Silent Speaker

A Lip-reading Model Using Deep Learning



Faria Karim Porna-1620424042

Nazmul Hauqe- 1530862642

Md. Sazzad Hossain Bhuiyan- 1521643042

Hira- 1522085642

Department of Electrical & Computer Engineering

North South University

Date: 23.12.2019

Course Number & Name: CSE499A (Senior Project Design I)

Instructor’s Name

Dr. Mohammad Ashrafuzzaman Khan

Table of Contents

[1 Introduction 3](#_Toc28244563)

[1.1 Objectives 3](#_Toc28244564)

[1.2 Challenges 4](#_Toc28244565)

[1.2 Application of Lip Reading 4](#_Toc28244565)

[2 Related Works and Our Contribution 5](#_Toc28244568)

3 Methodology 21

[3.1 Dataset 22](#_Toc28244564)

3.2 STCNN 22

[4 References 23](#_Toc28244572)

# 1 Introduction

The current era is becoming increasingly dependent on information technology. Today, new technologies are dependent on data. At present, there is a lot of work on language processing model. Lip reading is one of them. For a skilled lip reader, it is very difficult to read lips but when a machine wants to read lip then it is very difficult and challenging to train that machine especially when it is a difficult language like Bengali.

A lot of work has been done on English, French and Spanish, but there is no work on lip reading based on Bengali language. Because of presence of similar pronunciation techniques for different alphabets, lip- reading in Bengali is particularly challenging and thus avoided by most. And working with Bangla language is also hard because of lack of dataset and language model as well.

## 1.1 Objectives

1. If a skilled lip reader can understand what the speaker is saying through the movement of the lips, then it possible to train a machine and make it a lip reader by creating some model and algorithm

2. It was found that a lot of research has already been done on lip-reading based on English words and some amount on few other languages like Chinese, Arabic, Hindi and Tibetan. However, no information could be found on lip reading in Bangla. Therefore, development of a system to lip-read Bangla is very important.

3. Most of the algorithms are 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.

## 1.2 Challenges

1. The basic technique of lip reading is to observe the change in the shape of the lips as well as to keep the word sequence consistent. But when we have to train a machine, it will be very difficult because people move their lips in different ways while talking and it is very difficult to track lip shape.

2. Secondly, in most cases the speaker’s lips are obscured or not totally visible. In those cases, the result is not correct or a precise detection is not possible.

3. A more difficult challenge is in recognizing, extracting and categorizing the geometric features of the lips during speech.

