VISUAL SPEECH RECOGNITION BASED ON LIP MOVEMENT FOR BENGALI ALPHABET

MD. RAKIBUL ISLAM

K. M. ABIR MAHMUD

MD. SHAMIUL ISLAM

A THESIS SUBMITTED FOR THE DEGREE OF BACHELOR

OF SCIENCE

DEPARTMENT OF ICT

BANGLADESH UNIVERITY OF PROFESSIONALS

2018

CERTIFICATE OF APPROVAL

The thesis titled "VISUAL SPEECH RECOGNITION BASED ON LIP MOVEMENT FOR BENGALI ALPHABET" Submitted by Md. Rakibul Islam (Student ID: 150142), K. M. Abir Mahmud (Student ID: 150122), and Md. Shamiul Islam (Student ID: 150152), Session: 2014-15 is completed under my supervision, meets acceptable presentation standard and has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science (B. Sc.) in Information and Communication Engineering from the Department of ICT, Bangladesh University of Professionals.

_____

Dr. Mohammad Abu Yousuf
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

We hereby declare that this thesis is our original work and it has been written by us in its entirety. We have duly acknowledged all the sources of information which have been used in this thesis.

Md. Rakibul Islam
Dept. of ICT
Bangladesh University of Professionals
21 December, 2018

K. M. Abir Mahmud
Dept. of ICT
Bangladesh University of Professionals
21 December, 2018

Md. Shamiul Islam
Dept. of ICT
Bangladesh University of Professionals
21 December, 2018

SUMMARY

In recent days, Visual Speech Recognition (VSR) is one of the most important topics in research area. Visual Speech Recognition plays a vital role in human-computer interaction. There are many important applications that have been established based on VSR like crime-fighting, helping the hearing impaired, etc. Our thesis is based on visual speech recognition, i.e. it requires only recognizing speech without any type of auditory signal. So far, in this sector most of the work has been done in lip reading for English, Chinese, Germany, French and Indian languages. Previously some little work was done for Bengali language. Our work is based on VSR for Bengali language and provides a new approach for recognizing Bengali alphabets using multilayer deep neural network. We have used facial landmark detection for feature extraction of lip parameters. This method is robust to detect the shape of the lip at different angles and also works for low resolution images. The main aim of our work is to develop Visual Speech Recognition in a better way so that it can bring revolution for Bengali speech recognition.

ACKNOWLEDGEMENTS

There are a number of people who have been important in completion of this dissertation, both academically and personally. This work, and the time that the authors have spent in research work, would have been much poorer without them. The authors owe them with a great deal.

The authors wish to thank Dr. Mohammad Abu Yousuf for his inspiring and invaluable guidance throughout the course of this research work. The authors are highly indebted to their supervisor for their personal care and affection and for making their stay in BUP, a memorable one. The authors also express their deep sense of gratitude to their supervisor for initiating them into the field of image processing and machine learning. He took the authors work seriously and contested their ideas and expressions vigorously whenever necessary, and thus has certainly made it a better one.

The authors express their heartfelt gratitude to Nuruzzaman Faruqui for providing valuable guidance for detecting problems and finding solutions throughout the work. He is an experienced researcher in this field and the selfless support he has given to the authors for completing their work regardless of his personal time is really praiseworthy. The authors are truly grateful to him.

The authors would also like to thank Dr. Kazi Abu Taher for providing permissions for extra academic facilities including lab & classroom as well as mental support and guidance which helped the authors to complete their work successfully. His overall assistance made the authors' work easier.

The authors gratefully acknowledge the help and friendship received from their friends and teachers during the course of the study. Finally, the authors are highly indebted to their parents, who had support them at every stage and brought them to this level and hence their special thanks are due to them. The authors would like to thank their family for their understanding, encouragement and patience that they have shown during the course.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction

1.1 Visual Speech Recognition

Automatic Speech Recognition is one of the most important part of Artificial Intelligence. In human-computer interaction Automatic Speech Recognition performs a vital role. It applied in different type of important applications like crime-fighting, hearing-impaired, biometric person identification etc. Automatic speech recognition is divided into two parts.

i) Audio Speech Recognition (ASR)

ii) Visual Speech Recognition (VSR)

Speech recognition is an open research area that requires continuing research effort for further advancement. Although one can obtain high recognition rates in audio-only speech recognition in controlled environments, recognition accuracy degrades in noisy environments. For such cases, visual information is a commonly recommended approach in the literature. This approach is known as Visual Speech Recognition. In many works of audio speech recognition, speech is not detected perfectly but if we add visual speech recognition in this system then the system may be detected the speech perfectly. Because in the audio speech recognition system the speaker is audible but in the visual speech recognition system the motion of lips and other facial features like gestures etc. are available. A visual speech recognition system does not need to know context, language or residual hearing but definitely needs any type of information about the above-mentioned features. In visual domain the most important technique is viseme. A viseme interprets a particular sound by using a generic facial image. The hearing impaired can easily understand the sound using viseme.

1.2 Background of the Study

Visual Speech Recognition system can be divided into some steps

- Face detection
- Lip detection
- Lip segmentation
- Feature extraction
- Viseme recognition

First, we detect the lip from a real time video by using a proper algorithm. Then we go for lip segmentation. There are many techniques for lip segmentation.

- Image based technique
- Model based technique
- Clustering method
- Hybrid method

We used a technique for lip segmentation. Then we used an algorithm for feature extraction. In the viseme recognition part we used a classifier. Some classifiers are

- Artificial Neural Network (ANN)
- Hidden Markov Model (HMM)
- Convolutional Neural Network (CNN)
- Deep Neural Network (DNN)
- Recurrent Neural Network (RNN)

This classifier also used some algorithm like Backpropagation algorithm etc.

1.3 Objectives

The main objectives of our work are:

- We developed a system which is based on Visual Speech Recognition with a new approach in lip reading Bengali alphabets.
- Our system will provide accuracy in Bengali speech recognition through visual information over traditional audio speech recognition system. So, the hearing-impaired persons easily communicate with other peoples and to engage in any social activity.
- We will develop an efficient image processing and motion processing algorithm to recognize speech with limited lag and more accuracy.

1.4 Challenges and Limitations

Lip reading is an interesting art of inspection, inference and complicated guesswork. Different words are pronounced by the lip in the same position. For that reason, a most skilled lip reader cannot detect every word accurately. Many English sounds come from the same lip pattern or same position of our mouth, like “glue” and “blue”, when we pronounce these words our lip position is same. This is the same for the Bengali language. In Bengali alphabets some of the alphabets come from same lip pattern or same position of our mouth like “ক”, “গ” and “উ”, “ঊ” when we pronounce these letters the lip patters are same. But for lip reading the lip pattern must be different for every word.

If the system is speaker dependent, then the system is intended for use of a single speaker who is identified by the system. If the system is speaker independent, then multiple user can use the system. This is more difficult to implement. When a person reads a speech fluently which is prepared before then it can be easy to recognize. But if someone use

spontaneous speech then it is difficult to recognize the speech for the disfluencies. In visual speech recognition, every word cannot be detected accurately.

1.5 Applications

- Helping people with physically impairment and Dumb people: People with hearing impaired can benefit from visual speech recognition programs. People who 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. Recently Kavya et al. [1] proposed a system for physically impaired and dumb people.
- Crime fighting: At present, all the area is covered by CCTV. We can get any videos from the camera that are recorded and use them as input to lip reading system. Then we can identify any suspicious person.
- Lip reading system in computer: a lip-reading system can easily be used as an alternative keyboard for a physically disabled person.
- Biometric Person Identification: Automated lip reading may add to biometric identification by replacing password-based identification.

1.6 Contributions

Many works have been done on this topic. But a little work is done in lip reading Bengali language. Our contribution is that we design efficient model for lip reading Bengali alphabets. We are designed this platform with an efficient approach. We use facial landmark detection for feature extraction and use multilayer deep neural network as a classifier.

Chapter Two

Literature Review

2.1 Lip Segmentation

In visual speech recognition, lip segmentation or contour finding techniques may be arranged in image-based or model-based methods. Image based techniques can be divided into two techniques, color-based techniques and subspace-based techniques. Mainly image based techniques work on the pixel information of an entire image. Color based methods are used mainly by transforming the image into different color spaces like RGB, HSL, YCbCr or CIELUV spaces. Subspace based techniques basically use a subspace of the image to represent the image in a small number of dimensions. Model based techniques consist of Active Contour Model (ACM), Active Shape Model (ASM), and Active Appearance Model (AAM) [2].

2.1.1 Image-Based Techniques

Image based methods mainly work on the pixel information of the whole image directly. It can extract lip only for visual speech recognition [3]. The advantage of the image-based techniques is computationally less expensive than model-based techniques. Naz et al. [3] wrote in their paper that image-based techniques are restricted to illumination, mouth rotation and dimensionality. Image based methods are divided into Color based method and Subspace based methods.

