupled HMM (CHMM) for the audio visual integration and compare our approach with the multistream HMM [12].

2. RELATED WORK

Several of the current AVSR systems were evaluated and results were presented in [11]. Since the acoustic features used in speech recognition are now well understood [13], the main issues remain the choice of the visual features and the choice of the fusion model for the audio and visual data.

The visual features are often derived from the shape of the mouth [9] [4], [2]. 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 or 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 were described for principal component analysis [2], linear discriminant analysis, two-dimensional DCT and maximum likelihood linear transform [11]. Methods that combine shape and appearance modeling were presented in [7] and [11].

The audio and visual fusion techniques investigated in previous work include feature fusion, model fusion, or decision fusion. In feature fusion, the combined audio-visual feature vectors are obtained by the concatenation of the audio and visual features, followed by a dimensionality reduction transform [11]. The resulting observation sequences are then modeled using one HMM [13]. A model fusion system based on multistream HMM was proposed in [12]. The multi stream HMM assumes that audio and video sequences are state synchronous but allows the audio and video components to have different contribution to the overall observation likelihood. However, it is well known that the acoustic features of speech are delayed from the visual features of speech, and assuming state synchronous models can be inaccurate. Dupont and Luettin [7] proposed an audio visual model that uses a product HMM. The audio visual product HMM can be seen as an extension of the multi-stream HMM that allows for audio-video state asynchrony. Decision fusion systems model independently the audio and video sequences using two HMMs, and combine the likelihood of each observation sequence based on the reliability of each modality ([11]).

3. THE VISUAL FEATURE EXTRACTION

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 [5]. The lower half of the detected face (Figure 1a) is a natural choice for the initial estimate of the mouth region.

Next, linear discriminant analysis (LDA) is used to assign the pixels in the mouth region to the lip and face classes (Figure 1b).

LDA transforms the pixel values from the RGB chromatic space into an one-dimensional space that best discriminates between the two classes. The optimal linear discriminat space [6] is computed off-line using a set of manually segmented images of the lip and face regions.

The contour of the lips is obtained through the binary chain encoding method [3] followed by a smoothing operation. The refined position of the mouth corners is obtained by applying the corner finding filter $w[m,n] = \exp(-\frac{|m^2+n^2|}{2\sigma^2})$, $\sigma^2 = 70$, $-3 < m,n \leq 3$, in a window around the left and right extremities of the lip contour. The result of the lip contour and mouth corners detection is illustrated in Figure 1c.

(a) (b) (c)

Figure 1: (a) The original estimate of the mouth region. (b) Segmented lip region (black) using LDA. (c) The lip contour and the corners of the mouth

The lip contour and position of the mouth corners are used to estimate the size and the rotation of the mouth in the image plane. Using the above estimates of the scale and rotation parameters of the mouth, a rotation and size normalized grayscale region of the mouth (64 × 64 pixels) is obtained from each frame of the video sequence. However, not all the pixels in the mouth region have the same relevance for visual speech recognition. In our experiments we found that the most significant information for speech recognition is contained in the pixels inside the lip contour. The masking variable shape window, used to multiply the pixels values in the grayscale normalized mouth region, is described below:

$$w[i,j] = \begin{cases} 1, \ if \ i,j \ are \ inside \ the \ lip \ contour, \\ 0, \ otherwise \end{cases} \quad (1)$$

Figure 2 illustrates the result of the rotation and size normalization and masking steps.

Figure 2: The masked, size and rotation normalized mouth region

Next, the normalized and masked mouth region is decomposed in 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 × 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 features extraction cascade the multi class LDA [6], 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.

Table 1 compares the video-only recognition rates for several visual feature techniques and illustrates the improvement obtained by using the masking window and the use of the block 2D-DCT coefficients instead of 1D-DCT coefficients as described in [11]. In all the experiments the video observation vectors were modeled using a 5 state, 3 mixture left-to-right HMM with diagonal covariance matrices.

