Automatic Lip Reading: Classification of Words and Phrases using Convolutional Neural Network

Nikhil Kesarkar[1], Poornachandra P. Kongara[2], Manthan Kothari[3],Asst. Prof. Suresh Mestry[4]

[1]*B.E. Computer Science, Rajiv Gandhi Institute of Technology, Mumbai University, Maharashtra, India*
[2] *B.E. Computer Science, Rajiv Gandhi Institute of Technology , Mumbai University, Maharashtra, India*
[3] *B.E. Computer Science,* Rajiv Gandhi Institute of Technology, *Mumbai University, Maharashtra, India*
[4] *Assistant Professor, Dept. of Computer Science Engineering,* Rajiv Gandhi Institute of Technology, *Mumbai University, Maharashtra, India*

---***---

Abstract - *Previous studies on human-machine interaction have determined that visual information can augment the speech recognition accuracy especially in noisy surroundings. Here we show a model for predicting words from video data without audio. Although already existing models have succeeded in incorporating visual data into speech recognition, all of them contained some or the other deficiency. To overcome this, we have pre-processed the data by using the haar-cascade model [2] to detect and crop around the subject's face in all frames of the video data and then use the sequence of frames as input to the model. Our proposed lip-reading classification model is unique in its usage for all ranges of speakers.*

Key Words: **deep learning, automatic lip-reading, machine learning, Convolutional neural network, Image processing.**

1. INTRODUCTION

Lip-reading is basically understanding lip movement in order to understand the spoken words when there is no access to sound. Human beings identify speeches of a speaker using multimodal information. Besides speech audio, visual data such as lip and tongue movement can also aid in understanding speech. By using visual information that is watching lip movements and understanding what a subject is saying is referred to as lip-reading. Lip-reading can help to understand speeches especially in noisy environments or in environments with no noise. This paper probes into the task of recognizing speech from video data without audio. We propose a neural network with a definite amount of success in this classification task. The input which we provide to our algorithm is a sequence of images taken from a video data. We have used a Convolution neural network (CNN) model to get the output of one of 10 words and 10 phrases that are spoken by the subject.

Given that video traffic is growing at a pretty steady rate on the internet, the proposed model could help extract the data and process it to find missing information or gain insights into the action or topics that occurs in the video. Applications of lip-reading classification include profanity detection on social media sites to a live lip-reading mobile application.

In the past also, several researches and efforts have been made but they were more focussed on gesture recognition, making this project a relatively and very exciting field to explore. Also, there are few systems and applications out there for lip-reading but most of them do not use neural networks but instead they have used different machine learning techniques. More advanced visual speech recognition models such as Google's DeepMind Lip-Net[3] network were published only a few months ago.

2. RELATED WORKS

This section consists of the existing work that has been done in the field. Machine Learning methods are mostly used in many of the approaches, the concept of deep learning has emerged in the last few years. The Effect of Coarticulation and visual features diversity are two main challenges involved in the lip-reading process.

This paper [6] developed different methods for prediction of words and phrases from videos which doesn't include the presence of audio files. They also discussed that visual lip reading process is important in Human Machine Interaction that could replace the audio speech recognition technology as the machine finds it difficult in noisy environments & also due the different accents of the people. Researchers used a fixed number of images to concatenate on the pre-trained VGG model. The nearest neighbour interpolation method was used to normalize the sequence of images and the extracted features are fed to the LSTM & RNN by VGG-net for word classification.

Rathee [7] called recognition of lip movement patterns while speaking as lip-reading .Speech recognition system faces major problems due to noisy environments and it will add as a help for the physically hearing impaired for communication with normal people .The algorithm proposed passes by two main steps which include extraction of features and word classification. The extraction of features passes from five steps: Video Acquisition, Face and Mouth Detection, Intensity Equalization, Key point Extraction and Geometric Feature Extraction. The classification of words is done using Learning Vector Quantization neural network.
In the lipreading system of [8], proposed a convolutional neural network-based feature extractor. The speaker's

mouth region images along with the phoneme labels are used for network training. In this paper six different speakers with six different independent CNN models are used because of which the average phoneme recognition is 58% for 40 phonemes over 6.

In [9], the processing pipeline is based purely on neural networks which succeeds the lip reading process. A single structure is formed by stacking Long Short Term Memory. The inputs are in the form of raw mouth images, the performance of such network is experimentally calculated and compared to a standard SVM classifier. Evaluation are performed on data from 19 speakers of Grid corpus. Using end to end neural network-based architecture The best word accuracy reported from this paper is 79.6%.

