Department of Computer Science and Engineering
BRAC University

Pre-Thesis 1 Report

The Articulation of Speech: A Deep Learning based Lip and Voice Recognition

Submitted by:

Name	ID
Munia Shaheen	20101050
Akib Zabed Ifti	20101113
Junaed Hossain	20101196
Ariful Hassan	20301259

Supervisor: ***Dr. Muhammad Iqbal Hossain***

Co-supervisor: ***Mr. Rafeed Rahman***

Date of Submission: 25/5/2023

Table of Contents

Abstract

Understanding speech just through lip movement is known as lipreading. It is a crucial component of interpersonal interactions. The majority of the previous initiatives attempted to address the English lipreading issue. However, our goal is to build up a deep neural network for the Bangla language that can produce comprehensible speech from silent videos just by capturing the speaker's lip movements. Despite the fact that there is research on this topic in various languages, Bangla does not currently have a study or a suitable corpus to conduct research. Therefore, our research aims to investigate the performances of the state-of-the-art deep learning models on the Bangla corpus and build our own deep learning model suitable for Bangla lip-reading.

1 Introduction

Lipreading is the technique to comprehend what someone is saying only by their lip movement. Humans have long been able to read lips manually. However, automating lipreading is a significant goal since human lipreading performance is weak and constrained. Initially, systems for reading lips were created using traditional machine-learning techniques. Deep learning applications' prominence, particularly in recent years, has caused this topic to be studied more than in the past. The main components of automated lip-reading include face recognition, lip localization, feature extraction, classifier training with corpus, and word/sentence recognition by lip movement.

Nevertheless, due to challenges in the extraction of spatiotemporal features, generalization across speakers, and environmental noise, lip-reading is an extremely difficult task. Additionally, there are a lot of homophones, which implies that the same lip movement can produce a variety of different characters. Furthermore, some of the phonemes are created inside the throat and mouth and cannot be identified by simply observing a speaker's lips. Also, lip-reading from video without audio or text data depends on a variety of factors, including lighting, recording distance, and the speaker's speaking style. The endeavor becomes more challenging by external noises, mumbling sounds, and guttural sounds. Therefore, lipreading is a challenging task for both humans and machines.

Even though lip reading is a challenging technique to process, it is essential in a variety of fields, including accessibility, noise-sensitive communication, human-computer interaction, multimodal communication, security, and numerous more fields. Lip reading and voice recognition technologies can greatly improve accessibility for the deaf. Additionally, in noisy situations where speech could be hard to hear, such as crowded public spaces or industrial settings, lip reading and voice recognition technology might improve communication. These technologies enable individuals to understand speech even in challenging acoustic environments, improving the effectiveness and precision of communication in commonplace applications. Recent years have seen an increase in enthusiasm for studies

on the security of merging lipreading with biometric data, such as biological fingerprints or facial recognition. Studies that use lip movement monitoring for individuals include printing messages to smartphones in loud circumstances using visual data instead of auditory data and implementing various security measures utilizing visual silent passwords[27-29]. The applications of lip-reading technology are mentioned below:

- *Helping hearing − impaired people.*
- *Transcribing old silent movies or videos.*
- *Facilitating the alert for public safety.*
- *Improving audio in existing videos.*
- *Enabling video conversation in noisy or silent places, likelibraries.*
- *Identifying utilizing biometrics.*
- *Synthesizing simultaneous voice from*
- *multiple speakers. Enhancing automatic speech recognition's overall*
- *Dictating information or messages into a phone while surrounded*

2 Problem Statement

In today's world lip-reading is needed in various fields and both humans and machines are able to do some sort of lip reading. However, human lipreading performance is poor and limited. For example, studies showed that people with hearing impairments only attain an accuracy of 17±12% for a small subset of 30 monosyllabic words, and 21±11% for 30 complex words. Thus, automating lipreading is a key objective. While classical machine learning techniques can be used to read lips, they are slower and less accurate than deep learning techniques. These days, with the use of deep learning, it is liable to translate lip motions into relevant words.

Even though there have been several research on lipreading in English, currently, there is no known research on lipreading and speech recognition in the Bengali language. Even

there is no known corpus in Bengali suitable for this task. There are 265 million native and non-native speakers of the Bangla language worldwide, making it the fifth most widely used language. A significant majority of them are unable to speak English. Because of this, it will be challenging for them to use present technologies that require English. In our literature review, we will see most of the research used on English lip-reading becomes useless when trained on different language datasets. Therefore, research in lip reading and voice recognition in the Bengali language is crucial for the development of this field.

