SPEAKER INDEPENDENT AUDIO-VISUAL CONTINUOUS SPEECH RECOGNITION

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

Microcomputer Research Labs, Intel Corporation
Santa Clara, CA, 95052
{luhong.liang, xioaxing.liu, yibao.zhao, xiaobo.pi, ara.nefian}@intel.com

ABSTRACT

The increase in the number of multimedia applications that require robust speech recognition systems determined a large interest in the study of audio-visual speech recognition (AVSR) systems. The use of visual features in AVSR is justified by both the audio and visual modality of the speech generation and the need for features that are invariant to acoustic noise perturbation. The speaker independent audio-visual continuous speech recognition system presented in this paper relies on a robust set of visual features obtained from the accurate detection and tracking of the mouth region. Further, the visual and acoustic observation sequences are integrated using a coupled hidden Markov (CHMM) model. The statistical properties of the CHMM can model the audio and visual state asynchrony while preserving their natural correlation over time. The experimental results show that the current system tested on the XM2VTS database reduces by over 55% the error rate of the audio only speech recognition system at SNR of 0db.

1. INTRODUCTION

The success of 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 (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 [11] emerged as an attractive solution to speech recognition under less constrained environments. The visual features are often derived from the shape of the mouth [9] [3], [2]. Although very popular, these methods rely exclusively on the accurate detection of the lip contours, 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 [5] and [11].

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 [11]. The resulting observation sequences are then modeled using one HMM [15]. The multi-stream HMM proposed in [14]), 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. The audio visual product HMM introduced in [5] can be seen as an extension of the multi-stream HMM that allows for audio-video state asynchrony. Decision fusion systems independently model the audio and video sequences using two HMMs, and combine the likelihood of each observation sequence based on the reliability of each modality [11].

The speaker independent audio-visual continuous speech recognition system presented in this paper starts with the detection and tracking of the mouth region (Section 2) followed by the extraction of a robust set of visual observations from the mouth region (Section 3). The audio observations ([15]) are then combined with the visual observations using a coupled hidden Markov model ([10]) as described in Section 4.

2. MOUTH DETECTION AND TRACKING

The mouth detection approach presented in this paper (Figure 1) begins with the detection of the user's face (Figure 2a) using the neural network-based approach described in [1].

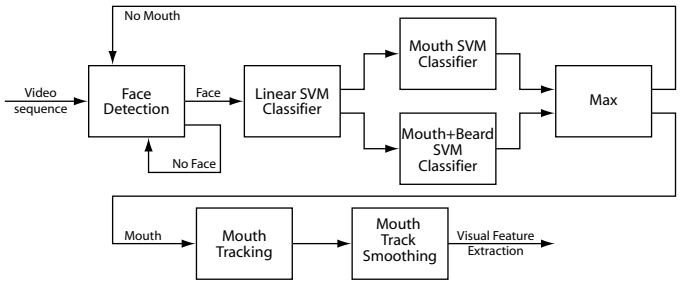


Figure 1: The mouth detection and tracking system.

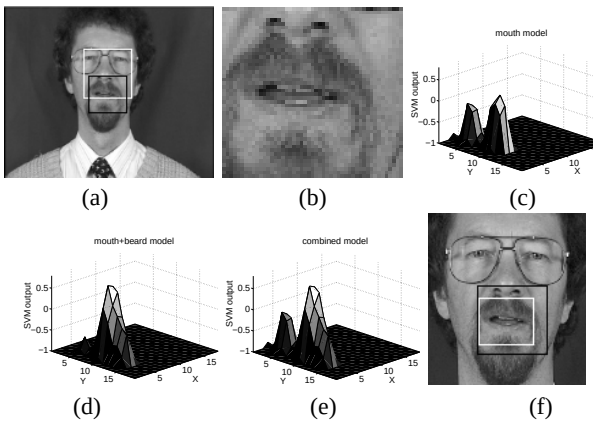
As the search area for the mouth is significantly reduced by the results of the face detector, we use a cascade of support vector machine (SVM) classifiers [17] to locate the mouth within the lower region of the face (Figure 2a). The SVM cascade is designed to insure both the real-time and accuracy requirements of the overall AVSR system. To adapt to scale variations, a multi-scale search in an estimated range is employed by repeatedly resampling the source region image by a constant factor. A pre-processing step normalizes each test pattern with respect to variations in illumination via histogram equalization and gradient illumination correction [16]. Next, a SVM filter of size 16×14 with linear kernel was applied to the lower region of the face to determine the coarse location of the mouth. Finally, two SVM classifiers with Gaussian kernel of size 32×28, trained on examples of mouth regions