In [10] this paper the authors proposed a method for Deep Convolutional Neural Network classifier along with VGG modal. The Vid TIMIT database is used which include 43 people speaking short sentences with head movements and suitable amount of delay. Accuracies for different classifiers used are FRS-93.53%, ARS-80.66%, Fusion-97.33%.

In [11] authors proposed a methodology such that deep belief network along with conventional HMM and VPR model is used. The dataset used is CUAVE a dataset is used which includes 36 speakers uttering over 7000 connected and isolated digits.

Accuracies for different classifiers used are PER-69.36%. When VPR is used accuracy increases to 45.63%.The advantages of this method is that it is simple as well as fast for interpreting and accurate results are produced when input is similar to the dataset . Accuracy for the system is very low for any of the classifiers used and Prediction becomes difficult when input is different in large amounts from the used dataset.

3. DATASET AND FEATURES

MIRACL-VC1 data set [1] is the dataset used here that consists of 15 speakers including 10 women and 5 men and they utter ten times a set of ten words and ten phrases.

The data set is a lip-reading data set which consists of both depth and colour images. It can be used for a variety of research fields like visual speech recognition, face detection, and biometrics. The dataset consists of a synchronized sequence of colour and depth images for each instance (both of 640x480 pixels) which was collected at 15 frames per second. For each word or phrase, there are approximately 4 to 27 image frames. The data set contains a total number of 3000 instances (15 speakers uttering 10 words and 10 phrases, 10 times i.e. 15*20*10).

The words and phrases in dataset are shown in Table1.

Words:	Begin, Choose, Connection, Navigation, Next, Previous, Start, Stop, Hello, Web
Phrases:	Stop Navigation, Excuse me, I am sorry, Thank you, Good bye, I love this game, Nice to meet you, You are welcome, How are you? , Have a good time

Table 1: Words and Phrases in dataset

In addition to the dataset present, we added 2 more speakers who utter ten times the same set of ten words and ten phrases to the dataset so that the model gets more data for training and validation. So in total we have 3400 instances(17 speakers * 20 object * 10 times). We split the dataset into two: training and testing. We included 80% of each speaker i.e. overall 2720 for training and 20% of each speaker i.e. overall 680 for testing.

4. METHODOLOGY

4.1 Pre-Processing

4.11 Face Extraction

Pre-processing was an important task performed on the data set. Our Pre-processing was to get the facial part of each speaker. This step was done with python library dlib and Open CV along with the use of a pretrained model i.e. haar cascade model[1] and all the points of the facial structure were obtained then we used these points to crop the speaker's face as shown in Figure 1.

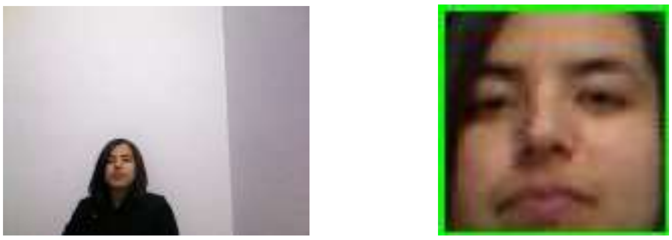

Figure 1: Elimination of background & Extraction of face

4.12 Concatenation and Resizing

Once we get the face of all the speakers, we concatenate them into a single image as shown below. This will help to get all the faces of speakers uttering a word in a single frame. Now this concatenation is as follows:

if there are 6 frames then we added 3 images in 1st row and 3 in 2nd row as shown in Figure 2.

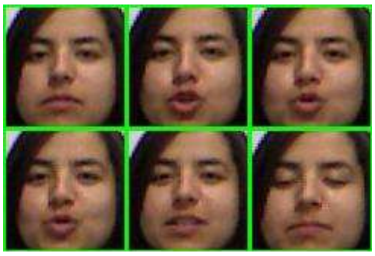

Figure 2: Concatenation of 6 Frames

If we have 7 frames then we added 6 frames as above and 7th frame in 3rd row. Now to fulfil this row completely we added 7th frame again twice in the 3rd row as shown in Figure 3.