Video Features	Recognition Rate
1D DCT + LDA	41.66%
Mask, 1D DCT + LDA	45.17%
2D DCT blocks + LDA	45.63%
Mask, 2D DCT blocks + LDA	54.08%

Table 1: A comparison of the video-only speech recognition rates for different video features extraction techniques

4. THE AUDIO-VISUAL MODEL

This paper introduces a novel model for audio-visual speech recognition, that uses a coupled hidden Markov model (CHMM). The CHMM [1] is a generalization of the HMM suitable for a large variety of multimedia applications that integrate two or more streams of data. A coupled HMM can be seen as a collection of HMMs, one for each data stream, where the discrete nodes at time t for each HMM are conditioned by the discrete nodes at time $t-1$ of all the related HMMs. Figure 3 illustrates a continuous mixture two-stream coupled HMM used in our audio-visual speech recognition system. The squares represent the hidden discrete nodes while the circles describe the continuous observable nodes. We will refer to the hidden nodes conditioned temporaly as coupled nodes and to the remaining hidden nodes as mixture nodes. The parameters of a CHMM are defined below:

$$\pi_0^c(i) = P(q_t^c = i) \quad (2)$$

$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i) \quad (3)$$

$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^0 = j, q_{t-1}^1 = k) \quad (4)$$

where q_t^c is the state of the couple node in the cth stream at time t. In a continuous mixture with Gaussian components, the probabili-

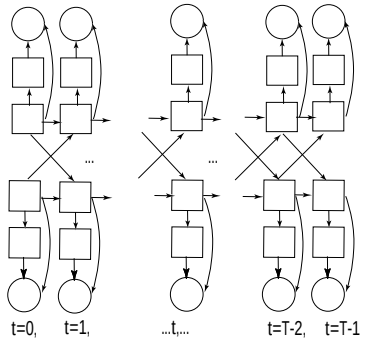

Figure 3: The audio-visual coupled HMM.

ties of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N(\mathbf{O}_t^c, \mu_{i,m}^c, \mathbf{U}_{i,m}^c) \quad (5)$$

where $\mu_{i,m}^c$ and $\mathbf{U}_{i,m}^c$ are the mean and covariance matrix of the ith state of a coupled node, and mth component of the associated mixture node in the cth channel. M_i^c is the number of mixtures corresponding to the ith state of a coupled node in the cth stream and the weight $w_{i,m}^c$ represents the conditional probability $P(s_t^c = m|q_t^c = i)$ where s_t^c is the component of the mixture node in the cth stream at time t.

5. TRAINING

The maximum likelihood (ML) training of the dynamic Bayesian networks in general and of the coupled HMMs in particular, is a well understood technique [8]. However, the iterative maximum likelihood estimation of the parameters 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 the initialization of the ML training that uses a Viterbi algorithm derived for the coupled HMM. The Viterbi algorithm determines the optimal sequence of states for the coupled nodes of the audio and video streams that maximizes the observation likelihood. The following steps describe the Viterbi algorithm for the two stream coupled HMM used in our audio-visual system. An extension to a multi-stream coupled HMM is straightforward.

- Initialization

$$\delta_0(i,j) = \pi_0^a(i)\pi_0^v(j)b_t^a(i)b_t^v(j) \quad (6)$$
$$\psi_0(i,j) = 0 \quad (7)$$

- Recursion

$$\delta_t(i,j) = \max_{k,l}\{\delta_{t-1}(k,l)a_{i|k,l}a_{j|k,l}\}b_t^a(k)b_t^v(l) \quad (8)$$
$$\psi_t(i,j) = \arg\max_{k,l}\{\delta_{t-1}(k,l)a_{i|k,l}a_{j|k,l}\} \quad (9)$$

- Termination

$$P = \max_{i,j}\{\delta_T(i,j)\} \quad (10)$$
$$\{q_T^a, q_T^v\} = \arg\max_{i,j}\{\delta_T(i,j)\} \quad (11)$$

- Backtracking

$$\{q_t^a, q_t^v\} = \psi_{t+1}(q_{t+1}^a, q_{t+1}^v) \quad (12)$$

The segmental K means algorithm for the coupled HMMs is described by the following steps:

Step 1 For each training observation sequence r, the data in each stream is uniformly segmented according to the number of states of the coupled nodes and an initial state sequence for the coupled nodes $\mathbf{Q} = q_{r,0}^{a,v}, \ldots, q_{r,t}^{a,v}, \ldots q_{r,T-1}^{a,v}$ is obtained. For each state i of the coupled nodes in stream c the mixture segmentation of the data assigned to it is obtained using the K-means algorithm [6] with M_i^c clusters. Consequently the sequence of mixture components $\mathbf{S} = s_{0,r}^{a,v}, \ldots, s_{r,t}^{a,v}, \ldots s_{r,T-1}^{a,v}$ for the mixtures nodes is obtained.