3 Research Objectives:

1. Creating a corpus that can be used in further research in the Bengali language.
2. Finding the feature extraction method for the Bangla corpus.
3. Examining the results of the current best models used in the English corpus using our Bangla corpus.
4. Establishing the ideal lip-reading model for the Bangla language.

4 Literature review:

In this literature review, we will be discussing the previous research in this domain of lip-reading. Most of the research used English speakers in their corpus. However, there are few types of research in German, Urdu, Korean, Chinese, and Turkish which gave state-of-the-art performance in their own language datasets. In our literature review, we will review lip-reading in English, Urdu, Chinese, Turkish, and German. As there is no known research on lip-reading in Bangla, we will review a few papers on Bangla voice recognition systems on the existing technologies.

4.1 Language: English

The first hand-segmented phone-based sentence-level lip-reading research[2] was released in 1997 and used Markov models (HMMs) in a small dataset. Later, using an HMM in conjunction with hand-engineered features in the IBM ViaVoice dataset, another researcher conducted the first sentence-level audiovisual speech recognition[3]. Additionally, the authors train an LDA-transformed version in an HMM/GMM ssystem[4] using the GRID corpus, outperforming the prior state-of-the-art with an accuracy of 86.4%. However, in these automated methods, extraction of motion information and generalization across speakers are both regarded as significant issues[5].

In automated lip reading, deep learning is needed as such tasks necessitate for intensive preprocessing for an image or video frame extraction or other forms of manually created vision pipelines[6][7]. However, previously deep learning is seldom used in lipreading and only performs classification, not sentence-level sequence prediction, in most research.

The paper [11], is the first known sentence-level lipreading model that uses the GRID corpus [12] consisting of 1000 sentences of audio and video by 34 speakers. In this paper, the DLib face detector, iBug face landmark predictor [13], and affine

transformation were used for preprocessing. Here, they introduced a new architecture called, LipNet. In the LipNet architecture, the input is processed by 3 layers of STCNN (Spatiotemporal convolutional neural networks) which is followed by a spatial max-pooling layer. Two Bi-GRUs then follow the retrieved features. Then at each time step, a linear transformation is applied before performing a softmax over the vocabulary augmented by the CTC (Connectionist Temporal Classification) blank, followed by the CTC loss. The authors evaluated the performance of the LIPNET model using the word error rate (WER) and character error rate (CER). The LipNet model is the first sentence-level lip reading which gives 95.2% accuracy at the sentence level outperforming a human lipreading baseline, and exhibiting better performance than the word-level state-of-the-art in the GRID corpus. However, this paper has some limitations, the model only works for sentences consisting of 6 words and to improve their research they need larger datasets.

The paper [14], used the same GRID corpus[12] used in the paper LipNet[11]. To begin with, they trained an autoencoder using a 128-frequency bin auditory spectrogram of the training audio files. A CNN-LSTM architecture is connected to a decoder of a pre-trained autoencoder in order to reconstruct the auditory spectrogram, while the main network uses a 7-layer 3D convolutional structure to extract spatiotemporal information from the input video sequence. The core lip reading network was trained with batches of 32, and data augmentation was carried out by either flipping images horizontally or adding reasonable amounts of Gaussian noise to the data. For evaluation, they compared their result with Vid2Speech[15] in terms of PESQ(Perceptual Evaluation of Speech Quality)[16], Corr2D, and STMI(Spectro Temporal Modulation Index)[17] and gave better results. This paper showed ways to improve the audio quality for better prediction with a 98% correlation. However, they only used the GRID dataset, which contained only 6 sentences with a small word probability, to evaluate their model. For future work, more train data needs to be gathered, speech reconstructions need to include emotions, and an end-to-end framework needs to be built.

Next, in the paper [30], the author introduces a novel method for text identification from a silent lip movement video where they used the GRID corpus[12]. Here, A different technique for data augmentation is applied to the dataset for each epoch: by randomly inserting photos from the dataset video and adding modest quantities of Gaussian noise, we can simulate real-world conditions. The auditory MFCC features are converted in the architecture to a variable-length sequence of video frames via the visual-to-audio feature architecture. Second, the text information is distinguished from the audio feature by the architecture of the audio feature to text. A 7-layer 3D convolutional network (CNN) is used to recover the spatial and temporal characteristics of the video stream. Their research revealed that the model suggested in this article is capable of 92.76% validation accuracy. This paper has the same limitations as the first two papers as they used GRID corpus.