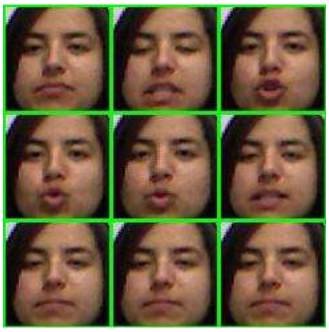

Figure 3: Concatenation of 7 Frames

Due to concatenation, the size of individual concatenated frames varies. So we have to reshape them into one single size for all frames. For our model, we set size to (224,224).

4.2 CNN implementation

In this section we described the implementation of the CNN model that we used for classification of the words and phrases which were included in the dataset. Our CNN model consists of 4 convolutional layers and 2 fully connected layers. The two fully connected layers used a soft max activation function layer to produce the probabilities of each and every word and phrase of which we will take the highest. The CNN model structure is as shown in Figure 4.

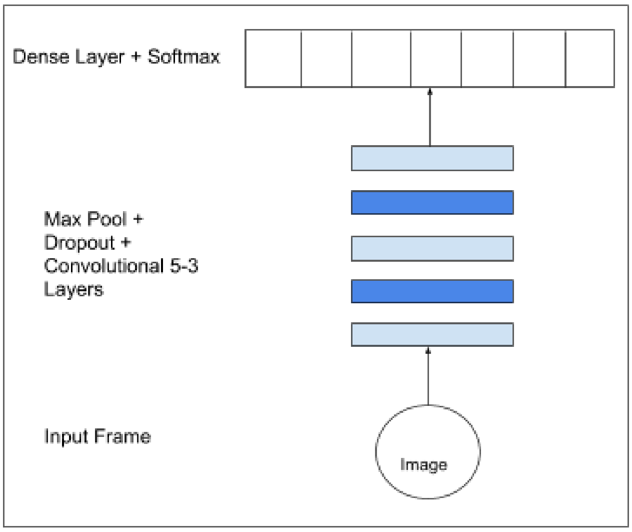

Figure 4: CNN Model Architecture

Each individual concatenated frame is fed to the Convolutional Neural Network.

Once we complete training and validation, the CNN model is saved to disk for prediction.

4.3 Workflow

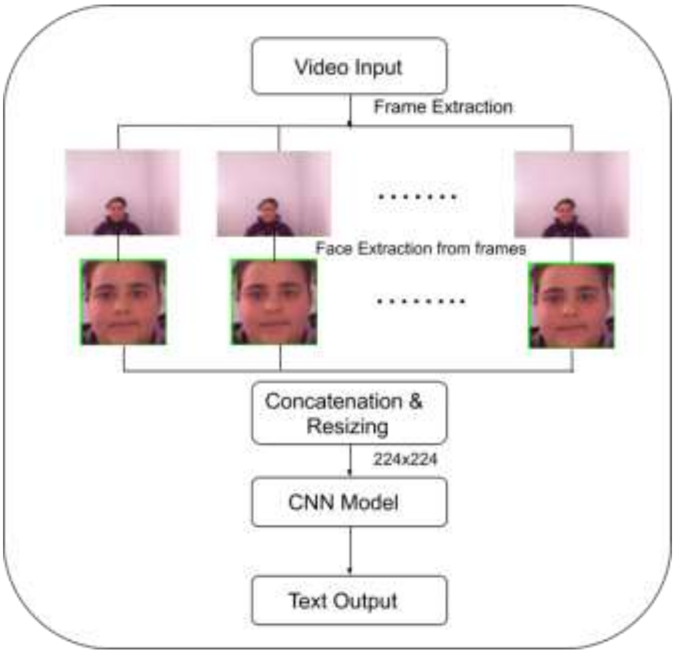

Figure 5: Workflow

Figure 5 explains the workflow. The prediction for lip reading involves the input in the form of video from which the frames are extracted. The frame rate we set is 15 fps. The face is augmented from the frames extracted using haar-cascade model[9].

Concatenation and resizing of the images is applied to the augmented frames of faces. Then the concatenated images are resized to 224x224 pixels.

The resized images are then given as input to the CNN model for classification. The result obtained is the classified word or phrase.

5. RESULT AND ANALYSIS

For this paper we have considered accuracy as a prime result analysis. Along with it we used confusion matrix for understanding errors. We achieved a maximum of 99.26 training accuracy and 80.44 validation accuracy.

The prediction breakdown can be seen using confusion matrix in Figure 6.

