A Viseme Recognition System using Lip Curvature and Neural Networks to Detect Bangla Vowels

By

Nahid Akhter

A THESIS

Submitted in partial fulfillment of the requirements for the degree
MASTER OF SCIENCE

Department of Computer Science and
Engineering,
BRAC University
66
Mohakhali, Dhaka
-
1212
Bangladesh

Certificate of approval

The thesis titled "A Viseme Recognition System using Lip Curvature and Neural Networks to Detect Bangla Vowels" is completed under my supervision, meets acceptable presentation standard and can be submitted for partial fulfillment of the requirement of the degree MSc in CSE from the department of Computer Science & Engineering, BRAC University.

_____

Dr. Amitabha Chakrabarty
Assistant Professor
Department of Computer Science & Engineering
BRAC University.

Declaration

I hereby certify that this material, which I now submit for assessment on the programme MSc in CSE is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:
ID number: 14166001
Date:

ABSTRACT

Automatic Speech Recognition plays an important role in human-computer interaction, which can be applied in various vital applications like crime-fighting and helping the hearing-impaired. It consists of two domains – Audio Speech Recognition and Visual Speech Recognition. This thesis is based on Recognition of Speech in the visual domain only, i.e. it involves recognizing speech without the presence or support of any auditory signal. So far, a lot of research has been done on lip-reading in English and some amount on French and Chinese, as well as few other languages, but not much research has been done on lip-reading in Bengali. This thesis work provides a new approach to lip reading Bengali vowels using a combination of the curvature of the inner and outer lips and Neural Networks. The method uses a more robust and faster algorithm to detect the lip contour than conventional methods used so far, such as Active Contour Model, Active Appearance Model and Active Shape Models. The method used for feature extraction is also new. It makes use of coefficients of the curves of the inner and outer lips. This way, it makes use of a lesser number of parameters to represent the shape of the lip when pronouncing a vowel. Moreover, the method is also robust to alignment of lips at different angles and can work with low resolution pictures also. Finally, for recognition of the viseme, a Backpropagation Neural Network is trained and simulated using gradient descent method.

ACKNOWLEDGEMENTS

I would like to express my heartfelt indebtedness and gratitude to my respected supervisor Dr. Amitabha Chakrabarty and faculty members for guiding me with continuous encouragement, technical suggestions and valuable instructions throughout the thesis work.

I would like to offer my special gratitude Dr. Jia Uddin, Dr. Md. Khalilur Rahman and Dr. Md. Zahidur Rahman for their kind guidance and valuable suggestions and to the committee members for their patience and time.

Thanks to all those who were kind enough to offer their time in contributing to my dataset with their lip images, especially to Mrs. Halima Zaman, who helped me make it all happen.

And last but not the least, thanks to my family members and parents for their prayers and blessings and the highest gratitude to my husband, Md. Saiful Haque Khan for his constant support, encouragement, guidance and for going out of his way every time to help with any issue that came up. This was not possible without you. Special thanks to my son, Abdul Muyeed Khan for patiently putting up with the times he had to manage on his own to allow me to finish my research work. You are truly an understanding, kind and responsible son.

List of Publications

1. Akhter, N. and Chakrabarty, A. (2016) A Survey-based study on lip segmentation techniques for lip-reading Applications, International Conference on Advanced Information and Communications Technology (ICAICT), June 2016.

2. Akhter, N. and Chakrabarty, A., Viseme Recognition using lip curvature and Neural Networks to detect Bangla Vowels

CONTENTS

List of Figures

List of Tables

Chapter 1: Introduction

Human-computer interaction is a research area that has fascinated scientists and engineers for a very long time. Within this arena, automatic speech recognition is of special interest as it forms the basis for important human applications, like teaching people with hearing or speech impairment to speak and communicate effectively. Moreover, a visual speech recognition system can help intelligence agencies track a remote conversation by using a camera, where auditory input or support is not available. Visemes are used by the hearing- impaired to view sound visually, thus effectively lip reading the entire human face [1]. Some applications of lip reading include crime fighting potential for computerized lip-reading, speech recognition systems in cars and lip reading systems in computer as an alternative to keyboard.

So far, a lot of research has been done on lip-reading in English and some amount on French [4] and Chinese [2], as well as few other languages [3] [5], but not much research has been done on lip-reading in Bengali. Lip reading Bengali vowels especially is challenging because of its slight variations in certain similar sounding vowels like আ, ও and ঔ as well as ই and ঈ and ঈ and ঊ. This thesis provides a new approach to lip reading Bengali vowels using a combination of the curvature of the inner and outer lips and Neural Networks.

The thesis has been divided into several chapters. The first chapter is this introductory chapter, which will describe what is meant by Visual speech recognition (VSR) using machines, the purpose and scope of VSR. It will also discuss the challenges of VSRs. The Chapter will then go on to give an overview of the steps involved in lip-reading and then include a rundown of various techniques used in lip reading and their evolution. The chapter will end with objectives of the research and its specific contributions.
In chapter 2, we will deal with a literature review of algorithms and technology normally used in contour-finding, lip feature extraction and viseme recognition. A detailed discussion about the proposed algorithm and dataset will be described in Chapter 3. It will also include a comparison between traditional methods of contour finding and the proposed contour-finding algorithm and explain the method used by the proposed system for viseme recognition.

In Chapter 4 we will present observations and results of experiments done using the system and the thesis ends with Chapter 5 which will have the conclusion, challenges and future endeavours.

1.1 Visual Speech Recognition

Automatic Speech recognition is a part of the field of Artificial Intelligence and consists of two domains – Audio Speech Recognition and Visual Speech Recognition. Visual Speech Recognition (VSR) or lip-reading is a technique of understanding speech by visually interpreting the movements of lips, face and tongue using the information provided (if any) by the context, language, and any residual hearing. It is different from speech recognition because in speech recognition the speaker is audible, but in lip-reading only the motion of lips and other facial features like gestures etc is available. Although a lip-reading system does not necessitate the availability of knowledge about context, language or residual hearing, but any information about the above three features is definitely exploited by such a system. The basic unit of speech in visual domain is known as Viseme. A Viseme is a generic facial image that can be used to describe a particular sound. Using Visemes, the hearing impaired can view sound visually, thus effectively lip reading the entire human face [7].

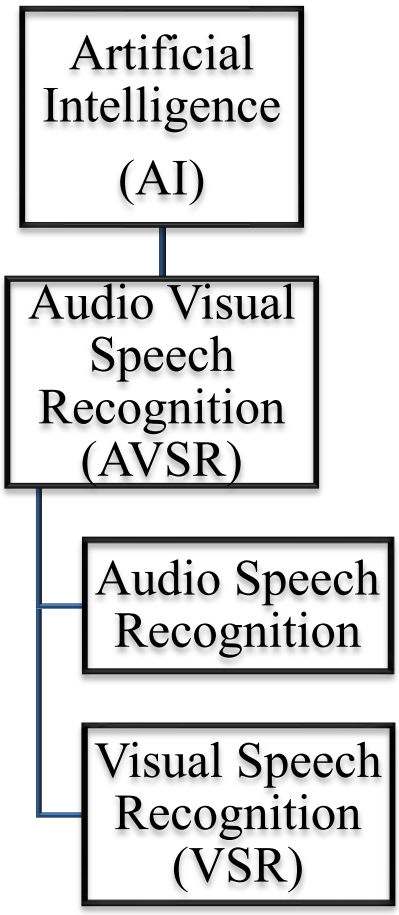


Figure 1.1: Hierarchy of Visual Speech Recognition (VSR)

1.1.1 Motivation

Having a hearing loss makes speech sound quieter, distorted or both. This is where lipreading is invaluable. For those who have a slight or moderate loss, lip-reading helps considerably in receiving the message. The following factors make lip-reading even more important:

a) **Lip-reading is not affected by acoustic environment and noise** – It is difficult to perceive a particular speaker in a noisy environment or when more than one speaker is speaking at the same time. Lip-reading is possible in such a scenario if a comprehensive view of speaker‟s facial features is available.

b) **Vocal sounds of two different words may be same** – Sometimes it is difficult to distinguish two different words only from their vocal sound because they sound same. For e.g. vocal sounds of „pa‟ and ‟ga‟ are same but they can be distinguished from each other as the motion of lips in each case is different. Similarly, voiced consonants /b/, /d/ and nasal consonants/m/, /n/ sound same even with different lip movement.

1.1.2 Challenges

Lip-reading is a complex art of observation, inference and inspired guesswork. Even the most skilled lip reader cannot accurately identify every word, because different sounds can be made with the lips in the same position: inferring likely words from the context is often necessary. Fast speech, poor pronunciation, bad lighting, faces turning away, hands over mouths, moustaches and beards, all these make lip reading more difficult or even impossible. Moreover, we form many English sounds in the middle of our mouth. Others come from the back of our mouth and even in our throat. There are numerous homophones in English. Words as different as "queen" and "white" look the same on a person's lips. Other challenges include:

a) Speaker dependence vs. independence: A speaker-dependent system is intended for use by a single speaker, while a speaker-independent system is intended for use by any speaker. The latter is more difficult to implement.

b) Read vs. Spontaneous Speech: When a person reads it's usually in a context that has been previously prepared, but when a person uses spontaneous speech, it is difficult to recognize the speech because of the disfluencies (like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter) and limited vocabulary[6].

c) Knowledge of the subject: Finally, In order for speech reading to be effective we have to know the subject being discussed.

1.1.3 Applications of Lip reading

Lip reading, in spite of the challenges it faces, has several applications:

a) Crime fighting potential for computerized lip-reading: Lip reading system is used in defense applications that require voice-less communication. Now a day‟s CCTV cameras are installed in areas of public gathering, markets and theaters. These cameras can capture videos to provide inputs to a lip reading system. This way we can determine what a group of suspicious persons is discussing and thus can help catch criminals or avert a terror attack.

b) Speech recognition systems in cars: In-car navigation systems use advanced speech recognition and text-to-speech capabilities that can identify spoken street and city names that exist across the entire continent. This allows drivers to speak all street addresses represented in the navigation system database and receive turn-by-turn voice guidance to their destinations. Installing a camera phone on the dash board of a car for in-car speech recognition systems can help driver to concentrate on driving and thus avert accidents.

c) Lip reading systems in computer - A lip reading system can be used as an alternative to keyboard to provide input to the computer and thus can be of significant help to physically disabled persons. This will minimize hardware requirement and improve speech based computer control in noise-filled environments.

d) Helping people with hearing impairment- Having a hearing loss makes speech sound quieter, distorted or both. For those who have a slight or moderate loss, lip-reading helps considerably in receiving the message. For individuals that are Deaf or Hard of Hearing, speech recognition software is used to automatically generate a closed-captioning of conversations such as discussions in conference rooms, classroom lectures, and/or religious services [6].

e) Biometric Person Identification: Automated lip-reading may contribute to biometric person identification, replacing password-based identification [6] [7].

1.2 Steps involved in Lip-Reading

The entire process of lip reading can be broken down into a number of steps, each of which are absolutely necessary for and contribute to the quality of the lip reading. These steps have been listed in chronological order in Figure 1.2.

a) Image Acquisition from video by breaking into frames

b) Face detection

c) Lip detection
d) Image pre-processing (image quality and resolution modification)
e) Lip segmentation / Feature Extraction
f) Pattern Recognition
g) Word detection

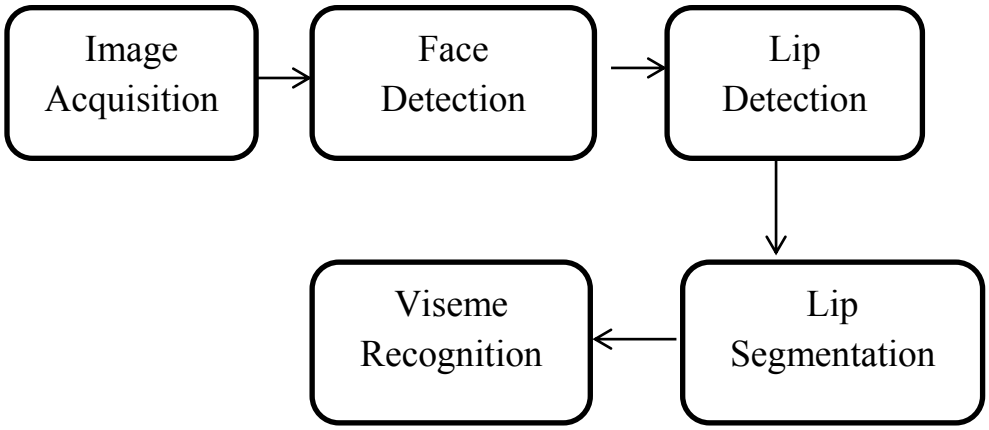


Figure 1.2: Sequence of steps in Lip-reading

1.2.1 Image Acquisition

A webcam or a camera is used to acquire the video of a person speaking in such a way that he utters each syllable as slow as possible, but continuously, without any sound. This utterance of words should be taken a few times, under varying brightness and backgrounds so that the best one can be chosen, such that it is easier to perform the subsequent steps. This video is then saved in avi or mpgeav format [1].

The acquired video is then broken down into frames or an image sequence, such that each video frame is now a separate image file. This is done using MATLAB‟s image processing toolbox.

1.2.2 Face Detection

The next step is to detect a human face in a given frame. For this Viola Jones algorithm is used in MATLAB [9].

