March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

Speech Recognition using Lip Reading

Sanika Gowaikar
Dept of Information Technology Sinhgad College of Engineering Pune, India
sanika.gowaikar@gmail.com

Ganesh Pathak
Dept of Computer Science & Engg.Sathyabama Inst. Of Sci. & Tech.(Deemed to be University)
Chennai, India pathak.gr@gmail.com

Shivani Devi
Dept of Information Technology
Sinhgad College of Engineering Pune, India shivanigdevi@gmail.com

Vaishnavi Gagare
Dept of Information Technology
Sinhgad College of Engineering Pune, India vaishnavi.gagare97@gmail.com

Kalyani Chavan
Dept of Information Technology Sinhgad College of Engineering
Pune, India chavankalyani87@gmail.com

M.S. Godwin Premi
Dept of Computer Science &Engg.sathyabama Inst. Of Sci. & Tech. (Deemed to be University)
Chennai, India

Article Info
Volume 83
Page Number: 2254 - 2260
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract

Visual Speech Recognition is the technique to understand what is being said by interpreting the lip movements of a person. Grasping the technique of lip reading is challenging for the hearing impaired as it is mostly based on trial and error methods. This paper proposes an approach to not only ease this process of learning but also help them in day to day life. The approach follows the steps: Lip detection using OpenCV, feature extraction using AutoEncoders and Long Short Term Memory neural network and classification using Softmax as a multi-class classifier. The output will be in the form of captions displayed below the video file.

Keywords- Lip reading, Long Short Term Memory, AutoEncoder, Softmax.

I. INTRODUCTION

Lip reading is a mechanism used to decipher the visual information like lip movement, face and tongue positions in the absence of audio input. Hearing impaired depend heavily on lip reading and people with normal hearing also use it to some extent when audio source is not enough [2]. Along with aid to the hearing impaired, lip reading also supports various applications such as improving the efficiency of speech recognition in noisy surroundings, advanced security system for determination of identity, a human-computer interaction method and also a system for providing reliable legal evidence [1].

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

Various works have been done in this field. In [3] WenJuan et al proposed a new method for lip detection and tracking by combining Adaboost and Haar features to get the lip Region of interest (ROI) by using the relative positions of the eyes and the face outline. Another approach to locate lip ROI is to consider the lower half of the detected face region. For extracting the features after lip detection, Mohammad Hasan Rahmani et al [2] proposed the use of Deep Neural Network and Hidden Markov Model to build the visual speech sequences. In [1] Mei-Li Zhu et al used AutoEncoders for feature extraction followed by a softmax layer for multi-class classification.

In this paper, the framework for lip reading consists of the following modules: frame extraction, lip ROI detection, neural network to understand the lip movement sequence and finally a multi-class classifier.

The rest of the paper is organized as follows: Section II provides the literature survey. Section III presents our proposed system. Section IV provides the implementation and algorithms details. Finally, the concluding remarks are presented in Section V.

II. Literature Survey

Many methods for a lip reading system have been designed. Mei-li Zhu et al [1], proposed a lip reading recognition method based on deep learning. It performed gray scale conversion on the dataset images and were given as an input to Deep Neural Network. Three layered auto encoders were used for learning features with softmax as a classification layer.

Yao WenJuan et al [3], introduced an improved lip location and tracking approach. They used OpenCV for face detection and lip localization. Face and eye regions were detected followed by locating mouth region marked with a rectangle. The focus was on analyzing distribution relationship between faces, eyes and mouth which resulted in effective region of interest (ROI) localization.

Mohammad Hasan Rahmani, Farshad Almasganj [2], extracted region of interest from video frames. Another type of feature vector they used was representing lips by some fixed points on inner and outer contour. A hybrid DNN-HMM was used to model the visual speech sequences. They performed all these processes on CUAVE dataset.

Binh T.H Nguyen et al [4], used an algorithm that locates four lip features in face images. With the help of this algorithm, left and right corners and the lower and upper lip centers were extracted. This was performed upon a condition that the position of the two eye centers are provided in advance. They tested the proposed algorithm on FERET dataset resulting in detection of lips.

Yang Pingxian et al [5] proposed a method that solves the problem of difficulty in detecting a lip due to changing lip shape. They used haar classifier to detect nose and face area in OpenCV. They used Adaboost algorithm to locate lip area in better way. They presented a lip detection method that adopts relative position of lip against to face and nose to detect lip.

In [6] Xingqun Qi et al show comparison between the two linear classifiers namely SVM and Softmax. SVM and Softmax in the specific structure and practical application are quite similar, but there are also some differences. Among the two classifiers mentioned above, Softmax is considered to have a unique advantage of dealing with n-dimensional vectors. Softmax classifier was used instead of SVM for the extraction purpose. Softmax determines probability of the extracted vector and then it classifies. The absolute value of final score obtained in SVM has no physical mean and this is one of its disadvantages. Although Softmax is complicated in calculation aspect as compared to SVM, but it is faster in training stage and consists of simple model. This is the main advantage for a Softmax to be used.

