Automatic Speech Recognition using Audio Visual Cues

Yashwanth H, Harish Mahendrakar and Sumam David

ABSTRACT: **Automatic speech recognition (ASR) systems have been able to gain much popularity since many multimedia applications require robust speech recognition algorithms. The use of audio and visual information in the speaker-independent continuous speech recognition process makes the performance of the system better compared to the ones with only the audio information. There has been a marked increase in the recognition rates by the use of visual data to aid the audio data available. This is due to the fact that video information is less susceptible to ambient noise than audio information. In this paper a robust Audio-Video Speech Recognition (AVSR) system that allows us to incorporate the Coupled Hidden Markov Model (CHMM) model for fusion of audio and video modalities is presented. The application records the input data and recognizes the isolated words in the input file over a wide range of Signal to Noise Ratio (SNR.) The experimental results show a remarkable increase of about 10% in the recognition rate in the AVSR compared to the audio only ASR and 20% compared to the video only ASR for an SNR of 5dB.**

1. INTRODUCTION

The variety of applications of Automatic speech recognition (ASR) systems, for human computer interfaces, for telephony, for robotics or for voice based commands has been the major driving force for the research in this field over the last decade. As a result ASR systems have undergone much advancement over the last few decades. During recent years the field of audio - visual speech recognition has received much attention. An audio - visual recognition system uses the video modality to aid the conventional audio modality to improve the performance of the ASR systems. Visual information is an essential constituent of speech recognition in humans. For example a person in a noisy environment like a subway station is able to perceive correctly what information other person is trying to convey based on the lip movements, which provides additional information to the impaired audio information (due to the background noise). The visual features provide robust information that is not corrupted by the ambient acoustic noise.

In the recent past, audio-visual speech recognition systems (AVSR) have outperformed the ASR systems. However there are essentially two problems that have to be considered. First there is a need to develop algorithms to extract lip information in real-time and the other is the integration of the audio and the visual features in order to develop a working AVSR model. Though several algorithms are available for the former, the later task is challenging and is open for research.

Yashwanth H, Harish Mahendrakar and Sumam David
Dept. of Electronics and Communication,
National Institute of Technology Karnataka Surathkal,
INDIA 575 025.{ece01450, ece0190, sumam} @ nitk.ac.in

0-7803-8909-3/04/$20.00 ©2004 IEEE

The AVSR system presented in this paper, starts with the speech integration using CHMM model followed by a discussion on audio and visual feature extraction. Finally the experimental goals and results are presented.

2. RELATED WORK

In AVSR systems the audio and visual modalities carry both complementary and supplementary information, i.e. sometimes the problem gets over-defined by relying on both audio and video (AV) modalities and sometimes it gets under defined by relying on only one modality. The AV integration is both inherently synchronous and asynchronous. During the onset of speech, the AV modalities tend to be asynchronous which make their integration a challenging one. In general, there have been two approaches to AVSR, early integration which assumes synchronous nature and the other late integration which assumes the two modalities to be asynchronous (which is not always the case)[1]. However it is well known that the acoustic features are delayed from the visual features of speech, and state synchronous models can be inaccurate. In most cases of early integration the visual features are interpolated to make them synchronous with audio features.

The audio and visual fusion techniques include feature fusion, model fusion, or decision fusion. In feature fusion, the combined audio and visual feature vectors are obtained by the concatenation of the audio and visual features followed by a dimensionality reduction transform. The resulting observation sequences have been modeled using one Hidden Markov Model (HMM) [2]. A model based fusion system based on multi stream HMM was proposed in [3]. Decision fusion systems model independently the AV sequences using two HMMs and combine the likelihood of each observation sequence based on the reliability of each modality.

3. AUDIO VISUAL SPEECH INTEGRATION USING CHMM

Coupled Hidden Markov Models (CHMM) have been used to implement the Audio Visual Speech Recognition System. The CHMMs are a subset of a larger class of networks known as Dynamic Bayesian Networks (DBN) [4]. A DBN is essentially a directed graph between a set of variables, with the edges in the graph defining the influence that each variable has on others. A CHMM can be imagined to be a combination of two HMMs whose state sequences are interdependent. The DBNs generalize HMM's where the hidden nodes are represented as state variables and allow these states to have complex interdependencies.

