Lip Reading using Convolutional Neural Network

Aditi Bhingarkar,

Department of Electronics
& Telecomunication
JSPM's Rajarshi Shahu College Of
Engineering, Pune, India

Siddhant Gardi,

Department of Electronics
& Telecomunication
JSPM's Rajarshi Shahu College Of
Engineering, Pune, India

Sushant Walokar

Department of Electronics
& Telecomunication
JSPM's Rajarshi Shahu College Of
Engineering, Pune, India

Ms. Shilpa Sonawane
Professor
Department of Electronics& Telecommunication
JSPM's Rajarshi Shahu College
Of Engineering, Pune, India

---***---

Abstract - Lipreading is that the task of understanding speech by analysing the movement of lips. instead, it can be represented because the method of secret writing text from visual info generated by the speaker's mouth movement. The task of perception depends conjointly on info provided by the context and data of the language. Lipreading, conjointly called visual speech recognition could be a difficult task for humans, particularly within the absence of context. many apparently identical lip movements will turn out totally different words, thus perception is associate inherently ambiguous drawback within the word level. Even skilled lipreaders come through low accuracy in word prediction for datasets with solely a couple of words. machine-driven perception has been a subject of interest for several years. A machine that may browse lip movement has nice utility in varied applications such as: machine-driven perception of speakers with broken vocal tracts, biometric person identification, multi-talker coincidental speech secret writing, silent-movie process and improvement of audio-visual speech recognition normally. Here we've used Convolutional Neural Network (CNN) from deep learning. The advancements in machine learning created machine-driven perception attainable. Also, we've used MIRACL-VC1 Dataset for our reference. The accuracy of our project came dead set as 76%.

Key Words**:** convolutional neural network, deep learning, lip reading, converting speech to text, open CV, keras.

1. INTRODUCTION

In our world there square measure such a large amount of completely different languages, therefore this task of lip reading isn't generic. skilled lip reading isn't a recent idea. it's been around for hundreds of years. Obviously, one in all the most important motivations behind lip reading was to produce folks with disablement the simplest way to grasp what was being aforesaid to them. Lip-Reading needs an excellent deal of concentration once done by somebody's. sensing is that the task of decipherment text from the movement of a speaker's mouth nonetheless, with the advancing technologies within the field of Computer Vision and Deep Learning, machine-driven lip reading by machines has become a true chance currently. Lip reading technology will increase the likelihood to permit higher speech recognition in clattering or loud environments.

Sounds square measure jazzy nodes created by vocal cords. every individual has a personal pattern which may discriminate him/her from others. within the case of imitation conjointly a personal wish to reproduce a pattern of sound that another person produces with simple. within the present work, we would like to grasp the pattern during which a personal creates a sound for expressing a selected letter. Further, a regular pattern is made which can be mental object for the comparison of the expressed sound with the mental object. Sounds are often created with the lips within the same position. quick speech, poor pronunciation, dangerous lighting, face dodging, hands over mouths, moustaches and beards build lip reading harder or perhaps not possible. The idea of image matching is preoccupied to grasp the expressed sound that makes it a novel method since it's associate understood parameter that no matter the individual's vocal band the movement of lips must be same to for saying the letters. This concept has been taken within the present project to bring out the various patterns for the pronunciation.

2. Literature Survey

There are various research papers that related to the study of Lip Reading. Based on the study of these pacers this project was proposed

A neural network-based lip-reading system has been developed to predict sentences covering a wide range of vocabulary in silent videos from people speaking. The system is lexicon-free, uses only visual cues represented by visemes of a limited number of distinct lip movements, and is robust to different levels of lighting.

We have discussed various deep learning, machine learning techniques and approaches for lip reading. As well as we discussed various types of available datasets. Deep learning can classify, cluster, and predict anything id we have data like images, videos, sound, text etc.

Various papers have been studied for lip's reading system techniques. From these papers, we have concluded that different techniques have different ways to recognize the lip's reading techniques. The colour transformation is used for the advancement in the quality of the lip segmentation and reduces computational complexity. The procedure of lip's feature extraction was continuously enhanced in the lip segmentation quality.

3. Block Diagram & Methodology

Fig -1 shows the breakdown stages of the lip reading.