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

$$\mu_{i,m}^{a,v} = \frac{\sum_{r,t}\gamma_{r,t}^{a,v}(i,m)\mathbf{O}_t^{a,v}}{\sum_{r,t}\gamma_{r,t}^{a,v}(i,m)} \quad (13)$$

$$\sigma_{i,m}^{2\ a,v} = \frac{\sum_{r,t}\gamma_{r,t}^{a,v}(i,m)(\mathbf{O}_t^{a,v} - \mu_{i,m}^{a,v})(\mathbf{O}_t^{a,v} - \mu_{i,m}^{a,v})^T}{\sum_{r,t}\gamma_{r,t}^{a,v}(i,m)} \quad (14)$$

$$w_{i,m}^{a,v} = \frac{\sum_{r,t}\gamma_{r,t}^{a,v}(i,m)}{\sum_{r,t}\sum_m \gamma_{r,t}^{a,v}(i,m)} \quad (15)$$

$$a_{i|k,l}^{a,v} = \frac{\sum_{r,t}\epsilon_{r,t}^{a,v}(i,k,l)}{\sum_{r,t}\sum_k\sum_l \epsilon_{r,t}^{a,v}(i,k,l)} \quad (16)$$

where

$$\gamma_{r,t}^{a,v}(i,m) = \begin{cases} 1, & if q_{r,t}^{a,v} = i, s_{r,t}^{a,v} = m, \\ 0, & otherwise \end{cases} \quad (17)$$

$$\epsilon_{r,t}^{a,v}(i,k,l) = \begin{cases} 1, & if\ q_{r,t}^{a,v} = i, \\ & q_{r,t-1}^{a} = k, q_{r,t-1}^{v} = l \\ 0, & otherwise \end{cases} \quad (18)$$

Step 3 At consecutive iteration the optimal state sequence $\mathbf{Q}$ of the coupled nodes is obtained using the Viterbi algorithm (Equations 7- 12). The sequence of mixture component $\mathbf{S}$ is obtained by selecting at each moment t the mixture $s_{r,t}^{a,v}$ such that:

$$s_{r,t}^{a,v} = \max_{m=1,\ldots,M_i^{a,v}} P(\mathbf{O}_t^{a,v}|q_{r,t}^{a,v} = i, m) \quad (19)$$

Step 4 The iterations in steps 2-4 are repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below the convergence threshold.

6. RECOGNITION

The word recognition is carried out via the computation of the Viterbi algorithm (Equations 7- 12) for the parameters of all the word models in the database. The parameters of the CHMM corresponding to each word in the database are obtained in the training stage using clean audio signals (SNR = 30db). In the recognition stage the influence of the audio and visual streams is weighted based on the relative reliability of the audio and visual features for different levels of the acoustic noise. Formally the observation probability at time t for the observation vector $\mathbf{O}_t^{a,v}$ becomes

$$\tilde{b}_t^{a,v}(i) = b_t(\mathbf{O}_t^{a,v}|q_t^{a,v} = i)^{\alpha_{a,v}} \quad (20)$$

where $\alpha_a + \alpha_v = 1$ and $\alpha_a, \alpha_v \geq 0$ are the exponents of the audio and video streams. The values of α_a, α_v corresponding to a specific acoustic SNR level are obtained experimentally to maximize the average recognition rate. Table 2 describes the audio exponents α_a used in our system.

