An Application to Convert Lip Movement into Readable Text

Saakshi Bhosale, Rohan Bait, Shivangi Jotshi, Rohan Bangera, Prof. Jinesh Melvin
Department of Computer Engineering,
PCE, Navi Mumbai, India - 410206

Abstract: **Lip-Reading is the process of interpreting spoken word by observing lip movement, without or with audio (Noisy). The input for the application will be frames from an entire video. The application needs to detect the face (lips) from the entire frame also, trace and understand the feature patterns of the lip moments over the parameter of time. This can be done using Computer Vision (feature extraction) and Deep Convolutional Neural Network (CNN) Model Lip-reading system is difficult to implement due to complex image processing, hard-to-train classifiers and long-term recognition processes. Automatic lip-reading technology is a very important component of human–computer interaction technology. It is very important for human language communication and visual perception. Traditional lip-reading systems usually consist of two stages: feature extraction and classification. For the first stage, a lot of methods use pixel values extracted from the mouth region of interest (ROI) represented as visual information. At present, deep learning has made significant progress in the field of computer vision (image representation, target detection, human behaviour recognition and video recognition). Therefore, automatic lip-reading technology has shifted from the traditional manual feature extraction classification methods to end-to-end deep learning architecture models.**

Keywords: Lip tracking, Phoneme, 3D CNN Model, facial points, active frames.

1. INTRODUCTION

With the increase in technological advancements, lip-reading systems have gained major attention. These systems are a great help to the hearing impaired as they convert audio into text that is easily readable. These systems have also proved to be extremely helpful in government operations as a spy camera.

The first step of a lip-reading system is to track facial points. The system must be able to highlight the lips from the entire face. The advantage of this method is that lips of different shapes and sizes can be tracked, allowing the system to get an exact set of points for the lips.

Once the lips have been tracked, the major challenge is to recognize the lip movement. Every person will have a different lip movement due to the many possible lip shapes and sizes. The similarity between these movements is that for a specific phoneme, the distance between the two lips is pre-defined. Phoneme is the sound created while pronouncing any syllable in any language. The distance that is measured is measured in terms of the Euclidean distance and this distance is specific to a syllable.

Hence, the classification of the lip movements will be done on the basis of this distance. This classification is a media-based classification hence, a 3D Convolutional Neural Network (CNN) Model is best suitable for this project.

2. THE CNN MODEL

This project uses the Three-Dimensional Convolutional Neural Network model for testing and training data and also predicting results. CNN is the go-to model for an image-based problem. It gives the most accuracy for multi-media problems. CNN is an unsupervised model and does not require human supervision to detect important features.

CNN is also computationally efficient. It uses operations like convolution and pooling and also performs parameter sharing. A CNN model can run on any device, making them universally usefully.

2.1 Architecture

Convolution-

Convolutional layer is the main building block of CNN model. Convolution is a mathematical operation that merges two sets of information together. A convolutional filter is applied to the input data and a feature map is produced. This process is performed by sliding the filter over the input. At every location, an element-wise matrix multiplication is done and the result is summed. This sum goes into the feature map.

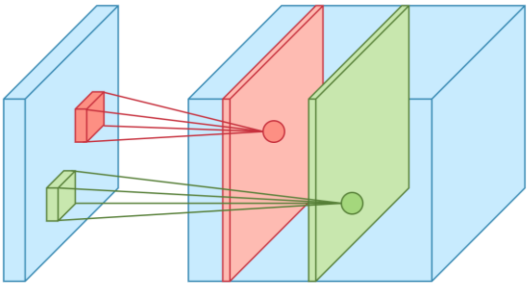

Fig. Convolution of multiple layers

Pooling-

After a convolution operation, a pooling operation is performed, mainly to reduce the dimensionality of the layers. This enables us to reduce the number of parameters, which both shortens the training time and combats overfitting. Pooling layers reduce the height and width of each feature map independently but keep the depth intact.

Pooling has no parameters, contrary to the convolution operation. It slides a window over its input, and simply takes the max value in the window, if it is max pooling. Here too, we specify the size and stride of the window.