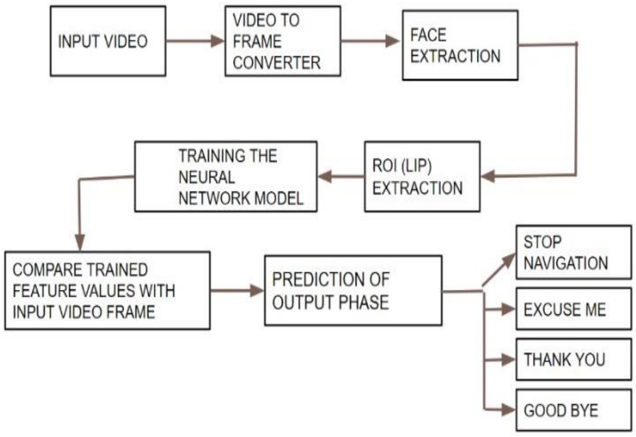

INPUT VIDEO
VIDEO TO FRAME CONVERTER
FACE EXTRACTION
TRAINING THE NEURAL NETWORK MODEL
ROI (LIP) EXTRACTION
COMPARE TRAINED FEATURE VALUES WITH INPUT VIDEO FRAME
PREDICTION OF OUTPUT PHASE
STOP NAVIGATION
EXCUSE ME
THANK YOU
GOOD BYE

Fig -1: Breakdown stages of how phrases are predicted from mute or silent video.

Dataset Pre-Processing: -

We had manually collected few lip motion datasets for the training of our model. We selected the 10 most frequently used phrase in English.

Here we have used MIRACL-VC1 Dataset for our reference. MIRACL-VC1 is a lip-reading dataset including both depth and colour images. It can be used for diverse research fields like visual speech recognition, face detection, and biometrics. Ten speakers (Four men and Six women) speak ten different phrases (Total ten video samples).

Step1: To extract the image from the raw video: We used OpenCV for converting our input video to frame. The frame was made after every 0.5 sec.

Step2: Cropping and resizing the data: Our Region of interest is lips so cropping and resizing of the extracted is to do and create the dataset.

Step3: Normalization of dataset: Data normalization is the process of rescaling one or more attributes to the range of 0 to 1

Input Video:

This is the first step of our project, where input is given in the form of video. Here a person speaks a phrase which must be predicted.

Video to frames converter:

In the case of continuous speech, when the speech rate tends to increase, so we need to convert video to frames. Hence the input video is converted to images at framerate of 0.25 sec each. We have used OpenCV for converting the video to frame.

Face extraction:

Face is extracted using the Bounding box algorithm. It is an imaginary rectangular box that contains an object or a set of points it refers to the border coordinates that enclose the face. A rectangular outline can be drawn over the images which include all important features in it. As per the project requirements the object can vary and the rectangular or any other shape can be drawn. Here we are interested in face detection. It reduces the range of search for the object features and thereby conserves computing. It not solely helps to classify the objects however additionally helps in object detection. It not solely helps to classify the objects however additionally helps in object detection.

Our interested object is face so we have used rectangular outline to indicates it as shown in figure 2. Region of interest (Using bounding box algorithm)as shown in Fig -2,

Fig -2: Image shows the region of interest.

Training the neural network model: -

To implement the deep Learning model of CNN, we used keras, an Open-Source Neural Network Library written in python. Keras provides an easy to use and understand API so that non-specialist can easily develop and utilize deep-learning models in their field. Keras has four different guiding principles i.e., Modularity, Minimalism, Easy Scalability and Python Based.

Step1: Create a Sequential model: Sequential model is like a list which stack the convolutional layers, so it is basically stack of layers where each layer has one input and one output.

Step2: Adding of layers: -The whole process of CNN training for image processing includes different operation such as convolution, activation function, pooling, flatten, and fully connected layers the output layer. we have used keras conv2D to create a convolution kernel it is a convolution matrix. Mandatory Conv2D parameter is that the numbers of filters that convolutional layers can learn from. The activation parameter to the Conv2D category is solely a convenience parameter that permits you to produce a string, that specifies the name of the activation operate you wish to use once playacting the convolution.

The Dropout layer at random sets input units to zero with a frequency of rate at every step throughout training time, that helps stop exceeding. Inputs not set to zero area unit scaled up by 1/ (1 - rate) specified the total over all inputs is unchanged. The convolution layer and pooling layer area unit region that feature extraction. And, the totally connected layer is region that classification. The flatten layer is found in between this region

Step 3: Activating the layer using 'SoftMax' activation function.

Compare trained feature value with input video frame: -

Optimizers are algorithms or methods used to change the attributes of your neural network such as weights and learning rate in order to reduce the losses. Keras uses fit () function to training the model. Data visualization is graphical representation that contains the information and the data.
By using visual components like charts, graphs, and maps, information image techniques give AN accessible thanks to see and perceive trends, outliers, and patterns in information.
Keras support .h5 format to save the train mode and use it to predict. After saving the model in .h5 format we need load that file. Keras provides a function load_model() to load the model. Then using predict () we get the feature values. This feature values was compared with the feature values of the dataset which was used to train the network.