This paper [31], proposes a deep-learning technique for speech enhancement. A well-established GRID [12,31] and ChiME3 [31,32] corpus have been used as datasets for this method. The authors used a speech enhancement framework to extract audio features from a noisy speech by Enhanced Visually-Derived Wiener Filter(EVWF). Then they enhanced this wiener filter by proposing LSTM-based filter bank estimation. After that, they worked on a regression model for lip reading where they used LSTM & MLP. From a video, they extracted audio and frames, and with the help of hamming window and object tracker, lip cropper they managed to get audio and video features successfully. With one frame the LSTM model MSE of 0.092. While, with 18 frames MSE of 0.058 has been achieved. Again, With one frame MLP-based lip reading model achieved MSE of 0.199 only and 0.209 MSE while working with eighteen frames. LSTM can safely find out audio features which are clean but vanilla neural network models can not accomplish this task as smoothly as LSTM. At low SNR the model can give a standard pl-score but at high SNR, it outperforms but still works better than conventional speech enhancement

approaches. However, the result and accuracy could be much higher, a context-aware AV algorithm can be used also and the dataset is much smaller to train as they have selected fewer frames while training.

The research [37] conducted small research where the two steps of the traditional lip-reading method are feature extraction and categorization. In this paper[37], they did not mention the dataset, however, they used Speech recognition using a Convolutional Neural Network. This model is trained to operate on large amounts of multi-dimensional data. After that, Dynamic lip videos, CNN, HMM, and decoding are used to extract features and appropriate words from HMM models. To compare artificial neural networks for voice identification, CNN and HMM were utilized for visual feature extraction and frame-level features, respectively. With an accuracy rate of 80%, the suggested architecture predicts words from a series of photographs of the lip region. RNN has certain drawbacks like explosive issues and it cannot process for a long sequence, which is the main restriction of this paper. RNN training is a challenging task as well. Research could lead to a speaker-independent recognition system in the future.

In this paper [22], the authors worked only with the micro-content of the English language which is the alphabet. They built the dataset having 20 letters and more than 2700 recorded videos which are a maximum of 2 seconds long. They prepared the data by converting videos into frames, detecting human faces, detecting mouths from the faces, and selecting the keyframes. The model used in this paper is CNN and stochastic gradient descent (SGD) where CNN is used for recognizing letters and also extracting features. In this paper[22], VGG19 pre-trained CNN has been used containing sixteen convolution layers, five max-pooling layers, and more than two fully connected layers. Finally, after testing, models can predict those twenty alphabets 95% of the time for the testing set whereas an accuracy of 98% for the testing set. Letters that have similarities with other letters were conducted with an accuracy of less than 99% and the letters of distinguished features managed to have over 99% accuracy. The authors only worked with letters only,

word or sentence prediction from videos was not in their concern in this dataset. Also, the dataset is much smaller for training which consisted of just 20 alphabets but at the same time they worked with so many videos so this could be much more in number.

In contrast to earlier published approaches for lip reading that are based on sequence-to-sequence architectures, a research paper[38] based on Deep Convolutional Neural Networks produced a hybrid lip-reading (HLR-Net) model from a video. In this paper GRID dataset[12] is used to conduct sentence-level lip reading. Here, HLR-Net was compared to LipNet, LCANet, and A-CTC models for unseen and overlapped speakers. The suggested model is composed of a CNN model, an attention layer, and a CTC layer. The proposed method is based on the attention deep learning model and converts a video of lip movement into frames using OpenCV before extracting the mouth component with the Dlip package. While the decoder employs attention, fully connected, activation function layers, the encoder makes use of inception, gradient, and bidirectional GRU layers. With a CER of 4.9

The authors of [23] attempted to compare all currently available deep learning architectures and additional techniques offered by other researchers in order to determine which could offer the best performance. The suggested research project uses the MIRACL-VC1[24] dataset, which contains 3000 occurrences of words and phrases, with 150 examples in each folder. 10 words and 10 phrases are spoken by five men and ten women in each piece of data and each word/phrase was spoken 10 times. The Dlib facial detector and Dlib mouth detector were used in this research to extract the image's mouth and identify the speaker's face. AlexNet, VGG16, Inception V1, ResNet50, Auto-Encoder, Q-ADBN, and EfficientNetBO are used to train the model. EfficientNetBO will serve as the model's foundation in this design, to which the Attention block and LSTM layer are added to produce a finer performance. The research had an accuracy of 80.133% using EfficentNetBO and 86.67% with EfficientNet and Attention Block. Finally, using LSTM on top of that gives an accuracy of 91.13%. For their future work, lip movement detection

can be improved to include additional words and phrases from different speakers.

