Audio-Visual Speech Recognition using LIP Movement for Amharic Language

Mr. Befkadu Belete Frew
College of Electrical and Mechanical Engineering,
Department of Software Engineering,
Addis Ababa Science and Technology University,
Addis Ababa, Ethiopia

***Abstract* - Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to a written text. In recent years, there have been many advances in automatic speech reading system with the inclusion of visual speech features to improve recognition accuracy under noisy conditions. By identifying lip movements and characterizing their associations with speech sounds, the performance of speech recognition systems can be improved, particularly when operating in noisy environments.**

In this study, for face and mouth detection we use Viola-Jones object recognizer called haarcascade face detection and haarcascade mouth detection respectively, after the mouth detection ROI is extracted. Extracted ROI is used as an input for visual feature extraction. DWT is used for visual feature extraction and LDA is used to reduce visual feature vector. For audio feature extraction, we use MFCC. Integration of audio and visual features are done by decision fusion. As a result of this, we used three classifiers. The first one is the HMM classifier for audio only speech recognition, the second one is HMM classifier for visual only speech recognition and the third one is CHHM for audio- visual integration.

In this study, we used our own data corpus called AAVC. We evaluated our audio-visual recognition system with two different sets: speaker dependent and speaker independent. We used those two evaluation sets for both phone (vowels) and isolated word recognition. For speaker dependent dataset, we found an overall word recognition of 60.42% for visual only, 65.31% for audio only and 70.1 % for audio-visual. We also found an overall vowels (phone) recognition of 71.45% for visual only, 76.34% for audio only and 83.92 % for audio-visual speech. For speaker independent dataset, we got an overall word recognition of 61% for visual only, 63.54% for audio only and 67.08% for audio-visual. The overall vowel (phone) recognition on the speaker independent dataset is 68.04% for visual only, 71.96% for audio only and 76.79 % for audio-visual speech.

Keywords: Amharic, Lip-reading, visemes, appearance-based feature, DWT, AAVC

1. INTRODUCTION

Automatic speech recognition (ASR) is a technology that an integral part of future human computer interfaces that are envisioned to use speech, among other means, to achieve natural, pervasive, and ubiquitous computing. [1]. Today's trend is to make communication and interaction between humans and their artificial partners easier and more natural. Speech recognition technology has reached a maximum of performance and good recipes for building speech recognizers have been written. However, the major problems of background noise and reverberations due to the environment are still insurmountable. Therefore, inspecting other sources, other than sound, for complementary information which could alleviate these problems, is a necessity [2].

It is well known that both human speech production and perception are bimodal process in nature. Visual observation of the lips, teeth and tongue offers important information about the place of pronunciation articulation. A human listener can use visual cues, such as lip and tongue movements, to enhance the level of speech understanding. The process of using visual modality is often referred to as lip-reading which is to make sense of what someone is saying by watching the movement of his lips [3]. A Visual speech recognition (VSR) system refers to a system which utilizes the visual information of the movement of the speech articulators such as the *lips, teeth* and somehow *tongue* of the speaker. The advantages are that such a system is not sensitive to ambient noise and change in acoustic conditions, does not require the user to make a sound, and provides the user with a natural feel of speech and dexterity of the mouth [4].

Speech command based systems are useful as a natural interface for users to interact and control computers. Such systems provide more flexibility as compared to the conventional interfaces such as keyboard and mouse. However, most of these systems are based on audio signals and are sensitive to signal strength, ambient noise and acoustic conditions [4]. To overcome this limitation, speech data that is orthogonal to the audio signals such as visual speech information can be used. The systems that combine the audio and visual modalities to identify utterances of approaching phonetics. One approach studies the physiological mechanisms of speech production. This is known as *articulatory phonetics*. The other, known as *acoustic phonetics*, is concerned with measuring and analyzing the physical properties of the sound waves we produce when we speak. According to articulatory phonetics, organs of articulation are divided into movable articulators and stationary articulators. Movable articulator is the articulator that does all or most of the moving during a speech gesture. The movable articulator is usually the lower lip, some part of the tongue and jaws. A stationary articulator is the articulator that makes little or no movement during a speech gesture. Stationary articulators include the upper lip, the upper teeth, and the various parts of the upper surface of the 63are known as audio-visual speech recognition (AVSR) system.

There are two ways oral cavity, and the back wall of the pharynx [5, 6]. Those articulators movement dose not affected by noise. Thus, visual speech information from the speaker's mouth region will be improve noise robustness of automatic speech recognizers.

2. LITERATURE REVIEW

Phonetics and Phonology

The human perception of the world is inherently multi-sensory since the information provided is multimodal. In addition to the auditory information, there is visual speech information provided by the facial movements as a result of moving the articulators during speech production [16, 17]. The use of visual speech information has introduced new challenges in the field of ASR. These are robust face and mouth detection, extraction and tracking of a visual region of interest (ROI), extraction of informative visual features from the ROI, the integration of audio and visual modalities and the provision of suitable classifiers [18].

In order to understand the link between the audio signal and the corresponding visual signal that can be detected on the mouth/lips of the speaker, we need to have some understanding of how speech is produced. Phonetics and phonology are the two fields of grammar which deal with the study of the sounds of human language. Phonetics studies the actual speech sounds of the language including the way how the sound is produced, transmitted, and perceived. Phonology on the other hand is the systematic study of how speech sounds are organized to form sound systems. Phonetics is related to the science of acoustics in that it uses much of the techniques used by acoustics in the analysis of sound [7].

Speech Recognition for Amharic Language

Automatic speech recognition for Amharic was conducted in 2001 by Solomon Berhanu [10].The author developed isolated Consonant-Vowel syllable Amharic recognition system which recognizes a subset of isolated consonant-vowel (CV) syllable using HTK (Hidden-Markov Modeling Toolkit). The author selected 41 CV syllables of Amharic language out of 234 and the speech data of those selected CV syllables were recorded from 4 males and 4 females with the age range of 20 to 33 years. The average recognition accuracies were 87.68% and 72.75% for speaker dependent and independent systems, respectively.

Kinfe Tadesse [11] developed a sub-word based isolated Amharic word recognition systems using HTK (Hidden Markov Model Toolkit). In this experiment, phones, triphones, and CV-syllables were used as the sub-word units and selected 20 phones out of 37 and 104 CV syllables for developing the system. The speech data of those selected recorded from 15 speakers for training and 5 speakers for testing. Average recognition accuracies of 83.07% and 78% were obtained for speaker dependent phone-based and triphone-based systems respectively. With respect to speaker independent systems, average recognition accuracies of 72% and 68.4% were obtained for phone and triphone-based speaker independent systems respectively.