The Viola-Jones face detection algorithm runs a detector or a window several times through the same image – each time with a new size. This technique is known as cascading. The detector in each cascading stage tries to detect whether the portion of the image in the detector window is a face object or not. In this way sub windows are applied in each cascading stage, discarding sub windows that do not have a face and

passing those that do to the next cascading stage. Final stage is the one which is considered to have a high percentage of face objects.

1.2.3 Lip Detection

After the face has been detected, it is time to locate the lips area in the face. This can be done by the Adaboost algorithm using haar features [8]. It is known that lip area is placed at the bottom half of face image. So, in order to reduce the computational complexity and improve the efficiency, the upper half of face image is removed.

The Adaboost algorithm of Freund and Schapire was the first practical boosting algorithm, and is still very popular in lip detection applications. Boosting is a kind of machine learning algorithm that can create a highly accurate prediction rule based on relatively weak and inaccurate rules.

Adaboost algorithm is used in conjunction with Haar-like features to detect lips. Haar-like features are equal rectangles used to calculate pixels in between adjacent regions, so that they can describe the connection between parts of an object [9]. While training, the adaboost algorithm updates weights of each classifier such that successive classifiers can concentrate on samples that the previous classifiers didn‟t recognize well. In the end, these weak classifiers will be combined to create a stronger classifier, thereby detecting the lip effectively.

1.2.4 Image Pre-processing

Image preprocessing is preparation of the image for subsequent processing. Image processing can consist of one or more of three basic operations – enhancement, restoration or compression [1]. Enhancement may consist of adjusting brightness and/or contrast of the image according to need. Restoration is the process of removing artifacts such as noise from the image. Compression is representation of the image by as fewer numbers as possible without distorting the quality of the image too much.

1.2.5 Lip Segmentation / Feature Extraction

Once the image of the lip has been processed according to requirement, lip segmentation is done. Lip segmentation is the method used to detect or trace the contour or edge of the lips so that we can get the shape of the lip at various instances or frames.

An efficient color space is needed to separate lip pixels from skin pixels. There are various color spaces such as RGB (Red-Green-Blue), HSI (Hue-Saturation-Intensity), and YCbCr (Luma, Blue chroma and Red chroma) [10].

After getting the contour of the lips, important features of the lip need to be extracted so that they can be used to represent the entire lip properly and efficiently. Some research papers use pixels of the whole lip image or the just the inner pixels of the mouth as inputs. Other papers advocate the use of certain points on the lip such as the centre of the upper lip, distances between corners of the lip and a few pairs opposite points on the edges of the lip. Whichever technique is used, the end product should be a set of numbers that can be efficiently used as input for a recognition algorithm.

1.2.6 Pattern Recognition

Various pattern recognition and machine learning tools are available to help recognize the viseme spoken. These include Artificial Neural Networks, Support Vector Machines (SVM), KNN classifiers, HyperColumn Model (HCM) and an online tool named WEKA. In our experiments, we have used the Artificial Neural Network tools that come with the MATLAB package.

Neural Network models have been utilized in various applications like Association, Clustering, Classification, Pattern Completion, Regression and Generalization, Forecasting, Optimization etc.[11][12] In image processing it has been used in classification, pattern recognition, character, symbol or object recognition as well as prediction. It mainly relies on using a system of neurons, somewhat similar to the nervous system of the human body to make learned decisions. The NN model uses Auto-Associative memory for training. The model reads the image in the form of a matrix, evaluates the weight matrix associated with the image. After training process is done, whenever the image is provided to the system the model recognizes it appropriately. The weight matrix evaluated here is used for image pattern matching.

Thus, given a set of values that represent important features of the mouth as input, the ANN will look through sample sets of similar features already learned before and find the set with the closest similarity to the input image. In this way, it will perform recognition of the spoken viseme.

Tools such as the HTK (Hidden Markov Model ToolKit) are available to recognize patterns of words. Once the extracted features from a video sequence are fed to it, it HMM algorithms learning algorithms like Viterbi or Baum-Welch to recognize the word uttered.

1.3 Evolution of Lip-reading techniques

1.3.1. Simple Image Thresholding

In 1984, Petajan [13] used simple image thresholding to extract binary mouth images, height, perimeter, area and width as visual features to produce their speech reading system.

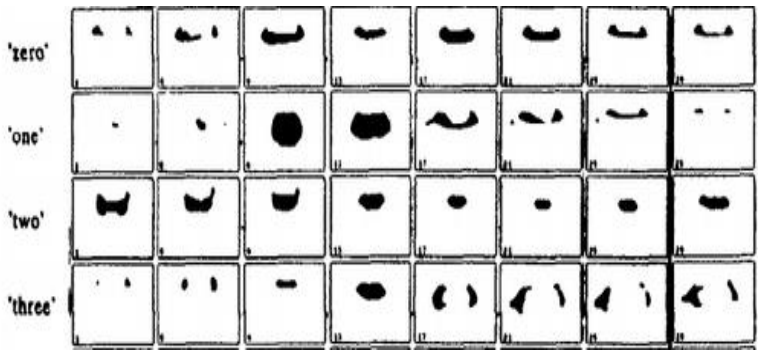


Figure 1.3: Petajan"s Simple Image Thresholding

1.3.2 Snakes

In 1988, Michael Kass et al [14] proposed the concept of a snake, which used an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges, as illustrated in Figure 1.4. These snakes lock onto nearby edges, localizing them accurately. Snakes consider different features for image energies like: color of pixels or sharpness of specified area, etc. and provide an account of visual problems like detection of edges, lines and subjective contours; motion tracking and stereo matching. They are guided by user-imposed constraint forces that navigate the snake near and around features of interest. This model is also called an ACM (Active Contour Model)

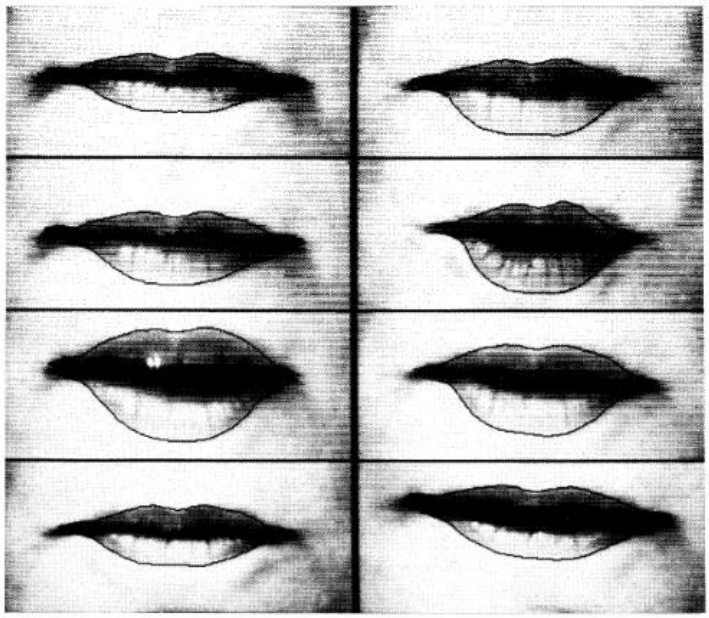


Figure 1.4: Working of Kass‟s Snakes algorithm

This diagram shows selected frames from a 2-second video sequence using snakes for motion tracking. After being initialized to the speaker‟s lips in the first frame, the snakes automatically track the lip movements.

1.3.3 Sampled Active Contour Model

To solve energy minimizing crisis snakes were found to require long computational time and large amount of calculations that made it unfeasible in stand-alone system to extract area function. So, Hashimoto et al [15] introduced variation of the Active Contour Model known as Sampled-ACM. This model assumes area extraction problems as force balancing problems of sample points on the closed curves, which are controlled by three local forces: attraction Fa , Pressure Fp , and repulsion Fr. By calculating the sum of these three forces on each contour point, sampled-ACM can extract the area more rapidly.

Another problem was when the snake is not contacting the object boundary it enters into the object region and then the repulsion force doesn‟t work. So, Sughara et al. [16] introduced a new force called vibration factor, to improve accuracy against noises in image and combined hardware circuits in FPGA (Field Programmable Gate Array) [17] with the S-ACM vibration factor. This helped improve the area extraction function in standalone systems, but with only one given image.

In 2006, Toshio [18] proposed the Sampled-ACM with splitting characteristics. The proposed Sampled ACM reduced the number of memory accesses required and increased processing speed. Moreover, the splitting characteristics allowed segmentation of more than one object in an image, as shown in Figure 1.5.

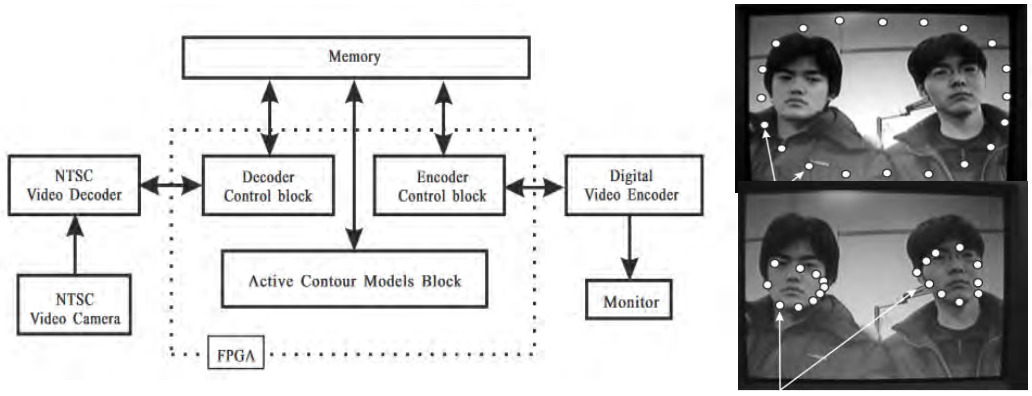


Figure 1.5: Toshio"s Sampled ACM with Splitting Characteristics

1.3.4 RASTA-PLP / Eigenlips

Later, in 1994, using the RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction) method of Bregler and Konig [19] , a system was built which was robust against any kind of distortions. The pattern recognizer is based on the first n principal components of a 24x16 gray-level matrix centered and scaled around the lips coded to get the outer boundary fairly with "Eigenlips" in regard to the similar approach of Turk and Pentaland"s "Eigenfaces" [20].

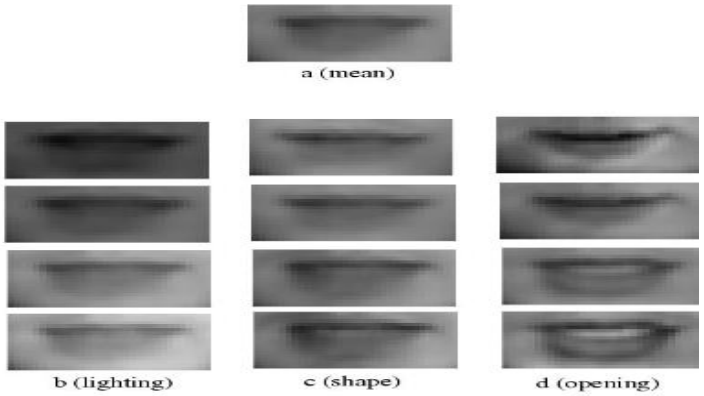


Figure 1.6: Bregler and Konig"s RASTA-PLP method: a) Mean vector, b) Variations along first principal axis, c) Variations along the second principal axis, d) Variations along the third principal axis

Then it leads on to find the mutual information for audio-visual lip-reading using an MLP (Multi-Layer Perceptron) Artificial Neural Network (ANN)

1.3.5 Neural Networks

Beale and Finaly [21] were the first to apply Neural Networks to the application of lip reading. Lippmann et al [22], in 1986 used a back propagation system to train a neural network. This went on to be known as a BP network. A BP network can be used to learn and store a great deal of mapping relations of an input-output model. Its learning rule is to adopt the steepest descent method in which the back propagation is used to regulate the weight value and threshold value of the network to achieve the minimum error sum of square. This led to the evolution of systems like NETtalk (Sejnowski and Rosenberg, 1987) for pronunciation of English sentences.

1.4 Objectives and Contributions

1. It has been observed that it is possible for trained people to recognize or understand speech by simply looking at the shape of the mouth while speaking. Thus, it should be possible to create an algorithm that uses the power of machine learning to read lips with some amount of effectiveness. Moreover, as human beings observe the shape of the mouth for reading lips, it was decided to use information on how much the mouth curves while speaking to be used in the machine learning process.
2. It was found that a lot of research has already been done on lip-reading English words and some amount on few other languages like Chinese, Arabic, Hindi and Tibetan. However, no information could be found by the author on lip reading in Bangla. Therefore, development of a system to lip-read Bangla visemes was desirable.
3. Most algorithms found on lip segmentation or contour extraction were iterative, which means using them in tracking videos would lead to a lot of time lag, so it was important to develop an algorithm with a good amount of accuracy that would read lips without the need of any iterative function to allow for quick lip reading.

Research on Lip-reading suggests that the field of Lip-reading is still in its infancy. A completely accurate and efficient lip-reading algorithm is yet to be developed. The objective of this research is to provide its contribution to this developing field with an algorithm that saves time and provides at least an above average accuracy. Contributions of this thesis are in three areas of lip reading. These are Lip Segmentation, Feature Extraction and Viseme Recognition.

Chapter 2: Literature Review

In this chapter, we will discuss some of the algorithms and technology normally used in contour-finding, lip feature extraction and viseme recognition, which are crucial parts of the lip-reading process.