In this paper [25], the authors developed their model by extracting the mouth area and segmenting the area from a frame by using a hybrid model. For this, they have used the LRW dataset [26] from BBC TV broadcasts having audiovisual speech segments. Each video has 29 frames from which 22 frames have been selected. Their suggested method of segmentation has been split into two different stages. First, detecting the mouth and then applying an improved hybrid model. The models they used for lip reading are Bi-GRU which has two hidden layers and 2D CNN. First, they wrap the entire CNN input model in the TD function, then pass it into the Bi-GRU layer for extracting comprehensive information from previous layers' characteristics. Afterward, the outputs have been infiltrated into a pooling layer, and then all the outputs are connected into a softmax function to produce every word's possibilities. Moreover, by using ReLU activation functions two convolutional layers have been added. That's how they omit the problem of vanishing gradient. After 6 epochs loss approaches 0 in training data and 0.4 in testing data. On the other hand, with segmentation lip reading gives 90.385% accuracy and without segmentation, it gives 85% accuracy. So, their proposed architecture is doing great work on classifying lip motion sequences on words. However, in this paper, they have not worked with sentences. Also, the authors have not mentioned the sequence of the words.

4.2 Language: Urdu

Few studies on lip reading in other languages have been published. The paper[18] shows lip-reading for the Urdu language that uses a corpus containing ten words. As there was no available corpus for Urdu lip-reading, the authors created a corpus of 10 words repeated 10 times by 10 participants. The recorded videos are preprocessed by cropping them to a standard size, applying the Viola-Jones face detector, and a mouth detector,

and then cropping each video. To train the dataset, at first, they used LipNet[11], which failed to train due to CTC loss. Then they used RNN and LSTM to train the model on Urdu words and digits and got better accuracy from the LSTM model which is 62 % and 72 % respectively for words and digits. To show how effectively audio-visual lipreading performed in noisy conditions, the researchers in this study individually trained two different models for audio and video. However, they were unable to combine both networks. Additionally, they presented a small corpus of Urdu words which worked as a classification task with no sequence. Finally, they hope to create a model in the future that can predict text sequence and is akin to lipreading.

4.3 Language: Chinese

The first sentence-level Chinese lipreading was introduced in the paper [19]. In this study, they create their own dataset from CCTV News and Logic Show, labeling it with Hanyu Pinyin (a phonetic interpretation of Chinese), and end up with 349 classes and 1705 characters. They present a unique two-step network for sentence-level Mandarin lipreading where predicting the probability of Hanyu Pinyin is the first step. A max-pooling layer is applied after a 3D convolution that suggests a series of frames as input. Then, more visual features are extracted using DenseNet. The visual features are processed using a two-layer resBi-LSTM, which is subsequently accompanied by linear and softmax layers. The complete network is then trained using the CTC loss function. The second phase involves using a stack of multi-head attention to translate Hanyu Pinyin into Chinese characters. The suggested network has an absolute 13.91% and 14.99% rise in Hanyu Pinyin and Chinese character accuracy compared to LipNet[11], whereas ResNet[20] technique has an increase of 4.68% and 4.74%. The problem that remains is that for different words Hanyu Pinyin might be the same, but Chinese characters would be different due to the different contexts which indicates that Hanyu Pinyin is unable to capture context.

4.4 Language: Turkish

In the paper[21], two new datasets were created for the Turkish language, one with 111 words and the other with 113 sentences, using image processing techniques. In the paper[21], the Media-pipe framework determines the lip position and removes it from the image. Using a convolutional neural network (CNN), features are first retrieved to perform lip detection and then the classification process is completed using bidirectional long short-term memory(Bi-LSTM). The results of the experiment demonstrate that, for words and sentences, respectively, ResNet-18 and Bi-LSTM pair offers the best results, with accuracy scores of 84.5% and 88.55%. However, the limitation of the paper is that each speaker is positioned at 1.5 m in these photos, and the video clips taken to generate the dataset were captured in identical lighting and ambient conditions.