Asratu Aemiro [70] developed two types of Amharic speech recognition (ASR) systems, namely canonical and enhanced speech recognizers. The canonical ASR system is developed based on the canonical pronunciation model which consists of canonical pronunciation dictionary and decision tree. The canonical pronunciation dictionary is prepared by incorporating only a single pronunciation for each distinct word in the vocabularies. The canonical decision tree is constructed by only considering the place of articulations of phonemes as it was commonly used by the previous Amharic ASR researchers. On the other hand, the development of enhanced speech recognition system takes enhanced pronunciation model which consists of enhanced pronunciation dictionary and enhanced decision tree where both are designed by considering the patterns we identified based on the co-articulation effects of phonemes. The construction of the enhanced pronunciation model incorporates alternative pronunciations for each distinct word according to the identified patterns. Finally, the author evaluated the recognition accuracy of the two ASR systems by introducing the enhanced pronunciation model into Amharic ASR systems, and obtained an improvement of 14.04% and 13.93% at the word and sentence level, respectively using enhanced ASR system. Furthermore, the author tested the recognition accuracy of the two ASR systems with different parameters and the test results are reported. Accordingly the author recommend that incorporating the enhanced pronunciation model in to Amharic ASR systems would be important in order to improve the recognition accuracy of the recognizers.

3. DESIGN OF AUDIO-VISUAL AMHARIC SPEECH RECOGNITION

Figure 1 depicts the overall system architecture developed in this work. The system architecture shows how these components interact to accomplish the recognition process. The system starts by acquiring the audio speech signal of a speaker through a microphone as well as the video frames of the speaker's face by means of a camera. The audio and visual streams are then ready for analysis at a signal level.

In contrast to audio-only speech recognition systems, where only the audio stream of information is available, here there are two streams of speech information, these are audio stream and the video stream. In this architecture, audio and video signals are separated and independently processed to extract relevant features. Once the relevant information is picked from each signal; they are fused together and then used for an improved speech recognition system. The system implementation consists of three main stages these are design of the front-end processing system for both audio and visual, integration of audio and visual vectors and the training of the recognizer.

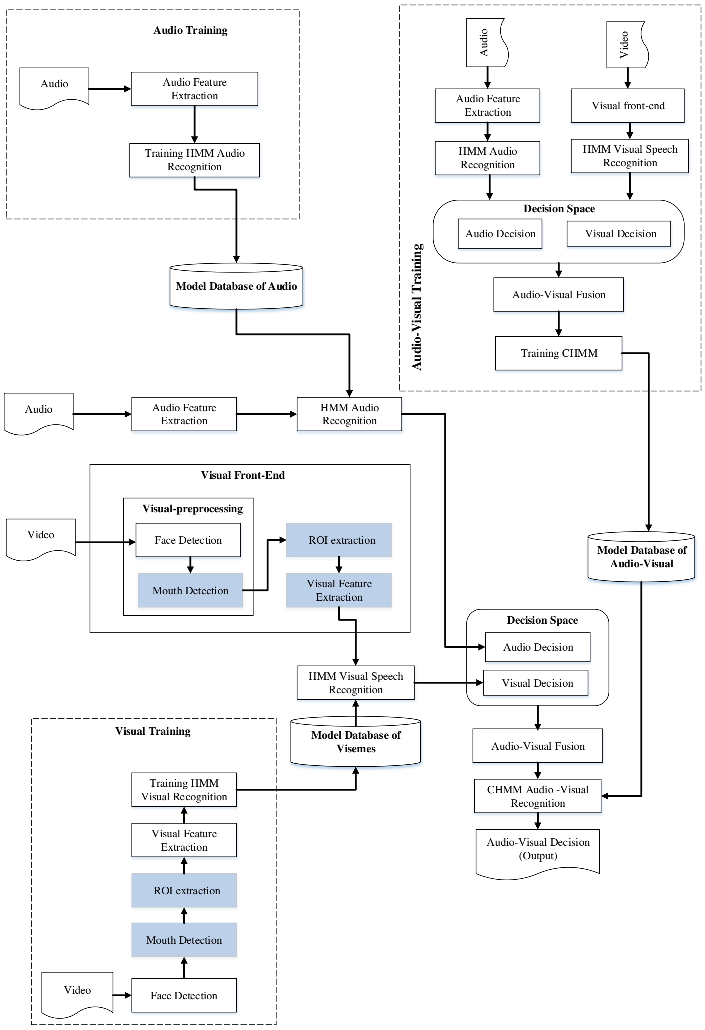


Figure 1: System Architecture

3.1. Visual Front-End Component

As the videos of speakers contain information not related to the speech itself, such as the identity and background, the visual front-end needs to remove this superfluous information leaving only that related to speech. The mouth region of the speakers is identified and a region of interest (ROI) is isolated prior to the extraction of visual speech features. Front-end processing transforms speech into a parameter vector suitable for subsequent processing and consists of the preprocessing of video sources, followed by feature extraction.

The visual front-end identifies the portion of the speaker's face following the mouth that contains the most speech information and extracts that information in a parametric form suitable for processing by the recognizer. The front-end component can be divided into three sub-tasks: visual preprocessing, region of interest (ROI) extraction and feature extraction. Though often considered separately, the three tasks are largely interdependent.

i. Visual Preprocessing

Before being applied to the recognizer for training or recognition purposes, visual streams need to be preprocessed to remove data irrelevant to speech and to enhance certain characteristics that help to improve speech recognition performance.

The first step of visual preprocessing is face detection followed by mouth detection and ROI extraction. The image which acquired is by the camera is RGB image. Before applying visual preprocessing on the input frame image the image should be change to the grey-level image.

Gray-level images are referred to as monochrome, or one-color image. They contain brightness information only. The typical image contains 8 bit/ pixel (data, which allows us to have (0-255) different brightness (gray) levels. The 8 bit representation is typically due to the fact that the byte, which corresponds to 8-bit of data, is the standard small unit in the world of digital computer.

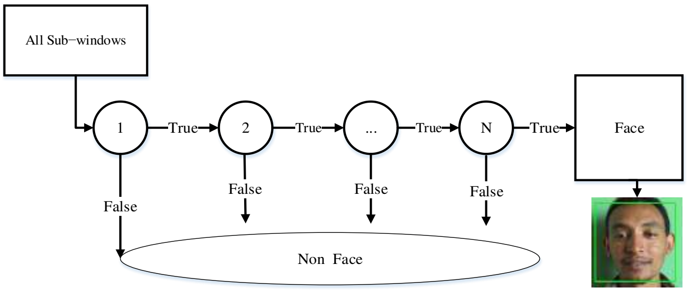


Figure 2: Results of Face Detection Using Viola-Jones Object Recognizer.

After detecting the face as shown in figure 2 then we divide the face into upper-face and lower-face to simplify the next process. When applying mouth detection on the full face, the probability of detection of false mouth become high. Thus, to reduce the detection of the false mouth we divide the detected face into two parts as illustrated the figure 3.

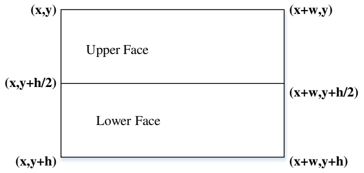


Figure 3: Coordinates of Detected Face.