The discrete nodes at time't' for each HMM are conditioned by the discrete nodes at time't-1' of all the related HMMs. We refer to the hidden nodes conditioned temporally as coupled nodes and to the remaining hidden nodes as mixture nodes. Fig 1 shows a sample Audio Video Speech Integration (AVSI) setup using CHMM model, containing two streams, where circles represent the hidden discrete nodes and the squares represent the continuous observable nodes.

Taking into account that the AV modalities are both asynchronous and synchronous, modeling the system using CHMM proves to be an efficient technique. In our AVSR model the CHMM models each viseme-phoneme pair.

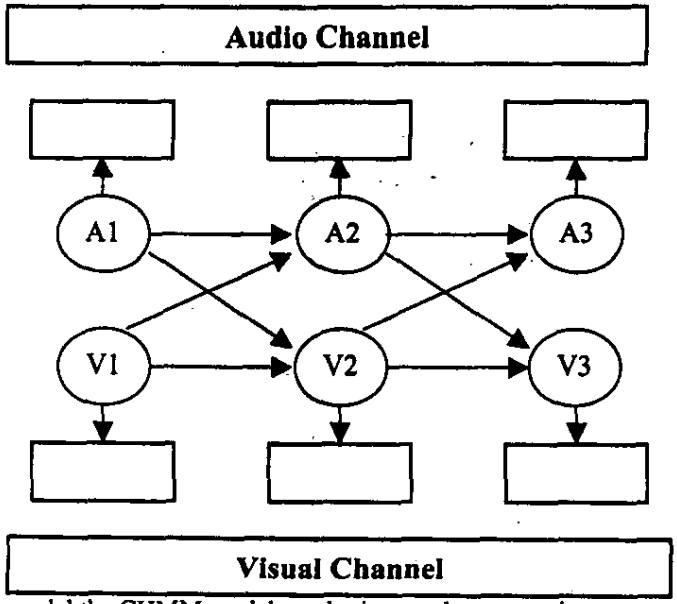


Figure1: Coupled Hidden Markov Model (CHMM) [5]

The parameters of a CHMM are defined below:

$$\pi_o^c(i) = P(q_t^c = i) \quad (1)$$

$$b_t^c(i) = P(O_t^c \mid q_t^c = i) \quad (2)$$

$$a_{i|j,k}^c = P(q_t^c = i \mid q_{t-1}^0 = j, q_{t-1}^1 = k) \quad (3)$$

where q is the state of the couple node in the qth stream at time t.

4. THE AUDIO AND VISUAL FEATURE EXTRACTION

The extraction of audio features of the input speech was done using Mel Frequency Cepstral Coefficients (MFCC). The extraction of the visual features starts with the detection of the speakers' face in the video sequence as in Fig 2. The face detector used here is the Haar face detector (provided by open CV) [6].The mouth region in the lower half of the detected face is extracted and the Region of Interest (ROI) is set accordingly.

Linear Discriminant Analysis (LDA) is used to assign the pixels in the mouth region to the lip and face classes which transforms the pixel values from the RGB space into a one dimensional vector that best discriminates between the classes [7]. Binary chain encoding method is used to extract the contour of the lips. The lip contour and the 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 x 64 pixels) is obtained from each frame of the video sequence.

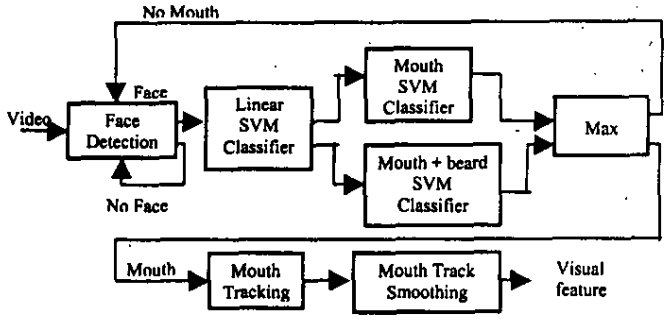


Figure 2: The Mouth Detection and Tracking System [7]