2.1. Lip Segmentation

Lip segmentation or contour finding techniques may be broadly classified as image-based or model based. Image-Based techniques use the pixel information of the entire image directly. These can further be subdivided as Colour-based methods and Subspace-based methods. Colour-based methods work mainly by converting the image to various colour spaces like HSV, YCbCr or CIELAB spaces. Whereas, Subspace-based methods use a subspace of the information, like principal components, Wavelet-frequency, etc. in the image to represent the entire image. Model-based techniques are based on prior knowledge of the lip shape. Model-based methods include Active Contour Models (ACM), Active Shape Models (ASM) and Active Appearance Models. In this Chapter, both these models will be discussed in detail, along with few other techniques that could not be put under any one of these two categories. At the end of the chapter, an organized summary of all these techniques with their pros and cons will be presented.

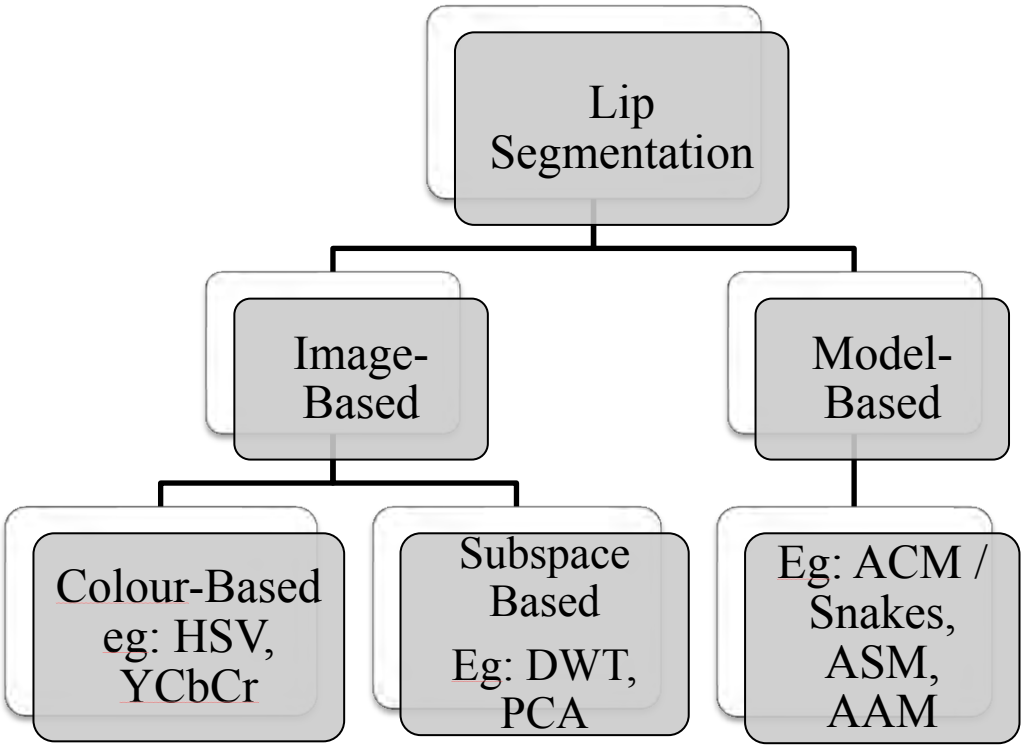


Figure 2.1: Hierarchy of Lip Segmentation/ Contour Extraction Techniques

2.1. 1. Image-Based Techniques

Image based techniques use the pixel information directly, the advantage is that they are computationally less expensive. However, they are restricted to illumination, mouth rotation and dimensionality [23]. Image-based techniques may be classified into Colour-based methods and Subspace-based methods.

2.1.1.1 Colour-Based Techniques

As discussed before, colour-based techniques work by converting the image to different colour spaces available, so that appropriate information from the image can be extracted efficiently. Some of the colour-spaces available include RGB, HSI, HSV, YCbCr, CIELAB and CIELUV. Images are typically in the RGB format.

Colour-based techniques base the detection of lips directly on the colour difference between the lip and skin. It was found in [24] that difference between red and green is greater for lips than skin and it was proposed to have a pseudo hue as a ratio of RGB values. [25] have also proposed a RGB value ratio based on the observation that blue color plays a subordinate role so suppressing it improves segmentation. In [26] it was proposed that Pixels of lip area have stronger red component and weaker blue component than other facial regions. Therefore, the chrominance component Cr in YC_bC_r colour space has greater value

than the C_b in the lip region. Badura in [27] suggested that Saturation component in HSV colour space provides a good base for lip segmentation.

Color clustering has also been suggested by some, based on the assumption that there are only two classes i.e. skin and lips. However, if facial hair or teeth are visible, then this does not hold true [28].

In some cases, the image is converted to a contrast-enhanced black and white image [29]. For better lip recognition, the histeq algorithm is used to enhance contrast so that the lip area appears much darker than the skin.

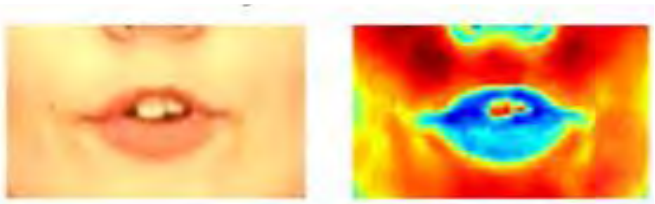


Figure 2.2: Colour Transform

In this way, and ROI (region of Interest) of the lip image from among the image of the entire face is extracted for further processing.

2.1.1.2 Subspace-based Techniques

Subspace-based techniques make use of a subspace of the original image to get a representation of the image in fewer dimensions. Some subspace-based methods include DCT (Discrete Cosine Transform) [31] [3] and DWT (Discrete Wavelet Transform) and Discrete Hartley Transform (DHT).

A Discrete Wavelet Transform (DWT) involves a transformation of an image by discretely sampling of pixels in the image by passing it through a series of filters. An advantage of this system is that it captures both frequency and location information. So it is a desirable method in lip-reading. In the end, it gives rise to features that are smaller in size.

The Discrete Cosine Transform (DCT) represents an image as a sum sinusoids of different magnitudes and frequencies [30]. An advantage of the DCT method is that it can be used to represent the most visually significant information in the image by just a few coefficients. T is for this reason that DCT is often used in image compression applications.

This way, the ROI (Region of Interest), i.e. the lip image can be represented in the form of a vector. However, even after these transformations, the dimensionality of this vector is generally too high to be used directly for statistic modeling. So, a dimensionality reduction usually needs to be performed subsequently, to get enough information for recognition purposes, while still retaining as much of the original speech information as possible.

In many lip-reading applications, DCT is found to be followed by a PCA transform to better compress the image. It is different from PCA in that it is the preferred method to differentiate frequencies while PCA is beneficial in selecting the most important representative components. PCA transforms data in such a way that the most of the variance in the data is contained to a small number of parameters called principal components.

In [32] a lip detector based on PCA was proposed. In this method, firstly outer lip contours were manually labelled on training data, PCA was then applied to extract the principal modes of contour shape variation, called Eigencontour, finally linear regression was applied for detection. LDA transforms data so as to maximize the discrimination between different classes. This has been illustrated in Figure 2.3.

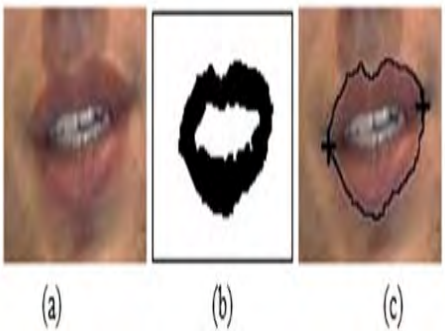


Figure 2.3: LDA applied on a PCA transformed space (a) The original estimate of the mouth region. (b) Segmented lip region (black) using LDA. (c) The lip contour and the corners of the mouth [32]

The Discrete Hartley transform, was introduced by R. N. Bracewell in 1983 [31]. Guyan applied Discrete Hartley Transform (DHT) to first enhance contrast between lip and skin, then applied a multi-scale

wavelet edge detection on the C3 component of DHT. It has been seen that around lip region C3 has maximum value.

Figure 2.4: Discrete Hartley Transform method [3]

2.1.2. Model-Based Techniques

It has been found that model based techniques are more widely used in most contour-finding applications currently. Model based techniques are based on prior knowledge of the lip shape. They learn the shape and appearance of lips from training data that has been manually annotated. Some model-based techniques are Active Contour Models (ACM), Active Shape Models (ASM) and Active Appearance Models (ASM) [33] [34]. In contrast to image-based methods, model-based methods can be quite robust. However, they tend to time-consuming and computationally expensive.

These techniques use a deformable template which dynamically changes form with each iteration to fit the outline or edge contour of the lip. ACM is basically an implementation Kass‟s Snakes algorithm [14], which has been explained in Chapter 1.

Active shape models (ASMs) are statistical models of the shape of objects which iteratively deform to fit to an example of the object in a new image. The method was developed by Tim Cootes and Chris Taylor in 1995 [35]. The shapes are constrained by the PDM (point distribution model),or a profile model which ensures that the model varies vary only according to a training set of hand-labeled, or „landmarked‟ examples. The shape of an object is represented by a set of points controlled by the shape model. The objective of the ASM algorithm is to match and deform the model to fit onto a new image.

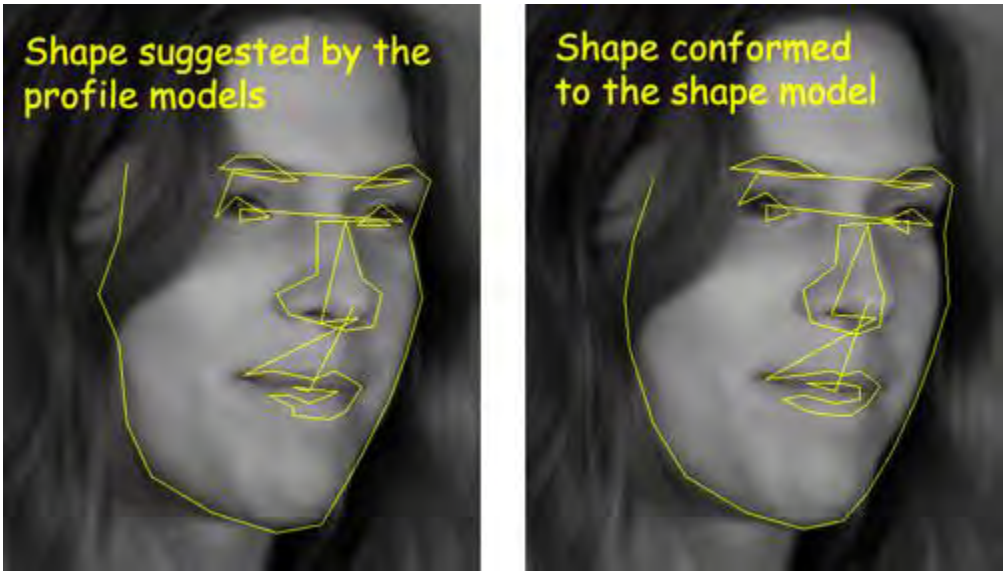


Figure 2.5: Operation of the shape model

The ASM works by an iteration of the following steps till a perfect match is achieved:

- Generate a shape by looking at each point to find a for a better position for the point in the new image. A profile model is used for this. After which, it looks for strong edges or uses the Mahalanobis distance to match a model template for the point.
- Conform the suggested shape to the point distribution model or "shape model" .

The ASM method is also known as a "Smart Snakes" method, since it is an analog to an active contour model which respects explicit shape constraints.

The AAM model was also introduced by Edwards, Cootes and Taylor, but in the year 1998. The method is widely used for matching and tracking faces and for medical image interpretation.

The algorithm uses the difference between the current estimate of appearance and the target image to drive an optimization process. It uses the least squares techniques to match to new images.

It is provides an improvement over ASM. The ASM only uses shape constraints (together with some information about the image structure near the landmarks), and does not take advantage of all the available information – the texture across the target object. However, AAM can be used to model both shape and texture of the image, thereby giving better accuracy. Table 2.1 summarizes the mail differences between ASM and AAM.

ASM	AAM
Considers only the shape attribute of the image	Considers both shape and illumination attributes
Objects with widely varying shapes, could not solved by ASM	Much more robust
Accuracy less	Accuracy more

Table 2.1: Active Shape Models vs. Active Appearance Models

Model-based techniques provide some advantages over image-based techniques by adding to their robustness. However, the prior process of manually fixing landmarks for the process makes using AAM and ASM a tedious and time-consuming task. This is followed by a training phase. Moreover, the contour-finding process itself takes place in iterations until a match is found, which again adds to the processing time, especially when large videos need to be processed. AAM also has the additional disadvantage of being too memory-intensive to run on an average smartphone or computer, as a good amount of RAM is needed to run it.

IMAGE-BASED	MODEL-BASED
Advantage: Computationally less expensive	**Advantage**: Can be quite robust
Disadvantage: Adversely affected by variations, such as illumination	**Disadvantage**: Require prior initialization and the iterations tend to be quite time-consuming and computationally expensive

Table 2.2: Image Based Techniques vs. Model-Based Techniques

2.1.3. Clustering Methods

Another method used for lip detection is Fuzzy clustering (FCM) [37]. This was applied in [36] by combining color information and spatial distance between pixels in an elliptical shape function. [32] have used expectation maximization algorithm for unsupervised clustering of chromatic features for lip detection in normalized RGB color space. As outlined by Leung et al [37], multiple clusters are adopted to model the background region sufficiently and a spatial penalty term is introduced to effectively differentiate the non-lip pixels that have similar color features as the lip pixels but located in different regions. Experimental results demonstrate that the proposed algorithm has good segmentation results over other segmentation techniques, However, this method is also based on iterations to find the correct mouth contour, which again adds to processing time.