2.1.1.1 Color-Based Techniques

Color based technique are used mainly by transforming the image into different color spaces to extract the proper information from the image efficiently. Hassanat et al. [4] proposed the color-based technique as a contour finding technique. There are many color spaces like RGB, YCbCr, CIELAB, CIELUV, HSL, HSI AND HSV.

2.1.1.1.1 RGB

Mainly images are in the RGB format. The RGB color model are the combination of three additive color Red, Green and Blue. This additive color system based on the tri-chromatic theory [5]. The tri-chromatic theory states that any visible color can be formed by combining these three separate color channels [5], [6]. The implementation of RGB is relatively simple and for this reason it is widely used over computer graphics for capture, storage and display. Akhter et al. [2] wrote in their paper about two major drawbacks of RGB color space: first one is luminance and chrominance information are not divided by RGB, and the other one is RGB is non-linear with visual per caption.

2.1.1.1.2 HSV

HSV stands Hue (H), Saturation (S) and Value (V). Hue is that color what human can actually identify from see that color. For example, a shirt is blue colored, actually we are identifying its hue. Saturation is when the amount of white light is combined with the hue. Value represents the intensity or brightness of the color. Hassanat et al. [4] said that the HSV color space is used cylindrical coordinate system which separates luminance (S, V) and chrominance (H).

Thejaswi et al. [7] wrote in their paper about the critical feature of HSV color space and the feature is the Hue (H). The hue value of pixels in the face is greater than the value of pixels in the lip and hue is a powerful discriminator. Sulistijono et al. [8] stated that the S and V is affected by light conditions for this reason we did not use them in the feature vector.

2.1.1.1.3 YCbCr

The RGB information is encoded by the YCbCr color space approach. RGB color model is transformed linearly by YCbCr. Thejaswi et al. [7] said that the Y in YCbCr is decoupled from the two chrominance components (Cb, Cr). So, Y is the luminance for YCbCr. Cr value is stronger than Cb value in the lip region, so lip pixels have strong red values and weak blue values [4].

2.1.1.1.4 CELUV

Ford et al. [6] said that in 1976 the CIE suggested CELUV (L*u*v*) color space as a try to achieve perceptual linearity. The logarithmic response of the eye is mimicked by the CIELUV non-linear color space. For that reason, it is designed. Wang et al [9] used the u* from CIELUV channel to identify the visibility of teeth and use the whole color space to extract color features for an AAM of the lip contour.

2.1.1.1.5 CIELAB

Ford et al. [6] said that in 1976 the CIE suggested CIELAB (L*a*b*) color space along with CELUV (L*u*v*) color space. Zhang et al. [10] said that CIELAB achieved perceptual linearity by mimicking the non-linear response of the eye as a second attempt where CIELUV represents the first attempt. The L* component approximately matches

the human perception of lightness, a* and b* are the color opponent dimensions. Liang et al. [11] and Juan et al. [12] use CIELAB for contour finding where a* component is used mainly. They report that a* component is the strongest to variability in skin and lip color.

2.1.1.2 Subspace-Based Techniques

Subspace based techniques basically use a subspace of the image to represent the image in a small number of dimensions. The subspace based methods are Discrete Cosine Transform (DCT) [13] [14], Discrete Wavelet Transform (DWT) and Discrete Hartley Transform (DHT).

Discrete Cosine Transform can be used to transform an image from the spatial domain to the frequency domain. It is relatively efficient in terms of space in the sense that it can represent required visual features of an image with minimum no. of co-efficient [15]. A discrete wavelet transform (DWT) is sampling the pixels of the image discretely by passing it through a series of filters for the transformation of image. It can take both frequency and location information. So, it is a very useful method in lip reading.

Guan [13] used the Discrete Hartley Transform which was invented by Bracewell in 1983. Using DHT, he could increase the contrast of different parts of the face image and differentiate between skin and lip. During this experiment it was found that C3 component of the DHT had the maximum value in the lip region. Thus, the edge of the lip could be detected using DHT.

2.1.2 Model-Based Techniques

At present model-based techniques are more broadly used in most lip segmentation applications recently. Model based techniques are based on earlier information of the lip shape. They take in the shape and appearance of lips from training data sat that has been manually explained [2]. Some model-based methods are Active Contour Model (ACM), Active Shape Model (ASM) and Active Appearance Model (AAM) [16] [17]. Model based techniques can be quite robust than image-based techniques and they also supposed to time consuming and computationally expensive [4].

2.1.2.1 Snakes

Kass et al. [18] was proposed the Snakes or Active Contour Model first. Snakes was proposed as a basic method for shape detection. Their development methods are based on splines. Their main methods are based on the splines. The idea is decreasing the energies of the spline frequently to fit local minima.

Delmas et al. [19], Barnard et al. [20], Eveno et al. [21] have been applied the snakes method efficiently for lips and mouth region detection and reported good lip localization representation results. Be that as it may, various issues can be related with the utilization of the snakes method for lips and mouth region identification. e.g. they can fit to the wrong feature, for example, the nose or the chin; particularly if the underlying position was far from the lip edges. Actually, spline do not bend sharply for that reason it is hard to locate sharp curves like the corners of the lips. Sometimes this method can’t work properly for the facial hair. After applying the snakes method, sometimes the tuning of snake parameters is very hard to determine and takes several seconds (a long time!).

2.1.2.2 Active Shape Model

Active Shape Model can be used more successfully for facial features identification especially for the lips and mouth region. Cootes et al. [22] proposed the ASM originally. ASM models which are known as Point Distribution Model, represent detected objects as sets of labelled points [22]. Cootes et al. [23] wished to derive a model to represent the shape of resistor as they appeared on a printed circuit board. It was known as Point Distribution Model (PDM). Cootes et al. [22] used PDM as a local optimizer to improve their estimation of position, orientation, scale and shape parameters of an example of the object of an image, this model was known as ASM.

ASMs work on the historical data of the shape of the object. In digital image it frequently adjusts the shape parameters to fit to the detected object. A shape is introduced by a set of n-labelled “landmarks”.

ASMs create a statistical shape model by using a training data set land-marked object in images. To fit model to new occurrence of the trained object by using this model. Luettin et al. [24] said that PCA is used to build this model. Recently some extensions are added to the ASMs for facial features identification by Milborrow et al. [25]. For some of the extensions the two-dimensional landmark templates are used instead of one-dimensional landmark templates and also fit more landmark. This version gives better performances.
ASM consider only the shape attribute of the image. The objects with widely varying shapes, could not solved by ASM.

2.1.2.3 Active Appearance Models

Cootes et al. [26] said that Active Appearance Model considers both the shape information and the gray scale information of the whole image. Good lip detection results illustrated using AAM by Mathews et al. [27] after 15 iterations. AAM considers both shape and illumination attributes. AAM detects the objects more accurately than ASMs. AAM works on the historical information about the shape of the object but they are not so fast because of handling more parameters. For this reason, AAM assembles using a small number of iterations. Recently Aubrey et al. [28] did a work involving lip segmentation using AAM. They used AAM because it was more consistent for the identification of the non-speech part that contain complex lip movements.

2.1.3 Clustering Methods

Leung et al. [29] used Fuzzy Clustering Method (FCM) for lip detection. Chandran et al. [30] used this model by combining color information and spatial distance between pixels in an elliptical shape function. Lip detection in normalized RGB color space expectation maximization algorithm for unsupervised clustering of chromatic features used by Lucey et al. [31]. To differentiate the non-lip pixels with the lip pixel that have similar color features and located in different regions a spatial penalty term is introduced and multiple clusters are selected to structure the background region sufficiently. This model proposed by Leung et al. [29]. The proposed algorithm has good segmentation result than other segmentation techniques. To find the correct mouth contour this method was used because it was based on iterations and it again adds to the processing time.

2.1.4 Hybrid Methods

Hybrid Methods are combining of both image-based and model-based techniques. This method applied the model-based technique to extract proper lip contours but before using model-based techniques it used color-based approach to determine the rough estimation of the candidate lip regions quickly.

Edge based and region-based detection methods are used to carry out lip segmentation. But for achieving better result Saeed et al. [32] proposed a "fusion" of edge based and region-based detection methods. For example, assumed that a human face is ready and already detected. Here there are three steps to find the final outer lip contour. First step is to select the mouth Region of Interest (ROI) and the next step is to detect the outer lip contour where the mouth ROI is issued to the edge and region-based methods. In the third step, the result of the two methods are fused to determine the outer lip contour.

2.2 Feature Extraction

Feature extraction is one of the most important factors in visual speech recognition. The system becomes inefficient if the data set of features extracted is too large and also important to ensure that the extracted features give the proper output. So, it depends on the choice of which features to extract is very important. For viseme recognition, many algorithms use geometrical features like height and width of mouth, area etc. whereas a lot of lip-reading algorithms use principal components as features.