III. Proposed System

This section describes the overall working of the proposed system.

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

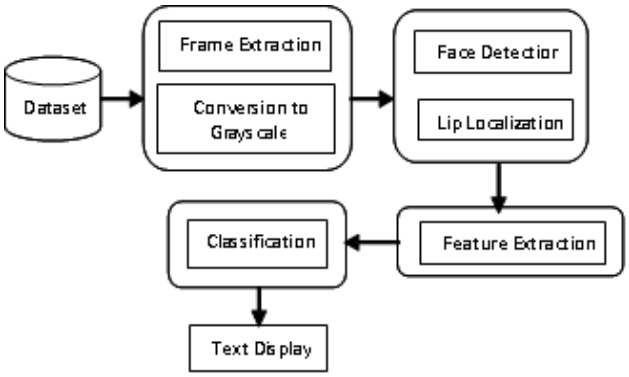

Fig. 1: System Architecture

Fig 1 shows the detailed view of the proposed system. The input to the system is a dataset consisting of 9 different speakers, pronouncing 19 isolated words. Each word is uttered 10 times by everyone resulting in a total of 1710 videos. The words in the dataset are given in Table I.

Table I. Words in the dataset

Zero	One	Two	Three	Four
Five	Six	Seven	Eight	Nine
Hello	Bye	Left	Right	Front
Back	In	Out	Welcome	

Following are the phases of the system:

1. Frame extraction and conversion to grayscale:

Every fifth frame is extracted and then converted to grayscale using OpenCV. These resulting frames are provided as an input to the next stage for lip localization.

Fig 3: Extracted frame and its conversion to grayscale of 2 speakers saying "left"

2. Face detection and lip localization:

After obtaining the pre-processed frames, HaarCascades, a machine learning algorithm for object detection is used for face detection. Further, the lip ROI is localized by considering the lower half of the previously detected face. This lip ROI is then passed to the feature extraction module.

3. Feature Extraction:

For feature extraction, AutoEncoders along with Long Short Term Memory (LSTM) are used. AutoEncoders take an image as the input, compress it and then reconstruct the input using the compressed data. The compressed data represents the features extracted. The output is further sent to LSTM to memorize the relations between consecutive frames. The output layer of this neural network is the Softmax classifier.

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

4. Classification:

Each word is considered as a class for classification. The classifier used is Softmax. It is an activation function used at the output layer. It transforms the input given by the LSTM into a probability distribution i.e. it calculates the likelihood of the output given by the LSTM belonging to each class. The class with the highest likelihood becomes the predicted word which is then displayed as the caption of the video, which is the final output of the system.

IV. Implementation and Algorithms

This section gives the important algorithms required for the system.

1. AutoEncoders:

AutoEncoder is an unsupervised algorithm used to compress the input into a compact representation. The input can be reconstructed using the compressed data.

Unlike Principal Component Analysis, which is used for dimensionality reduction for linear data, AutoEncoders can learn non-linear data. Thus, they are used for dimensionality reduction for non-linear data like images.

They introduce a bottleneck at the hidden layer to force it to learn the compressed version of the data. AutoEncoder is used for feature extraction.

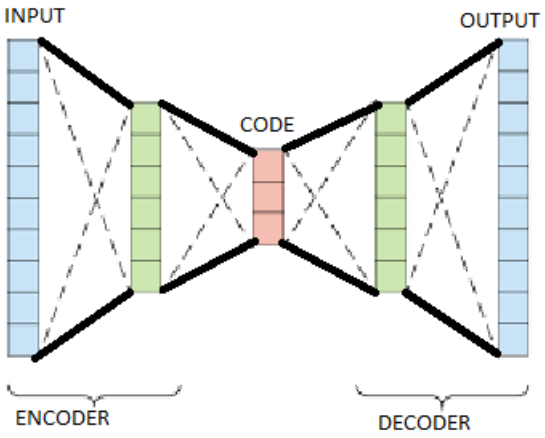

Fig. 2: AutoEncoder structure

Steps for the AutoEncoder network are as follows:

Step 1. Encoder:

It encodes the input image into a reduced representation. The result is a distorted version of the original input. It forwards this to the Code section.

Step 2. Code:

This is the bottleneck layer which gets forwarded to the decoder.

Step 3. Decoder:

The decoder reconstructs the input back to the original input.

The AutoEncoder compares the input and output to further learn better compression. After satisfactory encoder is constructed, the decoder can be removed.

2. LSTM

In image processing, the images which are sent through the network for processing are independent of other images that went through the neural network and those which will go through the neural network. However, the same case is not applicable for video processing.