From the detected face, we got the bounding box coordinate of the face. As shown in Figure 4.4, h= height of bounding box of the face, w = width of bounding box of the face, (x, y) = left-top the coordinate of bounding box of the face, (x+w, y) = right-top coordinate of the face, (x, y+h) = left-bottom coordinate of bounding box of the face and (x+w, y+h) = right-bottom coordinate of bounding box of the face. Figure 4.

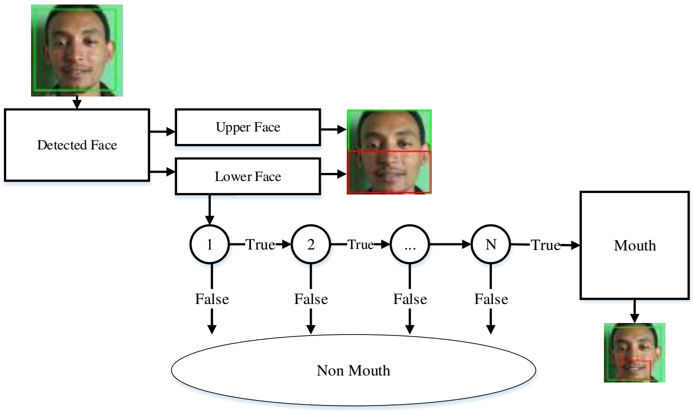


Figure 4: Results of Mouth Detection Using Viola-Jones Object Recognizer.

ii. Region of Interest (ROI) Extraction

The ROI provides the raw input data for visual feature extraction and thus the overall performance of an audio-visual automatic speech recognition (AVASR) system is greatly influenced by the accurate extraction of ROI. The identification of the ROI is made more difficult due to the high deformation of lip shape, as well as the variation in the content of the mouth region due to the presence or absence of tongue, teeth, and opening and closing of mouth during speech. ROI detection approaches are also often influenced by variations in lighting conditions and changes in the pose and orientation of the speakers. The presence or absence of a beard or moustache also affect ROI extraction.

After identification of speakers' mouth region the next stage is the extraction of the ROI as Shown in figure below for appearance based feature approaches a bounding box around the lower half of the face containing the mouth region is extracted as desired ROI. In our case, we use bounding box and with the size based on the detected mouth size and resize this bonding box to ROI size based on the center of the detected mouth as shown in the Figure 6 and 7 the coordinates of bounding box are selected in such a way that it contains the desired ROI in all the frames of utterance. This is used to create uniform image size for every utterances.

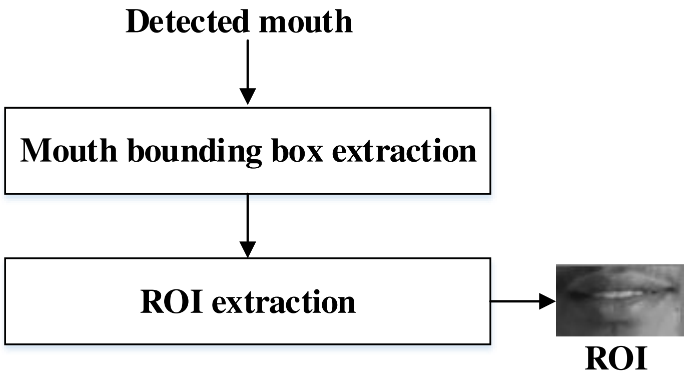


Figure 5: ROI Extraction form Single Frame.

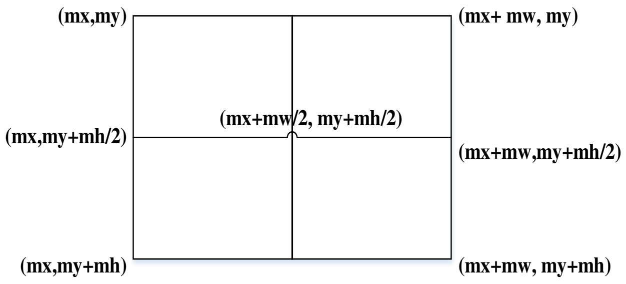


Figure 6: Coordinate of Detected Mouth and Center of Bounding Box.

As shown in Figure 6, mh is the height of bounding box of the detected mouth, mw is the width of bounding box of the detected mouth, (mx, my) is the left-top coordinate of bounding box of the detected mouth, (mx + mw, my) is the right-top coordinate of bounding box of the detected mouth, (mx, my+ mh) is the left-bottom coordinate of the detected mouth and (mx + mw, my + mh) is the right-bottom coordinate of the bounding box of the detected mouth. Therefore, the ROI coordinate is as shown in Figure 7 relative to the center of bounding box of the detected mouth (mx+mw/2, my+mh/2).

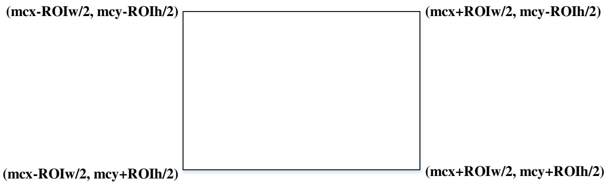


Figure 7: Coordinate of Region of Interest (ROI).

As discussed in Figure 6 the center of bounding box of the detected mouth is (mx+mw/2, my+mh/2). This is considered as the center coordinate of ROI to resize the bounding box of the detected mouth into a uniform size of ROI whose value is assigned to the (mcx, mcy) coordinate. Where mcx is the coordinate x value of the center of ROI, mcy is the coordinate of y value of ROI, ROIh is the height of the region of interest, and ROIw is the width of region of interest.

iii. Visual Feature Extraction

The purpose of feature extraction is to retain as much speech related information as possible from the original images of the speaker in a reasonably small number of parameters. In visual feature extraction, a range of transformation techniques, such as discrete cosine transform (DCT), discrete wavelet transform (DWT), principal components analysis (PCA) and linear discreminent analysis (LDA) are used.

Once we extracted the mouth region (the region of interest), there is a wide choice of algorithms that can be applied to extract visual features form the ROI. The most commonly used transforms in appearance-based feature extraction approaches for AVASR research are the DCT and the DWT. According to Lee *et al* [76], the DWT has many advantages over the DCT. First, DCT difference coding is computationally expensive. Second, wavelets do not cause blocking artifacts. Thus, for this work we used appearance-based feature extraction method called DWT. In this step the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 08, August-2019

extracted ROI image is used as an input for feature extraction.