Area of the dark region inside the mouth was known as a feature called DA (Dark Area) used in the French ALiFe system by Werda et al. [33]. Hassanat et al. [34] used a feature

set containing height, width, image quality value, presence of tongue and number of teeth pixels which is a combination of different descriptive features.

A variation of the red-exclusion technique for lip identification which was followed by a curve-fitting to detect the contour of the inner lips only, used in a feature extraction method by Chen et al. [35]. This feature extraction method was much faster than any other feature extraction techniques. The face and lips both are mostly red colored but the method focuses only the green and blue colors of the lip section of the image. So, the red and green colored values are by this given equation:

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

Lewis et al. [36] used the logarithm further enhances contrast.

By developing the AVSR system, most of the work done on VSR because the visual signals complete the audio signals and for this reason it increases the performance of these systems. This was written by Hassanat et al. [34]. Some work was properly done by using the visual only signal. There are two major steps Feature Extraction and Visual Speech Feature Recognition which is followed mostly by the lip-reading process. The feature extraction approaches are listed as bellow:

2.2.1 Geometric Features-Based Approaches

Petajan et al. [37] designed a lip-reading system to assist his speech recognition system by using geometric features-based technique and that was also the first work on VSR. Basically, his method was based on using geometric features like height and width, area, perimeter of the mouth.

Recently Werda et al. [33] proposed an Automatic Lip Feature Extraction (ALiFE) which contains lip localization, lip tracking, visual feature extraction and speech unit recognition. They proposed this prototype for detecting French vowels that was completed by multiple speakers (male and female) in natural conditions. Their system achieved 72.73% accuracy of French vowels,

2.2.2 Appearance-Based Approaches

Turk et al. [38] created a method. By inspiring this method, PCA is used by Eigenlips which are the compact representation of mouth Region of Interest. But the method was proposed by Bregler et al. [39] first. Arsic et al. [40] aimed to exploit the complementarity of audio and visual sources for this reason he investigated another Eigenlips based system. For the feature extraction process of the vertical lip motion and the mouth extension extracted by the lip-reading method using optical flow and a novel gradient-based technique respectively which is introduced by Belongie et al. [41].

Recently a speaker-independent audio-visual speech recognition (AVSR) system using a segment-based modelling strategy was developed by Hazen et al. [42]. By using a novel audio-visual integration mechanism this AVSR system collected the information from visual measurements of the speaker's lip region. This mechanism is known as a segment-constrained Hidden Markov Model (HMM).

Gurban et al. [43] developed a hybrid SVM-HMM system for AVSR where the lips being detected manually. To detect the pixel to pixel difference between consecutive frames the pixels of down-sampled images are size of 20 * 15 are coupled. Saenko et al. [44] developed as feature-based model for pronunciation variation to visual speech

recognition. To represent the future stream this model uses Dynamic Bayesian Network (DBN).

2.2.3 Image Transformation-Based Approaches

Viola et al. [45] described the pose estimation method. The visual feature extraction which is applied either on the front face, the left or the right face profile is determined by the pose estimation process. Lucey et al. [46] was designed to be posing invariant. They used the pose estimation method. Their method contains face detection and head pose estimation. To identify speech regardless of the pose of the head, they designed their audio-visual speech recognition. Jun et al. [47] used DCT for feature extraction from the mouth region to extract the most important features vector DCT coefficients.

2.2.4 Hybrid Approaches

Active Appearance Model (AAM) were projected onto 41-dimensional feature space using LDA and also visual features identify from Discrete Cosine Transform in the proposed audio-visual speech recognition system by Jun et al. [47].

Leszczynski et al. [48] compared three different algorithms for viseme 1. DFT + LDA, 2) MESH + LDA, 3) MESH + PCA in their book. There were two feature extraction procedures, first one was based on normalized triangle mesh (MESH), and the other one was based on Discrete Fourier Transform (DFT) where the classifiers designed by PCA and LDA.

Most researchers basically used dynamic time wrapping (DTW), e.g. Petajan et al. [37]. Artificial neural network (ANN), e.g. Yau et al. [49] and Werda et al. [33]. Dynamic

Bayesian Network (DBN), e.g. Belongie et al. [41]. Support vector machines, e.g. Gurban et al. [43], Saenko et al. [42]. Kandagal et al. [50] proposed in their paper that GLCM (Gray Level Co-occurrence Matrix) and Gabor Convolve algorithm used in feature extraction process.

2.2.5 GLCM for Feature Extraction

GLCM stands for Gray Level Co-occurrence Matrix. GLCM is known as the method of extracting the post order statistical texture feature. Hassanat et al. [51] proposed a system where four different spatial relationships 0°, 45°, 90°, 135° are specified and implemented.

Kekre et al. [52] wrote about extraction of four main properties. They are contrast, energy, entropy and correlation. Contrast measures the difference between intensity level of the adjacent pixels in a given image. Kekre et al. [52] said that correlation measures the level of correlations between pixels against the remaining pixels in the image. In the entire GLCM the energy measures the summation of the squared element. Information which requires 't' compress the given image and it contains lots of information from image for GLCM calculation is referred as entropy.

2.2.6 Gabor Convolve based Feature Extraction

The evaluation of Gabor wavelet algorithm determines to be very important and is massively accepted in the previous work by different authors. Texture aspects is identified by finding the mean and variant of the filtered image is used by the Gabor filters to recover the image.

2.3 Viseme Recognition

When the appropriate features have been extracted, then the process of viseme recognition starts. Nowadays 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), Deep Neural Network (DNN).

2.3.1 Artificial Neural Network

Artificial Neural Network (ANN) is an information processing system which is inspired by the biological nervous system, like brain that processes information. It consists of interconnected set of units which are known as neurons. There are two layers of a neural network, one layer is input layer and the other layer is output layer. Input layer take the input and use the sum of the products of those inputs with certain weights and then get the outputs by activating the activation function. There are two category of ANN first one is Unsupervised ANN where ANN itself can adjust the weight and the other one in Supervised ANN where the weights are specified by the users during the training process.

Perhaps these units or neurons are organized into layers but it depends upon the application basically. There are different types of ANNs available, like perceptron, Multilayer perceptron, Feed Forward Network (FFN), Hyper Column Model (HCM) and Self Organizing Maps (SOM). Ahmad et al. [53] said that the most popular architecture is the feed forward architecture with a single hidden layer.

To identify the viseme being spoken based on advanced learning of previous patterns have been used by ANNs. Nowadays different researchers are combined different

probability, statistics and ANN techniques to provide more accurate error free automatic Lip-reading systems

2.3.2 Hidden Markov Model

Ahmad et al. [53] Hidden Markov Models (HMMs) are statistical model which are used for pattern recognition of sequential data. There are many applications of Hidden Markov Model like reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, bioinformatics etc.

HMM can be used to model a system has finite number of internal states that generates a set of external observations. The outsider of the system cannot see the changes of internal states. HMM follows the Markov process so the current state of an HMM is always dependent on the immediate previous state only.

Yu et al. [54] proposed in their paper about an approach. The approach was known as a sentence systematic approach to lip reading full sentences which used HMMs integrated with grammar. A vocabulary of fundamental words is considered in this approach. A set of legal sentences is generated by identifying a grammar which is mainly based on the vocabulary. The each HMMs are integrated according to the rules of grammar and each word of the basic vocabulary id designed by an HMM.

Ahmad et al. [53] said that Artificial Neural Networks and Hidden Markov Models are the most popular classifiers for speech recognition. Basically, HMMs are the most used classifier because of their simplicity of implementation, better for training and computational efficiency.

2.3.3 Convolutional Neural Network

Basically, Convolutional Neural Network (CNN) is deep neural network that is used to classify image and group them homogeneity and execute object detection within scenes. It is a multilayer neural network. LeCun et al. [55] said about canonical structure of a CNN in their paper contains: 1. A given number of convolutional layers, four sub-tasks are divided for each like convolutional filtering, non-linearity, pooling and sub-sampling, 2. A set of fully connected layers with properties identical to that of classical neural network, 3. A softmax layer performed softmax function which outputs posterior probabilities for each class.

Chapter Three

Methodology

3.1 Overview

For Visual Speech Recognition, the first task would be to successfully detect the lips in an image. For this purpose, we need a running video or any video stream in real time. Whatever be the case, the idea is to extract the frames from the video so that we can get a set of static images. If we can detect the position of the lip in an image containing a face, then we can easily track the position of the lips in different frames of a running video. Finally, we have to generate a large set of training data for known outputs so that we can train a neural network model. In this case we have used supervised learning for training deep neural network so that it can finally predict the output when new data with unknown output is passed through the network.