2.1.4. Hybrid Methods

In addition, there are some hybrid techniques. These methods combine both image based and model based techniques. Majority of the hybrid techniques proposed in the literature use color based techniques for a quick and rough estimation of the candidate lip regions and then apply a model-based approach to extract accurate lip contours.

Usman Saeed and Jean-Luc Dugelay in [38] proposed a "fusion" of edge-based and region-based detection methods to carry out lip segmentation with comparatively better result than any of the two methods carried out individually. Here, given an image, it is assumed that a human face is present and already detected; the first step is to select the mouth Region of Interest (ROI) using the lower one third of the detected face. The next step involves the outer lip contour detection where the same mouth ROI is provided to the edge and region based methods. Finally the results from the two methods are fused to obtain the final outer lip contour.

2.2. Feature Extraction

Feature extraction is another important factor in lip reading. The choice of which features to extract is a very crucial one, because if the vector of features extracted is too large, the system becomes inefficient, whereas it is also important to ensure that the features extracted give as accurate an output as possible. A lot of lip reading algorithms use Principal Components as features for viseme recognition, while many others use geometrical features such as height and width of mouth, area, perimeter, etc.

The French ALiFe system in [4] used a feature called DA (Dark Area) which was area of the dark region inside the mouth. [39] had used a feature set consisting of a combination of a variety of descriptive features like Height and Width, Image Quality value, presence of tongue and Number of teeth pixels.

In 2006, Chen proposed in his paper [2] a feature extraction method for lips which used a variation of the red-exclusion method for lip detection, followed by a curve-fitting to get the contour of the inner lips only. This method was much faster than other contour-finding algorithms. The technique focuses on the green and blue colours of the lip image, rather than red, since both the face and lips are predominantly red. So any contrast would be better found in the red and blue ranges. So the red and green colour values are used as follows:

$$\log\left(\frac{G}{B}\right) \leq \beta$$

Where β is a threshold which is found manually. Using the logarithm further enhances contrast [40].

Table 2.3 gives a summary of all lip segmentation and extraction methods mentioned so far, along with pros and cons of each.

Method	Pros	Cons	
Petajan‟s simple thresholding	Simple	Low accuracy	Features extracted: height, width, perimeter & area of mouth
Image-based techniques	Computationally less expensive	Restricted to illumination, mouth rotation and dimensionality	
Kass‟s snakes	Robust if initialized properly	Long computational time, need for prior initialization, tends to converge to local minima	

Model-based techniques: ASM, AAM	Robust , good performance	Images need to be manually annotated and system needs to be trained. This becomes time-consuming and memory-intensive. Also, need for iterations adds to the processing time	
Leung‟s Fuzzy Clustering	Good segmentation results	Iterative, so time consuming	
AliFe system	Simple and fast	Average accuracy in recognizing vowels – 72%	Features extracted: Height, Width og mouth, Dark area inside mouth
Chen‟s curve-fiiting method	Simple and fast	Medium accuracy since only inner lip contours are extracted	Features extracted: Inner lip contour coefficients

Table 2.3: Summary of various available lip reading techniques

2.3. Viseme Recognition

Once the proper features have been extracted, the viseme classification needs to be done. At present, the most popular classifiers for speech recognition are Artificial Neural Networks (ANNs), PCA Classifier, KNN classifier and Support Vector Machines (SVM) and Hidden Markov Models (HMM).

2.3.1 ANN

Artificial neural networks (ANNs) are models that imitate the human brain activity. They consist of an interconnected set of units, called neurons, which take certain inputs, use the sum of products of these inputs with certain weights, and then, apply a function, called Activation Function to get outputs. These weights can either be adjusted by ANN itself (unsupervised ANN) or specified by the user (supervised ANN) during the training process..

These neurons can be organized into layers, depending upon the application. There are different types of ANNs available, like perceptron, Multilayer perceptron, FeedForward Network, Hypercolumn Model (HCM) and Self Organizing Maps (SOM). The most popular architecture is the feed-forward architecture with a single hidden layer [41].

ANNs have been used to detect the viseme being spoken based on advanced learning of previous patterns. Nowadays different researchers are combining different probabilistic, statistic and ANN techniques to provide appreciable and accurate error free automatic Lip-reading systems

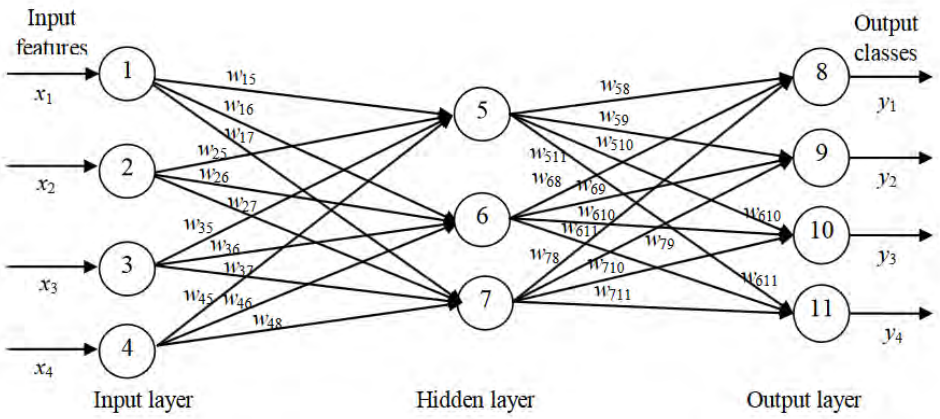


Figure 2.6 : A Feedforward Network Architecture with 1 hidden layer [41]

2.3.2. Hidden Markov Models

HMMs are statistical models which can be used for pattern recognition of sequential data [41]. Using HMMs, it is now possible to some extent, make predictions of words uttered by evaluating a series of visemes that are stochastically distributed. The first variable models the state transition probability between hidden states while the second models the probability of state output observation.

A model is normally produced for each of the speech units (phoneme/viseme) and these are concatenated to form an HMM for a word or a sequence of words. Normally an HTK (HMM TookKit) , which is easily available, is used in implementing HMM.

In [42], Yu, Jiang et al. proposed a sentence systematic approach to lip-reading whole sentences by using HMMs integrated with grammar. In this approach, a vocabulary of elementary words is considered. Based

on the vocabulary, they define a grammar that generates a set of legal sentences. Each word of the basic vocabulary is modeled by an HMM and the individual HMMs are concatenated according to the rules of grammar. Each HMM corresponding to one of the basic words consists of six states as shown in Fig 2.7. The HMMs are trained using forward-backward algorithm based on Baum-Welch formula.

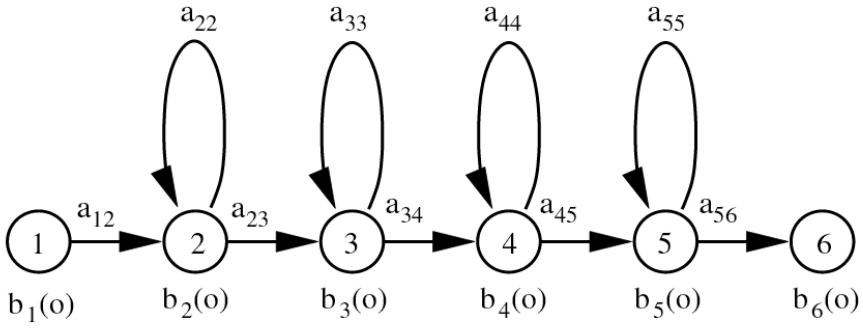


Fig 2.7: A six-state HMM. Here, O={o1, o2,…o6} is a visual observation of a word, represented by a sequence of feature vectors; A={aij} is an NxN matrix of state transition probabilities from state i to state j; and B is a set of observation probabilities bj(ot) for state j.

Simmons and Cox [46] developed an HMM based system that analyzed a small number of sentences to obtain several acoustic and visual training vectors. Then they created a fully connected 16 state discrete HMM, each state representing a particular vector quantized mouth shape, and producing 64 possible audio code-words. Subsequently, the trained HMM was employed in the Viterbi algorithm to generate the most likely visual state sequence, given the input audio observations.

At present, the most popular classifiers for speech recognition are artificial neural networks (ANNs) and hidden Markov models (HMMs) and their variants. Of the two approaches, HMMs are the more commonly used due to their simplicity of implementation, ease of training and computational efficiency [41].

ANNs are found to perform well when larger quantities of training data are available and in applications that require only a limited speech vocabulary. So, ANNs outperform HMMs on phoneme recognition and small vocabulary tasks, but when larger vocabularies are involved, the effective language modeling capabilities of HMMs are advantageous.

Chapter 3: Working Methodology

In this chapter, the proposed method will be discussed in detail, along with a presentation of the dataset used for the experiments

3.1 Dataset

Since this research primarily deals with viseme recognition of Bangla vowels, it was necessary to have a dataset consisting of images with Bangla vowels being spoken. Although there are a number of data sets for lip-reading available online, like the AVletters database, Tulips database and OuluVS database, all these datasets have images of English vowels being pronounced. Failing to find any existing dataset of Bengali visemes, it became necessary to create a Bangla viseme dataset of my own.

The dataset I used contained lip images of Bangla visemes being spoken. For simplicity, my dataset contained images of only three Bangla vowels, "আ", "অ" and "এ" being spoken. For each vowel, a total of 57 images were collected. That means, in all, the data set contained 57 x 3 = 171 images. These images were of different speakers, and they were taken under varying lighting conditions, during different times of the day and in different locations.

Following were the properties of the images used:

1. As can be seen from Table 3.1, the images were of varying sizes. This was done to show that the algorithm was independent of changes in image dimensions.
2. Mouth images were in varying angles. This was to show that the algorithm was robust against changes in mouth orientation with the camera tilted to different angles, as is common in amateur photography by smartphones.
3. Mouth images of different subjects were used to test the speaker-independence of the algorithm.
4. Images were taken under varying lighting conditions, for example, some were taken in natural light, while others were taken indoors, to test the robustness of the system to variations in illumination.

Of the 171 images, 120 of them were used for training, while the rest 51 were used as the test set.

	Viseme Spoken	Subject	Image Dimensions	Illumination
Image 1	অ	Speaker 1	297 x 313	Indoors
Image 2	আ	Speaker 2	344 x 317	Indoors
Image 3	আ	Speaker 3	465 x 417	Natural Light
Image 4	আ	Speaker 4	459 x 485	Natural Light
Image 5	আ	Speaker 5	312 x 280	Natural Light
Image 11	অ	Speaker 6	374 x 325	Natural Light
Image 12	অ	Speaker 7	456 x 354	Natural Light
Image 13	অ	Speaker 8	373 x 338	Indoors
Image 15	এ	Speaker 1	441 x 209	Indoors
Image 16	এ	Speaker 9	429 x 257	Natural Light
Image 17	এ	Speaker 2	545 x 232	Natural Light
Image 19	অ	Speaker 10	510 x 255	Natural Light

Table 3.1: Specifications of a portion of the dataset used

Figure 3.1: A sample of images in the dataset – 10 speakers uttering the Bangla visemes আ, অ and এ

3.2 Overview of Proposed Method

The method proposed by this thesis makes use of image-based techniques to roughly identify the lip area and then uses Chen‟s curve-fitting [2] to extract lip features to be used for describing the lip shape, which will aid in lip recognition. For better accuracy, both inner and outer lip curves have been found.

The algorithm is divided into three parts. The first part deals with extraction of the outer lip contour. Second part deals with extraction of the inner lip contour and third part deals with recognition of the uttered viseme. An overview of the process has been shown in flowchart form in Figure 3.2.