4. Another big challenge is to build own dataset.

## 1.3 Application of Lip Reading:

1. It will help deaf people as a hearing aid.

2. It will help others to recognize the speech of mute people.

3. It will also help to detect criminals in the detective branch.

4. It can be used for security purposes.

5. It can be used for video captioning.

6. This model will help to recognize speech in noisy areas.

7. It can be used in silent-movie processing.

8. It will help people to detect fake video.

# 2 Related Works and Our Contribution

*Lip-reading model:* In this paper [1], they have tried to extract the visual features of lip and with the help of saliency maps, they tried to track the behavior of the lip movement, finally they build a model that converted the motion of lip into textual form. In their model they have used Spatiotemporal Convolutional Neural Network as the feature extractor, 2 Bi-GRU as sequential model, the connectionist temporal classification loss to tackle the sequence problem of the character and finally they have used a SoftMax layer. They used GRID corpus dataset for their model. They got the accuracy of 95.2% in sentence-level. Another related work is [2]. In this work they presented an end-to-end visual speech recognition system which jointly learns to extract features directly from the pixels and perform classiﬁcation using LSTM networks. They acquired the classification accuracy of 84.5%. In the paper [3], they presented an end-to-end multi view lip reading system which jointly learns to extract features directly from the pixels and performs classification using BLSTM networks. The proposed model achieves state-of-the-art performance on the OuluVS2 without using external data for training or even data augmentation. The provided mouth ROIs are well cropped and this might not be the case when automatic tools for mouth ROI detection are used. The model can be easily extended to multiple streams. The best three-view model results in a 10.5% absolute improvement over the current multi-view state-of-the-art performance on OuluVS2, without using external databases for training, achieving a maximum classification accuracy of 96.9%. A novel approach fix speaker identification has been described in [4] which is based on ca spatial and temporal analysis of the mouth. Facial features are extracted from image sequences which represent the shape and intensity of the lips. The features are of low dimension and invariant to scale, translation and rotation. Another robust approach lip localization and tacking is mentioned in this paper [5]. They have used OpenCV based approach for the segmentation of lip and a component of Lab color space is proposed to accurately extract lip shape and track lip region. In [6] they developed the CAI application for the hearing-impaired student. This is basically a game-based platform where multiple choice questions are given to select the correct answer by comparing mouth shape. Another paper [7] describes an automatic lip-reading system consisting of two main modules 1) a preprocessing module able to extract lip geometry information from the video sequence and 2) a classification module to identify the visual speech based on dynamic lip movements. Lip geometry features including height, width, ratio, area, perimeter and various combinations of these features were evaluated to determine which performs the best when representing speech in the visual domain in the application of three separate classification methods, namely optical flow, Dynamic Time Warping (DTW) and a new approach termed Multi-Dimensional DTW. Experiments show that the proposed system is capable to show recognition performance of 68% just using lip height, lip width and the ratio of these features demonstrates that the system has the potential to be incorporated in a multimodal speech recognition system to use in noisy environments. In this paper, [11 ]they applied CNN as a visual feature extraction mechanism for VSR. They have trained a CNN with images of a speaker’s mouth area in combination with phoneme labels, the CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes. To recognize words from the phoneme label sequences generated by the CNN, monophony HMMs with 8, 16, and 32GMM components are utilized. Evaluation is conducted with the 84 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To compare with the baseline performance, two other visual features with similar dimensionalities are prepared. One feature has 36 dimensions, generated by simply rescaling the images to 6\*6 pixels, and the other feature has 40 dimensions, generated by compressing the raw images by PCA. Their proposed system is evaluated on an audio-visual speech dataset which is comprised of 300 Japanese words with six different speakers. The evaluation results of their isolated word recognition experiment demonstrate that the visual features acquired by the CNN significantly outperform those acquired by conventional dimensionality compression approaches, including principal component analysis. The average phoneme recognition performance is 58%. They reported that visual phoneme recognition works better for recognizing vowels than consonants. The result derives from the fact that, the mean recognition rate for all vowels is 60-100%, whereas the mean recognition rate for all other phonemes is 20-80%.In this paper, [12] feedforward and recurrent neural network layers (LSTM) are stacked to form a single structure which is trained by back-propagating error gradients through all the layers. The performance of such a stacked network was experimentally evaluated and compared to a SVM classifier using conventional computer vision features (Eigen lips and Histograms of Oriented Gradients). The LSTM lipreader with a single feed-forward network learns the features automatically together with training the LSTM sequence classifier, consistently achieved almost 80%-word accuracy in speaker-dependent lip-reading. They have used GRID corpus dataset which have 51 different words to classify. From their experiment they have recognized that the confusion on letters is far higher than on longer words. For this speaker and configuration, the accuracy on the letters is 69.8, the accuracy on the non-letter words is 93.4%. The total accuracy is 82.0%. They reported a best word accuracy on held-out evaluation speakers of 79.6% using the end-to end neural network-based solution (11.6% improvement over the best feature-based solution evaluated).

This paper [14] is concerned about the visualization of image classification models, learnt using deep ConvNets. They have considered two visualization techniques, based on computing the gradient of the class score with respect to the input image. The first generates an artificial image, which is representative of a class of interest. The second computes an image-specific class saliency map, highlighting the areas of the given image, discriminative with respect to the given class. They showed that such saliency map can be used to initialize Graph Cut based object segmentation without the need to train dedicated segmentation or detection models. Finally, they demonstrated that gradient-based visualization techniques generalize the DeconvNet reconstruction procedure. Their visualization experiments were carried out using a single deep ConvNet, trained on the ILSVRC-2013 dataset, which includes 1.2M training images, labelled into 1000 classes. Their weight layer configuration is: conv64-conv256-conv256-conv256-conv256-full4096-full4096-full1000, where convN denotes a convolutional layer with N filters, fullM – a fully-connected layer with M outputs. On ILSVRC-2013 validation set, the network achieves the top-1/top-5 classification error of 39:7%=17:7%, which is slightly better than 40:7%/18:2%, reported in for a single ConvNet. This paper [14] is concerned about an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss. They proposed a novel visual feature extraction approach that connects the lip image to audio features efficiently, and the use of CBN’s increases robustness with respect to speech fluctuations caused by hearing loss. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods. They have used utterances of one male person with hearing loss, where the text is the same as the ATR Japanese speech database A-set where 2,620 words are used as training data, and 216 words as test data. First, they prepared the input features for training a CBN from lip images and speech signals uttered by a person with hearing loss. For the audio signals, after calculating short-term Mel spectra from the signal, they obtained Mel-maps by merging the Mel spectra into a 2D feature with several frames, allowing overlaps. The visual signals of the eyes, mouth, nose, eyebrows, and outline of the face are aligned using the point distribution model (PDM) and its model parameter is estimated by constrained local model (CLM) and a lip image is extracted. For the output units of the CBN, they used phoneme labels that correspond to the input Mel-map and lip images. Audio and visual CBNs are separately trained, and the parameters of the CBN are trained by back-propagation with stochastic gradient descent, starting from random values. Following the training of CBNs, the input Mel-map and lip images are converted to the bottleneck feature by using each CBN. Then these features are concatenated, and used in the training of HMMs for speech recognition. In the test stage, they extracted features using each CBN, which tries to produce the appropriate phoneme labels in the output layer. Finally, extracted bottleneck audio and visual features are simply concatenated and used as the input features of HMMs to audio-visual speech recognition. Their proposed audio-visual feature outperforms the AV BNF I in the clean environment and SNR of 20dB, where the integrated features between the audio and the proposed visual bottleneck features improved 3.3% and 3.8% compared with the AV BNFs I, respectively. However, at the SNRs of 10dB and 5dB, the integrated feature using their proposed feature could not improve the accuracy in comparison with that of the AV BNF I.