The DWT transform decomposes the input image into a low-frequency sub band (known as the approximate image) and high-frequency sub-bands (known as detailed images), as shown in Figure 8. The LL region of the DWT transform in Figure 8 contains the low frequency contents of the image, the HL region contains the high-frequency horizontal details, LH the high-frequency vertical details and HH the high-frequency details for both the horizontal and vertical direction. The application of the DWT to an image results in high-pass and low-pass filtering of the image. Further refined details of an image can be extracted by applying higher levels of decomposition. This is achieved by the application of DWT to the sub-images obtained in the lower level, starting from the original input image. First-level decomposition means the DWT of the original image; second-level decomposition means the DWT of sub-images obtained in the first level and so on, whereas the low frequency components are known as approximate coefficients while the high frequency components are known as detailed coefficients.

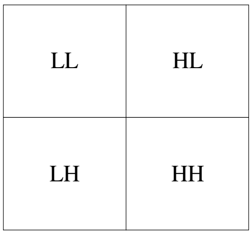


Figure 8: Single Level DWT Decomposition of an Image.

Generally, two dimensional wavelet transformation is applied on the image of ROI which results in four sub images, as shown in Figure 8, as average image(LL) and three detail images (HL, LH and HH). For the purpose of image classification, the three detail images are discarded and the average sub image is converted into a vector by concatenating the columns. This vector is used as image representation for the purpose of image classification. The average image which is in the form of vectors as ROI image feature, is reduced to 30 dimensions by applying LDA. Figure 9 shows the process of DWT on ROI image.

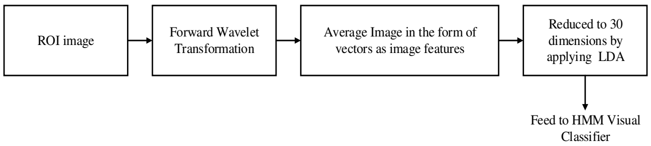

ROI image
Forward Wavelet Transformation
Average Image in the form of vectors as image features
Reduced to 30 dimensions by applying LDA
Feed to HMM Visual Classifier

Figure 9: DWT and HHM Visual Classifier.

3.2. Visual Speech Modeling

Prior to the visual feature extraction stage, visual speech modeling is required. This issue is very important to the design of audio-visual fusion. The basic unit of speech in the visual space is viseme. The concept of viseme is usually defined in accordance with the mouth shape and mouth movements. To represent a viseme, one should develop a method for representing the video sequence, further complicating the video processing stage. Fortunately, most of the visemes can be represented by stationary mouth images. The benefit of using such a representation is that it can be mapped directly to the acoustic speech units, which makes the integration process pretty easy. Therefore, for this thesis, we use visemes as visual speech units. For each phone and word we used 30-40 sequence of images.

To be able to design the visual front-end, it is desirable to define for each phoneme its corresponding viseme. This enables us to integrate the visual speech recognition system into existing acoustic-only systems. Unfortunately, speech production involves invisible articulatory organs, which renders the mapping of phonemes to visemes into many-to-one. Consequently, there are phonemes that cannot be distinguished in the visual domain. For example, the phonemes ፐ [pe], በ [be], ጰ [pe’], and መ [me] are all produced with a closed mouth and cannot be distinguished visually one phoneme from the other phonemes, so they will be represented by the same viseme. It is important also to consider the effect of the dual of the allophone, where the same viseme can be realized differently in the visual domain due to the speaker variability and the context. To our best knowledge, unlike the phonemes, there is no viseme set that is commonly used by all researchers for Amharic language.

For notational convenience, we shall identify the visemes by the names of the phonemes they represent. As our focus is on oral movement (place of articulation), we shall refer to the movement of mouth when voicing a particular phoneme as a viseme.

The clustering of the different mouth images into viseme classes is done based on the place of articulation and the manner of articulation manually on the base of visual similarity of these images. Accordingly, we obtain the viseme classes and the phoneme-to-viseme mapping for Amharic consonant are in Table 1 and phoneme-to-viseme mapping for Amharic vowels in Table 2. Table 3 shows sample sequences of image for Amharic vowels.

Table 1: Viseme Classes and the Phoneme-to-Viseme Mapping for Amharic Consonant.

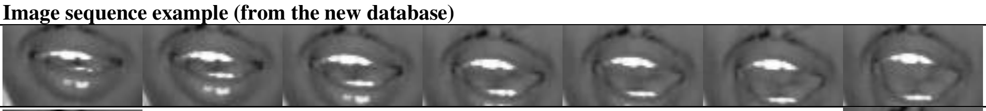
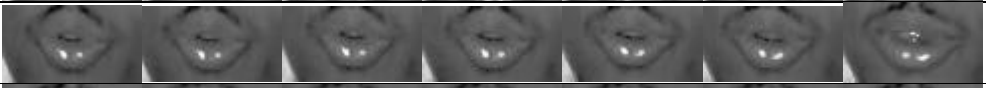
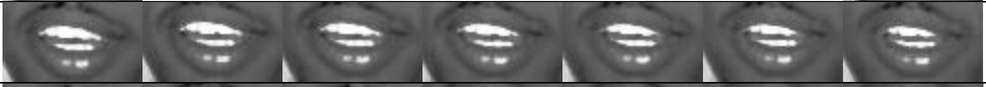
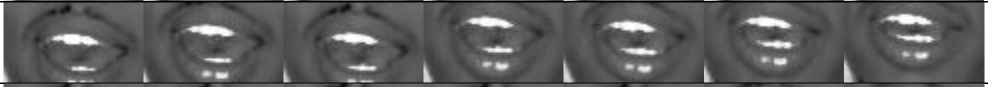
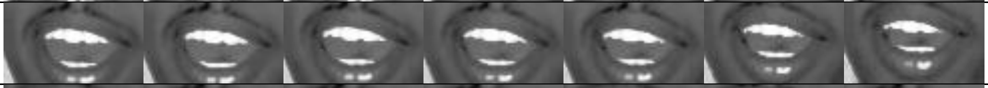
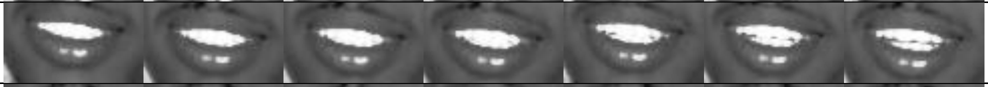
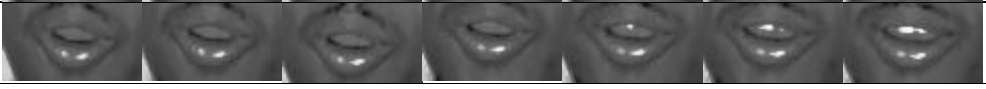
Viseme Number	**Group of phoneme**	**Place of Articulation**	**Manner of Articulation**	**Mouth description / Viseme description**
0	None	-	-	(silence and relax)
1	በ[b],ፐ[p],መ[m], ጰ[p’]	Bilabial	Stops	• These are sounds produced when the lips are brought together. • The nature of mouth: lips together.
2	ወ-[w]	Bilabial	Glides	• Small-rounded open mouth state.