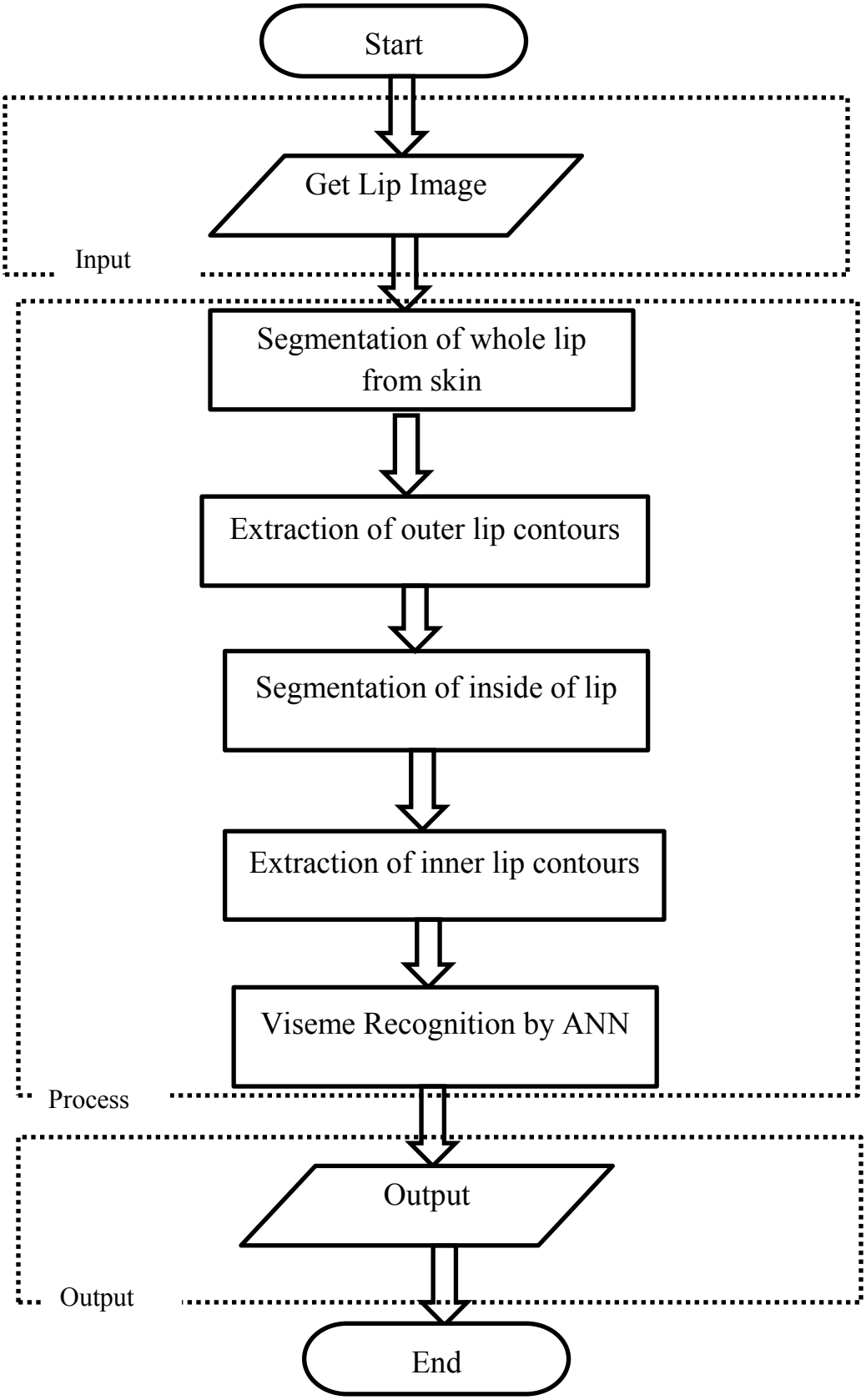


Figure 3.2: Overview Flowchart of proposed algorithm

3.3 Details of Proposed Algorithm

This section will explain the above flowchart step by step in detail.

3.3.1 Segmentation of whole lip from skin

1. In the first step, lip pixels are separated from skin pixels by using the red exclusion method.

    1.1 For this, the original image is first broken down into its RGB (Red, Blue and Green) components, and then the lip pixels are converted to the chromatic colourspace by computing the value of r for every pixel using the following equation:

$$r = \frac{R}{R\ G\ B}$$

    1.2 Computer the mean of r values over the whole image. Let this mean be x

    1.3 Compute the threshold Θ by the equation:

$$\Theta = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right)$$

    1.4 Threshold out all pixels that have r values less than Θ.

    1.5 A mask is thus obtained.

2. To further improve upon this mask, a second mask is found by converting the image to HSV space and masking out all the pixels that have saturation component less than the mean saturation.

3. Both masks obtained in steps 1 and 2 are combined by a logical AND operation and the final binary image obtained undergoes morphological cleanup by a dilation followed by filling up of any holes. Dilation is done using a disk structural element. This gives a preliminary outline of the outer lips. The lip is thus segmented from the image.

3.3.2 Extraction of outer lip contours

1. The mask of the segmented lip obtained is then reduced to an edge using canny edge detection [43].

2. The image obtained after edge detection is cropped to the edges and the left most and right most points of the mask are found. These two points correspond to the left and right corners of the lips.

3. Using a method described by [51] , the dipping point of the cupid‟s bow on the upper lip is found. This is denoted as „dp‟.Following is how it is found:

    3.1 If xt and xb be the two corners of the lips, the point xc that divides line xt–xb into equal segments is found.

3.2 Then, a boundary region (depicted by the shaded region in Fig. 3.3) that is within 20% distance to the left and right of point xc is found.

3.3 Within this 20% region, the lowest pixel is found. This pixel is the dipping point dp of the upper lip"s cupid"s bow, as shown in Fig3.3.

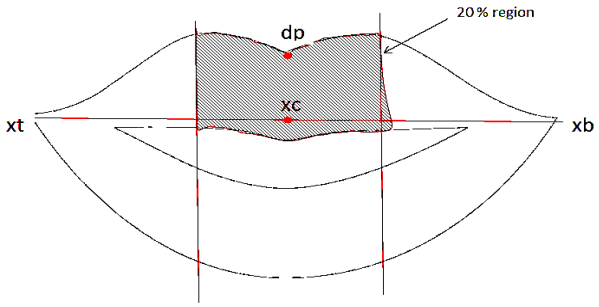


Fig 3.3 Selection of the dipping point of upper lip"s cupid"s bow

4. Now, sample points on the outline of the lip are found such that 8 points are found to the left of dp, 8 points to the right of dp and 16 points are found along the lower lip outline. That means, a total of 32 points are found along the outline of the outer lip. These 32 points give a rough estimate of the inner lip shape. However, the lip edge in reality is a smooth curve.

5. So, we interpolate these points using threes quadratic curves.

The upper right lip was interpolated by a quadratic curve:

$$a_1x^2+b_1x+c_1=0$$

The upper left lip was interpolated as:

$$a_2x^2+b_2x+c_2=0$$

and the lower lip was interpolated by:

$$a_3x^2+b_3x+c_3=0$$

Finally we use the vector of coefficients (a_1, b_1,c_1,a_2,b_2,c_2,a_3,b_3.c_3) to describe the outer-lip shape.
It has been found that quadratic polynomials are best in representing curves of the lip.

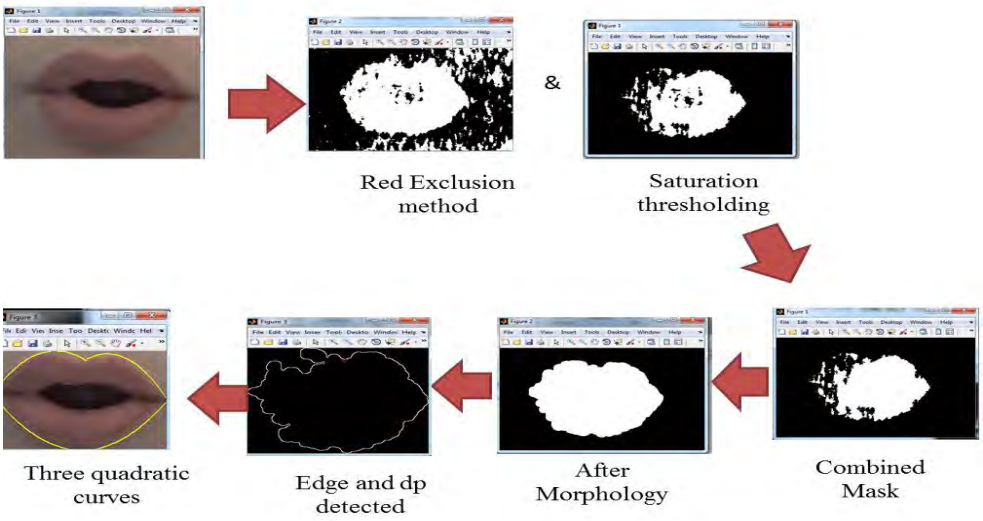


Fig 3.4: Extraction process of outer lip contours

3.3.3 Segmentation of inside of lip

1. The original image is first broken down into its RGB components. And the red component is thresholded using a threshold of 68. This masks out the inside of the lip.

2. If teeth are present, they need to be masked out too. So, as described in [24], the teeth are also masked through the following steps:

2.1 RGB image is converted to CIELAB colour space.

2.2 Mean of the A component is found, let it be Amean.

2.3 Standard Deviation of A component is found, let it be Astd

2.4 Find out Amask = Amean - Astd

2.5 RGB image is converted to CIELUV colour space.

2.6 Mean of the U component is found, let it be Umean.

2.7 Standard Deviation of U component is found, let it be Ustd

2.8 Find out Umask = Umean - Ustd

2.9 Both masks obtained in steps 2.4 and 2.8 are combined by a logical OR operation and the final binary image obtained undergoes morphological cleanup by removing any holes

3. Final mask is obtained by combining the teeth mask with the inside of the lip mask, i.e. masks obtained at the end steps 1 and 2. A final cleanup of this mask is done by morphological operations again.

3.3.4 Extraction of Inner Lip Contours

1. The binary image obtained in the last section is cropped to the edges of the mask

2. The left most and right most points of the mask are found to correspond to the left and right corners of the lips.

3. A total of 32 points are found along the edge of this inner lip mask, where 8 points each belong to the left side of the upper mouth, right side of the upper mouth, left side of the lower mouth and right side of the lower mouth.

4. These 32 points give a rough estimate of the inner lip shape. But we want a smooth contour. So, as in Step 5 of section 3.3.2 , we interpolate these points using three quadratic curves again:

$$a_4x^2+b_4x+c_4=0$$

$$a_5x^2+b_5x+c_5=0$$

$$a_6x^2+b_6x+c_6=0$$

So to describe inner lip shape, 9 more coefficients have been found, i.e. (a_4, b_4,c_4,a_5,b_5,c_5,a_6,b_6.c_6).

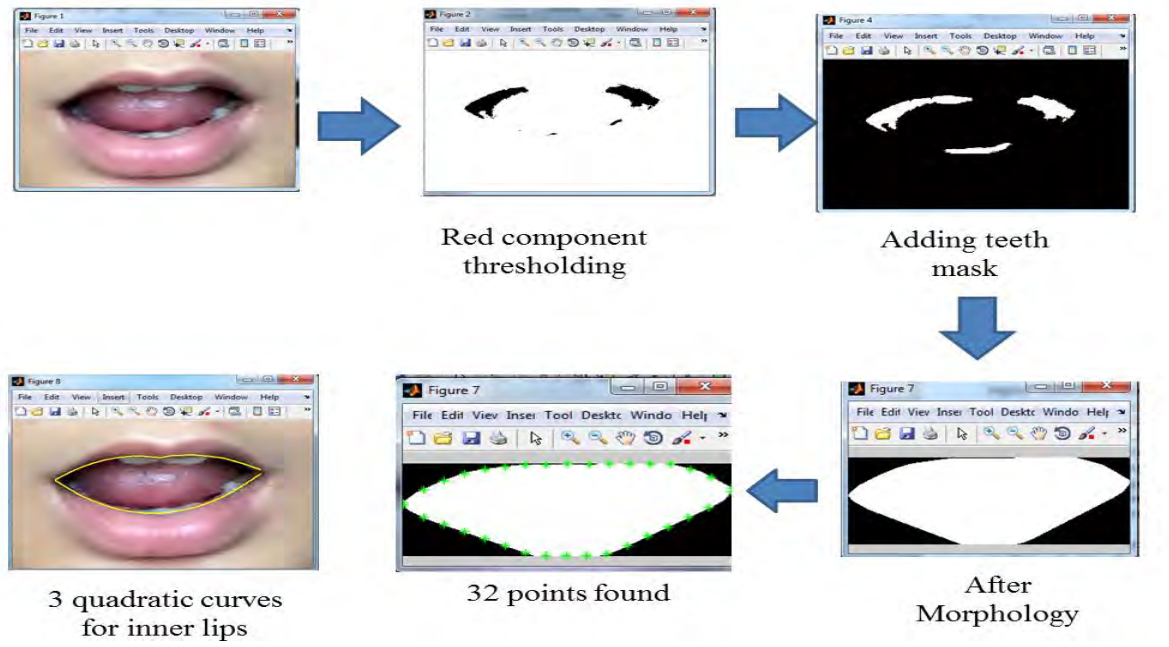


Figure 3.5: Extraction process of inner lip contours

This means each lip image can now be represented by a vector, or feature set of 18 coefficients – 9 representing outer lip and 9 representing inner lip contour. Figure 3.6 shows a lip image with the 6 contours found.

Figure 3.6: Screenshot of lip with all 6 contours showing

3.3.5 Viseme Recognition

For pattern recognition, various machine learning tools are available, like Support Vector machines, KNN classifier and WEKA. For our experiments, we have used the Artificial Neural Networks tool available with the MATLAB package.

The extracted vectors of 18 coefficients (a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3, a_4, b_4, c_4, a_5, b_5, c_5, a_6, b_6, c_6) thus found in the previous section is finally normalized to values between 0 and 1 by the following formula:

$$n(i) = \frac{c(i) - \min(i)}{\max(i) - \min(i)}$$

where the value of n(i) ranges from 0 to 1, and min(i) and max(i) denote the minimum and maximum value of the i-th vector, respectively.
The normalized vector for each lip image is then used as an input vector to the input layer of the single neural network system.

A Feedforward Neural Network was used with one hidden layer. The ANN had 18 input nodes, 10 hidden layer nodes and 3 output layer nodes (to identify 3 bangla vowels আ, অ and এ). The Neural Network was trained and simulated using gradient descent method to minimize the error between the output values and the target values. For both hidden and output layers, Tan-sigmoid („tansig") activation functions were used. The targets were set as follows:

আ: <1 0 0 >, অ: <0 1 0>, এ: <0 0 1>

The training parameters used were: Epochs: 5000, goal: 10e-5

There is no rule to determine the optimal number of neurons to be added in the Hidden Layer. However, through many experiments, it was found that the performance of the neural was optimal at 15 units. Adding any more neurons made no significant difference to the Mean Square Error.