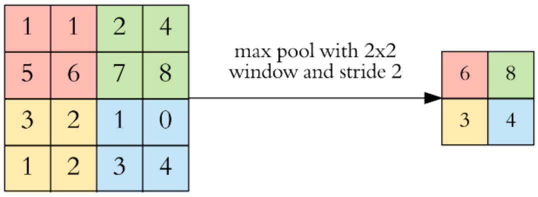

Fig. Max Pooling

Architecture of model in this Application-

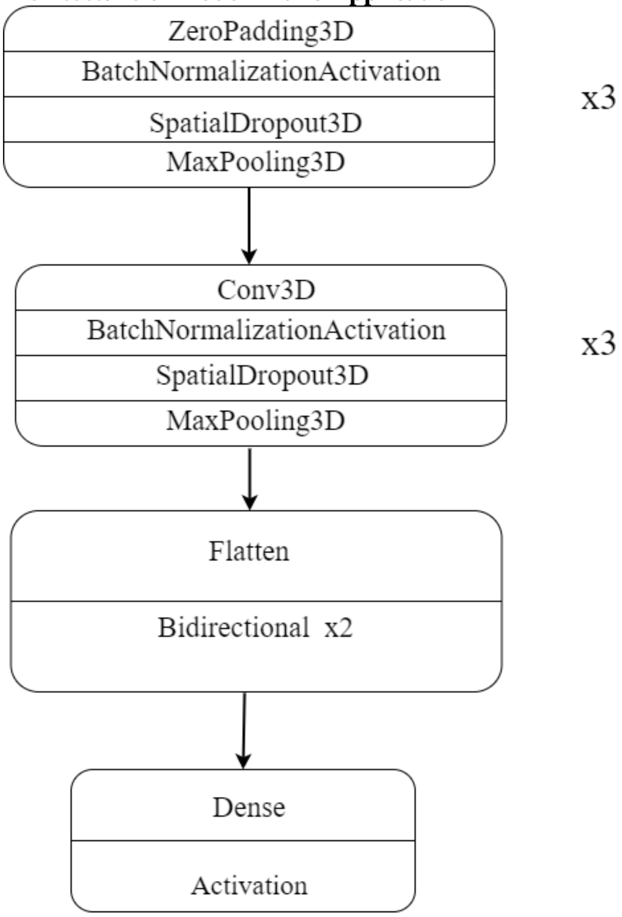

Fig. CNN Architecture

4. SYSTEM ARCHITECTURE

4.1 Video Input

The Lip-Reading Application will function on the basis of the input received from motion pictures or videos. These inputs will be in the form of vectors that have traced the lips and face structures in the input frame.

4.2 Tracking

The very first step is detecting the 62 facial points of the subject. According to these points the distance between the lips and the shape created by the lips can be calculated and therefore, each word can be predicted as spoken. This detection takes in place of vectors and the distances are measured and stored. It includes detection of lips, eyes, nose, eyebrows.

4.3 Data Pre-processing

In real-time face tracking and lip-reading systems, head pose, head movements, camera-to-head distance and head direction of the speaker are important parameters and these parameters affect lip reading performance and robustness of the system. It is necessary because of the following three main problems that may occur while collecting data:

□ People have different sizes of lip, different contour shape and different mouth height/width ratios, and these differences may cause different classifications.

□ The distance between the camera and the speaker affects pixel distances between lip points on the image and the classification algorithms based on the distances may suffer from those differences.

□ If a speaker turns his/her head to left or right slightly, the height/width ratio of the mouth may change and it may lead to incorrect classifications.

Therefore, data cleaning and preprocessing is very important before classification. Classification of data is done on the basis of lip distance for every phoneme. The aim here is to identify a sequence for every phoneme. This sequence refers to the exact movement made by the lips to produce that particular phoneme. Once this sequence has been identified for every phoneme in the English language, using a multi-dimensional array, every sequence will be added to its respective class.

4.4 Word Segmentation

In lip-reading systems, the first step is to determine the starting and ending points of a word in a speech video. Detecting where a word starts and ends is called word segmentation. Word segmentation takes place after lip activation has been detected.

A computer program is developed which displays a word to the speaker and the speaker reads that word while the camera records the data. Assume 10 frames are classified as active (1) or passive (0) in the order [0 0 0 0 0 1 1 1 1 1], thus 5 frames are passive followed by 5 active frames. The frame where 1 first occurs, i.e. the 6th frame, is assumed as the starting frame of the word. If the frame sequence is as such [1 1 1 1 1 0 0 0 0 0], i.e. 5 active frames followed by 5 passive frames, then the first occurrence of zero is assumed as the ending frame of the word. By using these starting and ending frames, the words can be segmented from the input video.

4.5 Classification

Classification of data is done on the basis of lip distance for every phoneme. The aim here is to identify a sequence for every phoneme. This sequence refers to the exact movement made by the lips to produce that particular phoneme. Once this sequence has been identified for every phoneme in the English language, using a multi-dimensional array, every sequence will be added to its respective class.