4.5 Language: German

In this paper[33], the researchers worked on the German language. They took the "Lip Reading in the Wild" dataset created by Chung Zisserman(2016). First of all, they used English for comparison and transfer learning. Then, they brought up the issue of copyright and data protection in Germany. The dataset consists of MPEG4 format videos which are 1.16 seconds long. In the processing part, they cropped the videos into 96*96 pixels only focusing on the lips of the speaker using the Glips model. As the videos have audio too, the video and audio files were stored in separate files and synchronized in the last step. For the subtitle part, they used WebMaus to synchronize with the audio. They did two experiments in this paper[33]. The validation accuracies of GLips and LRW in experiment 1 are respectively 78.4% and 77.8%. The validation accuracies of the transfer-learned models were 82.2% and 81.6% respectively. The LRW network's average validation accuracies in experiment 2 were 79.2% and 80.4%, respectively. The average validation accuracies of the GLips networks were 78.4% and 79.6%, respectively. The

fact that learning was successfully transferred between the two languages suggests that there are aspects of lip reading in both datasets that are linguistically independent and can be applied to different languages. We assume that the difference between the models trained on GLips and LRW lies in the quality of the data because the LRW-trained networks perform better.

4.6 Language: Bangla

In the Bangla language, there are numerous words or symbols that sound remarkably similar. In this paper[34], they investigated the DeepSpeech network for recognizing individual Bengali speech samples. To overcome this, they used ASR which is an algorithm to detect individual characters, words, and sentences. In this paper[34], almost two thousand speech samples which are mostly Bengali real numbers, and a total of 115 unique words are distributed around the data samples. The authors first tried to approximate the phone alignment in individual speech samples using a statistical approach to build their input set. Then, they fed these input samples to a two-layer unidirectional LSTM network and trained their model. But the problem here is recognizing words instead of full sentences. At first, to remove all the noise and distortion, they used Fourier transform on all individual audio samples and extracted highly 13 distinctive nonlinear mel frequency cepstral coefficients (MFCCs) from them. Next, they used Bi-Directional LSTM with CTC loss function for training their data and tested a Bengali real word speech dataset which gives an 8.20% WER and a 3% CER. The limitation of the paper is that it only works on word-level detection, each data sample has eight examples in the dataset and the dataset is very small.

In this paper[35], the authors suggested a convolutional neural network (CNN) architecture for Bangla short speech detection. As the dataset for Bangla is not widely available,