with and without facial hair, are applied to each test pattern and its rotated versions in the image plane. The highest mouth classification score among all rotated patterns and SVM classifiers is used to determine the refined location of the mouth.

The positive examples used for the training of the SVM filters consist of a set of manually labeled mouth images and a set of negative examples (facial regions other than the mouth). Using the labeled points in the positive set, including the corners, the upper and lower points, and the center of the mouth, the size of the training set is enhanced with several geometric variations such as scaling, horizontal mirroring and rotations in the image plane. Then, a pre-processing step consisting of scale normalization, histogram equalization and illumination correction [16] is applied to both the positive and the negative examples in the training set. The Gaussian kernel SVM filters are trained via bootstrapping as follows:

1. train the SVM filters using the positive and negative training sets, [13].
2. run the SVM filters on a validation set and enhance both the positive set with undetected mouth regions and the negative set with false alarms,
3. repeat step 1-2 until the mouth detector reaches the desired performance.

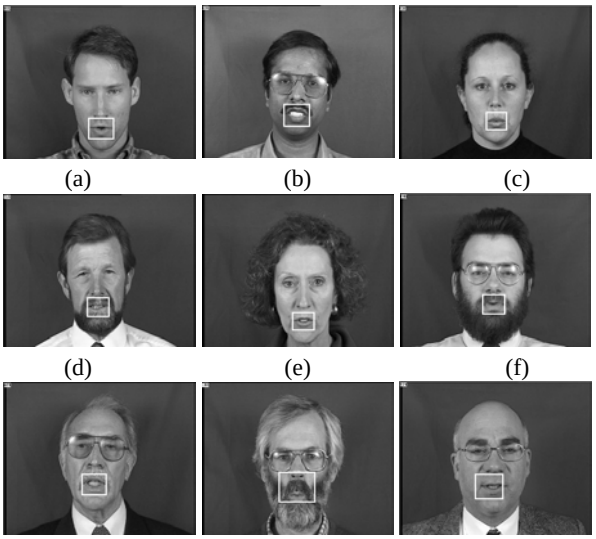
In our experiments, the training sets, obtained after the bootstrapping procedure, consist on approximately 8000 non-mouth, 9000 mouth and 6000 mouth-and-beard samples respectively. The mouth samples were obtained by mirroring, rotating, and re-scaling of 250 and 800 images of users with and without beards respectively.



(a) (b) (c) (d) (e) (f)

Figure 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 convolution result for the mouth SVM. (d) The convolution result for the mouth and beard SVM. (e) The combined convolution result. (f) The mouth detection result (white rectangle) from the initial region of search for the mouth (black rectangle).

Following the detection of the face and mouth region, the mouth position is tracked over consecutive frames. The center of the mouth is estimated from the previous frame, and the mouth detection algorithm is applied to a reduced area around the estimated center of the mouth. If all the test patterns in the search area fail to be assigned to the mouth region, the system re-initializes with the face and mouth detection algorithm, or the new mouth center is estimated and the mouth tracking continues. The mouth track is further smoothed using a median filter followed by a Gaussian filter. Figure 3 shows several results of the mouth detection and tracking system. The approach was tested on the "Clients" subset of the XM2VTS database [8] representing 190 sequences recorded from 95 speakers (Figure 3). The overall accuracy of the mouth detection and tracking system is 95.26%, with 86.67% for the 30 sequences of people wearing beards and 96.88% for the remaining sequences.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 3: Examples of accurate (a-g) and inaccurate (h,i) results of the mouth detection and tracking system

.