In this paper,[15] the author presented various methods to predict words and phrases from only video without any audio signal. They employed a VGGNet pre-trained on human faces of celebrities from IMDB and Google Images. The VGGNet is trained on images concatenated from multiple frames in each sequence, as well as used in conjunction with LSTMs for extracting temporal information. While the LSTM models fail to outperform other methods for a variety of reasons, the concatenated image model that uses nearest-neighbor interpolation performed well. They have used the MIRACL-VC1 dataset in their project. The dataset was created from 15 people who spoke each of ten words and ten phrases ten times leading to a total of 15 \* 20 \* 10 = 3000 instances. They proposed several new methods for performing visual speech recognition on sequences of color images with variable length. The initial methods concatenated the first k images of each sequence into a 2D grid, which was then classified by a VGGNet pre-trained on faces. One method attempted to train a smaller model on these concatenated images from scratch. The final model attempted to handle variable-length sequences with multiple LSTM layers which were given the feature vectors output from the VGGNet as input. Their best-performing model was the concatenated model that used interpolation. The model we trained from scratch did not perform well because the dataset we used is relatively small. Therefore, the pre-trained VGGNet was still able to perform well on these concatenated images. They also experimented with different parameter update strategies. They noticed that SGD was incapable of training the model in reasonable time. It gave no significant improvements even after 20 epochs. In comparison, Adam showed improvements right from the first epoch. They have achieved best validation accuracy of 76%, and the test accuracy of their best model is 44:5%. They noticed that phrases have a higher accuracy as compared to words. In this paper,[16] the authors have contrasted the performance of a machine-based lip-reading system with human lip-reading ability. They found that the automated system outperforms human lip-readers. For relatively simple tasks there is little improvement in recognition accuracy when adding full appearance features to the machine-based system, whereas for human lip-readers they observed significant improvements in performance. Finally, they measured the effect of ‘speaker training’ on human lip-reading ability and they found even very limited training is sufficient to improve performance. They also have looked at the contribution of shape and appearance to lip-reading ability and found that for the (6-class) task presented here, appearance is significant for human viewers, but not for the automated system. Generally, we find that automatic lip-reading systems outperform human viewers. However, an exception is recognized in full, isolated, real words. To train machine-based lip-reading systems, word-level HMMs were trained from both shape-only and full shape and appearance features using a leave-one-out cross-validation framework. The topology of the HMMs was optimized and the best performing topology used in the evaluation. This was a topology of 16 states each with 3 Gaussian mixture components. To measure the baseline performance of human lip-readers, 17 computing undergraduate students from a UK University volunteered to take part in the experiment. This provided a set of 60 utterances (six letters \* five speakers \* two stimuli type). In this experiment, participants were shown, in a randomized order, three repetitions of the 60 test movies (without audio), and they were each asked to circle on an answer sheet the letter they believed was being spoken. This was a closed test where all letters A–F letters were offered as possible answers to all utterances. At both the phoneme and viseme levels, the automated system performed significantly better than the human participants (p < 0:002) and achieved recognition rates of 80.27% and 91.6% (full shape and appearance) compared to 31.6% and 35.4% respectively. In this paper,[17] it is shown that the application of SAT appears to have considerable promise in speaker-independent lipreading. For data they have used an audiovisual corpus of twelve speakers, seven male and five female, each reciting 200 sentences selected from the Resource Management Corpus. Kaldi speech recognition toolkit was used to train the visual speech models (phonemes and visemes units) and decode the test data. The HMM/GMM systems that they built are: (i) monophone and monoviseme systems with and features (Mono), (ii) triphone and triviseme systems with LDA ((Tri:LDA)) (iii) triphone and triviseme systems with LDA+MLLT ((Tri:LDA+MLLT)), (iv) triphone and triviseme systems with LDA+MLLT+SAT (Tri:LDA+MLLT+SAT).In the experiment, firstly the visual features are considered in a block of 7 frames. They are then decorrelated and forced to a dimensionality of 40 using Linear Discriminant Analysis and further decorrelated using maximum likelihood linear transform. SAT is then applied using feature-space maximum likelihood linear regression of 40 \* 41. The 40-dimensional speaker adapted features are then spliced across a window of 9 frames and applying LDA to decorrelate the concatenated features and reduce dimensionality to 250. They found that word recognition accuracy is always higher when phonemes are used as the modelling units rather than visemes. This paper [18] presents a novel feature learning method for visual speech recognition using Deep Boltzmann Machines (DBM). The experimented method is able to explore both acoustic information and visual information to learn a better visual feature representation. During the test stage, only the videos are used to generate the missing audio features, and both the given visual and audio features are used to produce a joint representation. The experimental results show that the proposed techniques outperform the performance of handcrafted features and previously learned features. The data corpus used in this paper was collected