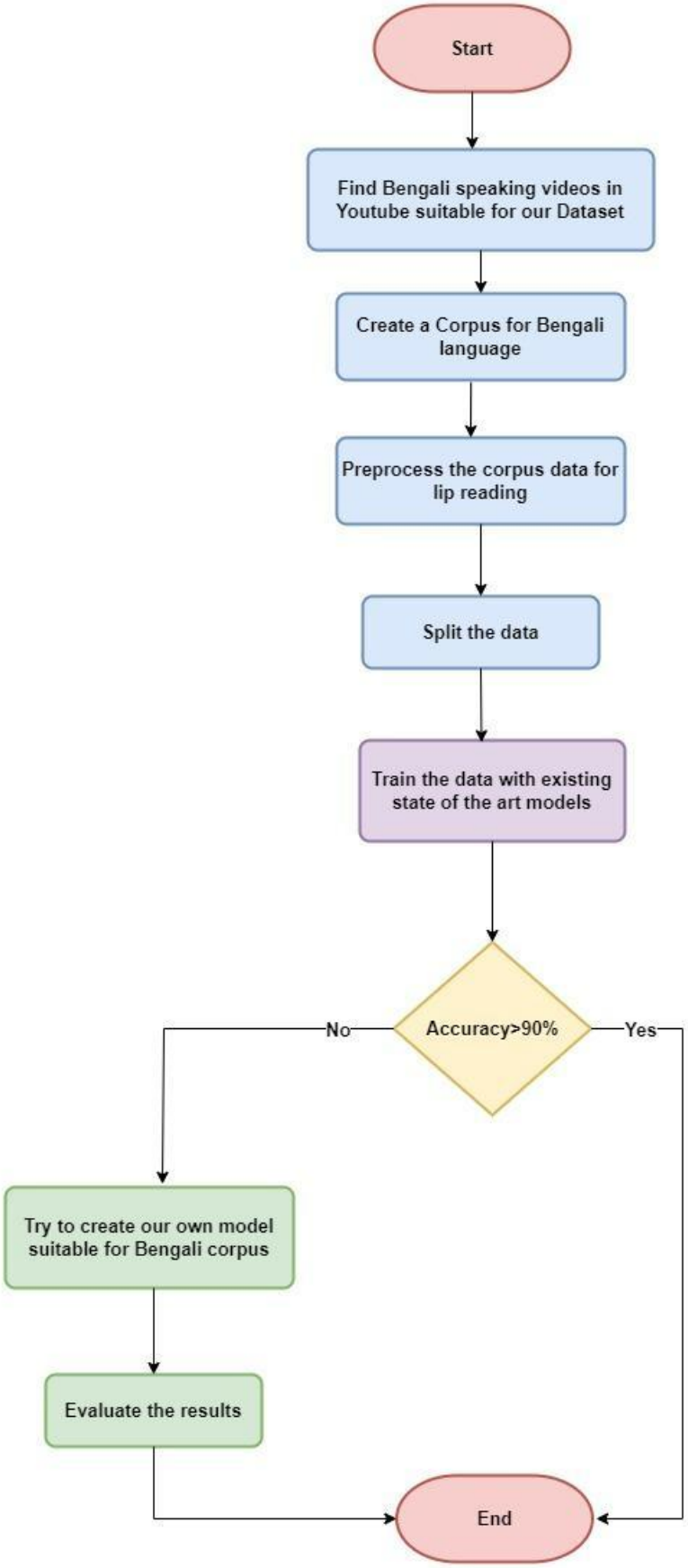
they made their own dataset for Bangla short speech commands. Their dataset contains 65,000 samples, considering one second for each word. They took Banglish as their data and then converted them to proper Bangla. They normalized their inputs before training their model with MFCC inputs. They then fed the raw audio data into a similar CNN architecture, which they call the raw model. Prior to training their data, they also pre-trained their model. They used three models for training and testing their dataset. The models are: MFCC, Raw, and Transfer and their training accuracies are respectively 85.44%, 69.08%, and 68.06%. Finally, their testing accuracies are respectively 74.01%, 71.44%, and 73%. In comparison to deep neural networks (DNN), they observed in their research that CNN architecture can lower the error rate on the TIMIT phone recognition dataset by 6-10%. Additionally, they conducted several tests utilizing the limited weight sharing (LWS) and full weight sharing (FWS) schemes and reported that LWS is more efficient since it can recognize feature patterns in many frequency bands. In this research, the main limitation was the dataset was relatively small. Finally, they proposed to expand the words in the dataset in the future so that they can find a better result from the used model MFCC.

In the research [36], they took the dataset from Google Bangla Speech Corpus which has a total duration of 229 hours with 508 native speakers and 50k unique words. Also, test Corpus: 10.6 million unique words and 602.5 million global words, 48.56 million sentences. In this paper[36], they claimed that their CTC-based Bangla ASR end-to-end deep CNN system surpasses the RNN-based DS2 system. For the first time in Bangla ASR research, they used deep learning techniques to analyze character-wise mistake rates in order to evaluate the corpus's quality. The CNN-based models, which frequently use acoustic characteristics in voice recognition and speaker identification, have been built using MFCC features. They used CTC-CNN and Deep Speech 2 models to train their data. The Deep Speech 2 model is an RNN -CTC-based model. The word error rate in the CNN-CTC model is 36.12% in the Deep Speech 2 model is 40.91%. The advantages of their model are they have a very large dataset, the error rate gets 10% when the

characters are found more than 100k times, and characters that are used frequently and have better performance. However, there are some limitations too. Characters found less than 1k times, found 50% error. The characters found less than 1k times, found 50% error. The letter 'R' arrived only 7 times. In this situation, the DS2 model gets 71.4% accuracy. On the other hand, the CNN model couldn't recognize the word at all. The same pronounced words like 2% cannot be differentiated by this model. The 10 Bangla digits are not very often used in this dataset. So, it has less accuracy than others. In the future, they will work with a large high-order N-gram model to improve Bangla LVCSR.