5. TRAINING AND RECOGNITION

Expectation Maximization (EM) algorithm is used to estimate the CHMM parameters. The process of training a DBN has been explained in [4]. The iterative maximum likelihood estimation of the parameters depends on the starting point and converges to local optima. The optimal weighting value and parameters P for speech at SNR of 30 dB are calculated as follows:

$$P = \arg\{ \max_{\lambda \in [0,1]} (WA \mid cleanspeech) \}$$

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 word recognition is carries out via the computation of the viterbi algorithm for the parameters of all the word models. In the recognition phase the influence of the audio and video streams is stream weighted, since the reliability on the audio and video streams is continuously changing which calls us to adapt our system to such changes. The stream weights are obtained so as to maximize the average recognition rate.

6. IMPLEMENTATION

An uncompressed, PCM recorded Audio data was processed using a Hamming Window, with a 10 ms sample size. For each frame 13 MFCC coefficients were calculated. The audio data was post processed while appending the delta and acceleration coefficients.

For visual features, a set of 10 coefficients, corresponding to the most significant generalized Eigen values of the LDA decomposition of mouth region are used as visual observation vectors. Principal Component analysis is used to map the gray level pixels into a 32 dimensional feature

space. The resulting vector of size 32 is up sampled to match the frequency of the audio features (100Hz) and standardized using the Feature Mean Normalization (FMN) described in [8].

6.1 Database Description

Since the system was an isolated word recognition system the database consisted of speakers uttering digits from zero to nine. It consisted of different speakers under different illuminations, varying background noise and under different resolutions. The Audio was PCM coded with 32 kHz sampling rate. The uncompressed video with resolutions 320 X 240, 640 X 480 and frame rates ranging from 10 to 25 was used.

6.2 Experimental Goals

We essentially had three experimental goals. The first was to check the effect of noise on the different recognition rates. Different samples were tested at different noise levels. Additive noise was added to the audio input data to vary the SNR, ranging from 0 (incomprehensible audio) to 30 (clean speech). The second was to compare the recognition rates obtained through audio-only ASR, video-only ASR and AVSR systems. The recognition rates of these three systems were evaluated for varying noise levels, illuminations, video resolution and frame rates. The effect of the variation in noise levels is presented in this paper.

An application that modeled the AVSR system was developed using the Intel OpenCV [9], an open source computer vision library that provides libraries for many audio and video processing functions. The audio-video input files supported by the current application are uncompressed AVI. The input video can be true color or grayscale. For increased accuracy, the face of the speaker in the video sequence must be frontal upright and of width larger than 128 pixels. The audio channel is16-bit, sampled at 32 kHz stereo.

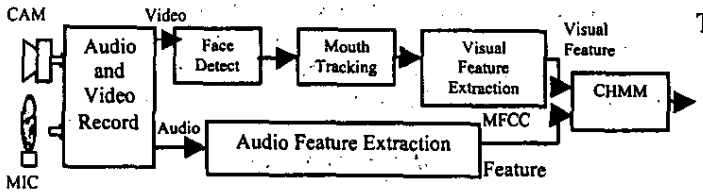


Figure 3: Schematic of the AVSR System

Our next goal was to integrate the recording process with the recognition process so that the overall system works real time. We were able to evaluate such a system with real-time recording, storage and recognition for various parameters listed above.