3.2 Facial Landmark Detection

Facial landmark detection is relatively new but an effective way for lip detection than other traditional systems. Facial landmark detection can be used for extracting different facial features. In this method different key points of a face are pre-defined as a trained model which is used later for detecting the positions of different face features. In our research, we have used two open-source libraries for facial landmark detection, namely- opencv and dlib which are licensed under the BSD License and the Boost Software License respectively.

Dlib has built-in algorithm for facial landmark detection. It uses a Histogram of Oriented Gradients (HOG) feature along with a linear classifier for detecting face. Dlib also

provide a handy 68 co-ordinates mapped model which we have used as our model for detecting co-ordinates for lip region.

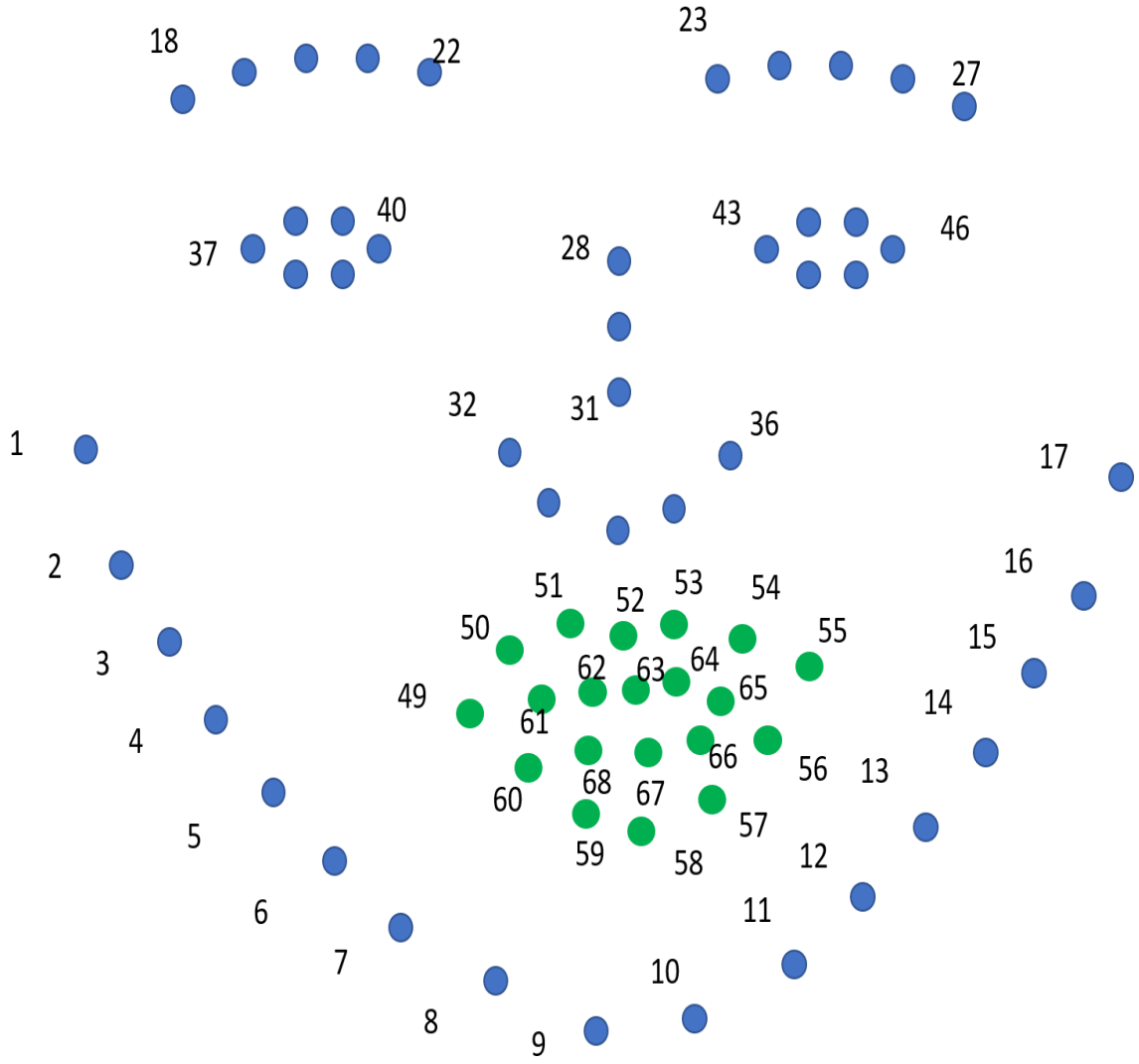

Figure-3.1: Facial Landmarks for 68 points

3.3 Artificial Neural Network

An Artificial Neural Network (ANN) is a system for processing data that works similar to the human brains. In its structure there remain a large number of interconnected nodes as like neurons in human brain. They process data together to solve complex real-world problems. ANN is used mostly for the complex scenarios like- pattern recognition, object classification, etc. where correct output can't be determined simply using an algorithm. Artificial Neural Network tries to learn from the provided environment by the passage of time and generates a model for any given scenarios. The more accurately a model can be trained, the more accurately it can predict any unknown data of similar pattern among the trained data. This type of learning is also known as supervised learning where model is created from given correct input and correct output.

In our case of visual speech recognition, if we can successfully generate the co-ordinates of lip positions, the rest part is to find a pattern for classification of different alphabets as it is spoken. So, an ANN has been developed in our research which satisfies our given condition to recognize different alphabets. We have incorporated supervised learning for this purpose. Two important concepts of ANN are also used in our research to detect alphabets accurately, which are - Multilayer Feed Forward Network and the Back-Propagation Algorithm.

3.3.1 Multilayer Feed Forward Network

In Feed-forward system, the signals travel from input to output which is sometimes referred to as top-down or bottom-up approach. The main idea for this system is that the output of any layer can't affect the same layer. In multilayer system, there remain multiple layers in between input layer and output layer which make the model more complex but

at the same time the accuracy is increased in a greater amount. This approach is being used actively in deep learning to train a deep neural network. As, we are working on data set with large numbers of input nodes having varied values, deep neural network has been used in our research.

3.3.2 Back-Propagation Algorithm

During the training phase, each unit is assigned some weights and bias values. The neural network is constructed by tuning those values to one that gives maximum accuracy. One popular method used for tuning the weights to reduce error between target output and actual output is this Backpropagation algorithm. Our proposed model also used this Back-Propagation technique to train the neural network.

3.3.3 Activation Function

Activation function is an important term in the context of Artificial Neural Network. Activation function is used in neural networks to make it more powerful and to handle data of more complex pattern as it uses non-linear functions. Using proper activation function, we can tune the weight and bias value more effectively. There are many popular activation functions, such as- Sigmoid, Hyperbolic tangent, ReLu, etc. In our research, we have used ReLu activation function as it fits best for our purpose of visual speech recognition. ReLu stands for Rectified Linear Units. The formula for ReLu function is:

$$R(x) = \max(0, x)$$

That means if $x < 0$, $R(x) = 0$ and if $x >= 0$, $R(x) = x$

It is so efficient that most of the deep learning models use ReLu activation function now-a-days. But, there is one problem for using ReLu function, i.e. it can only be used for

hidden layers. So, to solve this problem, we have used another Softmax function for the output node in our research work of visual speech recognition.