3	ቭ[v],ፍ[f]	Labiodental	Fricatives	• Almost closed mouth state; upper teeth visible; lower lip moved inside. • Lower lip is raised towards the upper front teeth.
4	ድ[d],ጥ[t’],ት[t]	Alveolar	Stops	• Medium open, not rounded, mouth state; teeth fully visible, tongue partially visible.
5	ዝ[z], ስ[s] ,ጽ[s’]	Alveolar	Fricatives	• Medium open, not rounded mouth state, teeth fully visible but the tongue is not visible.
6	ን[n]	Alveolar	Nasals	• Medium open, not rounded, mouth state; teeth visible.
7	ል[l]	Alveolar	Liquids	• Tip of tongue behind open teeth, gaps on sides.
8	ዥ[ž],ሽ[š],	Palatal	Fricatives	• The upper and the lower teeth closed together • longitudinal open mouth state
9	ጅ[j],ች[c], ጬ [c’]	Palatal	Affricates	• The upper and the lower teeth closed together • open mouth state
10	ኝ[N]	Palatal	Nasals	• The upper and the lower teeth closed together with little gap between. • open mouth state
11	ር[r]	Palatal	Liquids	• Open mouth state, the tip of tongue close to the upper teeth.
12	ይ[y]	Palatal	Glides	• Open mouth state, the upper and the lower teeth closed together with little opening.
14	ግ[g],ክ[k],ቅ[k’]	Velar	Stops	• Slightly open mouth with mostly closed teeth
15	ጓ[g^{w}],ኳ[k^{w}],ቋ[k’^{w}]	Labiovelar	Stops	• Started with round lip state, with small open and lip pic, and end with wiled mouth open.
16	ዕ [?] ,ህ[h] ,ኋ[h^{w}]	Glottal	Fricatives	• The lip is static but slightly open to pass the internal air to outside.

Table 2: Viseme Classes and the Phoneme-to-Viseme Mapping for Amharic Vowels.

Visemes Number	Group of phoneme	Mouth description / Viseme description
1	ኢ [i], ኤ [e]	• For ኢ [i], open mouth state, the middle of tongue at the lower teeth. • For ኤ [e], Wild open mouth state, visible tongue and the tip of the tongue at lower teeth.
2	ኦ[o], ኡ[u],	• Round lip, with small open, and the lip pic
3	እ[I], ኧ[A], ኣ [a]	• Longitudinal open mouth state; tongue visible.

Table 3: Sample Viseme Image for Amharic Vowels.

Phoneme	Image sequence example (from the new database)
ኣ[a]	
ኡ[u]	
ኢ [i]	
ኧ[A]	
ኤ [e]	
እ[I]	
ኦ[o]	

3.3. HMM Visual Speech Recognition

The last processing stage of visual speech (lip-reading) is feature classification. For the classification, process the HMM is used due to its popularity that has followed from many successful applications in the statistical modeling of audible speech and packages availability in python programing for implementation HMM (e.g., hmmlearn, scikit-learn).

Building HMM Visual Classifier
At this stage, it is crucial to define the basic structure of the HMM developed for viseme-based visual speech recognition, so that we can understand the general framework in which our proposed visual features will be integrated.

Training the HMM is one of the basic tasks of the recognition process. Now, each visual word or visual phone is represented with a sequence of symbols. For a single training video, we have a symbol vector. In training HMM, the basic inputs are the output sequence (which is the symbol matrix in our case), the initial transition and emission matrices. Before the preparation of the initial transition and emission matrices deciding the type of the model and number of states is a very critical task.

Depending on the type of the problem, various model types and number of states can be selected. In this work, two different types of HMM architectures were used based on word and phone models. As shown in Figure 10, for each *viseme (phone)* in the database, a 3 states HMM was designed, and the output (most likely) visemes sequences is recognized as a phone by means of HMM. As shown in Figure 11, for each *word* in the database, an HMM with a different number of states is designed, the number of the states being the number of the visemes (phone) appearing in a specific word.

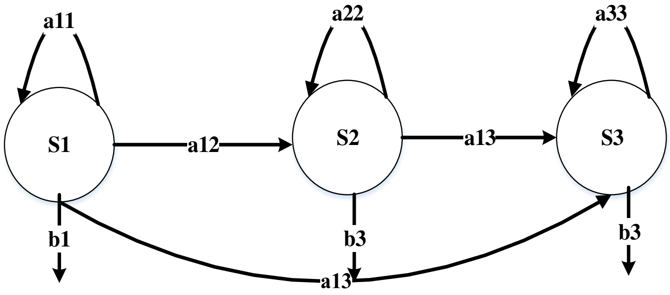


Figure 10: Three State HMM Topology for Phone

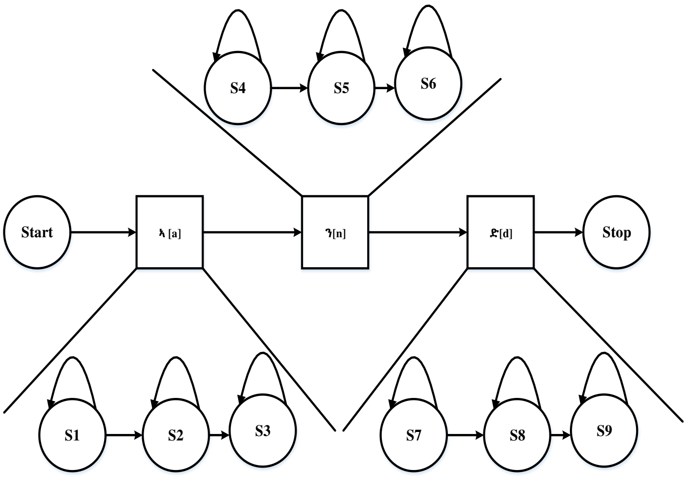


Figure 11: HMM Topology for the Word One [ANID]/ አ ን ድ.

To train the HMM, in addition to the observed symbol sequence and the decision of the number of states, we need initial state transition and emission matrices. We will prepare these matrices with random numbers. The state transition matrix is an NXN matrix where N is the number of states. The emission matrix is an NXM matrix where N is the number of states and M is the size of the observable symbols which is found empirically during the vector quantization process.
Now we can start training our model and the final HMM parameters will be produced by iterative process. We used the Baum Welch algorithm [77] to train the HMM. The algorithm updates the parameters of the HMM iteratively until convergence following the procedure below.

Finally, we will have our HMM for a given visual word or visual phone. The HMM model will be represented by the state transition and emission matrices produced after training. These matrices will be retained in the database with their respective visual word or visual phone id to use them later for recognition process.

Visual Speech Recognition
Recognition is the process of finding the most probable HMM from a set of HMMs (which were produced during the training phase) that can produce a given observed sequence. For this process we used the forward and backward algorithms to compute the likelihood that a model produced for a given observation sequence.