SNR(db)	30	26	20	16
α_a	0.9	0.8	0.5	0.4

Table 2: The optimal set of exponents for the audio stream α_a at different SNR values of the acoustic speech

7. EXPERIMENTAL RESULTS

We tested the speaker dependent audio-visual word recognition system on the 36 words in the CMU database [5]. 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. The average audio-only, video-only and audio-visual recognition rates are presented in Figure 4 and Table 3. For audio-only

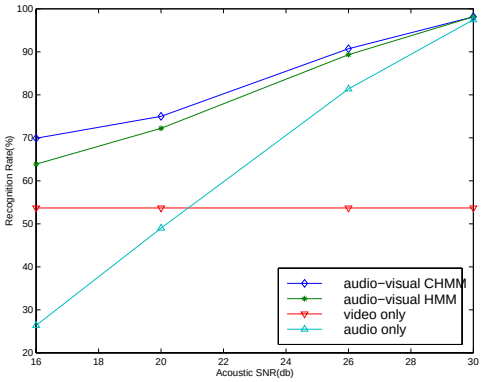

Figure 4: The recognition rate of the audio-visual speech recognition system.

speech recognition, the acoustic observation vectors (13 MFCC coefficients extracted from a window of 20ms) are modeled using a HMM with the same characteristics as the one described in Section 3 for video-only recognition. For the audio-video recognition, we used a coupled HMM with five states for the coupled nodes in both audio and video streams, no back transitions, and three mixture per state. Our experimental results indicate that the CHMM-based audio-visual speech recognition rate increases by 45% the audio-only speech recognition at SNR of 16db. Compared to the multistream HMM [11], the coupled HMM-based audio-visual recognition systems shows consistently better results with the decrease of the SNR reaching a nearly 7% reduction in word error rate at 16db. The audio-visual multistream HMM has the same characteristics as the HMMs used for video-only and audio-only recognition.

8. CONCLUSIONS

This papers presents a speaker dependent audio-visual word recognition system that uses a two-stream coupled HMM to model the audio and video observation sequences. Unlike the HMM, the CHMM allows for asynchrony in the audio and visual states, while preserving the natural dependency of the audio and video signals. In addition, with the coupled HMM, the audio and video sequences are treated separately and is no need for the concatenation of the audio and video observation that is often a challenging problem.

SNR(db)	30	26	20	16
V HMM	53.70%	53.70%	53.70%	53.70%
A HMM	97.46%	80.58%	50.19%	28.26%
AV HMM	98.14%	89.34%	72.21%	63.88%
AV CHMM	98.14%	90.72%	75.00%	69.90%

Table 3: A comparison of the speech recognition rate at different levels of acoustic SNR using a HMM for video only features (V HMM), a HMM for audio only features (A HMM), a HMM for audio-visual features (AV HMM), and the coupled HMM for audio visual features (AV CHMM)

The above advantages of the CHMM are reflected by our experimental results. The coupled HMM-based system outperforms the multistream HMM-based recognition system, and shows increasingly better recognition rate than the multistream HMM with the degradation of the acoustic SNR. Furthermore, with a CHMM the observation probabilities for the audio and video streams, can be obtained independently, making this model very attractive for machines that allow parallel computation. We believe that the CHMM, and the efficient training algorithm for CHMM described in this paper can be applied to a variety of multimedia system involving two or more data streams. Future work will include the investigation of CHMMs for continuous audio-visual speech recognition.

9. REFERENCES

[1] 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.

[2] C. Bregler and S. Omohundro. Nonlinear manifold learning for visual speech recognition. In *IEEE International Conference on Computer Vision*, pages 494–499, 1995.

[3] K. R. Castleman. *Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1996.

[4] T. Chen. Audiovisual speech processing. *Signal Processing Magazine*, 18:9–21, January 2001.

[5] Advanced Multimedia Processing Lab, CMU, `http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/`.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley Sons Inc., New York, NY, 2000.

[7] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2:141–151, September 2000.

[8] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press Limited, London, UK, 1998.

[9] J. Luettin, N.A. Thacker, and S.W. Beet. Speechreading using shape and intensity information. In *IEEE International Conference on Spoken language*, volume 1, pages 58–61, 1996.

[10] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, pages 746–748, September 1976.

[11] C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio visual speech recognition. In *Final Workshop 2000 Report*, 2000.

[12] G. Potamianos, J. Luettin, 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.

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