AVSR Through HMM and Lip Tracking Method

Mayank Bhatnagar  
IPIT  
J.P. Nagar, UP, INDIA  
Email: Er.mayank2009@gmail.com

Kshitij Shinghal  
Department of E and C Engineering  
MIT, Moradabad, UP, INDIA  
Email: kshinghal@gmail.com

Prabhat Kr. Gupta  
IPIT  
J.P. Nagar, UP, INDIA  
Email: pkgie08@gmail.com

Sanjay Gurjar  
Assistant Professor  
Bhagwant University, Ajmer, Rajasthan, INDIA  
Email: ersanjay86@yahoo.in

ABSTRACT

This paper focuses on combining audio-visual signals and Hidden Markov Model for Audio Visual Speech Recognition in conditions of the highly disturbed audio speech signal. Recognition of audio-visual speech was based on combined hidden Markov models (CHMM). The described methods were developed for a single isolated command; nevertheless their effectiveness indicated that they would also work similarly in continuous audio visual speech recognition. The problem of a visual speech analysis is very difficult and computationally demanding, mostly because of an extreme amount of data that needs to be processed. Therefore, the method of audio-video speech recognition is used only while the audio speech signal is exposed to a considerable level of distortion. There are proposed the authors own methods of the lip edges detection and a visual characteristic extraction in this paper.

Moreover, the method of fusing speech characteristics for an audio-video signal was proposed and tested. A significant increase of recognition effectiveness and processing speed were noted during tests – for properly selected CHMM parameters and an adequate codebook size, besides the use of the appropriate fusion of audio-visual characteristics. The experimental results were very promising and close to those achieved by leading scientists in the field of audio-visual speech recognition.

Keywords: Coupled hidden, Markov models, Audio-visual speech recognition and Lip reading.

I. INTRODUCTION

At present, the most effective approach for achieving robustness of environment focuses on obtaining a clean signal through a head-mounted or hand-held directional microphone. However, this is neither tether-free nor hands-free, and it makes speech-based interfaces Very unnatural. Moving the speech source away from the microphone can degrade the speech recognition performance due to the contamination of the speech signal by other extraneous sound sources. For example, using monitor microphones for far-field input can severely degrade performance in the presence of noise.

It is well known that humans have the ability to lip read. We combine audio and visual information in deciding what has been spoken, especially in noisy environments. This fact has recently motivated significant interest in the area of audio-visual speech recognition (AVSR), also known as automatic lip reading or speech reading. Work in this field aims at improving automatic speech recognition by exploring the visual modality of the speaker’s mouth region, in addition to the traditional audio modality. Not surprisingly, automatic speech reading has been shown to outperform audio-only ASR over a wide range of conditions.

There were selected as recognition targets four languages with 20 words recorded for each of them. The proposed analysis of a feature trajectory based on three shapes (features, area with aspect ratio of internal lip region and area of intraoral region) provided the highest recognition rates of 93.6% in comparison with traditional methods and other regions. It was hard to achieve real-time lip-reading. In order to solve this problem they proposed the lip area detection method and feature extraction method suitable for smart phone environment. To find the accurate lip area the face area was detected by means of face colour information and eyes were located to detect the lip area with the geometrical relation. Then there were applied histogram matching, lip folding and KALMANN filter-to extract the outstanding features of the lip.
area in terms of light changes according to the surrounding. Then extracted features were used during the recognition process. They showed that the changes were recognized almost in real-time, and 30 out of 50 words were recognized. That indicated about 60% recognition rates.

The authors presented an integrated AVSR system, where noise tolerance was improved through enhancing the performance of three main components of the system. First, subsystem visual performance was improved by means of stochastic optimization methods for the hidden Markov models. Second, a new method of speech dynamic analysis was proposed, which improved acoustic efficiency. Third, an efficient integration of both signal streams was used to determine final robust recognition results by utilizing neural networks.

AAM is an efficient and robust method of tracking the motion of deformable objects in a video sequence. AAMs model variations in shape and texture of an object of interest. In contrast to MFCCs, AAMs require prior training before being used for feature extraction. In order to build an AAM it is necessary to provide sample images with the shape of the object annotated.

The shape of an appearance model is given by a set of \((x, y)\) coordinates represented in the form of a column vector:

\[
s = (x_1, y_1, x_2, y_2, ..., x_n, y_n)^T
\]

The coordinates are relative to the coordinate frame of the image. Shape variations are restricted to a base shape \(s_0\) plus a linear combination of \(N\) shape vectors:

\[
s = s_0 + \sum_{i=1}^{N} p_i s_i
\]

Where \(p_i\) are called the shape parameters of the AAM.