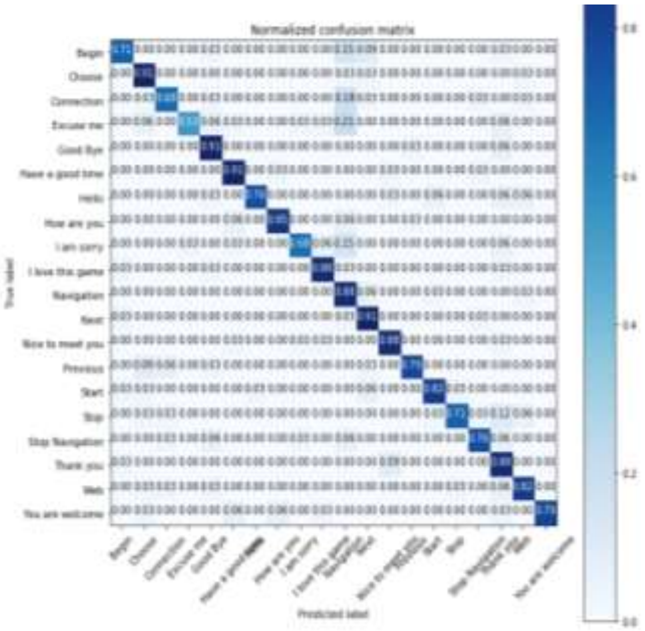

Figure 6: Confusion Matrix

The model was trained for 50 epochs with a learning rate of 0.0001. It's validation accuracy was very less at the beginning of about 9 to 10% at the beginning but then jumped to 80.44% till the end of 50 epochs as shown in Figure 7.

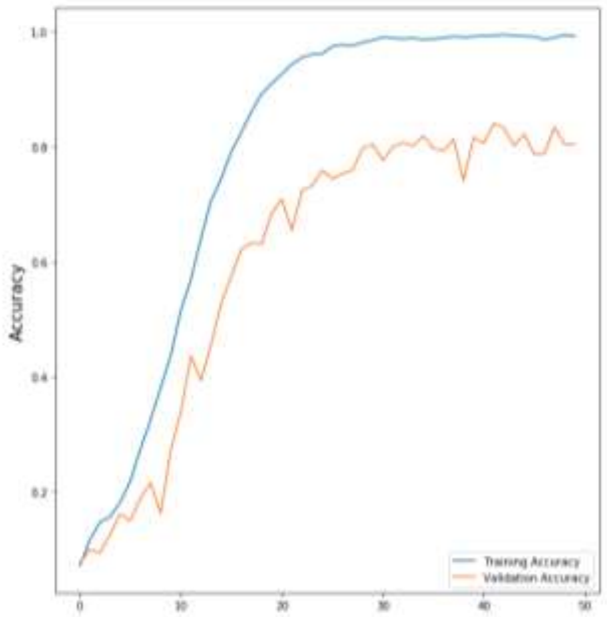

Figure 7: Model Accuracy

The validation loss of 3.0576 at the beginning and was reduced to about 0.7504 till the last epoch. It's training loss started with approximately 3 and was almost 0.04 at the end as shown in Figure 8.