5 Working Plan:

Our study seeks to evaluate the effectiveness of the latest deep learning algorithms on the Bangla corpus and develop our own deep learning model appropriate for lip-reading Bangla. For this, at first, we need a Bangla corpus for lip-reading. As there is no known Bangla corpus for lip-reading, we will collect some Bangla videos from youtube suitable for our task and make our corpus. Secondly, we will use preprocessing to extract the features for our task. After that, we will split our dataset into train and test data. To train our data, we will use some of the models with the highest success used in lip-reading in other languages, in our own dataset. If those architectures give an accuracy of less than 90, then will try to build a unique model which might provide better accuracy for lip-reading in Bangla.


Start
Find Bengali speaking videos in Youtube suitable for our Dataset
Create a Corpus for Bengali language
Preprocess the corpus data for lip reading
Split the data
Train the data with existing state of the art models
Accuracy>90%
No
Yes
Try to create our own model suitable for Bengali corpus
Evaluate the results
End

6 Conclusion:

In conclusion, lipreading is the skill of understanding a speaker's words solely from the movement of their lips which is crucial in many different disciplines like accessibility, noise-sensitive communication, human-computer interaction, multimodal communication, security, and a great deal more. Despite the fact that the English language has been the subject of numerous studies in this area, the Bangla language has not yet been the subject of any studies. CNN, Bi-LSTM, Bi-GRU, and CTC were frequently employed for English lipreading. However, after reviewing some studies, we discovered that the lipreading models utilized for English were inadequate for other languages. Therefore, it is imperative that we carry out research on Bangla lip-reading. However, there isn't yet a corpus of Bangla that can be used for lipreading. In our study, we'll build our own corpus and train cutting-edge lip-reading methods for English to test how well they function on Bangla datasets. Finally, we'll attempt to create a custom model that works with Bangla lip reading.

References:

1. D. Easton and M. Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32(6): 562-570, 1982.

2. J. Goldschen, O. N. Garcia, and E. D. Petajan. Continuous automatic speech recognition by lipreading. In Motion-Based recognition, pp. 321-343. Springer, 1997.

3. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.

4. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa. Dynamic stream weighting for turbodecoding-based audiovisual ASR. In Interspeech, pp. 2135-2139, 2016.

5. Zhou, G. Zhao, X. Hong, and M. Pietikainen. A review of recent advances in visual speech decoding. "Image and Vision Computing, 32(9):590-605, 2014.

6. 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, 24(2):198-213, 2002.

7. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7):1254-1265, 2009.

8. S. Chung and A. Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision, 2016.

9. Wand, J. Koutnik, and J. Schmidhuber. Lipreading with long short-term memory. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6115-6119, 2016.

10. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

11. Assael, Y. M. (2016, November 5). LipNet: End-to-End Sentence-level Lipreading. http://arXiv.orgarXiv.org. https://arxiv.org/abs/1611.01599https://arxiv.org/abs/1611.01599.

12. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421-2424, 2006.

13. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In IEEE International Conference on Computer Vision Workshops, pp. 397-403, 2013.

14. Akbari, H. (2017, October 26). Lip2AudSpec: Speech reconstruction from silent lip movementsvideo.http://arXiv.orgarXiv.org.https:/arxiv.org/abs/1710.09798 https://arxiv.org/abs/1710.09798

15. Ariel Ephrat and Shmuel Peleg, "Vid2speech: Speech reconstruction from silent video," arXiv preprint arXiv:1701.00495, 2017

16. AW Rix, J Beerends, *M* Hollier, and A Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation, vol. 862, 2001. 17. Mounya Elhilali, Taishih Chi, and Shihab A Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," Speech communication, vol. 41, no. 2, pp. 331-348, 2003.

17. Faisal, M. (2018, February 15). Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language. http://arXiv.orgarXiv.org. https://arxiv.org/abs /1802.05521https://arxiv.org/abs/1802.05521

18. Chen, X., Du, J., & Zhang, H. (2020). Lipreading with DenseNet and resBi-LSTM. Signal, Image and Video Processing, 14(5), 981-989. https://doi.org/10.1007/s11760-019-01630-1https://doi.org/10.1007/s11760-019-01630-1

19. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lipreading. In: conference of the international speech communication association, pp. 3652 − 3656 (2017)