The base shape \(s_0\) is the mean of the object annotations in the training set, and the shape vectors are the \(N\) singular vectors corresponding to the \(N\) largest singular values of the training shape data matrix. Figure 4 shows an example of a base mesh and the first three shape vectors corresponding to the three largest singular values. The appearance of an AAM is defined with respect to the base shape \(s_0\). Appearance variation is restricted to a base appearance plus a linear combination of \(M\) appearance vectors:

\[
A(x) = A_0 + \sum_{i=1}^{M} \lambda_i A_i(x) \quad \forall x \in s_0
\]

The mechanism used a bank of amplitude-frequency filters with characteristics similar to human hearing. Besides that, twenty-dimensional MFCC (Mel Frequency Cepstral Coefficients) were used as the standard audio features for acoustic speech recognition. In addition, own methods were created to determine the beginning and the end of isolated words in audio speech signal. Finally, automatic methods of face, eyes and region of mouth detection were used for visual feature extracting. The visual features were: the corners, outside edges of the lips, and the visible tongue. In this paper, we present a novel method to directly retrieve clean speech features by exploiting visual features. As shown in Figure 2, we first generate multiple speech candidates from captured mouth movements, and then evaluate the consistency of the candidates using simultaneously captured audio signals. Finally, the clean speech signal estimated from using this method can be used as the input for subsequent ASR or AVSR classifiers. This study focuses on the steps for the estimation of clean speech signals in this framework.

In particular, we need to carry out the following issues:

1. Reduce the number of generated candidates because one mouth movement usually corresponds to several sounds.
2. Generate continual and smooth speech signals; this is required to check the consistency of the candidate sequence with the input audio signal.
3. Manage the temporal differences between the mouth movements and speech sounds because they are sometimes synchronized loosely.

To satisfy these requirements, we use a hybrid dynamical system (HDs) as a model for each audio and visual signal. HDSs are integrated models of discrete-event systems (e.g., HMMs) and dynamical systems.
II. CANDIDATE GENERATION BASED ON HDSS

Here, Figure 3 shows an overview of the learning and candidate generation phases. In the learning phase, we extract feature sequences of speech signals and mouth movements under a low noise environment. We train two HDS models from each of the speech and video feature sequences. In the following, we use HDSs and HDSv to denote the trained HDSs from speech and visual features, respectively. As a result of training of HDSs, the captured multimedia signals are partitioned and represented by pairs of interval sequences.

Figure 3: Flow for Generating Clean Speech Candidates

Let, \[ I_s^{(s)} = [b_k^{(s)}, e_k^{(s)}] \] and \[ I_v^{(v)} = [b_k^{(v)}, e_k^{(v)}] \] be overlapped intervals that appear in the speech and video interval sequences. Let \( m_1^{(s)}; m_2^{(s)} \) be labels of LDSs that represent the dynamics in the intervals. Assuming that HDSs and HDSv comprise sets of LDSs, \( D_s^{(s)} = \{D_1^{(s)}, D_{K_s}^{(s)}\} \) and \( D_v^{(v)} = \{D_1^{(v)}, D_{K_v}^{(v)}\} \), respectively, we calculate the following distributions for all the pairs of LDSs \( (D_s^{(s)}, D_v^{(v)}) \):

\[
P(m_k^{(s)} = D_s^{(s)}, m_k^{(v)} = D_v^{(v)} | I_k^{(s)} \cap I_k^{(v)} \neq \emptyset),
\]

\[
P(b_k^{(s)} - b_k^{(v)}, e_k^{(s)} - e_k^{(v)} | m_k^{(s)}, m_k^{(v)}, I_k^{(s)} \cap I_k^{(v)} \neq \emptyset)
\]

The first distribution represents the co-occurrence of LDSs, and the second distribution represents the possible degree of temporal gaps between two LDSs. We refer to the set of these distributions as the timing structure model and use to represent all the model parameters.

