AUDIO-VISUAL CONTINUOUS SPEECH RECOGNITION USING A COUPLED HIDDEN MARKOV MODEL

Xiaoxing Liu, Yibao Zhao, Xiaobo Pi, Luhong Liang and Ara V. Nefian

Microcomputer Research Labs, Intel Corporation
{xioaxing.liu, yibao.zhao, xiaobo.pi, luhong.liang, ara.nefian}@intel.com

ABSTRACT

With the increase in the computational complexity of recent computers, audio-visual speech recognition (AVSR) became an attractive research topic that can lead to a robust solution for speech recognition in noisy environments. In the audio visual continuous speech recognition system presented in this paper, the audio and visual observation sequences are integrated using a coupled hidden Markov model (CHMM). The statistical properties of the CHMM can describe the asyncrony of the audio and visual features while preserving their natural correlation over time. The experimental results show that the current system tested on the XM2VTS database reduces the error rate of the audio only speech recognition system at SNR of 0db by over 55%.

1. INTRODUCTION

The success of currently available speech recognition systems is restricted to relatively controlled environments and well defined applications such as dictation or small to medium vocabulary voice-based control command (hands free dialing, etc). In recent years, together with the investigation of several acoustic noise reduction techniques, the study of systems that combine the audio and visual features emerged as an attractive solution to speech recognition under less constrained environments. A number of techniques have been presented to address the audio-visual integration problem, which can be broadly grouped into feature fusion and decision fusion methods. In an audio-visual feature fusion system, the observation vectors are obtained by the concatenation of the audio and visual observation vectors, followed by a dimensionality reduction transform [8]. The resulting observation sequences are then modeled using one hidden Markov model (HMM) [11]. However, this method cannot model the natural asyncrony between the audio and visual features. Decision fusion systems on the other side model independently the audio and video sequences and enforce the synchrony of the audio and visual features only at the model boundaries. These systems fail to capture entirely the dependencies between the audio and video features. The feature fusion system using a multi-stream HMM proposed in [10] assumes the audio and video sequences are state synchronous, but allows the audio and video components to have different contributions to the overall observation likelihood. The audio visual product HMM introduced in [2] can be seen as an extension of the multi-stream HMM that allows for audio-visual state asynchrony. The coupled hidden Markov model (CHMM) based audio-visual continuous speech recognition (AVCSR) system presented in this paper is an extension of the decision fusion system at phone level. The CHMM can model the audio-visual state asynchrony and preserve at the same time the natural audio visual dependencies over time. The formal definition of the audio-visual model is given in section 2. In sections 3 and 4, we describe the framework of CHMM training and decoding for AVCSR respectively. The experimental results are presented in section 5.

2. THE AUDIO-VISUAL CHMM

A CHMM can be seen as a collection of hidden Markov models (HMM), 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. Figure 1 illustrates a continuous mixture two-stream coupled HMM used in our audio-visual speech recognition system. The squares represent the hidden discrete nodes (backbone and mixture nodes) while the circles describe the continuous observable nodes. Unlike the independent HMM used for audio-visual data, the CHMM can capture the interactions between audio and video streams through the transition probabilities between the backbone nodes. In the AVCSR system presented in this paper, the audio visual CHMM allows for asynchrony in the audio and visual states but forces them to be synchronized at the model boundaries. In addition, with the coupled HMM, the audio and video observation likelihoods are computed independently, significantly reducing the parameter space and complexity of the model compared to the models that require the concatenation of the audio and visual observations [8].

The parameters of a CHMM are defined below:

$$\pi_0^c(i) = P(q_t^c = i) \quad (1)$$
$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i) \quad (2)$$
$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^0 = j, q_{t-1}^1 = k) \quad (3)$$

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,

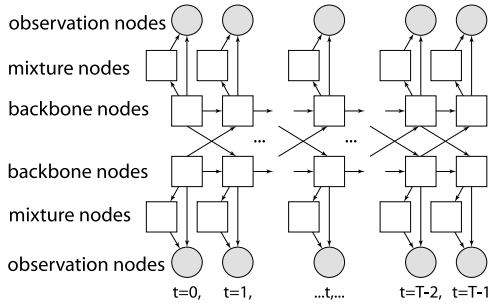

Figure 1: The audio-visual coupled HMM.

the probabilities of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N(\mathbf{O}_t^c, \mu_{i,m}^c, \mathbf{U}_{i,m}^c) \quad (4)$$

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. Unlike isolated word audio-visual speech recognition where one CHMM is used to model each audio-visual word, in audio-visual continuous speech recognition, each CHMM models one of the possible phoeneme-viseme pairs as defined in [8].