through an Australia wide research project called AusTalk. It is a large-scale audio-visual database of spoken Australian English, including isolated words, digit sequences, and sentences, recorded at 15 different locations in all states and territories of Australia. In the experiment, the visual feature learned by the Deep Boltzmann Machine (DBM) is concatenated with Discrete Cosine Transform (DCT) feature vector, followed by a Linear Discriminant Analysis (LDA) to decorrelate the feature and reduce the feature dimension. Then, the Gaussian Mixture Model- Hidden Markov Model (GMM-HMM) is used as a classifier for visual speech recognition. Their proposed method showed the accuracy of 69.1%. This paper [19] proposed the technique to localize lip region for the Myanmar consonants recognition. The experimental system demonstrates that this technique performs lip motion sequences in video. They are localizing all of the test lip movement successfully and the results were perceived to be acceptable for lip reading. Therefore, this paper presents Myanmar consonant recognition based on lip movements towards lip reading by using CIELa\*b\* color transformation, Moore Neighborhood Tracing Algorithm and linear SVM classifier. The purpose of this study was to develop a visual training technique to accurately identify the characteristics of the lip’s movement for hearing impairment. In this paper,[20] they are employed three different evaluated feature sets for representing the spoken information found in the video frames recorded from the speakers, inside a speaker independent lip-reading mission. Two HMM-based visual models, including the conventional GMM-HMM and the young DNN-HMM hybrid were implemented to test and compare the introduced features. After testing the easy-to-access raw gray level ROI of the speakers’ lips, the geometric specifications of the lips (the shape features) were employed which showed a lower error rate by 20.4% relative, on average. The DBNFs which benefit from the advantages of both the former feature sets were then employed and showed a relative improvement with an average of 15.4% in comparison to the shape features, over the test data. In this paper; [21] They are show that CNN architectures can be used to classify temporal sequences with excellent results. On the 333-word test set, they are achieved top-1 accuracy of 65.4%, which exceeds state-of-the-art on multiple datasets that have lexicon sizes that are orders of magnitude smaller, and a top-10 accuracy of 92.3%. They also demonstrate a recognition performance that exceeds the state of the art on a standard public benchmark dataset, Oulu VS. Next steps include extending to lip reading of proﬁle views, and combining the CNNs pre-trained using this approach with LSTMs trained with a language model in order to recognize sentences rather than individual words. Of course, the visual only speech recognition method developed here can also be combined with audio only speech recognition to both their beneﬁts. In this paper,[22] they have presented LCANet, an end-to-end deep neural network architecture for machine lipreading. To accurately predict visemes, LCANet introduced a cascaded attention-CTC decoder which effectively compensates the defect of the conditional independence assumption of the CTC approach. In addition, LCANet stacks highway net- work layers over 3D-CNN layers to further improve performance. In this paper, extensive experiment results show that LCANet achieves the state-of-the-art accuracy as well as faster con- vergence. The experimental results show the proposed system achieves a 1.3% CER and 3.0% WER on the GRID corpus database, leading to a 12.3% improvement compared to the state-of-the-art-models. the proposed cascaded attention-CTC model is better. they got Accuracy of 97.4% for AH-CTC(LCANet). This paper [23] presents a lip localization based visual feature extraction method to segment lip region from image or video in real time. In this paper the main goal are, they are implement a system for synchronizing lips with the input speech. To extract visual features i.e. visemes from input video frame or image they have used HSV and YCbCr color model along with various morphological operations. They have developed algorithm to work with normal lighting conditions and natural facial images of female and male. work is to implement a system for synchronizing lips with the input speech. This involves face detection followed by ROI extraction. Then, the lips of the speaker are tracked in consecutive frames of recorded video. Following these steps, and given an informative set of features, the visual front-end module can proceed with feature extraction. Four video sequences were taken from different subjects, each one having at most than 3X30 frames. Those sequences were analyzed using the proposed lip segmentation algorithm. This research paper [24] proposes a novel approach using Coordinate Based Super-Pixel Segmentation algorithm (CBSS) to improve the accuracy of mouth segmentation. The proposed CBSS algorithm is able to robustly segment the mouth region that belongs to a given mouth shape. For the extracted mouth region, Discrete Cosine Transform (DCT) is applied to segregate the crucial features. Then the visual lip features are trained using Support Vector Machine (SVM) to recognize the speech. The proposed work describes automatic Visual Speech Recognition (VSR) based on mouth detection. They are presenting an overview of the VSR model. They are also used Artificial Neural Networks (ANN) and Radial Basis Function (RBF) Knowledge-based method. The current work uses Support Vector Machine (SVM) as the classifier to recognize the spoken words. SVM is a supervised machine learning algorithm. They are using iBallface2face web camera. Framing of the video sequence is done at 25 frames per second with the resolution of 320 \* 240 pixels. The recordings are stored in AVI or MPEG file format. They got 90% accuracy for viola and jones model, 65% accuracy for RGB color model, 78% accuracy for Active Shape model , 95% accuracy for Threshold base model ,98% accuracy for CBSS.