In the candidate generation phase, we first extract feature sequences from novel audio and visual input. Let \( V = [v_1; \vdots; v_{TV}] \) be a captured visual feature sequence. We generate candidates of clean speech sequences \( S_c = [s(c)_1; \vdots; s(c)_{TV}] \) from \( V \). This generation technique is almost similar to the method described except that the generated sequences are multiple and the modality is inverted; that is, the original method generates a lip motion sequence from a given input audio signal. The following are the steps for generating a single candidate:

1. Partition the input sequence \( V \) into an interval sequence \( I(v) = [I(v)_1; \cdots; I(v)_{K_v}] \) by using the trained HDS v.
2. Generate an interval sequence \( I(s) = [I(s)_1; \cdots; I(s)_{K_s}] \) from \( I(v) \) based on the trained timing structure model.
3. Generate a speech feature sequence candidate \( S_c \) from the generated interval sequence \( I(s) \) by using the trained HDSs, where:

\[
\tilde{I}(s) = \arg \max_{I(s)} P(I(s) | I(v), \Phi)
\]

II. PRE-PROCESSING OF INPUT AUDIO SIGNAL

In case of recognition, after a preliminary filtration of a signal the next stage is to create a clean and proper audio signal, through removal silence at the beginning and at the end of a signal. An identification, which frames are to be rejected, is not simple to determine whether energy matches the frame condition. There can appear instantaneous power spikes before the beginning of a useful signal. In most cases they are related to interference caused by the environment in which the signal recording takes place. Therefore, a more complex search is needed to determine the beginning and the end of the audio signal. For this reason the system uses two more parameters of the LRP (the initial number of frames) and LRK (the final number of frames). The first one specifies the number of frames consecutively, which energy satisfies the condition. If that number of frames is found, then the first of these frames is marked as the beginning of the audio signal. The second parameter specifies the number of frames consecutively; the power does not satisfy the condition. If that number of frames is found, then the first of these frames is determined as the end of the audio signal. If the required number of frames is not found before reaching the last frame, then the end of the signal is assigned to the frame, the first condition is not fulfilled. As the speech signal is not stationary, what results from dynamic properties of human speech, next stage depends on the use of division of entrance signal onto stationary frame Boxes. Every such a stationary frame box was replaced by the symbol of observation in the process of creation the observation vectors. In the created system it was accepted the length of every frame box equal 30 ms. to keep the signal stationary a method of delaying next frame boxes was applied. As a result every next frame box is sewing on previous with delay.

Figure 4: The Sequence or the Way of Searching LRP and LRK
IV. MODELING AUDIO-VISUAL SPEECH USING DYNAMIC BAYESIAN NETWORKS

A Dynamic Bayesian Network (DBN) is an extension of Bayesian networks that allows for modeling variable-length (and potentially semi-infinite) sequences of hidden and observed random variables and their dependencies. Variables are represented by nodes in a graph, and dependencies are represented by directed arcs connecting the nodes. As with Bayesian networks, a DBN must constitute a directed acyclic graph (DAG). The term dynamic is used as DBNs are typically used to model dynamic systems.

A DBN graph constitutes a set identically structured time slices. The semantics of a DBN is defined by a prior distribution over the nodes in the initial slice.

\[ p(V_1) = \prod_{i=1}^{N} p(v_i^1 | pa(v_i^1)) \]

and a distribution over a two-slice temporal Bayesian network defining the transition from a slice to the next:

\[ p(V_t | V_{t-1}) = \prod_{i=1}^{N} p(v_i^t | pa(v_i^t)) \]

In above equation \( v_i^t \) is the random variable represented by the \( i \)th node in time slice \( t \) of the DBN and \( pa(v_i^t) \) is the set of variables representing the parents of the \( i \)th node in the graph. We restrict parent nodes to lie in the same time slice as node \( i \), or in the previous time slice. The set of random variables is typically partitioned into hidden and observed nodes \( V = (Z, X) \) where \( Z \) and \( X \) are hidden and observed variables, respectively. A well-known example of a DBN is the Hidden Markov Model (HMM). An HMM is probabilistic model defined by:

\[ p(X, Z | \theta) = p(x_1 | \pi) \prod_{t=1}^{T} p(z_t | z_{t-1}, A) \prod_{t=1}^{T} p(x_t | z_t, \phi) \]

Where \( \pi \) is the prior over HMM states, \( A \) is the transition matrix and \( \phi \) is the observation model parameters. From above equations we see that an HMM has the graphical representation shown in Figure 5. In speech recognition applications it is common to model observations as a mixture of Gaussians:

\[ b_t(x_t | z_t) = \sum_{j=1}^{M} w_{ij} N(x_t | \mu_{ij}, \Sigma_{ij}) \]

where \( w_{ij}, \mu_{ij} \) and \( \Sigma_{ij} \) are the weight, mean and covariance of the \( j \)th mixture component and \( i \)th HMM state, respectively. In fact, by introducing an additional multinomial hidden variable \( y_t \) indicating which mixture component is selected, we can model an HMM with a Gaussian mixture observation model as a DBN where each node in the graph has a distribution belonging to the exponential family [6] (we can show that GMMs are not in the exponential family). This property has important consequences for inference and learning, as it allows us to use general DBN inference and learning algorithms. The graphical representation of the HMM with GMM observations is shown in Figure 6.