4.6 Interpreting Dataset

According to the values received, we interpret a dataset. The dataset will include various words and the approximate distance of lips for pronouncing each word. This in turn will predict the word spoken. The dataset will work on word segmentation. Word segmentation is the parting distance between the lips required for pronouncing a specific alphabet in a word. On relating the values received by tracking to the dataset, we will get the spoken word which can be further built onto making a phrase.

According to the word segmentation, the words will be classified in the dataset. Depending upon the distance of the tracked lips, they will be classified into the specific group.

4.7 Prediction

The final step is a prediction of words or phrases. These predicted words will be displayed as captions under the video to enable reading simultaneously while watching.

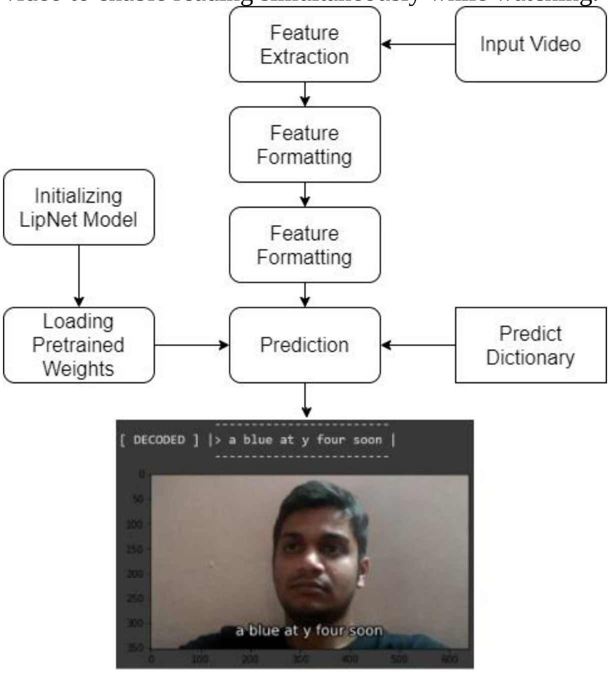

Figure 3.1 Flow Chart of Proposed System

5. DATASET

The dataset used in this project has been downloaded from GRID. GRID is a large multitasker audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 males, 16 female). Sentences are of the form “put red at G9 now”.

In each category of speaker i.e. male and female 4 types of file formats are available. The files include:

- Audio only
- Video (normal quality)
- Video (high quality)
- Transcript

Audio files were scaled on collection to have an absolute maximum amplitude value of 1 and down sampled to 25 kHz. These signals have been end pointed. In addition, the raw original 50 kHz signals are included below. Video files are provided in two formats: normal quality (360x288; ~1kbit/s) and high quality (720x576; ~6kbit/s).

5.1 A sample of dataset

bin green with y eight please	set white by f eight please
lay red in c four now	place blue at I five please
place blue at I one soon	set blue with l eight now
bin white by f six soon	place red by u one soon
set green with c zero again	lay white with d six soon

Table 4.1 Dataset Sample

6. DATA FLOW DIAGRAM OF SYSTEM

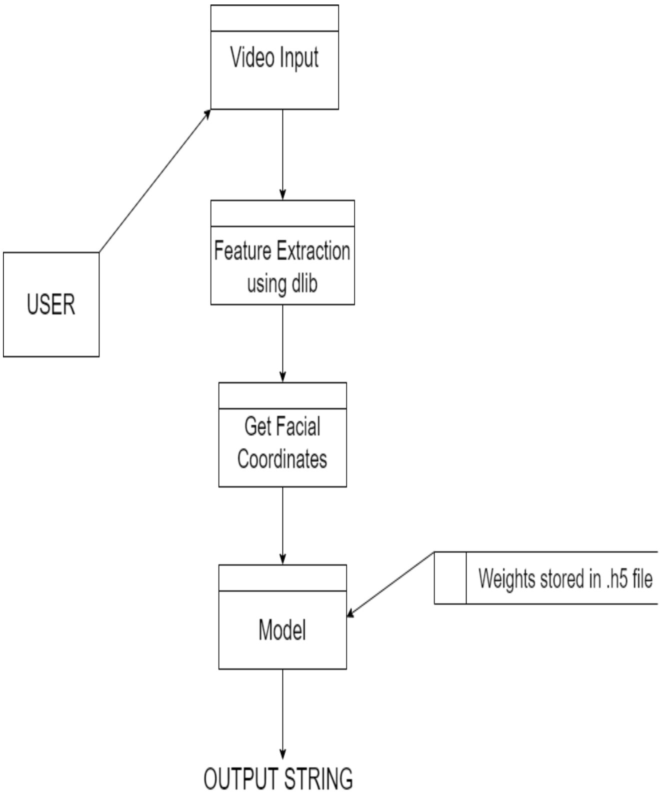

Fig. Data Flow Diagram

As depicted in the diagram, the User provides a video input with or without audio. The dlib library is used to extract features from the face. Features like lips, nose, eyes are extracted and facial coordinates are produced. These facial coordinates are passed through the Model that is stored in .json file along with weights stored in .h5 file. These weights are previously saved during the construction of the model. The final output of the model is an Output String that has the spoken sentence as text.