Our proposed model is related to the mentioned works, but not exactly the same. Most of the mentioned works are based on English or other languages like Japanese, Mandarin etc. None of them are based on Bangla. According to our knowledge, we are the first to introduce Bangla language-based lip-reading model. Some of the papers have used very limited variety of words during training session. So, the machine learned very less word. But in our model, we will try to add more variety of words so the model can learn more. Most of the related works have used dataset based on front-facing speaker but we will also use the dataset where speaker can move face while speaking but the model still can track the lip position accurately. Some of the works are strictly speaker-dependent. But we will build speaker-independent model. All the above-mentioned works are confined to only suitable model building. But unlike others, we will try to reach our model to the application level also.

*Bangla speech recognition model:* One of the notable works based on Bangla speech to text conversion is [8]. In the experiment, the speech engine is used for converting digital sound signal to text format. This text forms are written into text editor using keyboard handler function. Grammar rules are used by speech recognition (SR) process to analyze human speech input. In this process, this attempts to understand what a person is saying. In this research, the grammars are represented in a XML file in text form for the upcoming voice command. In the case of dictation, the grammars are used to indicate some words that are likely to be spoken. In this thesis paper [9], they tracked 3 vowel of Bangla language (আ, অ, এ). For doing this they followed Image acquisition, Face detection, Lip detection, Lip segmentation, Viseme detection. For face and lip detection they used same technique of [7]. In this thesis paper ANN is used for viseme recognition. Another Bangla speech-based work is [10]. In this paper, they have used a Microsoft based API which is SAPI, to convert the Bangla speech to text. Here, English is used as middle factor during conversion. The final Bangla word is got by matching the related English character to the respective Bangla character. When they tested the model on newspaper article, they got 78% accuracy. In this paper [25] they are presented a model for voice to text conversion method for Bengali language using various techniques to refine the data and adapt it for the complexities of implementing Bengali characters. An open source frame work called CMU Sphinx 4 was used to generate Bengali UNICODE font. A digital audio workstation called Audacity was used to manipulate the recorded data. The performance of the proposed model was tested using a dataset where both male and female voices were recorded. The proposed model showed around 75% accuracy for the tested dataset. This paper the best work done in this arena was using Microsoft SAPI which obtained a 78% accuracy but that detection was only on a word by word basis which was a major limitation and the accuracy was for a specific data. This system not only recognizes the word but also the sentence due to our better training model as well as the newer Sphinx 4 framework. In this paper [26] they are concentrated on the research and development of a Bangla Speech Recognizer using the appropriate technique and tools. This work is the first reported attempt to recognized Bangla speech using HMM Technique with the assist of stochastic language model. They are creating a regular grammar and convert it to an intermediate form of decoding network using the HParse tool. Networks are specified using the HTK Standard Lattice Format (SLF). In the grammar the legal word sequences explicitly for translation, automotive speech recognition, dictation, hands-free computing: voice command recognition computer user interface, home automation, interactive voice response, medical transcription, mobile telephony, pronunciation evaluation in computer-aided language learning applications and robotics. The isolated speech recognition for commands & control, data entry, mobile telephony and home automation task. On the other hand, continuous speech recognition can be used for speech to text conversion. Recognizing continuous speech with ANN classifier has average accuracy rate of 73.36% for three-layer Backpropagation Neural Network the maximum accuracy rate is 86.67% and spoken letter recognition by measuring Euclidian distance, which can recognize only the vowels, has an 80% accuracy rate In comparison, the recognizer presented in this paper has an average accuracy rate of 85%. In this paper, [27] they applied Building Acoustic model, Utterance, Speaker Dependence, Vocabularies, HMM model and language model. Speech recognition (SR) in terms of machinery is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. It is a broad term which means it can recognize almost anybody's speech, but to make the machine independent of voice, huge training data is required. There are basically two types of SR: 1. Isolated speech recognition - ISR 2. Continuous speech recognition - CSR. They tried to test the system by varying speaker, environment, microphone etc. Testing the system is done using audio inputs of two test speaker and live test from the microphone in different environment (speaker profiles are given in Appendix section). Tests are done using two different decoders: 1. Pocket Sphinx 2. Sphinx 4. Sampling rate of the audio: 16 kHz • Bit rate (bits per sample) : 16.Then after the process is completed they got an output like: MODULE: DECODE Decoding using models previously trained Decoding 529 segments starting at 0 (part 1 of 1) 0% This step had 2 ERROR messages and 45 WARNING messages. Aligning results to find error rate SENTENCE ERROR: 2.8% (15/529) WORD ERROR RATE: 1.0% (23/2319).