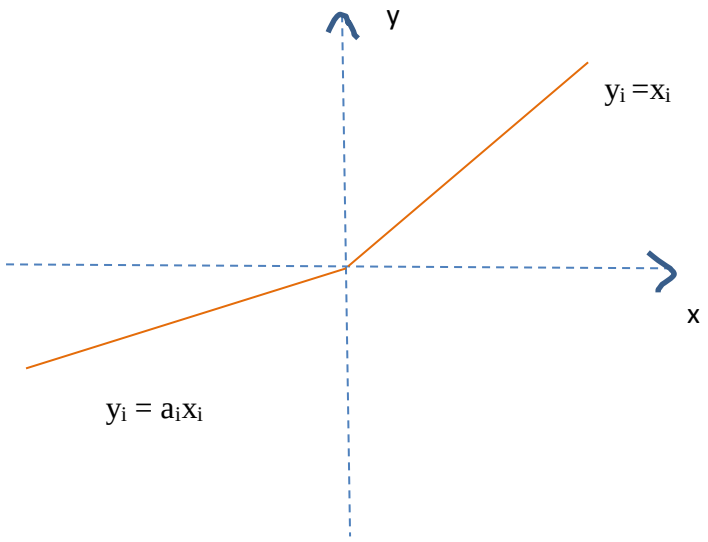

Figure-3.2: Leaky ReLU/PReLU

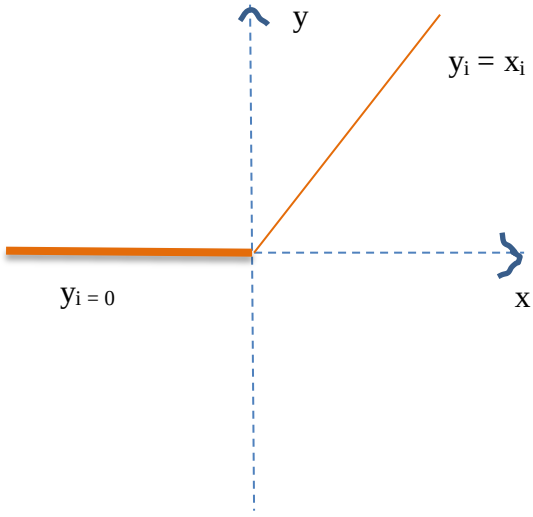

Figure-3.3: ReLU

3.4 Flowchart for Proposed Method

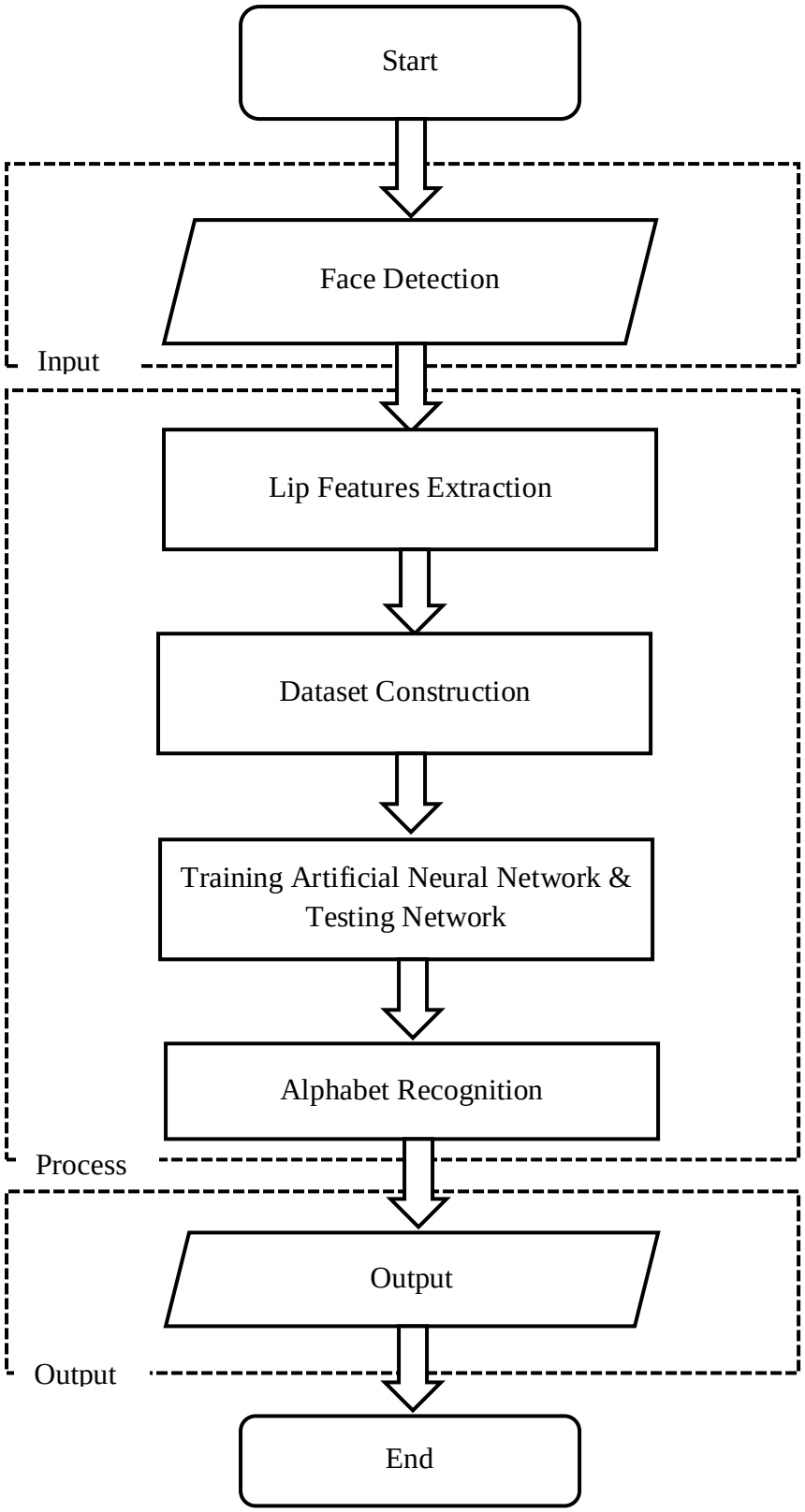

Figure-3.4: Flowchart for proposed method

3.5 Details of Proposed Method

3.5.1 Face Detection

As mentioned earlier, we have used two open source libraries, namely- opencv and dlib for face feature detection. Opencv uses Haar feature based cascade classifiers for face detection. The machine is trained from a lot of positive and negative datasets. The classifier is already provided with opencv which was trained using public dataset. So, using that cascade feature, face matrix in the image included with face can be extracted.

3.5.2 Feature Extraction

The dlib library has built-in algorithm for face feature extraction. We have used 68-point model in dlib. Different sections of human face are already denoted with 68 points (x, y co-ordinates). So, from this point, we can easily extract the lip features.

3.5.2.1 Inner Lip Contours Extraction

From the 68 points, we can get inner lip contour by connecting 61^{th} to 68^{th} points. We then keep those value in separate matrix for inner lip.

3.5.2.2 Outer Lip Contours Extraction

From the 68 points, we can get outer lip contour by connecting 49^{th} to 60^{th} points. We then keep those value in separate matrix for outer lip.

3.5.3 Dataset Construction

We get total 20 points (40 co-ordinates) for inner and outer lips. When we run the video or real-time stream through opencv, we extract 10 frames to detect a letter. So, for each

sample, we get 400 co-ordinates, which will be passed as the input node in our neural network. We also take one output as 401th parameter in each row of our dataset. The output node can have any values among 1, 2, 3, 4 or 5 which denotes '**অ**', '**আ**', '**ই**', '**উ**', '**এ**' respectively.

3.5.4 Training of Neural Network

After constructing the data set, we need to train model for our proposed system. For training neural network, we have used open-source software library TensorFlow developed by Google Brain Team which is licensed under Apache License 2.0. Using TensorFlow, we optimized our training process using first-order gradient-based optimization of stochastic objective functions through Adam Optimizer which is implemented by Kingma et al. [56]. Adam is an algorithm which estimates adaptively based on lower-order moments.

The algorithm for back-propagation used here is as follows [57]:

Steps:

- ◊ Initialize all weights in the network
- ◊ For each input layer unit j
- ◊ $O_j = I_j$
- ◊ For each hidden or output layer unit j
- ◊ $I_j = \sum_i w_{ij} O_i$
- ◊ $O_j = 1/(1+e^{-Ij})$
- ◊ For each unit j in the output layer
- ◊ $Err_j = O_j (1-O_j) (T_j – O_j)$
- ◊ For each unit j in the hidden layer, from the last to the 1st hidden layer
- ◊ $Err_j = O_j (1 - O_j) \sum_k Err_k w_{jk}$
- ◊ for each weight w_{ij} in network
- ◊ $\Delta w_{ij} = (\eta) Err_j O_i$
- ◊ $W_{ij} = w_{ij} + \Delta w_{ij}$

3.5.5 Accuracy Testing

From the Training model saved in the training phase the network is tested via some specified values to check the validation of the current neural network model, Accuracy is tested using TensorFlow testing method for given input testing datasets.

3.5.6 Recognition of Alphabet

Through the whole system we can recognize our expected outcome in a supervised learning method. We also get more accurate recognition as TensorFlow builds a more

optimized network. We give the input dataset from a real-time stream getting the input co-ordinates. And our trained AI network will recognize the Bengali letter.

3.6 Advantages over other methods

Firstly, for lip detection, we have used facial-landmark detection technique which can detect lip features more accurately than the color based or space-based technique, as the contrast of lip relative to surrounded area creates unwanted noise. As, facial-landmark detection is co-ordinate based, it is more accurate for lip detection.

Secondly, the Adam optimizer used in TensorFlow for Neural Network is very efficient and it is used for many neural network implementations now-a-days.

Chapter Four

Observation and Result

The Whole Simulation and Observation of the system has five sequent phases Detection, Dataset Modeling, Training, Testing and Recognition. Each phase consists of several tests with several inputs and output. Each phase is done one after another during the whole system model life-cycle.

4.1 Detection

In this phase the system detects the lip from the face object and marks out the lip area.