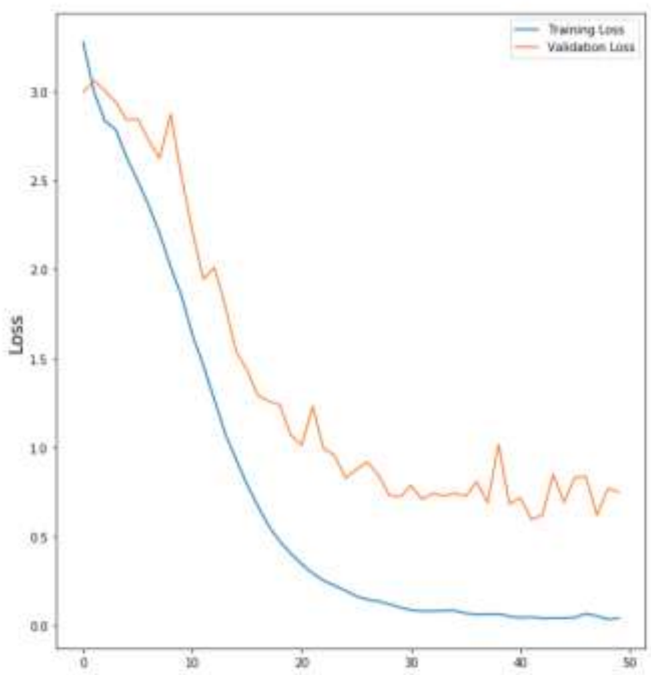

Figure 8: Model Loss

6. CONCLUSION

Over all we found that the model improved due to the inclusion of the pre-trained facial recognition model. The data augmentation proved helpful for seen people. With the help of this paper you cannot determine the words or phrases spoken by unseen people as the dataset was not that large. In this model, it was difficult to avoid overfitting with unseen people. Thus certain models and hyperparameters would fit better if we are working on seen/unseen people for testing and validation. The work inclusion increases if we use pre-trained models to reduce overfitting. The addition of properties like regularization would reduce the overfitting problem to a greater extent.

This project is easily extendible and raises the question in the mind of the project makers of how to perform visual speech recognition on a much larger or wider range of corpus (e.g. English Dictionary). How could we interpret the video as text by addition of audio data? The MIRACL_V1 dataset[0] includes phrase inputs which is an interesting area for exploration and in real life phrases are preferred over words.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr. Sanjay U. Bokade, Principal and Dr. Satish. Y. Ket , H.O.D of Department Computer Engineering of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on - Automatic Lip Reading: Classification of Words and Phrases using Convolutional Neural Network. This project bears on imprint of many peoples. We sincerely thank our project guide Prof. Suresh R. Mestry for his guidance and encouragement in carrying out this synopsis work. Finally, we would like to thank our colleagues and friends who helped us in completing project work successfully.

REFERENCES

[1] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. A new visual speech recognition approach for RGB-D cameras. In Image Analysis and Recognition - 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II, pages 21–28, 2014.

[2] Padilla, Rafael & Filho, Cicero & Costa, Marly, “Evaluation of Haar Cascade Classifiers for Face Detection”, (2012)

[3] Yannis M. Assael, Brendan Shillingford, Shimon White son, and Nando de Freitas. Lipnet: Sentence-level lipreading. CoRR, abs/1611.01599, 2016.

[4] Ziad Thabet, Amr Nabih, Karim Azmi, Youssef Samy, Ghada Khoriba and Mai Elshealy, “Lip Reading using a Comparative ML Approach", 978-1-5386-5083-7/18/978-1-5386-5083-7/18/31.00 c. 2018 IEEE

[5]Amirsina Torfi, Seyed Mehdi Iranmanesh, Nasser Nasrabadi and Jeremy Dawson, 3D Convolutional neural networks for cross audio-visual matching recognition" 2169-3536 c. 2017 IEEE.

[6] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, “Towards better analysis of deep convolutional neural networks,” IEEE transactions on visualization and computer graphics, vol. 23, no. 1, pp. 91–100, 2017.

[7] N. Rathee, “A novel approach for lip reading based on neural network,”in Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on, pp. 421–426, IEEE, 2016.

[8] K.Nodal ,Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata ,“Lipreading using convolutional neural network.’’ In Fifteenth Annual Conference of the International Speech Communication Association . 2014.

[9]M. Wand,J.Koutn et al.,“Lipreading with long short-term memory,’’ in 2016 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).IEEE, 2016,pp.6115-6119.

[10]Vegad, Sagar, Harsh Patel, Hanqi Zhuang and Mehul R. Naik. “Audio-Visual Person Recognition Using Deep Convolutional Neural Networks.” (2017).

[11] F. Vakhshiteh, F. Almasganj, Lip-reading via deep neural network using appearance-based visual features, in 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME) (IEEE, 2017), pp. 1–6.

BIOGRAPHIES

Nikhil Kesarkar, Undergraduate Student, BE Computer Engineer, MCT Rajiv Gandhi Institute of Technology, Mumbai University, Mumbai

Poornachandraprasad Kongara, Undergraduate Student, BE Computer Engineer, MCT Rajiv Gandhi Institute of Technology, Mumbai University, Mumbai

Manthan Kothari, Undergraduate Student, BE Computer Engineer, MCT Rajiv Gandhi Institute of Technology, Mumbai University, Mumbai

Mr. Suresh R. Mestry, Assistant Professor in Department of Computer Engineering. Having teaching experience of 13 years. Specialization areas are Machine Learning , Artificial Intelligence, NLP