Test results with Pocket Sphinx

TOTAL Words: 2292 Correct: 2276 Errors: 16 TOTAL Percent correct = 99.30% Error = 0.70% Accuracy = 99.30%

TOTAL Words: 142 Correct: 134 Errors: 9 TOTAL Percent correct = 94.37% Error = 6.34% Accuracy = 93.66%

TOTAL Words: 392 Correct: 381 Errors: 12 TOTAL Percent correct = 97.19% Error = 3.06% Accuracy = 96.94%

Average Accuracy Rate = 90.65%

Input Type: Microphone-

TOTAL Words: 48 Correct: 43 Errors: 5 TOTAL Percent correct = 89.58% Error = 10.42% Accuracy = 89.58%

TOTAL Words: 128 Correct: 118 Errors: 11 TOTAL Percent correct = 92.19% Error = 8.59% Accuracy = 91.41%

The described above system is in its preliminary level, all the tools, data used to train and build the system has been discussed throughout the whole report. To make the system more natural, lots of improvement in case of data, tools and its parameters is required, but again no speech recognizer till now has 100% accuracy. In this work [28], an automatic recognizer of connected Bangla digits has been developed using BPNN and MFCC feature extraction method. Neural networks are very sensitive classifiers. A small amount of changes in the network architecture may cause significant change in the output. One of the major goals of this research experiment was to optimize the network based on network parameters such as number of hidden layers, learning rate, error threshold and number of epochs. It is also evident from the result is that the recognition accuracy varies over digits due to their phonetic traits. To acquire persistent performance over digits, BPNN based recognition system can be improved either by employing hybrid classifier and/or incorporating robust features. BPNN learning algorithm is used to train the network. The required time to train the network, number of hidden layers, error threshold and number of epochs are considered while training the network to reach the best possible recognition accuracy. This proposed system has been implemented using object-oriented programming and the achieved recognition accuracy is very much satisfactory and consistent. The network has been tested for three different setups and the best recognition accuracy achieved for digit dataset is 98.46%. The paper [29] presents the capability of an HMM-based TTS system to produce Bengali speech. In this synthesis method, trajectories of speech parameters are generated from the trained Hidden Markov Models. A final speech waveform is synthesized from those speech parameters. They are experimented, spectral properties were represented by Mel Cestrum Coefficients. Both the training and synthesis issues are investigated in this paper using annotated Bengali speech database. Experimental evaluation depicts that the developed text-to-speech system is capable of producing adequately natural speech in terms of intelligibility and intonation for Bengali. The speech database employed here for training purpose is originally developed by C-DAC. The output of the Bengali-HTS is also compare with previously developed Epoch Synchronous Non-Overlap Add (ESNOLA) based concatenated speech synthesis technique. The total average score for the original sentences is 4.66 and the ESNOLA based synthesis sentence is 2.34HTS is 3.6.

All of these above-mentioned works are related to Bangla speech recognition. But none of them are directly related to lip-reading and text extraction from lip movement. On the other hand, some of the works are based on some specific alphabet. But in our approached model we will try to cover all the possible Bangla alphabets.

# 3 Methodology

We have not decided yet which of the machine learning algorithms we are going to use in our model. But as we are dealing with video-based dataset so we are assuming that spatiotemporal convolutional neural network will be the best option for extracting features of the lip movement from the video. As we are going to use sequential data and text conversion approach so we will need sequential language model. We can use models like LSTM, Bi-Directional GRU, RNN for this purpose.

## 3.1 Dataset

We are collecting Bangla language-based video dataset from news videos. Firstly, we will collect news videos of a specific news reporter. Here we are going to use the videos clips of “Mithila Islam” who is a renowned news presenter in “Bangla Vision News”. Then we will try to generate Bangla caption of those videos. After that we will cut the video into word by word using video editing tool. Finally, we will convert those short video clips into multiple images.