3. TRAINING

The training of the CHMM parameters for AVCSR is performed in two stages and is an extension of the training used in audio-only continuous speech recognition [3]. In the first stage, the CHMM parameters are estimated for isolated phoneme-viseme pairs. In this stage, the training sequences are labeled using an audio-only speech recognition system, and the phoeneme-viseme correspondence tables [8]. The parameters of the isolated phoeneme-viseme CHMMs are estimated first using the Viterbi-based initialization described in [7] followed by the estimation-maximization (EM) algorithm [4]. 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. In the second stage, the parameters of the CHMMs, estimated individually in the first stage, are refined through the embedded training of all CHMM from continuous audio-visual speech. In this stage, the labels of the training sequences consist only on the sequence of phoeneme-viseme with all boundary information being ignored. In a way similar to the embedded training for HMMs [3], 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 phoeneme-viseme synchrony at the model boundaries.

The embedded training follows the steps of the EM algorithm for continuous audio-visual speech, and is described by the following:

E step. The forward probability $\alpha_t(i,j) = P(\mathbf{O}_1, \ldots, \mathbf{O}_t, q_t^0 = i, q_t^1 = j)$ and the backward probability $\beta_t(i,j) = P(\mathbf{O}_t + 1, \ldots, \mathbf{O}_T | q_t^0 = i, q_t^1 = j)$ are computed. Starting with the initial conditions

$$\alpha_1(i,j) = \pi_1^0(i)\pi_1^1(j)b_1^0(i)b_1^1(j),$$

the forward probabilities are computed recursively from

$$\alpha_t(i,j) = b_{t-1}^0(i)b_{t-1}^1(j)\sum_{l,k} a_{i,j|l,k}\alpha_{t-1}(l,k)$$

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

$$\beta_T(i,j) = 1$$

the backward probabilities are computed recursively from

$$\beta_t(i,j) = \sum_{l,k} b_{t+1}^0(l)b_{t+1}^1(k)a_{l,k|i,j}\beta_{t-1}(l,k)$$

for $t = T-1, T-2, \ldots, 1$ where i, j are the states of the audio and video chain respectively and $a_{i,j|k,l} = a_{i|k,l}a_{j|k,l}$ is the transition probabilities between the set of audio visual states i, j and k, l. The probability of the rth observation sequence $\mathbf{O}^r = [\mathbf{O}_1^r, \ldots \mathbf{O}_{T_r}^r]$ is computed as $P_r = \alpha_{T_r}(N, M) = \beta_1(1,1)$, where N, M are the number of states in the audio and video chain respectively and T_r is the length of the observation sequence $\mathbf{O}_r$.

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

$$\tilde{\mu}_{i,m}^c = \frac{\sum_r \sum_t \gamma_t^{r,c}(i,m)\mathbf{O}_t^r}{\sum_r \sum_t \gamma_t^{r,c}(i,m)}$$

$$\tilde{\mathbf{U}}_{i,m}^c = \frac{\sum_r \sum_t \gamma_t^{r,c}(i,m)(\mathbf{O}_t^r - \mu_{i,m}^c)(\mathbf{O}_t^r - \mu_{i,m}^c)'}{\sum_r \sum_t \gamma_t^{r,c}(i,m)}$$

$$\tilde{w}^c_{i,m} = \frac{\sum_r \sum_t \gamma^{r,c}_t(i,m)}{\sum_r \sum_t \sum_m \gamma^{r,c}_t(i,m)}$$

where

$$\gamma^{r,c}_t(i,m) = \frac{\sum_j \frac{1}{P_r}\alpha^r_t(i,j)\beta^r_t(i,j)}{\sum_{i,j}\frac{1}{P_r}\alpha_t(i,j)\beta_t(i,j)} \frac{w^c_{i,m}N(\mathbf{O}^r_t, \mu^c_{i,m}, \mathbf{U}^c_{i,m})}{\sum_m w^c_{i,m}N(\mathbf{O}^r_t, \mu^c_{i,m}, \mathbf{U}^c_{i,m})}$$

The state transition probabilities can be estimated using:

$$\tilde{a}^{0,1}_{i|k,l} =$$

$$\frac{\sum_r \frac{1}{P_r}\sum_t \alpha^r_t(k,l)a_{i|k,l}b^{0,1}_t(i)\sum_j \beta^r_{t+1}(i,j)b^{1,0}_t(j)}{\sum_r \frac{1}{P_r}\sum_t \alpha^r_t(k,l)\beta^r_t(k,l)}$$

Assuming that $a^{0,1}_{i|k,l} = P(q^{0,1}_t = i|q^{0,1}_t = k)P(q^{0,1}_t = i|q^{1,0}_t = l)$, the re-estimation of the transition probabilities can be simplified. For example, $P(q^0_t = i|q^1_t = k)$ can be estimated as:

$$P(q^0_t = i|q^1_t = k) =$$

$$\frac{\sum_r \frac{1}{P_r}\sum_t \sum_j \sum_l \alpha^r_t(k,l)a_{i,j|k,l}b^0_t(i)b^1_t(k)\beta^r_{t+1}(i,j)}{\sum_r \frac{1}{P_r}\sum_t \sum_j \sum_l \alpha^r_t(k,l)\beta^r_t(k,l)}$$