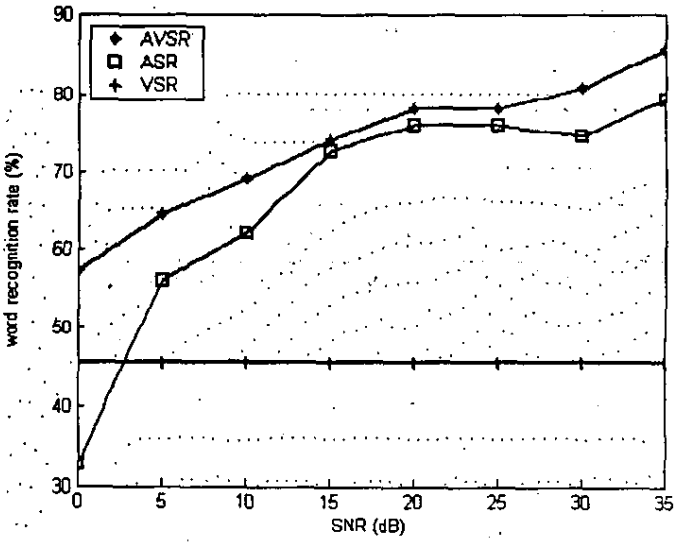
7. EXPERIMENTAL RESULTS

We tested the speaker independent AVSR system on the available database. The database consisted of speakers speaking in different ambient conditions like background noise, noise of the vehicles. It consists of people uttering words from zero to nine in different orders repeatedly. The Audio-only ASR, Video-only ASR and AVSR average word recognition rates were obtained in each sample.

7.1 Dependence of Recognition Rates on Additive Noise

Shown in Fig 4 are the recognition rates of the three recognition systems, which allow us to compare the performance of the three systems under different SNR. The noise was added using a noise addition algorithm which adds noise to recorded audio stream corresponding to the required SNR. The stream weight for evaluating the CHMM parameters were fixed depending on the SNR of audio data. The acoustic observation vectors consist of 13 MFCC coefficients extracted from a window of 10 ms. The visual features were obtained as explained previously.

Figure 4: Comparison of the recognition Rates for the CHMM based AVSR model, audio-only ASR and Video-only ASR.



The obtained results are tabulated below:

SNR(dB)	00	05	10	15
AVSR	57.04	64.52	69.05	74.05
A-ASR	32.71	55.81	62.00	72.66
V-ASR	45.31	45.31	45.31	45.31

SNR(dB)	20	25	30	∞
AVSR	78.22	78.22	80.72	85.61
A-ASR	76.00	76.00	74.66	79.33
V-ASR	45.31	45.31	45.31	45.31

Table1: A comparison of different recognition rates at different values of audio signal to noise ratio (SNR).

Our experimental results indicate indicates that the speech recognition based on CHMM shows a remarkable increase in the word recognition rate. It shows an increase of 7.05% over audio only at an SNR of 10 dB. It also shows an increase of 23.74% over video only along expected lines.

8. CONCLUSION

This paper presents a speaker independent audio-visual speech recognition system that significantly increases the recognition rates, thus providing a robust speech recognizer. This uses a two stream CHMM to model the audio and video observation sequences. The CHMM model is able to process the audio and visual modalities independently, still maintaining the natural dependency between them. The improved accuracy is due to the set of visual features obtained around the mouth region. Thus the application alleviates the problem of noise effects on speech recognition and thus outperforms the other systems. The above advantages are evident in our experiments. Further, we were able to record the audio and video data separately, and feed the streams independently to the feature extraction units, making this system very compatible at places using parallel computing. The system is computationally intensive as it involves processing of audio and video simultaneously. However this is price to be paid for obtaining the improved results.

9. ACKNOWLEDGEMENT

We would like to thank Dr Ashok Rao, IISc, for his constant support and encouragement. Also we profusely thank Microcomputer Research Labs, Intel Corporation for providing the open source library for developing the AVSR model.

REFERENCES

[1] V. I. Pavlovic, "Dynamic Bayesian Networks for information fusion with applications to Human Computer Interaction," PhD Dissertation, *University of Illinois, Urbana- Champaign,* 1999.

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

[3] 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.

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

[5] Subramanya Amarnag, Sabri Gurbuz, Eric Patterson and John N. Gowdy, "Audio-Visual Speech Integration Using Coupled Hidden Markov Models for Continuous Speech Recognition," *IEEE Intenational Conference on Acoustics, Speech and Signal Processing, HongKong, 2003.*

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

[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] Open Source Computer Vision Library, Intel Corp http://www.intel.com/research/mrl/research/opencv/