Prediction of output: -

We have selected 10 phrases for testing the output
And the phrases are as shown in the Table -1

SR NO	PHRASES
1	Stop Navigation
2	Excuse Me
3	I Am Sorry
4	Thank You
5	Good Bye
6	I Love This Game
7	Nice to Meet You
8	You Are Welcome
9	How Are You
10	Have A Good Time

Table -1: Phrases to be predict

The working and Classification of Deep Neural Network Input, hidden and output layers shown in the Fig-3,

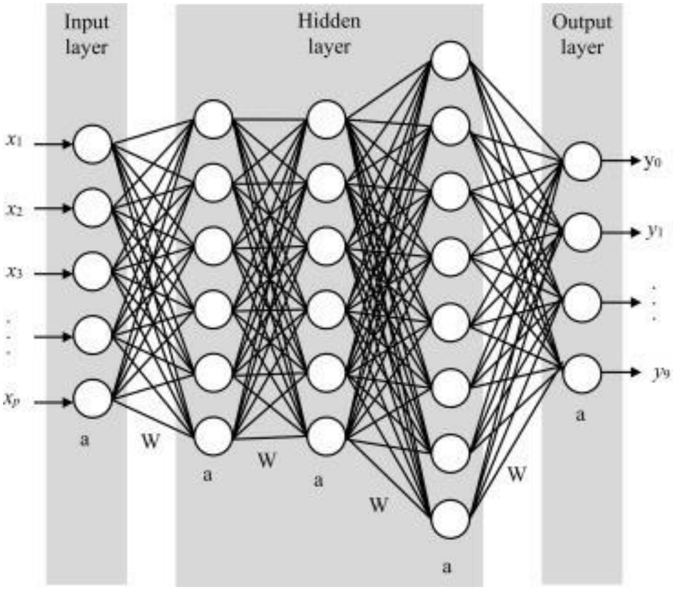


Fig -3: Deep-neural network and its layers.

Deep Neural Network:

Deep Learning

Deep learning is a subset of machine learning that uses multiple layers to progressively extract high-level feature from raw input. Deep learning neural network, or artificial neural network, attempts to mimic the human brain through a combination of data inputs, weights and bias.

Neural network

Neural Network is composed of layer and the layer has several neurons. There are three layers- input layer, output layer and hidden layer. Each neuron is fully connected as shown in diagram. If the number of hidden layers are exceeding, then the neural network is called as 'Deep Neural Network (DNN)'. And deep learning uses this deep neural network for learning model

The meaning of 'Training' or 'Learning' in the deep learning is adjustment of weights. There square measure many sorts of neural network or deep learning model. CNN and RNN that are used in our model.

Convolutional Neural Network (CNN)

CNN is primarily used to detect features and patterns within an image, enabling tasks like object detection or recognition. The whole process of CNN training for image processing includes different operation such as convolution, activation function, pooling, flatten, and fully connected layers the output layer.

4. Result

As a significant component of the Human Computer Interface (HCI), automatic lip reading is designed for understanding the content of speech by interpreting the movements of the lips. Although performance of automatic lip-reading system is easily affected by challenging conditions such as illumination and low resolution, enormous advancements in the relevant fields accompanied with enhancement in computer capability have improved the robustness of the system, making it more adaptable to the real environment. An intelligent system will be trained by giving user's lip-movement frames sequences as input and can determine lip movement and therefore the same word mistreatment either visual info or each audio and visual info.The most important step is to evaluate the performance of the neural network, which receives video samples and produces a sequence of frames for prediction for each of the 10 phrase samples. we assume 10 video samples for each phrase for testing the output. We run the evaluation process on the trained tested dataset and obtained the accuracy results of the networks shown in Table -2 Result of test cases,

Sr. no	**Phrase**	**No of inputs**	**No of O/P correctly matched**	**No of O/P failed to match**
1	Stop Navigation	10	8	2
2	Excuse Me	10	7	3
3	I am Sorry	10	9	1
4	Thank You	10	9	1
5	Good Bye	10	8	2
6	I Love this game	10	7	3
7	Nice to meet you	10	7	3
8	You are welcome	10	7	3
9	How are you	10	8	2
10	Have a good time	10	6	4