The following sigmoid function was used as the unit function for neurons in the hidden and output layers:

$$\text{Tansig(n)} = \frac{2}{1+e^{-2n}} - 1$$

A diagram of the constructed network structure has been given in Fig 3.7.

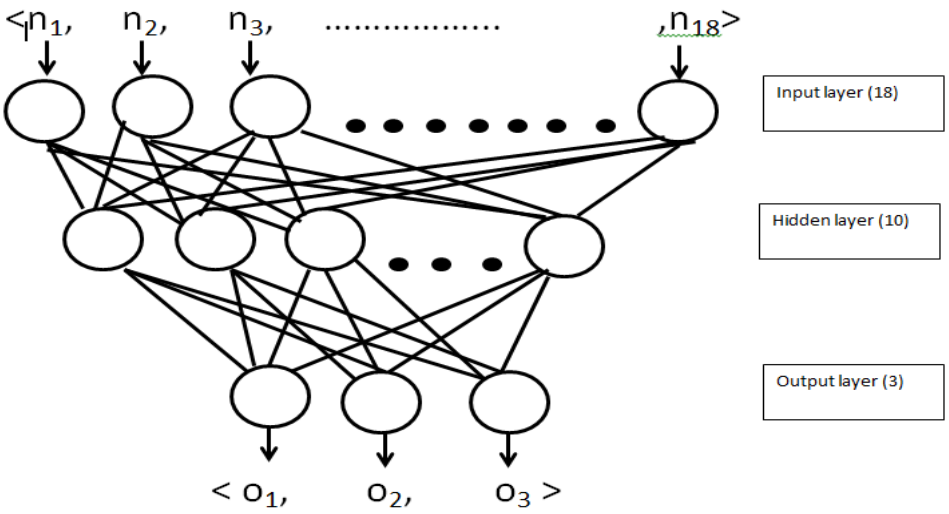


Figure 3.7. Architecture of the three layer FeedForward ANN used for the proposed method

The training set contained 5 images of each viseme. That means a total of 18 images were used in training of the Neural Network. The test set consisted of a total of 45 images, among which 27 were new images, which were not part of the training set.

3.4 Advantages over other methods

The proposed method improves upon some of the drawbacks of the existing methods of contour extraction

It adds robustness and accuracy to image-based algorithms. Since it extracts the curvature of the lips, the results are independent of the size or quality of the picture, illumination or mouth rotation. However, the images have to be front-facing.

As for model-based methods like ACM, ASM and AAM, the proposed method does away with the need to initially add manual landmarks to the image as well as the need to train the contour-extractor. This saves a lot of processing time, memory resources and the possibility of wrong initialization by the user. This makes the proposed method ideal to be used on low performance machines and simple smartphones.

The method does not require high quality or high resolution images. It does well with images taken on simple smartphone cameras or even images with small amount of noise.

In the same way, it does away with the need to iterate as in fuzzy clustering systems and saves time. The proposed method is based largely on Chen‟s method [2]. However, Chen‟s method extracts the features of the inner mouth only, whereas this method uses information of both inner and outer mouth to give a better representation of the mouth. Moreover, the preliminary steps of lip localization in this system used Chen‟s version of the red-exclusion method in combination with a conversion to HSV space to ensure a better extraction. Most importantly, the paper only describes how to extract features of the inner lip, but not how to recognize the viseme. A flowchart of Chen‟s method, as described in his paper has been given in Figure 3.8.

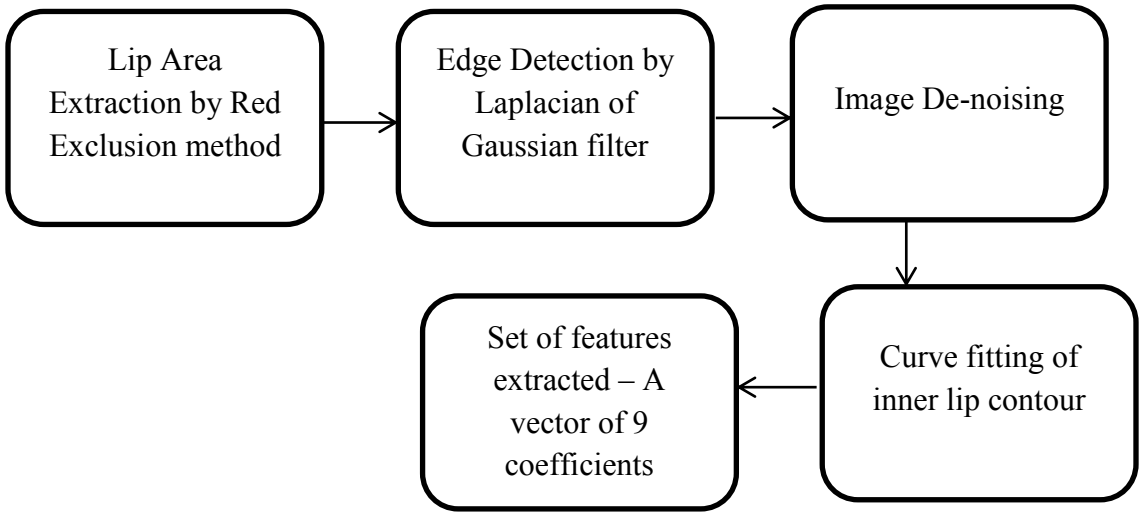


Figure 3.8: An overview of Chen‟s algorithm in [2]

Chapter 4: Results and Observations

In this chapter, the different experiments conducted along with the results obtained will be discussed. The proposed method was mainly developed as an improvement of Chen‟s algorithm in [2], and its adaptation to viseme recognition. So a major portion of this chapter will be dealing with comparisons between the two methods.

Our experiments were divided into three parts – A lip segmentation part, a contour extraction part and viseme recognition part.

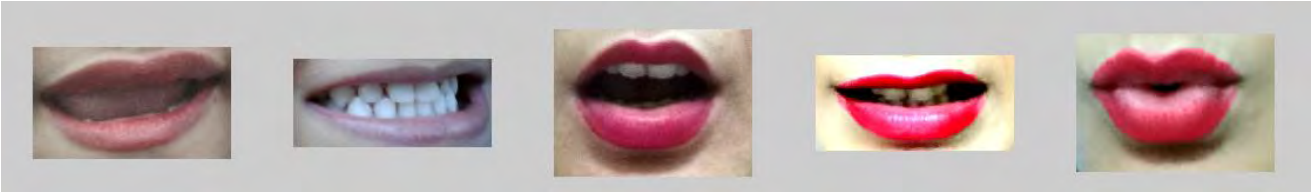
To demonstrate how conversions to different colour spaces affect the success of lip segmentation, an experiment was conducted by converting few images to different commonly used colour spaces for lip reading.

Experiments with different contour extraction algorithms were also performed, like ACM, ASM, AAM and curve fitting.

Finally, for viseme recognition, Artificial Neural Networks were used. To evaluate our results, two ANNs were designed. One with 9 input nodes, to test the accuracy of Chen‟s algorithm, and another with 18 inputs to test the accuracy of the proposed method.

4.1 Lip Segmentation

To demonstrate how conversions to different colour spaces affect the success of lip segmentation, an experiment was conducted by converting few images to different commonly used colour spaces for lip reading, like RGB, HSV and YC_bC_r. The examples of the ROIs obtained, represented in these spaces, are shown in Figure 4.1.



(a) Original Images

(b) R channel

(c) G channel

(d) B channel

(e) Y Channel

(f) C_b channel

(g) C_r channel

(h) H channel

(i) S channel

(j) V channel

Figure 4.1: ROI representations in different colour spaces

From Figure 4.1, we can clearly see that the Cr channel of YC_bC_r space and the S channel of the HSV space are most suitable for lip segmentation. However, on further experimentation, it was found that a combination of Saturation component and the Red Exclusion method gave the best results. Figure 4.2 shows a few examples.

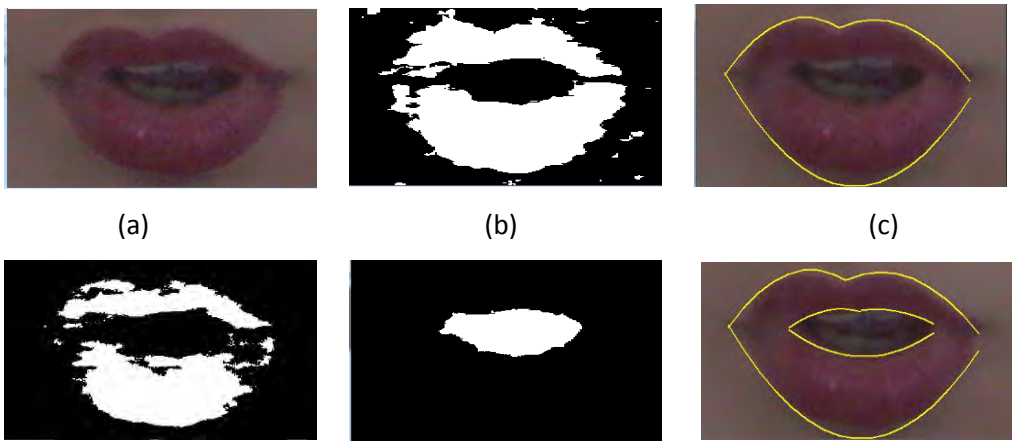


Figure 4.2: (a) Original Image (b) Mask obtained after YC_bC_r segmentation (c) Outer contour obtained by YC_bC_r segmentation (d) Mask of whole lip obtained after proposed method segmentation (e) Mask of inner lip after proposed method segmentation (f) Inner and outer contours obtained after proposed method

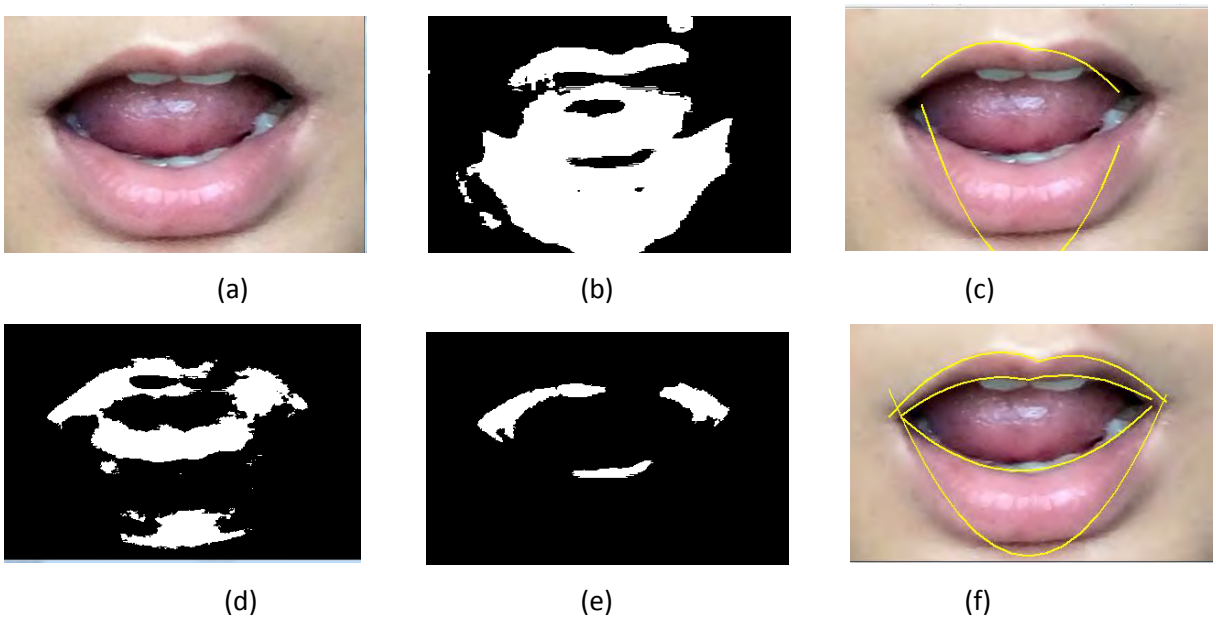


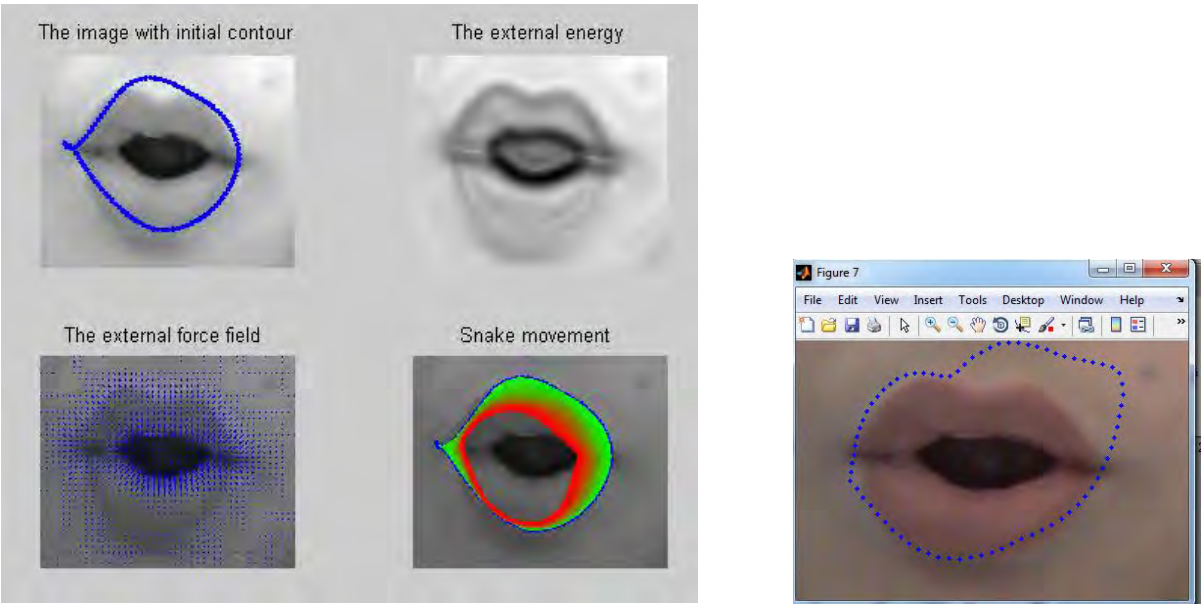
Figure 4.3: (a) Original Image (b) Mask obtained after YC_bC_r segmentation (c) Outer contour obtained by YC_bC_r segmentation (d) Mask of whole lip obtained after proposed method segmentation (e) Mask of inner lip after proposed method segmentation (f) Inner and outer contours obtained after proposed method

Figure 4.2 and 4.3 show two cases where the YC_bC_r segmentation was compared to the proposed method"s segmentation. Figure 4.2 shows a case where both methods produced the same result, whereas 4.3 shows a case where the proposed method performed better in lip segmentation and contour finding

than YC_bC_r segmentation. Similarly, it has been found that in many cases both methods provide same result, but in some cases the proposed method provides a better result.