The likelihood can be effectively calculated using dynamic programming by forward algorithm which reproduce the observation through HMM and backward algorithm which back trace the observation through HMM.

The same steps will be followed as the training part to have a set of feature vectors that represent the frames and vector quantization process which enabled us to substitute a sequence of feature vector values to a set of distinct symbols. But, for testing the recognition system we used different data from the ones that we used for training purpose.

A given sequence of images for a visemes of word or phone (Amharic vowels phone) will be checked against each trained HMM and the model with a largest probability will be selected.

3.4. Audio feature extraction
Audio-only ASR solutions, perhaps due to their relative maturity, have generally settled on the use of a single set of feature types, namely the Mel-frequency Cepstral Coefficients (MFCC). In this study, we use MFCC for audio feature extraction. Before applying feature extraction the audio file type change to wav file. To implement audio feature extraction we use python package called librosa.

3.5. Audio-Visual Fusion
The main difference between audio-only and audio-visual ASR lies in the design of the front-end, as two input streams (the audio stream and the video stream) are now available. Additionally, at some stage in the recognition process, the streams of information from the audio and visual modalities need to be fused.

During fusion the issue where the fusion of the data takes place should be addressed. Feature fusion integrates data on the feature level, where audio and visual features are used simultaneously and equally to identify the corresponding speech unit, thus, feature-level fusion algorithms train a single classifier on the concatenated vector of audio and visual features. Decision fusion, on the other hand, takes place after the independent identification of each stream and is thus an integration of identification results.

In this study, for audio-visual recognition we used a decision fusion architecture because different comparisons showed superior performance of the decision fusion compared to the other fusion architectures. By using this techniques, we combined the likelihoods of single-modality (audio- only and visual-only) HMM classifier decisions outputs to recognize audio-visual speech. Thus, this isolated word or phone speech recognition, we implemented by calculating the combined likelihood for the acoustic and the visual observation for a given word or phone model.

As shown in Figure 12 based on decision fusion architecture in this study, there are two recognizers working independently for the audio and the video/visual channel respectively. The combination is performed at the output of each recognition process. Therefore, for each word in the model two different probabilities will be provided one for each modality $P\ (W_j|\ \ O^A\ )$ and $P\ (W_j|\ O^V)$. This is the reason why it is also called decision fusion. The final solution will be the word that maximizes the combined probability $\underset{W_j}{arg\,max}\{\ P\ (W_j|\ O^{AV}\ )\}$.

In order to obtain the two probabilities $P\ (W_j|\ \ O^A\ )$ and $P\ (W_j|\ O^V)$, *an* acoustic model and a visual model must be found. Each of these independent models will be defined by a set of parameters that will be obtained in two independent training processes one for the audio and for the video.

The identification results in our case are the a posteriori probabilities of the observation vectors. Finally, the audio and visual features are combined and the resulting is used for training and testing. Figure 12 shows an overall diagram of our fusion system.

As shown in the Figure 12, we use weighted Bayesian fusion to combine the two complementary features (audio and visual), originating from audio and visual modalities, in order to maximize information gather and to overcome the impact of noise in each individual stream.

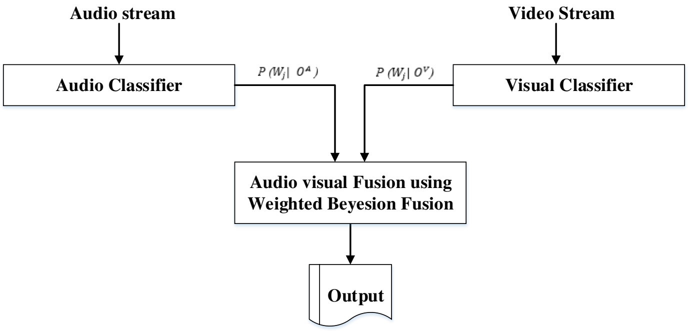


Figure 12: Block Diagram of the Multimodal (Audio-Visual) Fusion.

3.6. Audio-Visual Recognition

After deciding the methods of integration audio and visual, the last step of AVSR system is tanning of the integration features and audio-visual recognition. There are many audio-visual recognition models, such as product HMM, factorial HMM and coupled HMM. In this work, the coupled HMM is used which is the work of Ara *et al* [78]. The CHMM is a generalization of the HMM suitable for a large variety of multimedia applications that integrate two or more streams of data.

In this study, a two-stream CHMM used for our audio-visual speech recognition system. In our experiments, the bimodal speech recognition system employs a two-chain CHMM, with one chain being associated with the acoustic observations, the other with the visual features.

CHMM Training

In this work, the training of the CHMM parameters is performed in two stages. In the first stage, the CHMM parameters are estimated for isolated phoneme-viseme pairs. These parameters are determined first using the Viterbi-based initialization [78], followed by the expectation-maximization (EM) algorithm [42]. In the second stage, the parameters of the CHMMs, estimated individually in the first stage, are refined through the embedded training of all CHMMs. In a way similar to the embedded training for HMMs, each of the models obtained in the first stage are extended with one entry and one exit non- emitting states.

Recognition

The word and phone (vowel) recognition is carried out via the computation of the Viterbi algorithm [79] for the parameters of all the word and phone (vowel) models in the database. The parameters of the CHMM corresponding to each word and phone (vowel) in the database are obtained in the training stage. 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.

4. CONCLUSION

Audio is used as principal source of speech information in automatic speech recognition systems, but their performance degrades in presence of noise. Not only this, some phones are acoustically ambiguous. To compensate, a number of approaches have been adopted in the ASR literature, of which the use of the visually modality is probably the most suitable candidate being supported by both human speech perception studies and the work reported on AVSR systems.

The purpose of this study is to develop an automatic audio-visual speech recognition for Amharic language using the lip movement which include face and lip detection, region of interest (ROI), visual features extraction, visual speech recognition and integration of visual with audio. The architecture of the system that we adopted in our study is the decision fusion architecture. As a result of this architecture, we used three classifiers. The first one is the HMM classifier for audio only speech recognition, the second one is HMM classifier for visual only speech recognition and the third one was CHHM for audio-visual integration.

For implementation we use python programing language and OpenCV. For face and mouth detection we use Viola-Jones object recognizer called haarcascade face detection and haarcascade mouth detection respectively, after the mouth detection ROI extracted. Extracted ROI used as an input for visual feature extraction. Appearance-based (low-level) DWT is used for visual feature extraction and LDA is used for reduce visual feature vector. For audio feature extraction we use MFCC this is implemented by using a python package called librosa.

The system has been tested using the videos and audios which were captured for testing and training purpose. We used two main evaluation criteria for both phone (vowels) and word recognition, these are speakers dependent and speakers' independent. Based on the first evaluation criteria (speaker dependent) we found overall word recognition 60.42% on visual only, 65.31% on audio only and 70.1% for audio-visual. We also found overall vowels (phone) recognition 71.45% on visual only, 76.34% on audio only and 83.92 % on audio-visual speech based on speakers' dependent evaluation criteria.