3. THE VISUAL FEATURE EXTRACTION

The set of visual observation vectors used in our AVSR system is extracted from a region of size 64×64 around the center of the mouth using a cascade algorithm. First, the gray level pixels in the mouth region are mapped to a 32-dimensional feature space using the principal component analysis (PCA). The PCA decomposition is computed from a set of approximately 200,000 mouth region images obtained from the tracking system described above. Figure 4 shows the 32 eigenvectors used in our PCA decomposition. The

Figure 4: The first 32 eigenvectors corresponding the mouth region

resulting vector of size 32 is upsampled to match the frequency of the audio features (100Hz) and standardized using the feature mean normalization (FMN) described in [11]. Next, blocks of

N visual observation vectors are concatenated and projected on a 13 class linear discriminant space ([4]), obtaining a new set of visual observation vectors of size 13. The class information used in the linear discriminant analysis (LDA) corresponds to the 13 English visemes ([11]). The visual feature extraction technique described in this paper was tested in a speaker independent visual-only continuous speech recognition system tested on the XM2VTS database. In our experiments, the English visemes are modeled by hidden Markov models with 3 states, 12 mixture of Gaussian components per state and diagonal covariance matrix. The training and testing sets consists of over 700 sequences from the "Clients" set and 139 digit enumeration sequences spoken by 70 speakers from the "Impostors" set respectively. Figure 5 and Table 1 show the decrease in visual-only word error rate (WER), obtained using PCA instead of DCT coefficients [11], followed by LDA with N concatenated frames.

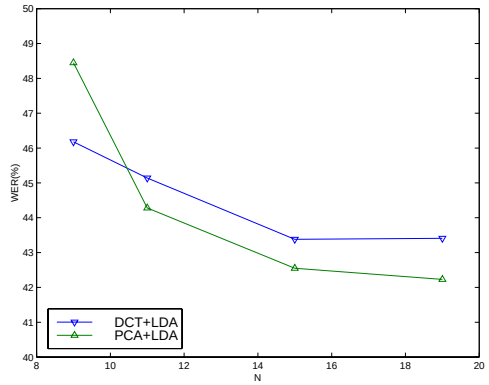


Figure 5: The visual-only word error rate.

N	9	11	15	19
DCT+LDA	46.18%	45.14%	43.38%	43.41%
PCA+LDA	48.45%	44.28%	42.55%	42.23%

Table 1: The word error rate of the visual-only speech recognition system using PCA and DCT coefficients followed by LDA using N concatenated frames.

4. THE AUDIO-VISUAL MODEL

The audio and visual observation vectors described above are further integrated using a coupled hidden Markov model (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 $t-1$ of all the related HMMs. Figure 6 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. The CHMM allows for asynchrony in the audio and visual states, while preserving the natural dependency of the audio and video sequences. In addition, with the coupled HMM, the audio and video observation likelihoods are computed independently significantly reducing the parameter space and complexity

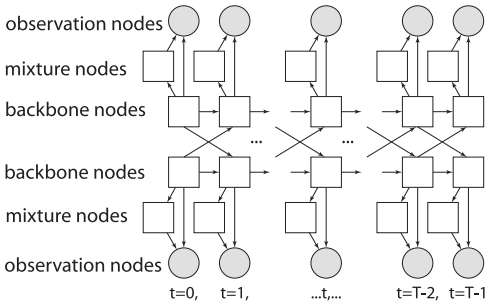


Figure 6: The audio-visual coupled HMM.