4.2 Contour Extraction

Experiments with different contour extraction algorithms were also performed, like ACM, ASM, AAM and curve fitting. In the end, curve fitting was found to be ideal, as it was simpler and computationally less expensive. Moreover, there was no need for prior initialization by the user with landmarks, as was necessary for the former 3 algorithms.



(a) (b)

Figure 4.4: Example of contour-finding using (a) ACM method and (b) ASM method

As can be observed in Figure 4.4, the initialized contour for the ACM method could not properly find the right side contour of the lip image, as the snake failed to find the edge and entered into the object region. In Figure (b), the shape model failed to converge to the proper lip contour. These kinds of problems were observed for quite many images in the dataset.

However, the proposed algorithm was able to detect proper contours for almost 90% of the dataset images. Some examples of correctly detected contours have been shown in Figure 4.5. However, for very few cases, contours could not be correctly found. One example has been shown in Figure 4.6. The reason for wrongly detected contours is mostly due to certain darkish patches on lips or uneven tone of lip colour.

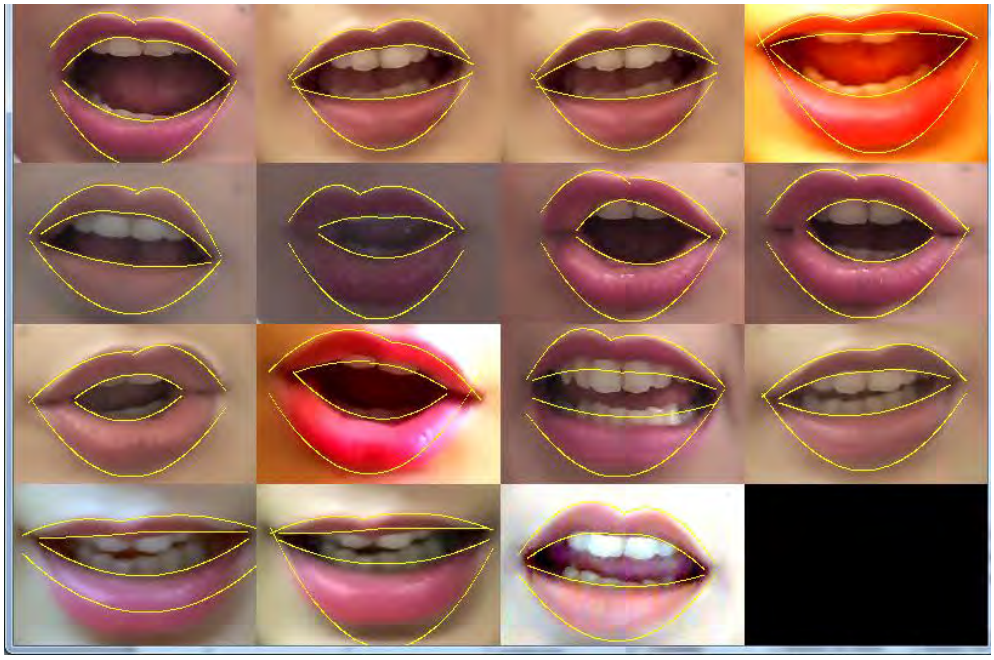


Fig 4.5: A montage of some lip images and contours found

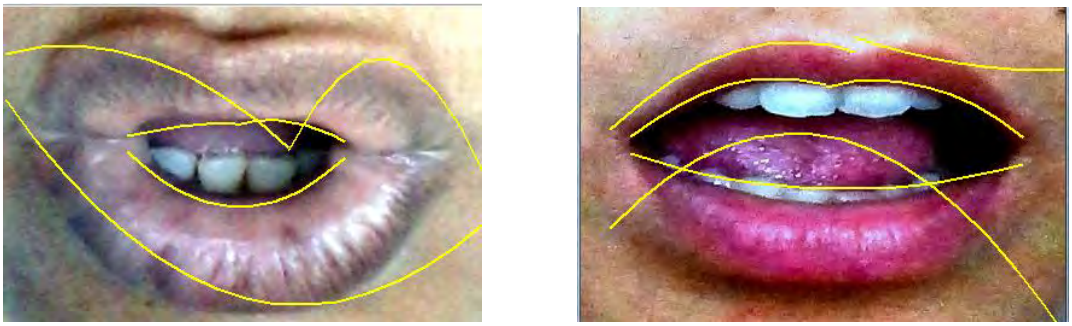


Figure 4.6: Few cases where contours could not be properly found.

4.3 Feature Extraction and Viseme Recognition

4.3.1 Design Issues

For viseme recognition, two ANNs were constructed, one to see the result of using only inner mouth contours (a 9 feature vector) , and another to see the result of using both inner and outer mouth contour (an 18 feature vector). ANN1 was used for the former and ANN2 was used for the latter.

So ANN1 was constructed with 9 input nodes, 15 hidden layer nodes and 3 ouput nodes, while ANN2 was constructed with 18input nodes, 15 hidden layer nodes and 3 output nodes, as illustrated in Figure 3.7

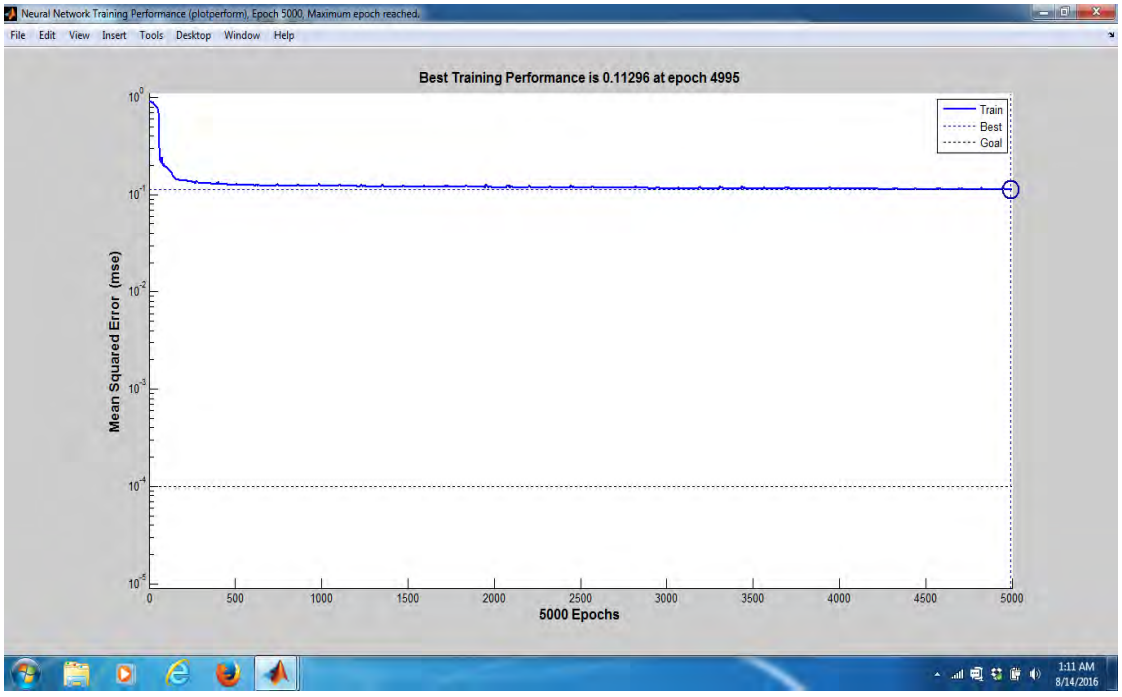
in Chapter 3. All other parameters were kept the same. Table 4.1 summarizes the parameters of two neural networks.

Neural Network Parameters	ANN 1	ANN 2
Size of input layer	9	18
Size of hidden Layer	15	15
Size of output Layer	3	3
Epochs	5000	5000
Goal	10e-5	10e-5
Function used	Tansig	Tansig

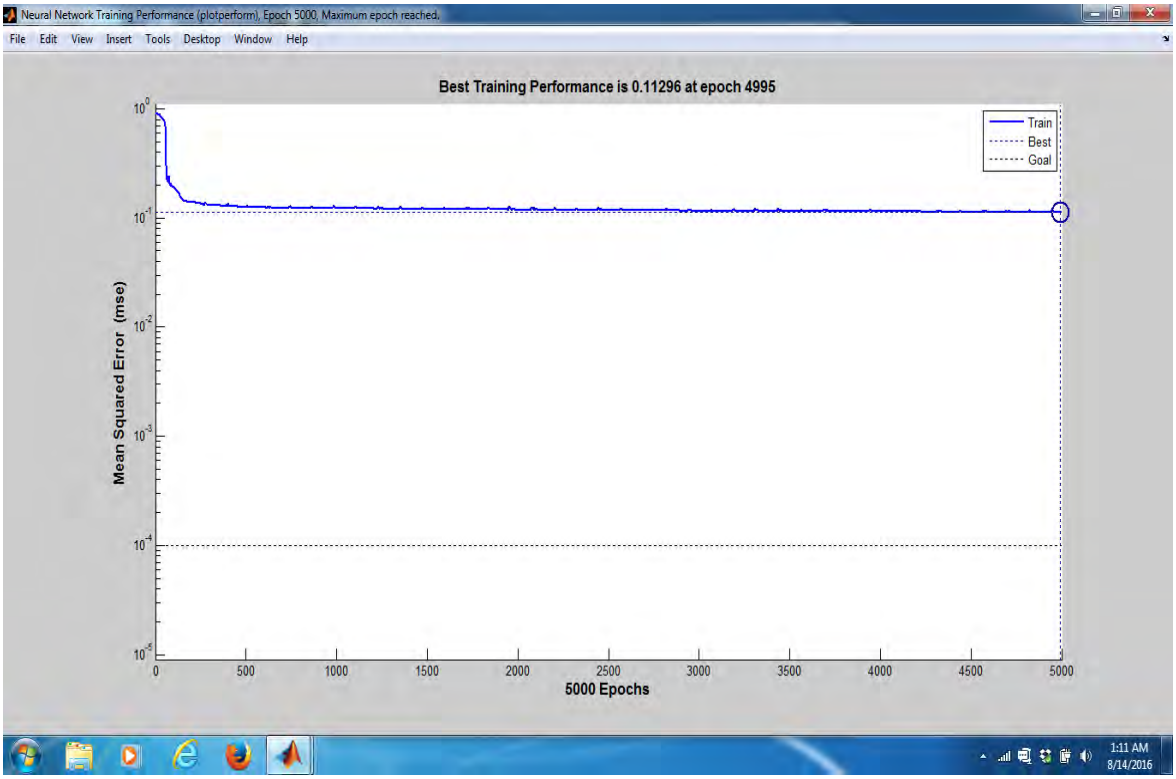
Table 4.1: Parameter settings for the Neural Networks

4.3.2 Training

It was found that the error value during training for ANN1 was 0.112 at the end of 5000 iterations, whereas for ANN 2, the error was 0.072. This shows how using both inner and outer lip coefficients gives better results than only using inner lip coefficients.



(a)



(b)

Figure 4.7: Training Performance graphs of (a) ANN 1 (b) ANN 2

5.3.3 Simulation and Testing

A comparison was drawn between the accuracies of the two methods over the training and validation data sets. Results of this experiment have been given in Table 4.2. Time complexity for Segmentation, feature extraction and training for 120 images in first case was 123s, while that for second case was 165s. Therefore, not much time difference could be found in the two methods.

Viseme	Ratio of true positives using ANN 1	Ratio of true positives using ANN 2
আ	25 / 42	36/42
অ	19/42	37/42
এ	9/42	37/42
Accuracy	53/126 = 0.421	110/126 = 0.873

Table 4.2: Results obtained during Testing phase

From Table 5.2, we find that the result from Chen"s method after Neural Network simulation was 42.1% while that achieved from the proposed method was 87.3% .