Table -2: Result of test cases

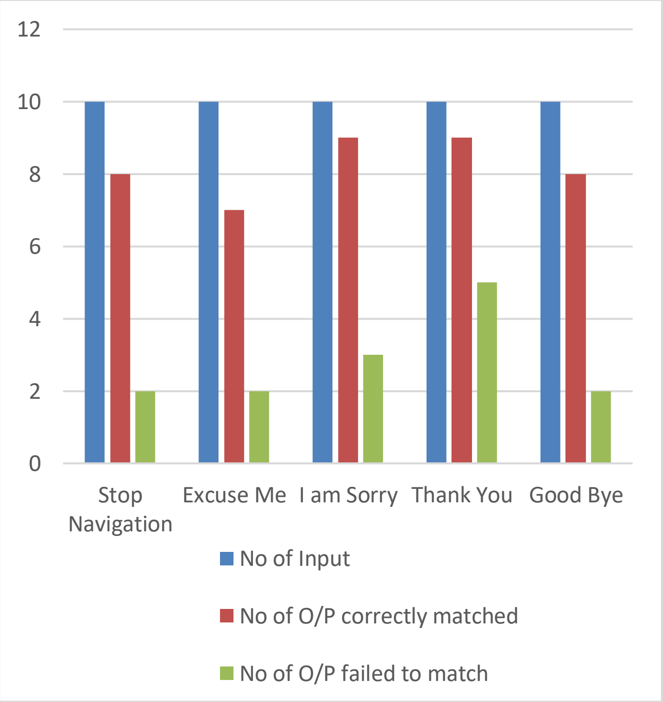
$$Accuracy\ in\ percentage = \frac{\sum No\ of\ Output\ Correctly\ match}{\sum Total\ no\ of\ inputs} \times 100$$

$$= 0.76 \times 100$$

$$= 76\ \%$$

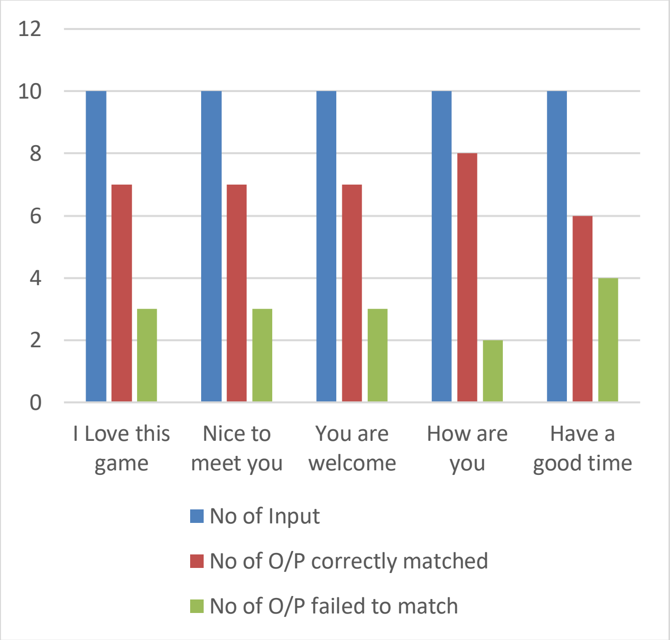
Accuracy in percentage = 76%

Chart-1 represents the bar graph of the phrase stop navigation, excuse me, I am sorry, thank you, good bye



Charts -1 Bar graph represents result of test cases.

Chart -2 Represents the bar graph of the phrase I love this game, nice to meet you, you are welcome, how are you, have a good time,



Charts -2 Bar graph represents result of test cases.

5. CONCLUSIONS

A neural network-based lip-reading system has been developed to predict sentences covering a range of vocabulary in silent videos from people speaking. This work investigates on the development of lip identification and recognition, which achieves significantly improved performance over previously proposed approaches. Automatic words detection and recognition approach from different lip movements. We present an approach for identifying and recognizing different lip expression of the human. We are working on dataset collection part. After evaluating the test case the total accuracy was calculated by dividing the sum of no of output correctly matched to total number of inputs, which came out as 76%.

REFERENCES

1 – Sanaullah Manzoor, Muhammad Faisal, “Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language”, Published in ArXiV 2018.

2 – Parth Khetarpal, Riaz Moradian, Shayan Sadar, Sunny Doultani, Salma Pathan, “Lip Vision: A Deep Learning Approach”, International Journal of Computer Applications (0975 – 8887) Volume 179 – No.8, December 2017.

3 - Amit Garg, Jonathan Noyola, Sameep Bagadia, “Lip reading using CNN and LSTM”, 2016

4 – Y. Li, Y. Takashima, T. Takiguchi, Y. Ariki, “Lip reading using a dynamic feature of lip images and convolutional neural networks,” in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–6, June 2016.

5 – S. Agrawal, V. R. Omprakash, Ranvijay, “Lip reading techniques: A survey,” in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 753–757, July 2016.

6 – François Chollet. 2015. Keras documentation. keras. io (2015).

7 – Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong and R. Bowden, "Improving visual features for lip-reading", AVSP, pp. 7-3, 2010.