## 3.2 STCNN

## Spatiotemporal convolutional neural networks (STCNNs) can process video data by convolving across time, as well as the spatial dimensions. It will extract the feature from video using filtering and spatial pooling.

# 4 References

[1] Yannis M. Assael1, Brendan Shillingford, Shimon Whiteson1 & Nando de Freitas, “Lipnet: end-to-end sentence-level lipreading”, 16 Dec 2016. [online]. Available at: <https://arxiv.org/abs/1611.01599> [Accessed 25- Sep- 2019].

[2] Stavros Petridis, Zuwei Li, Maja Pantic,” End-To-End Visual Speech Recognition With LSTMs”, 20 Jan 2017. [online]. Available at: <https://arxiv.org/abs/1701.05847> [Accessed 30- Sep- 2019].

[3] Stavros Petridis, Yujiang Wang, Zuwei Li, Maja Pantic,” End-to-End Multi-View Lipreading 1 Sep 2017. [online]. Available at: <https://arxiv.org/abs/1709.00443> [Accessed 30- Sep- 2019].

[4] Juergen Luettin, N.A. Thacker, Steve Beet, ” Speaker identification by lipreading”, 1996. [online]. Available at: <https://www.researchgate.net/publication/3703270_Speaker_identification_by_lipreading> [Accessed 01- Oct- 2019].

[5] Yao WenJuan, Yaling Liang, Du MingHui, ”A real-time lip localization and tacking for lip reading”, 2010. [online]. Available at: <https://ieeexplore.ieee.org/document/5579830> [Accessed 01- Oct- 2019].

[6] Warunee Nittaya, Kaskaew Wetchasit, Kittasil Silanon, “Thai Lip-Reading CAI for Hearing Impairment Student”, 2018. [online]. Available at: <https://ieeexplore.ieee.org/document/8523956> [Accessed 30- Sep- 2019].

[7] M. Z. Ibrahim, D. J. Mulvaney,” Geometry based lip reading system using Multi Dimension Dynamic Time Warping”, 2012. [online]. Available at: <https://ieeexplore.ieee.org/abstract/document/6410805> [Accessed 30- Sep- 2019].

[8] Nazia Hossain, Dilruba Shareen, Tarin Kazi,”Instant Bangla Speech to Text Conversion”, 2014. [online]. Available at: <https://www.researchgate.net/publication/326722992_Instant_Bangla_Speech_to_Text_Conversion> [Accessed 30- Nov- 2019].

[9] Nahid Akhter, Amitabha Chakrabarty, ” A Viseme Recognition System using Lip Curvature and Neural Networks to Detect Bangla Vowels”, 2016. [online]. Available at: <http://dspace.bracu.ac.bd/xmlui/handle/10361/8366> [Accessed 30- Nov- 2019].

[10] Shaheena Sultana, M. A. H. Akhand, Prodip Kumer Das, M. M. Hafizur Rahman, ”Bangla Speech-to-Text Conversion using SAPI”, 2012. [online]. Available at: <https://www.researchgate.net/publication/261312925_Bangla_Speech-to-Text_conversion_using_SAPI> [Accessed 30- Nov- 2019].  
[11] Noda, Kuniaki / Yamaguchi, Yuki / Nakadai, Kazuhiro / Okuno, Hiroshi G. / Ogata, Tetsuya (2014): "Lipreading using convolutional neural network", In *INTERSPEECH-2014*, 1149-1153. [online]. Available at:  
<https://www.isca-speech.org/archive/interspeech_2014/i14_1149.html>