Based on the second evaluation criteria called speaker independent we got overall word recognition 61% for visual only, 63.544% for audio only and 67.08 % for audio-visual. We also found the overall vowels (phone) recognition 68.04% for visual only, 71.96% for audio only and 76.79 % for audio-visual speech.

5. REFERENCES

[1] C. Vimala, V. Radha, "A Review of Speech Recognition Challenges and Approaches", *Journal of World of Computer Science and Information Technology (WCSIT),* Vol. 2, No. 1, pp. 1-7, 2012.

[2] Alin G and Leon J. M., "Visual speech recognition: automatic system for Lip reading of Dutch" *collaborative information system ICIS project supported by the Dutch Ministry of Economic Affairs*, BSIK03024, Mar 2009.

[3] Guoying Zhao, Mark Barnard and Matti Pietik¨ainen, "Lip-reading with Local Spatiotemporal Descriptors*", IEEE Journal of Transactions on Multimedia,* Vol. 11, No. 7, November 2009.

[4] Ayaz A. Shaikh, Dinesh K. Kumar, Wai C. Yau, and M. Z. Che Azemin, "Lip Reading using Optical Flow and Support Vector Machines", in *Proceedings of 3rd International Congress on Image and Signal Processing*, Yantai, China, 2010.

[5] Peter Roach, "English Phonetics and Phonology a practical course", Fourth edition, Cambridge university press, New York, 2009.

[6] http://clas.mq.edu.au/speech/phonetics/phonetics/consonants/place.html, Last accessed on 10/31/2017.

[7] P.Ladefoged, *A Course in Phonetics*, Harcourt Brace Jovanovich College Publishers, Third edition, 1993.

[8] Sharma R., Pavlovic V., Huang T. S., "Toward Multimodal Human-Computer Interface", in *Proceeding of IEEE*, Vol. 86, No. 5, pp. 853-869, 1998.

[9] Namrata Dave and Narendra M. Patel," Phoneme and Viseme based Approach for Lip Synchronization" ,*International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 7, No. 3, pp.385-394, 2014.

[10] Solomon Berhanu, "Isolated Amharic Consonant-Vowel (CV) Syllable Recognition. An experiment using the Hidden Markov Model", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, 2001.

[11] Kinfe Tadesse, "Sub-word based Amharic speech recognizer: An experiment using Hidden Markov Model (HMM)," MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, June 2002.

[12] Martha Yifru. "Automatic Amharic Speech Recognition System to Command and Control Computers", Masters Thesis, School of Information Studies for Africa, Addis Ababa, 2003.

[13] Molalegn Girmaw. "An Automatic Speech Recognition System for Amharic", Masters Thesis, Department of Signals, Sensors and Systems, Stockholm, Sweden: Royal Institute of Technology, 2004.

[14] Solomon Teferra, "Syllable-Based Speech Recognition for Amharic", in *Proceedings of the 5th Workshop on Important Unresolved Matters, Association for Computational Linguistics: Prague,* Czech Republic. pp. 33–40, 2007.

[15] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, "Audio-Visual Speech Recognition Using Deep Bottleneck Features and High-Performance Lipreading" in *Proceedings of APSIPA Annual Summit and Conference*, Asia-Pacific, December 2015.

[16] A. Davis, M. Rubinstein, N. Wadhwa, and William T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video*", Journal of ACM Transactions on Graphics (TOG)*, Vol. 33, No. 4, July 2014.

[17] Vibhanshu Gupta, and Sharmila Sengupta, "Automatic speech reading by oral motion tracking for user authentication system", *International Journal of Software Engineering Research & Practices*, Vol. 3, No.1, April, 2013.

[18] Ming-Hsuan Yang, David J. Kriegman and Narendra Ahuja, "Detecting Faces in Images: A Survey", *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, pp. 34-58, January 2002.

[19] Solomon Teferra, "Automatic Speech Recognition for Amharic", PhD Thesis, der Universität, Hamburg, 2006.

[20] Baye Yimam, *አጭርና ቀላል የአማርኛ ሰዋሰው (ačirna qälal yä'amrña säwasw/Short and Simple Amharic Grammar)*, Addis Ababa: Alpha, 2010.

[21] Getahun Amare, *ዘመናዊ የአማርኛ ሰዋሰው በቀላል አቀራረብ (Zemenäwi yä'amrña säwasw beqäläl akärärb/Modern Amharic Grammar Simple presentation*), Addis Ababa: Alpha , 2016.

[22] Simon Luce, "Audio-visual Speech Processing", Published PhD Thesis, School of Electrical & Electronic Systems Engineering, Queensland University of Technology, Brisbane, April 2002.

[23] Hussien Seid and B. Gambäck, "A Speaker Independent Continuous Speech Recognizer for Amharic", Published Masters Thesis, Computer Science & Information Technology, Arba Minch University, 2005.

[24] Timothy J. Hazen, "Visual model structures and synchrony constraints for audiovisual speech recognition", *Journal of IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 3, pp. 1082-1089, 1 May 2006.

[25] Stéphane Dupont and Juergen Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition", *IEEE Journal of Transactions on Multimedia*, Vol. 2, No. 3, September 2000.

[26] Paul Duchnowski, Martin Hunke, Dietrich Busching, Uwe Meier and Uwe Meier, "Toward movement-Invariant Automatic Lip-Reading and Speech Recognition", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, Michigan, USA, May 08-12, 1995.

[27] T. Chen and R. Rao, "Audio-visual integration in multimodal communication", in *Proceedings of the IEEE*, Vol. 86, No. 5, pp. 837–852, May 1998.

[28] G. Potamianos, H. P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lip-reading" in *Proceeding of International Conference on Image Processing*, Vol. 3, pp. 173-177, 1998.

[29] Alan Wee-Chung Liew and Shilin Wang, *Visual Speech Recognition: Lip Segmentation and Mapping*, Medical Information science reference, Hershey, New York, 2009.

[30] Jesús Fernando Guitarte Pérez , "Improvements in Speech Recognition for Embedded Devices by taking Advantage of Lip Reading Techniques", PhD Thesis, Departamento De Ingeniería Electrónica y Comunicaciones, Universidad De zaragoza, March 2006.

[31] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Journal of Transactions on Acoustics, Speech, and Signal Processing* ,Vol. 28, No. 4, pp 357-366, Aug 1980.

[32] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America,* Vol. 50, pp. 637–655, 1971.

[33] Farooq, O., and Datta, S., "Speech recognition with emphasis on wavelet based feature extraction", *IETE Journal of Research*, Vol. 48, No. 1, pp. 3-13, 2002.

[34] Junda Dong, "Designing a Visual Front-End in Audio-Visual Automatic Speech Recognition System", Published Master Thesis, Faculty of California Polytechnic State University, USA, June 2015.