Figure 6: An HMM with GMM Observation Model

Several asynchrony DBN models for AVASR are proposed. From the seven models considered, it was found that the coupled HMM gives the lowest word error rate on the AVASR task. In the audio-visual coupled HMM (AV-CHMM) the observation and state nodes of the audio and visual streams are separate, but coupled at the state level. The AV-CHMM model is shown in Figure 7. Note that, to avoid clutter we have omitted the node labels and the details of the observations model. The level of asynchrony is constrained by limiting the number of states that the two streams are allowed to desynchronize.

Figure 7: Audio-visual Coupled HMM

V. EVALUATION RESULTS

To evaluate our proposed system, we tested it indifferent noisy environments with different persons, by the Intel AVCSR application. The Intel AVCSR allows us to add noise to the input video file concurrently in various acoustic SNR. Our video recording files are completely without noise, so we can review and analyze the various conditions imposed by external noise in our system. The facial and environmental conditions are very important in these tests and will lead to changes in test results. The time between saying a word and showing it in AR environment directly depends on the power of AVSR engine. The Intel AVCSR engine needs approximately 10 seconds to process video file and get results, but if we use a real-time AVSR engine in the future, this time decreases to less than 2 or 1 seconds.
We classified our tests according to different conditions. These tests are classified to: Without Noise, 20 dB SNR, and 0 dB SNR. Each of these conditions simulates a particular noisy environment for our system. We collect and review the results of Word Error Rate in each test and observed that our proposed system can work successfully and performs good results in all tests.

The following will explain the test results and how our proposed system will work in different noisy condition. After opening and processing video file with the Intel AVCSR application, it wrote the output results of each engine separately in the text file that will be used for testing our proposed system. These tests have been done on different aspects and and the results have been shown in Figure 8. It can be seen form Figure 8 that in different noisy condition (i.e. without noise, 20 dB SNR, and 0 dB SNR) the AVSR system has lower word error rate compared to ASR and VSR only.

Therefore, our system can work fine on average in different noisy environments by using Intel AVCSR application. All words that the Intel AVCSR application wrote in the text file, after processing video file that take about 10 seconds, were shown immediately. We compared the average word error rate of each test in the Intel AVCSR application in equal conditions and the results have shown in Figure 9. The result shows that our system (AVSR System) and the Intel AVCSR application still has lower word error rate in 0dB, 10 dB, 20dB, 30dB SNR and without noise compared to ASR and VSR.

As it can be seen from Figure 9, with noise reduction feature the percentage of word error rate decreases. Therefore, our system only depends on how much the AVSR engine is powerful and advances in this area have a direct impact on the system and makes our system more powerful in the future. It also shows that our system can be very useful for people in different noisy places.

**Figure 8: Word Error Rate of Three Persons with Average, Under Various Acoustic SNR.**

**Figure 9: Word Error Rate in the AVCSR Engines**

**VI. CONCLUSION AND FUTURE SCOPE**

Based on the tests there has been shown that the method of audio-visual speech recognition worked properly and it could work in the systems in the real world. Test result shows the accuracy of speech recognition that a large impact on the proper identification was affected, disturbed or not, the environment. The results shown also that this method should be developed. There are plans to expand the method of automatic detection of the position of the tongue, for each of the spoken video phonemes. Further work will also build a system for Polish speech recognition, based on an analysis of individual phonemes. Such an approach would allow for continuous speech recognition. The method of audio-visual recognition of Polish speech was used in the system to control the camera movement using voice commands. To increase the efficiency of the method FPGA (Field Programmable Gate Array) can be used. As a consequence the system is enabled to work properly in a real time as well as the hardware level is supported. An advantage of the proposed method is the satisfactory effectiveness created by the lip-tracking procedures, and
the simplicity and functionality by the proposed methods, which fuse together the audio and visual signals. A decisively lower level of mistakes was obtained in audio-visual speech recognition, and speaker identification, in comparison to only audio speech, particularly in facilities, where the audio signal is strongly disrupted.

REFERENCES