20. Atila, Ü., & Sabaz, F. D. (2022). Turkish lip-reading using Bi-LSTM and deep learning models. Engineering Science and Technology, an International Journal, 35, 101206. https://doi.org/10.1016/j.jest ch.2022.101206https://doi.org/10.1016/j.jestch.2022.101206

21. Ali, N. H. M., Abdulmunem, M. E., & Ali, A. E. (2021). Constructed model for micro-content recognition in lip reading based deep learning. Bulletin of Electrical Engineering and Informatics, 10(5), 2557-2565.https://doi.org/10.11591/eei.v10i5.2 92https://doi.org/10.11591/eei.v10i5.292

22. A Novel Method for Lip Movement Detection using Deep Neural Network. (2022). Journal of Scientific & Industrial Research, 81(06). https://doi.org/10.56042/jsir.v8 1i06.53898https://doi.org/10.56042/jsir.v81i06.53898

23. Afouras T, Chung *JS* & Zisserman A, Deep Lip Reading: a comparison of models and an online application, ArXiv, (abs/1806.06053) (2018)

24. Miled, M., Messaoud, M. B., & Bouzid, A. (2022). Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications, 82(1), 551-571. https://doi.org/10.1007/shttps://doi.org/10.1007/s 11042-022-13321-0

25. Chung *JS*, Zisserman *A* (2016) Lip reading in the wild. Asian Conference on Computer Vision, pp 87-10

26. L. McQuillan, Is lip-reading the secret to security?, Biometric Technol Today 2019 (2019) 5-7, https://doi.org/10.1016/S0969-4765(19)30085-2https://doi.org/10.1016 /S0969-4765(19)30085-2.

27. F.S. Lesani, F.F. Ghazvini, R. Dianat, Mobile phone security using automatic lip reading, in: 9th Int Conf e-Commerce Dev Ctries With Focus e-Business, ECDC 2015, 2015, https://doi.org/10.1109/E
CDC.2015.7156322https://doi.org/10.1109/ECDC.2015.7156322.

28. A. Hassanat, Automatic lip reading for security, in: 1st Mosharaka International Conference on Biomedical Engineering, Electronics and Nanotechnology (MIC-BEN 2011), 2011, pp. 11-16

29. Text Recognition from Silent Lip Movement Video. (2018, July 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/8600535 https://ieeexplore.ieee.org/document/8600535 31. Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. M. (2021). Lip-Reading Driven Deep Learning Approach for Speech Enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(3), 481-490. https://doi.org/10.1109/tetci.2019.2917039https://doi.org /10.1109/tetci.2019.291703 9

30. J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015, pp. 504-511.

31. Schwiebert, G., Weber, C., Qu, L., Siqueira, H., & Wermter, S. (2022). A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning. arXiv (Cornell University). https://doi.org
/10.25592/uhhfdm.10047https://doi.org/10.25592/uhhfdm.10047

32. Purkaystha, B., Nahid, M. M. H., & Islam, M. S. (2019). End-to-End Bengali Speech Recognition using DeepSpeech. ResearchGate. https://www.researchgate.ne t/publication/337940431$_{E}nd - to - End_{B}engali_{S}peech_{R}eco$

33. Bangla Short Speech Commands Recognition Using Convolutional Neural Networks. (2018, September 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/8554395
https://ieeexplore.ieee.org/document/8554395

34. Samin, A. M., Kobir, M. H., Kibria, S., & Rahman, M. M. (2021). Deep learning based large vocabulary continuous speech recognition of an under-resourced language Bangladeshi Bangla. Acoustical Science and Technology, 42(5), 252-260. https://doi.org/10.1250/ast.42.252https://doi.org/10.1250/ast.42.252

35. Soundarya, B., Krishnaraj, R., & Mythili, S. (2021). Visual Speech Recognition using Convolutional Neural Network. IOP Conference Series: Materials Science and Engineering, 1084(1), 012020. https://doi.org/10.1088/1757-899x/1084/1/012020 https://doi.org/10.1088/1757-899x/1084/1/012020

36. Sarhan, A. (2021). HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural

37. Soundarya, B., Krishnaraj, R., Mythili, S. (2021). Visual Speech Recognition using Convolutional Neural Network. IOP Conference Series: Materials Science and Engineering, 1084(1), 012020. https://doi.org/10.1088/1757-899x/1084/1/012020

38. Sarhan, A. (2021). HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural Networks. https://www.semanticscholar.org/paper/HLR-Net