[12] *Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6115-6119.* [online]. Available at:   
<https://www.semanticscholar.org/paper/Lipreading-with-long-short-term-memory-Wand-Koutn%C3%ADk/675419a4faaf71ab8178f79523308d3dd3392913>  
[13] Simonyan, Karen & Vedaldi, Andrea & Zisserman, Andrew. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. preprint. [online]. Available at:  
<https://www.researchgate.net/publication/259399927_Deep_Inside_Convolutional_Networks_Visualising_Image_Classification_Models_and_Saliency_Maps>  
[14] *Takashima, Y., Aihara, R., Takiguchi, T., Ariki, Y., Mitani, N., Omori, K., & Nakazono, K. (2016). Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss. INTERSPEECH.* [online]. Available at:  
<https://www.semanticscholar.org/paper/Audio-Visual-Speech-Recognition-Using-Bottleneck-a-Takashima-Aihara/563ed3af254a5ef4549a34155718df772ba7b8b6>  
[15] *amit, A.G., jnoyola, J.N., & sameepb, S.B. (2016). Lip reading using CNN and LSTM. .* [online]. Available at:  
<https://www.semanticscholar.org/paper/Lip-reading-using-CNN-and-LSTM-amit-jnoyola/5d80149e005894ab57f47e667f3e060e247d8e43>  
[16] *Hilder, S., Harvey, R., & Theobald, B. (2009). Comparison of human and machine-based lip-reading. AVSP.* [online]. Available at:  
<https://www.semanticscholar.org/paper/Comparison-of-human-and-machine-based-lip-reading-Hilder-Harvey/4217dfb670208ed33e2abb4587371c2de1315b55>  
[17] Almajai, Ibrahim & Cox, Stephen & Harvey, Richard & Lan, Yuxuan. (2016). Improved Speaker Independent Lipreading using Speaker Adaptive Training and Deep Neural Networks. 10.1109/ICASSP.2016.7472172.[online]. Available at:  
[https://ieeexplore.ieee.org/document/7472172/keywords#keywords](https://ieeexplore.ieee.org/document/7472172/keywords%23keywords)  
[18] Sui, Chao & Bennamoun, Mohammed & Togneri, Roberto. (2015). Listening with Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines. 154-162. 10.1109/ICCV.2015.26. [online]. Available at:  
<https://ieeexplore.ieee.org/document/7410383>  
[19] Thein, Thein & San, Kalyar. (2018). Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition. 123-127. 10.1109/INFOCT.2018.8356854. [online]. Available at:  
<https://ieeexplore.ieee.org/document/8356854>  
[20] Rahmani, Mohammad Hasan & Almasganj, Farshad. (2017). Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. 195-199. 10.1109/PRIA.2017.7983045. [online]. Available at:  
<https://ieeexplore.ieee.org/document/7983045>  
[21] Chung, Joon Son & Zisserman, Andrew. (2017). Lip Reading in the Wild. 87-103. 10.1007/978-3-319-54184-6\_6. [online]. Available at:  
<https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16/chung16.pdf>  
[22] Xu, Kai & Li, Dawei & Cassimatis, Nick & Wang, Xiaolong. (2018). LCANet: End-to-End Lipreading with Cascaded Attention-CTC. [online]. Available at:  
[https://www.researchgate.net/publication/323770816\_LCANet\_End-to  
End\_Lipreading\_with\_Cascaded\_Attention-CTC/citation/download](https://www.researchgate.net/publication/323770816_LCANet_End-toEnd_Lipreading_with_Cascaded_Attention-CTC/citation/download)

[23] Dave, Namrata. (2015). A LIP LOCALIZATION BASED VISUAL FEATURE EXTRACTION METHOD. Electrical & Computer Engineering: An International Journal (ECIJ). 4. 2015. 10.14810/ecij.2015.4403. [online]. Available at:  
<https://www.researchgate.net/publication/298133678_A_LIP_LOCALIZATION_BASED_VISUAL_FEATURE_EXTRACTION_METHOD/citation/download>  
[24] P. Sujatha & M. Radhakrishnan (2018). Mouth Segmentation Using Coordinate-Based Method for the Improvement of Visual Speech Recognition.[online]. Available at:  
<http://www.naturalspublishing.com/files/published/zw31qa4r4k7i14.pdf>  
[25] Nasib, Abdullah & Kabir, Humayun & Ahmed, Ruhan & Uddin, Jia. (2018). A Real Time Speech to Text Conversion Technique for Bengali Language. 1-4. 10.1109/IC4ME2.2018.8465680. [online]. Available at:  
<https://www.researchgate.net/publication/327817644_A_Real_Time_Speech_to_Text_Conversion_Technique_for_Bengali_Language/citation/download>  
[26] *Hasnat, M.A., Mowla, J., & Khan, M. (2007). Isolated and continuous bangla speech recognition: implementation, performance and application perspective.* [online]. Available at:  
<https://www.semanticscholar.org/paper/Isolated-and-continuous-bangla-speech-recognition%3A-Hasnat-Mowla/e6697a3bb8601ab533012473b6844b2121c7d7bd>  
[27] Chowdhury, Shammur. (2010). Implementation of speech recognition system for Bangla. 10.13140/RG.2.1.4658.4488. [online]. Available at:  
<https://www.researchgate.net/publication/277671513_Implementation_of_speech_recognition_system_for_Bangla/citation/download>  
[28] Ahammad, Khalil & Rahman, Md. Mahfuzur. (2016). Connected Bangla Speech Recognition using Artificial Neural Network. International Journal of Computer Applications. 149. 38-41. 10.5120/ijca2016911568. [online]. Available at:  
<https://www.researchgate.net/publication/308182498_Connected_Bangla_Speech_Recognition_using_Artificial_Neural_Network/citation/download>  
[29] Mukherjee, Sankar & Mandal, Das. (2014). A Bengali HMM Based Speech Synthesis System. [online]. Available at:  
<https://www.researchgate.net/publication/263163666_A_Bengali_HMM_Based_Speech_Synthesis_System/citation/download>