ACKNOWLEDGMENT

We would like to extend our deepest gratitude to our Project guide **Prof. Jinesh Melvin** for his exemplary guidance, monitoring and constant encouragement throughout this project which helped us improve our work.

We would also like to extend our gratitude to our Head of Computer Engineering Department **Dr. Sharvari Govilkar** for providing us with an opportunity and platform to carry out this project.

We are also extremely grateful to our Principal **Dr. Sandeep Joshi** who provided us with this golden opportunity as well as all the facilities needed to carry out this project successfully.

REFERENCES

[1] A Lip-Reading Application on MS Kinect Camera Alper Yargıç, Muzaffer Doğan Computer Engineering Department Anadolu University Eskisehir, Turkey {ayargic, muzafferd}@anadolu.edu.tr

[2] Lipreading Using a Comparative Machine Learning Approach Ziad Thabet Faculty of Computer Science MISR

INTERNATIONAL UNIVERSITY Cairo, Egypt Ziad1407174@miuegypt.edu.eg
Amr Nabih Faculty of Computer Science MISR INTERNATIONAL UNIVERSITY Cairo, Egypt Amr1410718@miuegypt.edu.eg
Karim Azmi Faculty of Computer Science MISR INTERNATIONAL UNIVERSITY Cairo, Egypt Karim1405338@miuegypt.edu.eg
Youssef Samy Faculty of Computer Science MISR INTERNATIONAL UNIVERSITY Cairo, Egypt Youssef1410209@miuegypt.edu.eg
Ghada Khoriba Faculty of Computers and Information Helwan University Cairo, Egypt ghada khoriba@fcih.helwan.edu.eg
Mai Elshehaly School of Computing University of Leeds Leeds, UK m.h.elshehaly@leeds.ac.uk

[3] A Review on Methods and Classifiers in Lip Reading
Leticia Ria Aran1, Farrah Wong2 and Lim Pei Yi3 Faculty of Engineering Universiti Malaysia Sabah Kota Kinabalu, Sabah, Malaysia 1latishaaran@gmail.com 2farrah@ums.edu.my 3lpy@ums.edu.my

[4] Lip Reading Techniques: A Survey Shreya Agrawal Computer Science and Engineering, MNNIT Allahabad Allahabad, India 211004 Email: cs134027@mnnit.ac.in Verma Rahul Om Prakash Computer Science and Engineering, MNNIT Allahabad Allahabad, India 211004 Email: cs134061@mnnit.ac.in Ranvijay Computer Science and Engineering, MNNIT Allahabad Allahabad, India 211004 Email: ranvijay@mnnit.ac.in

[5] Personal Computer Based Real Time Lip Reading System Kazunori SUGAHARA, Makoto KISHINO, Ryosuke KONISHI Faculty of Engineering, Tottori University 4-101 Koyama-cho Minami, Tottori-shi, Tottori, Japan 680-8552 Fax: +8 1-857-3 1-5246, e-Mail:sugahara@ele.tottori-u.ac.jp

[6] Lip Movements Recognition Towards An Automatic Lip Reading System for Myanmar Consonants Thein Thein Ph.D Student University of Computer Studies, Mandalay (UCSM) Mandalay, Myanmar theinthein.cmw@gmail.com Kalyar Myo San Faculty of Computer Systems and Technologies (FCST) University of Computer Studies, Mandalay (UCSM) Mandalay, Myanmar kalyar.myosan@gmail.com

[7] Lip Localization Technique Towards an Automatic Lip Reading Approach for Myanmar Consonants Recognition

[8] Thein Thein University of Computer Studies, Mandalay (UCSM) Mandalay, Myanmar e-mail: theinthein.cmw@gmail.com Kalyar Myo San Faculty of Computer Systems and Technologies (FCST) University of Computer Studies, Mandalay (UCSM) Mandalay, Myanmar e-mail: Kalyar.myosan@gmail.com

[9] A REAL-TIME AUTOMATIC LIPREADING SYSTEM S.L. Wang+ , W.H. Lau+ , S.H. Leung* and H. Yan+ +Department of Computer Engineering and Information Technology, *Department of Electronic Engineering City University of Hong Kong, 83 Tat Chee Avenue, Hong Kong