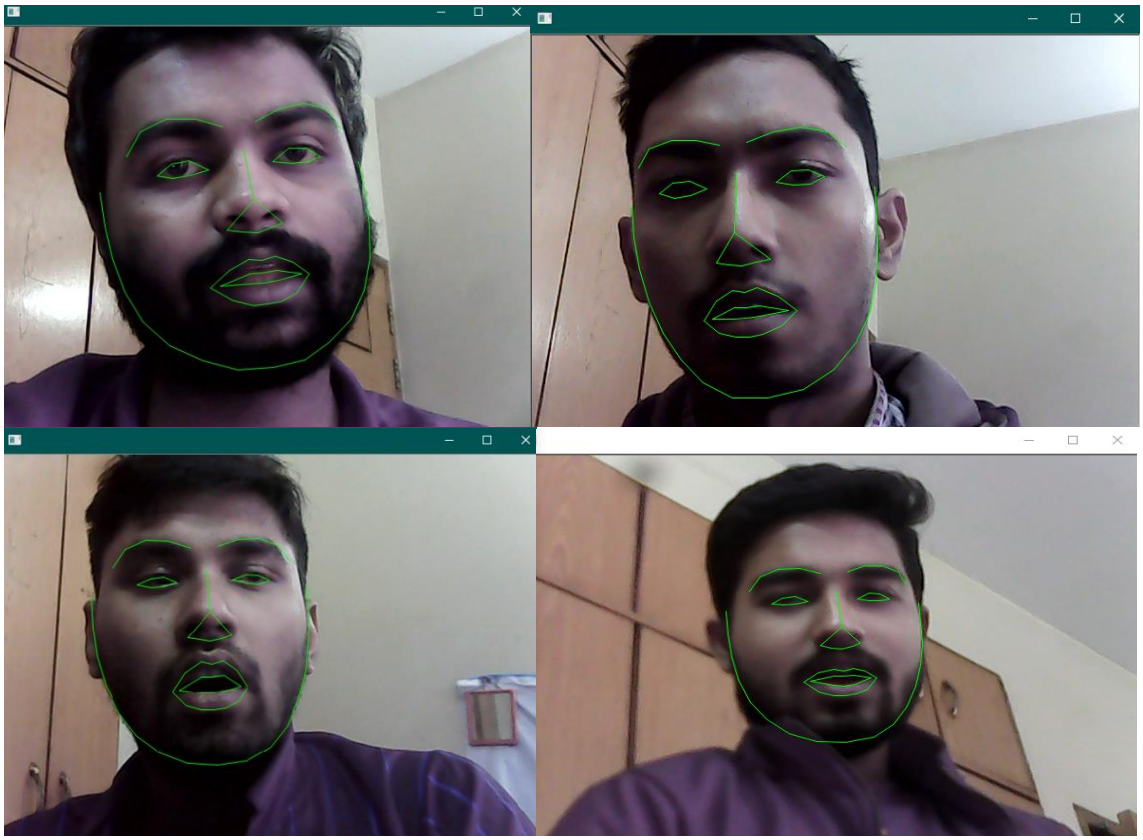

Figure-4.1: Lip Detection

After detecting the lip, the machine extracts the lip's feature based on X and Y axis points for each frame. In our system we take 20 points as feature.

Table-4.1: Table of XY points for Lip Positions

X	Y
399	282
413	274
426	269
426	269
435	271
444	269
456	272
468	279
457	290
446	294
436	294
426	295
413	292
405	281
426	277
435	278
444	277
462	279
445	280
436	281

4.2 Dataset Modeling

From the feature points we get in the detection phase we have made our input dataset combining 10 frames to record a single letter. Each letter is encoded as a Number input and it is decoded in the next phases.

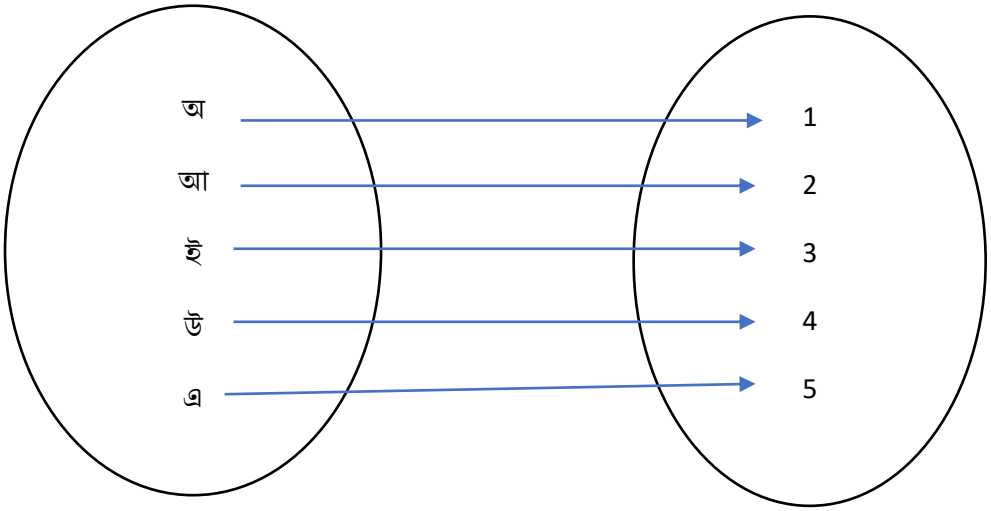

Figure-4.2: Alphabet Encoder

Each dataset has 401 parameters in our system where first 400 nodes are the XY parameters and last node is the encoded value of the letter.

Table-4.2: Dataset Model Layering

X_1	Y_1	……..	X_{200}	Y_{200}	Encoded Letter
392	282	……..	436	281	1

4.3 Training

In this phase the machine learns from the inputted dataset and generates a learn model. On the training session the number of epochs we used is 10000 and the console outputs the costs per 100 epochs.

Input Dataset we used to train the machine are listed in Table-4.3.

Table-4.3: Training Input Dataset

Set	X_1	Y_1	……	X_{200}	Y_{200}	Encoded Letter
1	399	282		425	293	1
2	420	294		447	284	1
3	414	291		441	299	1
4	417	295		448	296	1
5	399	305		427	300	1
6	400	293		424	296	1
7	415	300		443	306	1
8	401	300		429	296	1
9	389	304		414	307	1
……						
266	304	344		325	366	5
297	311	342		332	365	5
298	313	349		329	372	5
299	303	343		324	362	5
300	294	340		313	362	5

Output after the training session are listed in Table-4.4:

Table-4.4: Table of Epochs and Cost

Epochs	Cost
500	2743.3757
1000	422.4627
1500	3653.8381
2000	346.12448
2500	279.27844
3000	618.399
3500	94.56942
4000	159.1855
4500	942.3589
5000	44.68699
5500	6.562656
6000	13.34959
6500	8.245904
7000	12.765981
7500	93.47377
8000	60.80157
8500	32.377586
9000	4.119176
9500	10.757331
1000	4.734081

Costs per Epochs:

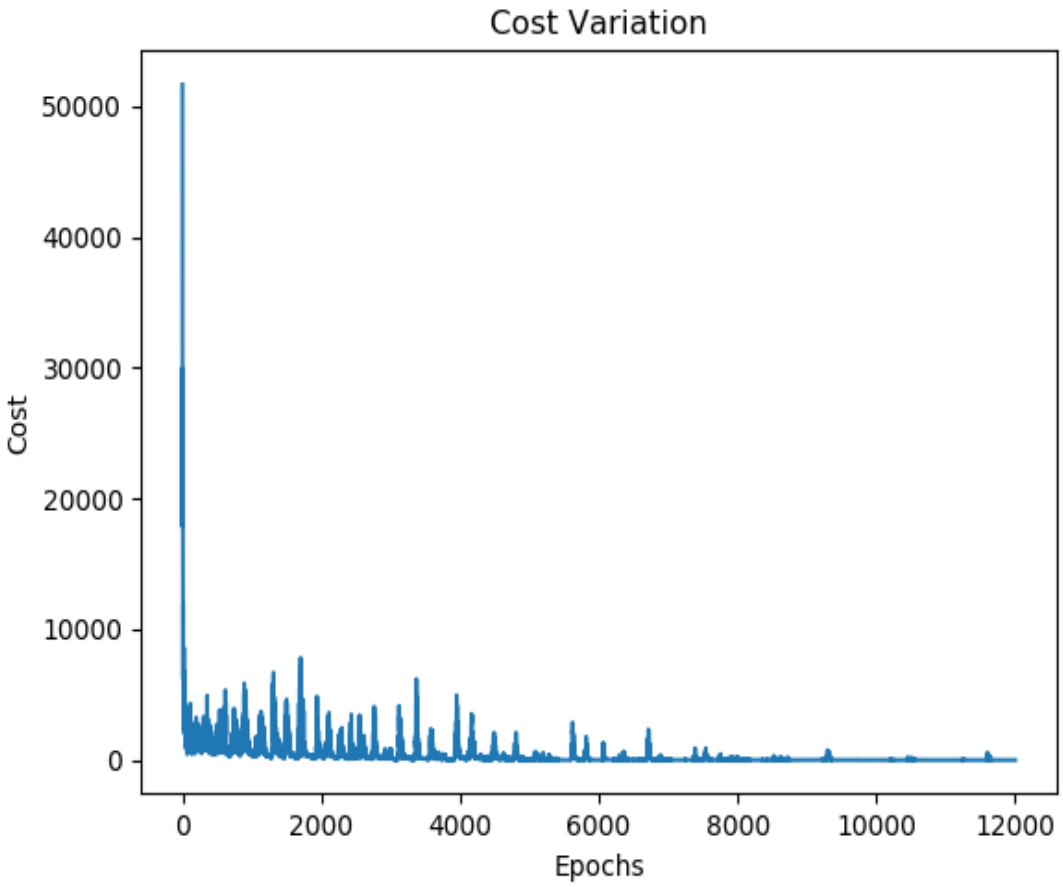

Figure-4.3: Cost per Epochs Graph

4.4 Testing

After the training the learn model is tested to analyze the validation of the model.