The transitions from a non-emitting entry state i to any pair of audio-visual states (k,l) is given by

$$a_{i|k,l} = \frac{1}{R}\sum_r \frac{1}{P_r}\alpha^r_1(k,l)\beta^r_1(k,l)$$

and the transitions from a state pair (k,l) to the exit non-emitting exit state o are given by

$$a_{k,l|o} = \frac{\sum_r \frac{1}{P_r}\alpha^r_T(k,l)\beta^r_T(k,l)}{\sum_r \frac{1}{P_r}\sum_t \alpha^r_t(k,l)\beta^r_t(k,l)}$$

4. RECOGNITION

The audio-visual continuous speech recognition is carried out via a graph decoder applied to the word network consisting of all the words in the test dictionary. Each word in the network is stored as a sequence of phoneme-viseme CHMMs, and the best sequence of words is obtained through an extension of the token passing algorithm [3], [9] to audio-visual data. To handle different levels of noise in the audio channel, the audio and video observation probabilities are modified such that $\tilde{b}^{0,1}_t(i) = b^{0,1^{\alpha_{0,1}}}_t$ where $\alpha_0 + \alpha_0 = 1$ and $\alpha_0, \alpha_0 \geq 0$ are the exponents of the audio and video streams respectively. The values α_0, α_1 corresponding to a specific acoustic SNR level are obtained experimentally to minimize the average word error rate.

5. EXPERIMENTAL RESULTS

We tested the audio-visual continuous speech recognition system described here on the XM2VTS database [6]. We used a set of 1450 digit enumeration sequences captured from 200 speakers for training and a set of 700 sequences from other 95 speakers for decoding. The training sequences are recorded with "clean" audio. The audio data of the testing sequences is corrupted with several levels of white noise to allow the study of AVSR under less constrained acoustic conditions. In our experiments the acoustic observation

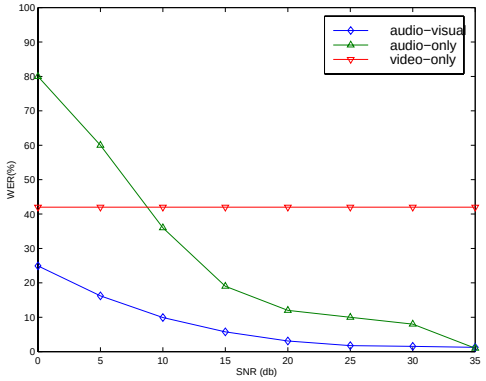

Figure 2: The word error rate of the audio-only, visual-only and audio-visual speech recognition system at different levels of SNR.

vectors consist of 13 MFCC coefficients, extracted from a window of 25.6 ms, with an overlap of 15.6 ms, with their first and second order time derivatives. The visual features are obtained from the mouth region through a cascade algorithm described in more detail in [5]. 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. The pixels in the mouth region are mapped to a 32-dimensional feature space using the principal component analysis. Then, blocks of 15 visual observation vectors are concatenated and projected on a 13 class, linear discriminant space [1]. 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, with 32 mixture per state, and diagonal covariance matrix.

Table 1 and Figure 2 compare the WER of our current AVSR system with an audio only speech recognition system. For fair comparison, in the audio-only speech recognition system, all phoenemes were modeled using a HMM with the same characteristics as the audio HMM in the audio-visual CHMM.

SNR(db)	0	5	10	15
WER(%)	24.62	15.71	9.47	5.13
SNR(db)	20	25	30	clean
WER (%)	2.95	1.86	1.59	1.14

Table 1: The word error rate of the audio-visual speech recognition system for several SNR levels.

6. CONCLUSIONS

This paper presents an audio-visual continuous speech recognition system that uses a CHMM for audio and visual feature integration. This audio visual model used in our system allows for the natural asynchrony between the audio and visual states while imposing state synchrony at the model boundaries. Furthermore, the CHMM models the audio and video state dependency,and therefore preserves the natural properties of audio-visual speech. The experimental results, tested on the XM2VTS database, show that our system improves the recognition rate of the audio only speech recognition system consistently at all SNR levels, achieving a WER reduction of over 55% at SNR of 0db.

7. REFERENCES

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

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

[3] S. Young et. al. *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, UK, 1995.

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

[5] 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*, 2002.

[6] J. Luettin and G. Maitre. Evaluation protocol for the XM2FDB database. In *IDIAP-COM 98-05*, 1998.

[7] A. V. Nefian, L. Liang, X. Pi, X. Liu, and C. Mao. An coupled hidden Markov model for audio-visual speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, 2002.

[8] 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.

[9] M. Oerder and H. Ney. Word graphs: an efficient interface between continuous-speech recognition and language understanding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 1993.

[10] 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.

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