[35] S. Gurbuz, Z. Tufekci , E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lip-reading for audio-visual speech recognition", Proc. ICASSP, pp. 177-180, 2001.

[36] S. Dupont, and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition", in *Proceeding of IEEE Transactions on Multimedia*, Vol. 2, No. 3, 2000.

[37] J.W. Picone, "Signal modelling techniques in speech recognition", in *Proceedings of the IEEE*, Vol. 81, No. 9, pp. 1215-1247, 1993.

[38] P.Viola, M. Jones, "Robust Real-Time Face Detection", *International Journal of Computer Vision* ,Vol. 57, No. 2, pp.137-154, 2004.

[39] J. Kamarainen and V.kyrki, "Invariance Properties of Gabor Filter-Based Features-Overview and Applications", Journ*al of IEEE Transaction on Image Processing*, Vol.15, No. 5,pp. 1088-1099, May 2006.

[40] Ivana Arsic and Jean-Philippe Thiran, "Mutual Information Eigenlips for Audio-Visual Speech Recognition", in *Proceedings of 14th European Signal Processing Conference (EUSIPCO),* Florence, Italy, September 4-8, 2006.

[41] A. B. Hassanat, "Visual Words for Automatic Lip-Reading", Published PhD Thesis, Department of Applied Computing University of Buckingham, United Kingdom, December 2009.

[42] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods", *Journal of Pattern Recognition*, Vol. 40, No. 3, pp. 1106 –1122, Mar. 2007.

[43] Syed Ali Khayam , *The Discrete Cosine Transform (DCT): Theory and Application*, Michigan State University, March 10, 2003.

[44] Zhanyu Ma and Leijon A. , "A Probabilistic Principal Component Analysis Based Hidden Markov Model for Audio-Visual Speech Recognition", in *Proceedings of 42nd Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, CA, USA, 26-29 Oct. 2008.

[45] L. Smith, "A tutorial on Principal Components Analysis", *Cornell University, USA,* February 26, 2002.

[46] Yu, H., and Yang, J., "A direct LDA algorithm for high-dimensional data with application to face recognition", *Journal of Pattern Recognition Society*, Vol. 34, pp. 2067-2070, 2001.

[47] J. Shlens. "A Tutorial on Principle Component Analysis", Systems Neurobiology Laboratory, University of California at San Diego, Version 2, December 2005.

[48] E. D. Petajan, "Automatic lip-reading to enhance speech recognition", in *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 44-47, 1985.

[49] Nasir Ahmad, "A Motion Based Approach for Audio-Visual Automatic Speech Recognition", Published PhD Thesis, Department of Electronic and Electrical Engineering Loughborough University, United Kingdom, May 2011.

[50] Ole Helvig Jensen, "Implementing the Viola-Jones Face Detection Algorithm", Published Master's thesis, Image Analysis and Computer Graphics, Department of Informatics and Mathematical Modeling, Technical University of Denmark, 2008.

[51] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition," *Journal of EURASIP* Applied Signal Processing, Vol. 2002, No. 11, pp. 1260–1273, 2002.

[52] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading", *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 24, pp.779-789, 2002.

[53] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision,* Vol. 1, No. 4, pp. 321–331, Jan. 1988.

[54] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Journal of* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23,No .6, pp 681 - 685, Jun 2001.

[55] E. Petajan , B. Bischoff ,and D. Bodoff, "Automatic lipreading to enhance speech recognition," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Washington, D.C., USA , May 15 - 19, 1988.

[56] C. Bregler and Y. Konig. "Eigenlips for robust speech recognition," in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94*, Adelaide, SA, Australia, 19-22 April 1994.

[57] G. Krone, B. Talle, A. Wichert, and G. Palm. "Neural architectures for sensor fusion in speech recognition," in *Proceeding of ESCA Workshop on Audio-Visual Speech Processing (AVSP'97),* Rhodes, Greece September 26-27, 1997.

[58] M. Gordan, C. Kotropoulos, and I. Pitas. "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP Journal on Applied Signal Processing*, Vol., 11, pp., 1248-1259, 2002.

[59] F.V. Jensen, *Introduction to Bayesian networks*. Springer -Verlag New York, Inc. Secaucus, NJ, USA, 1996.

[60] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Cambridge,* Entropic Ltd., 2002.

[61] J. Luettin, G. Potamianos, C. Neti, and A.S. AG, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01),* Salt Lake City, UT, USA, 7-11 May 2001.

[62] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop," in *Proceeding of IEEE Fourth Workshop on Multimedia Signal Processing*, Cannes, France , 3-5 Oct. 200.

[63] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, Vol.,2002, No.,1, pp., 1274-1288, January 2002.

[64] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, USA, 17-19 June 1997.

[65] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," in *Proceedings of the IEEE,* Vol., 91, No.,9, Sept. 2003.

[66] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Towards unrestricted lip reading," *International Journal of Pattern Recognition and Artificial Intelligence,* Vol., 14, No., 5, pp., 571-585, 2000.

[67] G. Potamianos, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, Ch 10, 2004.

[68] R. Goecke, "Audio-Video Automatic Speech Recognition: An Example of Improved Performance through Multimodal Sensor Input," in *Proceeding of NICTA-HCSNet Multimodal User Interaction Workshop*, Vol., 5, pp., 25-32 Sydney, Australia, 2005.

[69] Mustapha A. Makkook, "A Multimodal Sensor Fusion Architecture for Audio-Visual Speech Recognition", Published Master's Thesis, Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, 2007.

[70] Asratu Aemiro, " Pronunciation Modeling Based on Pattern Identification of Amharic Phonemes for Automatic Speech Recognition" Msc Thesis, Department of Computer Science Addis Ababa University, Ethiopia, 2015.

[71] Rodomagoulakis Isidoros, "Feature extraction optimization and stream weight estimation in Audio-visual speech recognition", A Thesis presented for the degree of Electronic and Computer Engineer, University of Crete, Chania, Greece, October, 2008.

[72] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno and Tetsuya Ogata, "Audio-visual speech recognition using deep learning*"*, Springer Science Business Media, New York, 2014.

[73] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proceedings International Conference on Image Processing 2002 (ICIP2002)*, Rochester, New York, USA ,2002.

[74] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119–139, Aug. 1997.

[75] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," *Microsoft Research*, Rep. MSR-TR-2010-66, 2010.

[76] H. Lee, Y. Kim, A. Rowberg, and E. Riskin, "Statistical Distributions of DCT Coefficients and their Application to an Inter frame Compression Algorithm for 3-D Medical Images," *IEEE Journal of Transactions of Medical Imaging*, Vol. 12, No.3 , pp. 478-485, 1993.

[77] Baum LE, Petrie T, Soules G, Weiss N. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The annals of mathematical statistics,* Vol.41 No. 1, pp. 164-171, Feb. 1, 1970.

[78] D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *Journal of American Scientist,* Vol. 86, No. 3, pp 236-244, 1998.