Table-4.5: Table of Testing Dataset

Set	X_1	Y_1	……	X_{200}	Y_{200}	Encoded Letter
1	336	308		360	318	1
2	332	303		359	308	1
3	334	306		360	312	1
4	343	302		367	307	1
5	360	299		382	310	1
6	357	293		384	300	1
7	349	300		376	304	1
8	361	293		378	312	2
9	346	298		376	334	2
10	366	288		391	297	2
…… …… …… ……						
46	329	309		358	302	4
47	340	323		359	326	4
48	347	319		309	360	5
49	288	343		316	364	5
50	291	344		316	365	5

Accuracy:

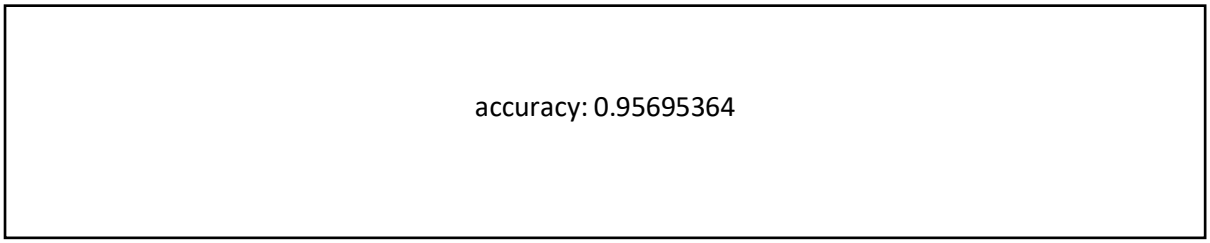

Figure-4.4: Output of testing Accuracy

4.5 Recognition

Dataset consists of 400 nodes are used to recognize the letter in a real-time scenario. Each node keeps the coordinate information of a random letter and our machine detects the letter from the trained model.

[336,308, ……. ,360,318]

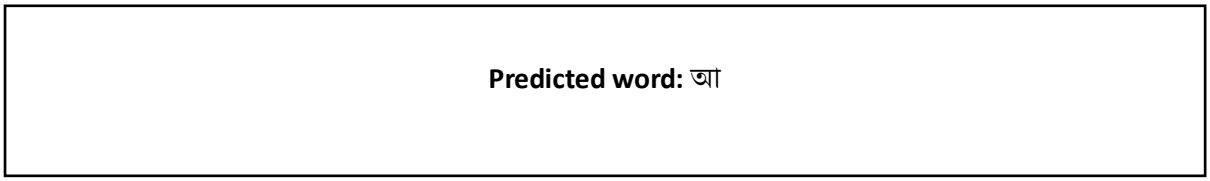

Figure-4.5 Predicted Value from Recognition Dataset

Chapter Five

Conclusion

In this thesis we are designed a platform for lip reading Bengali alphabets. We designed this platform by using multilayer neural network to identify visemes of Bengali alphabet being spoken. The success of multilayer pattern classification system depends on many factors. Most important factors are the choice of lip detection and lip segmentation algorithm, the choice of features extracted and pattern recognition system used.

Facial landmark detection technique is used to detect the lip's inner and outer curves and for feature extraction also. By using facial landmark detection, the system got total 68 points for the face. Then the method selects only 20 points of X and Y coordinates to detect only the lip shape. The system generates 400 points for a viseme, each frame consists of 40 points. After extracted the feature, we passed these to multilayer neural network. This method much faster and memory efficient than using Active Shape Model and Active Appearance Model.

The system overcome many challenges that comes with lip reading. The challenges include fast speech, bad lighting, poor pronunciation and low image quality.This system is easily implemented in embedded systems like Android or iOS to use when needed.

In future, this idea can be extended to detect more Bengali alphabets as well as the whole Bengali words and sentences.

References

[1] C. S. Kavya, N. H. Poornima, N. Sahana, K. V. Vidyashree, and G. R. Kiranmayi, "Conversion of LIP movement to speech: An aid to physically impaired and dumb people," in *Proc. Of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Oct. 3-5, 2016, Paralakhemundi, India.* pp. 1868-1871.

[2] N. Akhter and A. Chakrabarty, "Viseme Recognition using lip curvature and Neural Networks to detect Bangla Vowels," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, no. 4, pp. 7-11, Dec. 2017.

[3] S. Rahim and B. Naz, "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.

[4] A. B. A. Hassanat, M. Alkasassbeh, M. Al-awadi, and E. A. A. Alhasanat, "Colour-based lips segmentation method using artificial neural networks," in Proc. of the 6th International Conference on Information and Communication Systems (ICICS), April 7-9, 2015, Amman, Jordan. Pp. 188-193.

[5] E. Dahlman, S. Parkvall, J. Skold, P. Beming, A. C. Bovik, B. A. Fette, K. Jack, F. Dowla, C. DeCusatis, E. da Silva, L. M. Correia, P. A. Chou, M. van der Schaar, R. Kitchen, and D. M. Dobkin, *Communications Engineering Desk Reference*, Orlando, FL, USA: Academic Press, Inc., 2009.

[6] A. Ford and A. Roberts, "Colour space conversions," 1998. [Online]. Available: https://poynton.ca/PDFs/coloureq.pdf. [Accessed: July 15, 2018]

[7] S. Thejaswi and S. Sengupta, “Lip Localization and Viseme Recognition from Video Sequences,” in *Proc. of the Fourteenth National Conference on Communications, 2008*.

[8] I. A. Sulistijono, H. H. Baiqunni, Z. Darojah, and D. S. Purnomo, “Vowel recognition system of Lipsynchrobot in lips gesture using neural network,” in *Proc. of the 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), July 6-11, 2014, Beijing, China*. pp. 1751-1756.

[9] S. L. Wang, W. H. Lau, S. H. Leung, and H. Yan, “A real-time automatic lipreading system,” in *Proc. of the ISCAS-2004: IEEE International Symposium on Circuits and Systems, May 23-26, 2004, Vancouver, BC, Canada* [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 15 June 2018].

[10] J. Zhang, H. tao, L. Wang, Y. Zhan, and S. Song, “A real-time approach to the lip-motion extraction in video sequence,” in *Proc. of the ICSMC-2004: IEEE International Conference on Systems, Man and Cybernetics, Oct. 10-13, 2004, The Hague, Netherlands* [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 25 June 2018].

[11] L. Ya-ling and D. Ming-hui, "Lip extraction method based on a component of lab color space", *Computer Engineering*, vol. 37, no. 3, pp. 19-21,24, 2011.

[12] Y. W. Juan, L. Y. Ling, and D. M. Hui, “A real-time lip localization and tacking for lip reading,” in *Proc. of the ICACTE-2010: 3rd International Conference on Advanced Computer Theory and Engineering, Aug. 20-22, 2010, Chengdu, China* [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 27 June 2018].

[13] Y. P. Guan, “Automatic extraction of lips based on multi-scale wavelet edge detection,” *IET Computer Vision*, vol. 2, no. 1, pp. 23-33, March 2008.

[14] A. N. Mishra, M. Chandra, A. Biswas, and S. N. Sharan, "Hindi phoneme-viseme recognition from continuous speech," *International Journal of Signal and Imaging Systems Engineering,* vol. 6, no. 3, pp. 164-171, Jan 2013.

[15] "Discrete Cosine Transform," The MathWorks, Inc. [Online]. Available: http://www.mathworks.com/help/images/discrete-cosine-transform.html [Accessed: Nov. 9, 2018].