of this model compared to the models that require the concatenation of the audio and visual observations [11]. In our continuous audio-visual speech recognition system, a CHMM models each viseme-phoneme pair. The training of the parameters starts with Viterbi-based initialization [10] followed by the estimation-maximization (EM) algorithm [7] for isolated words. Next, the parameters of all CHMMs are estimated together from continuous audio-visual speech using the EM algorithm. The *embedded training*for HMM, similar to the training method described for in this paper CHMM, is extensively explained in [6]. The continuous speech recognition is carried out via a graph decoder using a single-pass Viterbi beam search [6], [12]. At the recognition stage, the audio and video observation probabilities are modified to handle different levels of noise.

$$\begin{aligned} \tilde{b}_t^a(i) &= b_t(\mathbf{O}_t^a|q_t^a=i)^{\alpha_a} \\ \tilde{b}_t^v(j) &= b_t(\mathbf{O}_t^v|q_t^v=j)^{\alpha_v} \end{aligned}$$

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.

5. EXPERIMENTAL RESULTS

We tested the speaker dependent audio-visual word recognition system on the XM2VTS database using the same testing and training sets used for visual-only speech recognition. In our experi-

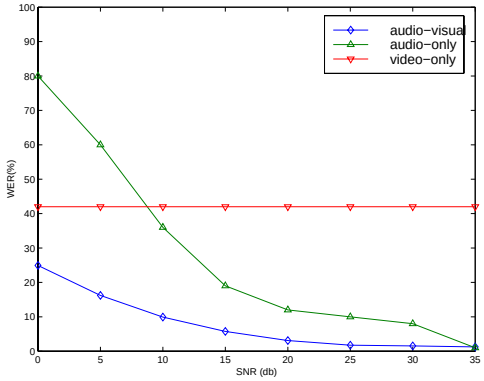


Figure 7: The word error rate of the audio-only, visual-only and audio-visual speech recognition system at different levels of SNR.

ments the visual features are obtained as explained in Sections 2 and 3, with $N = 15$. The acoustic observation vectors consist of 13 MFCC coefficients extracted from a window of 20ms, with an overlap of 15ms. For the audio-video recognition, we used a coupled HMM with three states per node in both the audio and the video streams, with no back transitions, with 32 mixture per state, and diagonal covariance matrix . The average word error rate (WER) of our audio-visual recognition system is presented in Table 2. Our experimental results (Figure 7) indicate that the audio-visual speech recognition rate increases by 55% the audio-only speech recognition at SNR of 0db.

SNR(db)	0	5	10	15
WER(%)	24.93	16.22	9.93	5.76
SNR(db)	20	25	30	clean
WER (%)	3.09	1.76	1.55	1.26

Table 2: The word error rate of the audio-visual speech recognition system for several SNR levels

6. CONCLUSIONS

This papers presents a speaker independent audio-visual continuous speech recognition system that significantly reduces the error rate of the audio-only system in noisy environments. The improved accuracy of the audio-visual system is achieved using a robust set of visual observation vectors obtained from the mouth region. The accurate detection of the mouth is obtained from the detected face region using a set of SVM classifiers trained for different mouth appearances. The visual observation vectors are obtained from a cascade algorithm that applies PCA and LDA to the mouth region. Further the audio and visual observation vectors are integrated using a coupled HMM. 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. The experimental results, tested on a subset of the XM2VTS database of numeric sentences, 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 0 db.

7. REFERENCES

[1] H.Z Ai, L. Liang, and G.Y.Xu. Face detection in template matching constrained subspace. In *IEEE International Conference on Artificial Intelligence*, volume 2, pages 603–608, 2001.

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

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

[5] S. Dupont and J. Luettin. 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] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press Limited, London, UK, 1998.

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

[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] 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.

[13] J.C. Platt. Sequential minimal optimization: a fast algorithm for training support vector machines. In *Technical Report: MSR-TR-98-14*, 1998.

[14] 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.

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

[16] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[17] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Berlin:Springer-Verlag, 1999.