Chapter 6: Conclusion & Future Work

This thesis proposed a neural network as a multi-class pattern classifier to identify visemes of Bangla vowels being spoken. The success of a viseme classification system depends upon many factors, most importantly, the choice of lip-localization and contour-finding algorithm, the choice of features extracted and the pattern recognition system used.

A combination of image-based techniques which involved conversions to different colour spaces like HSV, CIELAB and CIELUV were used to localize the lip‟s inner and outer curves. Features extracted were the curvature of the inner and outer lips. This proved much faster and memory-efficient than using Active Shape Models and Active Appearance Models.

Qing Cai Chen et al. in 2006 used coefficients of quadratic polynomials to describe the contour of the inside of the lips while pronouncing Chinese vowels. The paper dealt only with the extraction of inner lip contour, but not with recognition of the viseme spoken. The method described in this thesis extracts contour of not only the inner lip, but also the outer lip and goes on to attempt recognition of the viseme thus spoken. A total of 18 parameters were used to describe the curvature of a single lip viseme image. These parameters were the coefficients of quadratic curves that traced the inside and outside of the lip contour. Finally, for pattern recognition, an Artificial Neural Network with 18 input units, a hidden layer of 10 units and output layer of 3 units was used. The Feedforward ANN was based on tansig function for minimizing the mean square error.

Using this Neural Network, a comparison was drawn between using curvature of inner lip only as pointed out in Chen‟s paper and using both inner and outer lip curvatures as introduced in this paper. During training phase, at the end of 5000 epochs, the mean square error for the former was found to be 0.112, while that of the letter was found to be 0.072. Moreover, when tested over images of 3 different Bangla visemes, the former method showed an accuracy of 42.1%, while the latter showed an accuracy of 87.3%.

It is important to remember that lip reading comes with many challenges. Firstly, different sounds can be made with the lips in the same position. Secondly, some sounds are made in the middle of our mouth, others come from the back of our mouth and even in our throat. These latter are impossible to speech read so far. Moreover, there are numerous homophones in Bangla. Words as different as "Pori" and "Bhori" look the same on a person's lips. This accounts to very less accuracy in such a system. Other challenges include fast speech, poor pronounciation , dialects, bad lighting and smiling while speaking.

The proposed system can be easily implemented in embedded systems such as Android or iOS to use with smartphones and tablets so that it can be used as and when needed.

In my future endeavour, I would like to implement this system in identifying whole Bengali words or numbers spoken by tracking lip movements on video, so that it will have some substantial use.

References

[1] Mei, L.P. (2014) Interpretation Of Alphabets By Images Of Lips Movement For Native Language. Universiti of Teknologi, Malaysia, 2014.

[2] Chen, Q.C. et al. (2006) An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech. International Conference on Machine Learning and Cybernetics, August 2006.

[3] Mishra, A.N. and Chandra, M. (2013) Hindi Phoneme-Viseme Recognition from Continuous Speech. International Journal of Signal and Imaging Systems Engineering, Volume 6 , No. 3, pp. 164-171, January, 2013.

[4] Werda, S., Mahdi, W. and Hamadou, A.B. (2007) Lip Localization and Viseme Classification for Visual Speech Recognition. International Journal of Computing & Information Sciences, vol. 5, No.1, pp. 62-75, April 2007.

[5] Zhang, D. et al. (2013) The Lip Position Analysis of the main consonant /w/ in Tibetan Xiahe Dialect. International Workshop on Computer Science in Sports, 2013.

[6] Speech Recognition. Wikipedia, 2015. https://en.wikipedia.org/wiki/Speech_recognition

[7] Lip Reading. Wikipedia, 2016. https://en.wikipedia.org/wiki/Lip_reading

[8] Feng, W. (2012) A Novel Lips Detection method Combined Adaboost Algorithm and Camshift Algorithm. The Second International Conference on Computer Application and System Modeling, 2012.

[9] Hoai, B.L., Hoai, V.T. and Ngoc, T.N. Lip Detection In Video Using Adaboost and Kalman Filtering.

[10] Saini, N. and Singh, H. (2015) Comparison of two different approaches for multiple face detection in color images. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 3, Issue 1, pp 2321-2004.

[11] Zurada, J.M. (1992) Introduction to Artificial Neural Systems, 1992 edition, pp 1-21.

[12] Md. Khalilur Rahman (2005) Neural Network using MATLAB. BRACU, Dhaka, Bangladesh, Powerpoint Presentation, 2005.

[13] Petajan, E.D. (1984) Automatic lipreading to enhance speech recognition. Proceedings of Global Telecomm. Conf., Atlanta, GA, 1984, pp. 265–272.

[14] Kass, M. et al. (1987) Snakes: Active contour models. International Journal of Computer Vision, pp 321-331.

[15] Hashimoto, M., Kinoshita, H. and Sakai, Y. (1994) An Object Extraction Method Using Sampled Active Contour Model. IEICE Trans. D-II, Vol.J77-D-II, No.11, pp.2171-2178, 1994.

[16] Miyaki, T., Sughara et al. (2006) Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization. SICE-ICASE International Joint Conference 2006.

[17] Sughara, K., Shinchi, T. and Konishi, R. (1997) Active Contour Model with Vibration Factor. IEICE Trans. DII, Vol.J80-D-II, No.12, pp. 3232-3235.

[18] Toshio M. (2006) Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization. SICE-ICASE International Joint Conference, pp. 5723-5726, 18-21 October, 2006.

[19] Bregler C. and Konig, Y. (1994) Eigenlips for robust speech recognition. Proceedings of ICASSP94, Adelaide, Australia, pp. 669–672, April 19-22, 1994.

[20] Turk, M. and Pentland, A. (1991) Eigenfaces for Recognition. Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[21] Beale, R. and Finaly,J. (1992) Neural networks and pattern recognition in human-computer interaction . Neural networks and pattern recognition in human-computer interaction, pp. 460.

[22] Lippmann, R.P. (1990) Review of Neural Networks for Speech Recognition, Readings in Speech Recognition . Waibel and Morgan Kaufmann Publishers, pp. 374-392, 1990.

[23] Naz, B. and Rahim, S. (2011) B Audio-Visual Speech Recognition Development Era; From Snakes To Neural Network: A Survey Based Study. Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition Vol. 2, No. 1, 2011.

[24] Hulbert, A. Poggio, T. (1998) Synthesizing a Color Algorithm from Examples. Science, vol. 239, pp. 482-485 ,1998.

[25] Canzlerm, U., Dziurzyk, T. (2002) Extraction of Non Manual Features for Video based Sign Language Recognition. Proceedings of IAPR Workshop, pp. 318-321, 2002.

[26] Kalbkhani, H. and Amirani, M.C. (2012) An Efficient Algorithm for Lip Segmentation in Color Face Images Based on Local Information. J. World Electrical Engineering Technology vol. 1(1), pp 12-16, 2012.

[27] Badura, S. and Mokrys, M. (2012) Lip detection using projection into subspace and template matching in HSV color space. International Conference TIC, 2012.

[28] Wang, S.L. et al (2007) Robust lip region segmentation for lip images with complex background. Science Direct, pp 3481 – 3491, 2007.

[29] Contrast Enhancement Techniques, Mathworks website, http://www.mathworks.com/help/images/examples/contrast-enhancement-techniques.html

[30] Discrete Cosine Transform, Mathworks website, http://www.mathworks.com/help/images/discrete-cosine-transform.html

[31] Guan, Y.-P. (2008) Automatic extraction of lips based on multi-scale wavelet edge detection. IET Computer Vision, vol.2, no.1, pp.23-33, 2008.

[32] Lucey, S., Sridharan, S. and Chandran, V. (2000) Initialized Eigenlip estimator for fast lip tracking using linear regression. Proceedings of the 15th International Conference on Pattern Recognition, vol.3, pp.178-181, 2000.

[33] Badura, S., Mokrys, M. (2015) Feature extraction for automatic lips reading system for isolated vowels. The 4th International Virtual Scientific Conference on Informatics and Management Sciences, March 23, 2015.

[34] Mattehews, I. et al. A comparison of Active Shape Model and Scale Decomposition Based features for Visual Speech Recognition. School of Information Systems, University of East Anglia, Norwich, UK.

[35] Active Shape Model, Wikipedia, 2016. https://en.wikipedia.org/wiki/Active_shape_model

[36] Lucey, S., Sridharan, S. and Chandran, V. (2002) Adaptive mouth segmentation using chromatic features. Pattern Recognition Lett, vol. 23, pp. 1293-1302, 2002.

[37] Leung, S.-H., Wang, S-L., Lau, W.-H. (2004) Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. IIEEE Transactions on Image Processing, vol.13, no.1, pp. 51-62, 2004.

[38] Saeed, U. and Dugelay, J.L."Combining Edge Detection and Region Segmentation for Lip Contour Extraction", AMDO'10 Proceedings of the 6th international conference on Articulated motion and deformable objects, pp 11-20.

[39] Ahmad B.A. Hassanat. Visual Speech Recognition. IT Department, Mutah University, Jordan.

[40] Lewis, T. and Powers, D.M.W., Lip Feature Extraction Using Red Exclusion. http://crpit.com/confpapers/CRPITV2Lewis.pdf

[41] Ahmad, N. A Motion Based Approach for Audio-Visual Automatic Speech Recognition (2011), Thesis paper, Department of Electronics and Electrical Engineering, Loughborough University, U.K., May 2011.

[42] KerenYu, Jiang, X. and Bunke, H. Sentence Lipreading Using Hidden Markov Model with Integrated Grammar. Department of Computer Science ,University of Bern, Switzerland.

[43] Canny Edge Detector algorithm Matlab Codes, http://robotics.eecs.berkeley.edu/~sastry/ee20/cacode.html

[44] Rabiner , L.R and Juang, B.H. (1986) An Introduction to Hidden Markov Models. IEEE ASSP Magazine.

[45] Gurbuz, S., Patterson, E.K., et al. Lip-Reading from Parametric Lip Contours for Audio-Visual Speech Recognition. Department of Electrical and Computer Engineering, Cemson University, Clemson, SC 29634, USA.

[46] Lee, C.H. et al. (1999) Automatic Speech and Speaker Recognition-Advanced Topics , Springer, third Edition.

[47] Stork, D.G. , Wolf, G. and Levinet, E. (1992) Neural network lipreading system for improved speech recognition. IJCNN, 1992.

[48] Yuhas, B.P., Goldstein, M. H., et al. (1989) Integration of acoustic and visual speech signals using neural networks. IEEE Communications Magazine, 1989.

[49] Kabre, H. (1997) Robustness of a chaotic modal neural recognition network applied to audio-visual speech. Neural Networks for Signal Processing, pp. 607 - 616, September 1997.

[50] Cappe , O. and Moulines, E. (2005) Inference in Hidden Markov Models. Springer, pp 42, 2005.

[51] Govind. Introduction to Hidden Markov Models. Lecture 12, CEDAR, Buffalo (Powerpoint Presentation)

[52] Nefian, A. et al. (2002) A coupled HMM for audio-visual speech recognition. Proceedings of ICASSP, pp. 2013–2016, 2002.

[53] Eveno, N., Caplier, A., Coulon, P. (2004) Accurate and quasi-automatic lip tracking. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 706 – 715, 2004.

[54] Kaucic, R., Dalton, B., Blake, A.(1996) Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications. Proceedings of the 4th European Conference on Computer Vision, vol. II, 1996.

[55] Cootes, T. F. 9(2004) Statistical Models of Appearance for Computer Vision. Technical report, University of Manchester , 2004.

[56] Padilla, R., Costa Filho C. F. F. and Costa M. G. F. (2012) Evaluation of Haar Cascade Classifiers Designed for Face Detection. World Academy of Science, Engineering & Technology, Issue 64, pp 362, 2012.

[57] Sagheer A., Tsuruta, N. Taniguchi, R. Arabic Lip Reading System: A combination of Hypercolumn Neural Network Model with Hidden Markov Model.

[58] Alan C. Bovik (2009) Essential Guide to Video Processing, Academic Press, pp 720, 2009.

[59] Ukai, N. et al. GA Based Informative Feature for Lip Reading. Department of Information Science, Gifu University, Japan.

[60] Image Processing: Morphology-Based Segmentation using MATLAB with program code. www.code2learn.com/2011/06/morphology-based-segmentation.html.

[61] Stillittano, S., Girondel, V. and Caplier, A. (2013) Lip Contour Segmentation and Tracking compliant with lip reading application constraints. Machine Vision and Applications, vol. 24, Issue 1, pp. 1-18, January 2013.

[62] Kang, S.H., Song, S.H., Lee, S.H. (2012) Identification of Butterfly Species with a single Neural Network System. Journal of Asia-Pacific Entomology, v. 15(3), pp. 431-435, September 2012.

[63] Butt, W.R. and Lombardi, L. (2013) Comparisons of Visual Features Extraction Towards Automatic Lip Reading. 5th International Conference on Education and New Learning Technologies, Barcelona, Spain, March 2013.

[64] Liew, A.W.C., Wang, S. Visual Speech Recognition: Lip Segmentation and Mapping, Medical Information Science Reference.

[65] Luo, X. (2006) Algorithms for Face and Facial Feature Detection. Master of Science Thesis, Tampere University of Technology.

[66] Kumar, V., Agarwal, A. and Mittal, K. (2011) Tutorial: Introduction to Emotion Recognition for Digital Images. HAL Archives-Ouverte, February, 2001. https://hal.inria.fr/inria-00561918