[16] S. Badura and M. Mokrys, "Feature extraction for automatic lips reading system for isolated vowels," in *Conference on Informatics and Management Sciences: Proc. of the 4th International Virtual Scientific Conference on Informatics and Management Sciences, ICTIC 2015, March 23-27, 2015, Zilina, Slovakia*. vol. 4, no. 1, pp. 96-104.

[17] I. Matthews, J. A. Bangham, R. Harvey, and S. Cox, "A comparison of Active Shape Model and Scale Decomposition Based features for Visual Speech Recognition," in *Computer Vision — ECCV'98: Proc. of the European Conference on Computer Vision, ECCV 1998, June 2-6, 1998, Freiburg, Germany*. pp. 514-528.

[18] M. Kass, A. Witkin, and D. Teropoulos, "Snakes: Active contour models," *International Journal of Computer Vision,* vol. 1, no. 4, pp. 321-331, 1987.

[19] P. Delmas, P.Y. Coulon, and V. Fristot, "Automatic snakes for robust lip boundaries extraction," in *Proc. of the ICASSP-1999: IEEE International Conference on Acoustics, Speech, and Signal Processing, March 15-19, 1999, Phoenix, AZ, USA* [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 17 Aug. 2018].

[20] M. Barnard, E. J. Holden, and R. Owens, "Lip tracking using pattern matching snakes," in *Proc. of the ACCV2002: The 5th Asian Conference on Computer Vision*, Jan. 23-25, 2002, Melbourne, Australia.

[21] N. Eveno, A. Caplier, and P. Y. Coulon, "A parametric model for realistic lip segmentation," in *Proc. of the ICARCV-2002: 7th International Conference on Control, Automation, Robotics and Vision, Dec. 2-5, 2002, Singapore*. vol. 3, pp. 1426-1431.

[22] T. F. Cootes and C. J. Taylor, "Active Shape Models — 'Smart Snakes'," in *British Machine Vision Conference 1992: Proc. of the British Machine Vision Conference, Sep. 22-24, 1992, Glasgow*. pp. 266-274.

[23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Training Models of Shape from Sets of Examples," in *Proc. of the British Machine Vision Conference, Sep. 1992, London*. pp. 9-18.

[24] J. Luettin, N.A. Thacker, and S.W. Beet, "Speaker identification by lipreading," in *Proc. of the Fourth International Conference on Spoken Language Processing, ICSLP '96, Oct. 3-6, 1996, Philadelphia, PA, USA*. vol. 1, pp. 62-65.

[25] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," in *Computer Vision: ECCV-2008: Proc. of the 10th European Conference on Computer Vision, Oct 12-18, 2008, Marseille, France*. vol. 5305, pp.504-513.

[26] T. F. Cootes, G. Edwards, and C.J. Taylor, "A Comparative Evaluation of Active Appearance Model Algorithms," in *British Machine Vision Conference: Proc. of the 9th British Machine Vison Conference, 1998*. pp. 680-689.

[27] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198-213.

[28] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filtering," in *Proc.*

of the 15th European Signal Processing Conference, Sep. 3-7, 2007, Poznan, Poland. pp. 2409-2413.

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

[30] V. Chandran, S. Lucey, and S. Subramanian, "Adaptive mouth segmentation using chromatic features," *Pattern Recognition Letters,* vol. 23, pp. 1293-1302, 2002.

[31] S. Lucey, S. Sridharan, and V. Chandran, "Initialised eigenlip estimator for fast lip tracking using linear regression," in *Proc. of the ICPR-2000: 15th International Conference on Pattern Recognition, Sep. 3-7, 2000, Barcelona, Spain* [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 18 July 2018].

[32] U. Saeed and J. L. Dugeley, "Combining Edge Detection and Region Segmentation for Lip Contour Extraction," in Articulated Motion and Deformable Objects: *AMDO-2010: Proc. of the 6th International Conference on Articulated motion and deformable objects, July 7-9, 2010, Mallorca, Spain*. pp 11-20.

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

[34] A. B. A. Hassanat, "Visual Speech Recognition," in *Speech and Language Technologies*, 1st ed., I. Ipsic, Ed. United Kingdom: IntechOpen, 2011, pp.279-303.

[35] Q. Chen, G. Deng, X. Wang, and H. Huang, "An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech," in *Proc. of the 2006 International*

Conference on Machine Learning and Cybernetics, Aug. 13-16, 2006, Dalian, China [Online]. Available: IEEE Xplore, http://www.ieee.org. [Accessed: 14 July 2018].

[36] T. W. Lewis and D. M. W. Powers, "Lip Feature Extraction Using Red Exclusion," in *Proc. of the Selected papers from the Pan-Sydney Workshop on Visual Information Processing, VIP2000, Sydney, Australia*. vol.2, pp 11-20.

[37] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Doctoral Dissertation, University of Illinois at Urbana-Champaign Champaign, IL, USA. Jan. 1984.

[38] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neurosci*, vol. 3, no.1, pp. 71-86, 1991.

[39] C. Bregler and Y. Conig, "Recognising Spoken Words from Lip Movement," in *Proc. of the ICASSP'94 IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-22, 1994, Adelaide, SA, Australia*. vol. 2, pp. 669-672.

[40] I. Arsic and J. P. Thiran, "Mutual information eigenlips for audio-visual speech recognition," in *Proc. of the 2006 14th European Signal Processing Conference, Sep. 4-8, 2006, Florence, Italy*. pp. 1-5.

[41] S. Belongie and M. Weber, "Recognising Spoken Words from Lip Movement," Technical Report CNS/EE248, California Institute of Technology, USA, 1995.

[42] T. J. Hazen, K. Saenko, C. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: data collection, development, and initial experiments," in *Proc. of the 6th International Conference on Multimodal Interfaces, Oct. 13-15, 2004, State College, PA, USA*. pp. 235-242.

[43] M. Gurban and J. Thiran, “Audio-visual speech recognition with a hybrid SVM-HMM system,” in *proc. of the 2005 13th European Signal Processing Conference, Sep. 4-8, 2005, Antalya, Turkey*. pp. 1-4.

[44] K. Saenko, K. Livescu, J. Glass, and T. Darrell, “Production domain modeling of pronunciation for visual speech recognition,” in *Proc. of the (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, March 23-23, 2005, Philadelphia, PA*. vol. 5, pp. 473-476.

[45] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 8-14, 2001, Kauai, HI, USA*. vol. 1.

[46] P. Lucey and S. Sridharan, “A visual front-end for a continuous pose-invariant lipreading system,” in *Proc. of the 2008 2nd International Conference on Signal Processing and Communication Systems, Dec. 15-17, 2008, Gold Coast, QLD, Australia*. pp. 1-6.

[47] H. Jun and Z. Hua, “Research on Visual Speech Feature Extraction,” in *Proc. of the 2009 International Conference on Computer Engineering and Technology, Jan. 22-24, 2009, Singapore*. vol. 2, pp. 499-502.

[48] M. Leszczynski and W. Skarbek, “Viseme recognition - a comparative study,” in *Proc of the IEEE Conference on Advanced Video and Signal Based Surveillance, Sep. 15-16, 2005, Como, Italy*. pp. 287-292.

[49] W. C. Yau, D. K. Kumar, and S. P. Arjunan, “Voiceless Speech Recognition Using Dynamic Visual Speech Features,” in *Vision in Human-Computer Interaction: Proc. of the HCSNet Workshop on the Use of Vision in Human-Computer Interaction: VisHCI*

2006, Canberra, Australia, Nov. 2006, R. Goecke, A. Robles-Kelly, and T. Caelli, Eds. Australian Computer Society. vol. 56, pp. 93-101.

[50] A. P. Kandagal and V. Udayashankar, "Visual Speech Recognition Based on Lip Movement for Indian Languages," *International Journal of Computational Intelligence Research,* vol. 13, no. 8, 2017.

[51] A. B. Hassanat, "Visual Passwords Using Automatic Lip Reading," *International Journal of Sciences: Basic and Applied Research,* vol. 13, no. 1, pp. 218-231, 2014.

[52] H. B. Kekre, S. D. Thepade, T. K. Sarode, and V. Suryawanshi, "Image Retrieval using Texture Features extracted from GLCM, LBG and KPE," *International Journal of Computer Theory and Engineering,* vol. 2, no. 5, pp. 695-700, Oct. 2010.

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

[54] K. Yu, X. Jiang, and H. Bunke, "Sentence Lipreading Using Hidden Markov Model with Integrated Grammar," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 161-176, 2001.

[55] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436–444, May 2015.

[56] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of the 3rd International Conference for Learning Representations, San Diego, 2015.*

[57] M. Swain, S. K. Dash, S. Dash, and A. Mohapatra, “An Approach for Iris Plant Classification Using Neural Network,” in *Proc. of the International Journal on Soft Computing (IJSC), Feb. 2012*. vol.3, no.1.