Video is a form of a sequential data. Each frame in a video affects the subsequent frame and it itself is affected by the frame which came before it. Thus, the requirement arises to develop a neural network which has a capability to store or memorize the previous computations. It should understand the relation between the previous inputs that went through the network with the current one and continue doing so till the end of input arises.

For lip reading, the neural network should be able to learn the relations between positioning of the lips and tongue while a word is being spoken. Which position occurs after which one is essential to determine what the speaker is saying.

The ability of the Recurrent Neural Network (RNN) is to enable the processing of sequences of data where each output depends on the previous results of classification. RNN thus have the capability to memorize inputs.

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

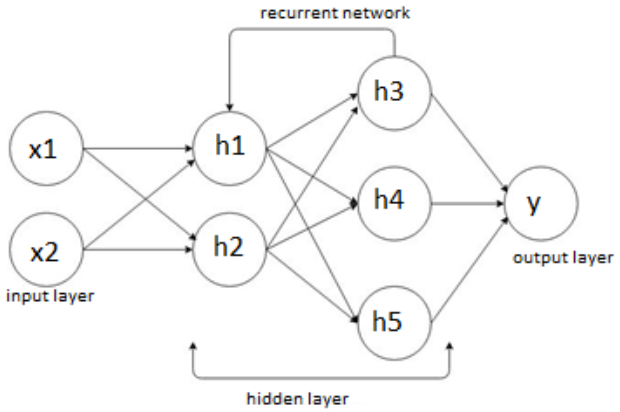

Fig 5. Recurrent Neural Network

However, the effectiveness decreases for long sequences. Thus, LSTMs are used to solve the problem of long term dependencies. LSTMs contain a network of loops which allows them to maintain the previously obtained information. In this approach, LSTM will memorize the sequence of lip movement for each word.

3. Softmax

Softmax is used for dealing with multiple classes. It is used as the activation of the output layer of the neural network. The softmax transformation transforms a set of input into a probability distribution. It provides the probabilities of output of the neural network for each class of words. It considers all the inputs to the neurons at a particular layer. The values given by softmax always add up to 1.

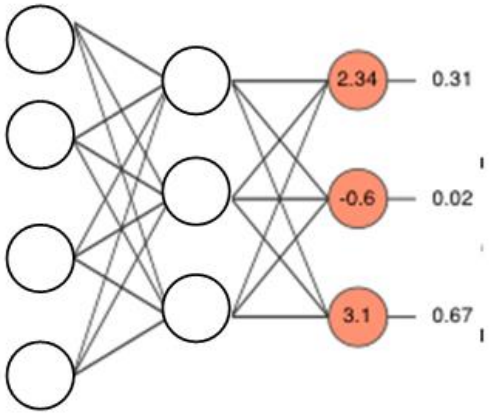

Fig 4. Softmax output.

The steps of the softmax layer as follows:

Step 1. Apply the activation function to the received input.

$$\sigma i(a) = \frac{e^{a_i}}{\Sigma_j e^{a_j}} \quad (1)$$

Where, e^{a_i} is the exponential of elements at position i,

$\Sigma_{i^e}^{a_j}$ is the sum of exponentials of all elements of the vector at layer j.

It will give the probability of the input belonging to each class.

Step 2. The result will be decided by choosing the class with the maximum likelihood.

The result will be the predicted word. This word will then be displayed at the bottom of the video as the caption for which the lip reading of the speaker is to be done.

V. Results and Analysis

In the interest of checking the workability of the proposed approach, a subset of the dataset consisting of 10 words is selected for testing. Following Table II represents the experimental results and analysis.

Table II. Experimental Results

Word Class	0	1	2	3	4	5	6	7	8	9
0	61	0	1	2	0	2	4	2	3	0
1	5	79	3	1	3	0	3	4	4	3

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

2	6	1	72	2	0	3	1	3	2	1
3	0	3	3	68	2	3	3	3	3	2
4	3	0	2	3	75	1	2	3	1	0
5	5	2	3	2	2	78	2	2	4	4
6	6	1	0	4	2	2	70	2	2	0
7	0	1	4	3	3	1	0	69	3	1
8	4	2	0	5	1	0	2	0	65	2
9	0	1	2	0	2	0	3	2	3	77
Correct Prediction	67.7%	87.7%	80%	75.5%	83.3%	86.6%	77.7%	76.6%	72.2%	85.5%

With a larger dataset, there is a potential for achieving better accuracies. Table 1 shows the accuracies for each class. Consider an example where the accuracy of word 0 is to be calculated. No of correct predictions were 61 and the incorrect predictions were 29 out of 90. The accuracy obtained in this case was 67.7%.

The graph below compares the accuracies between all the spoken words.

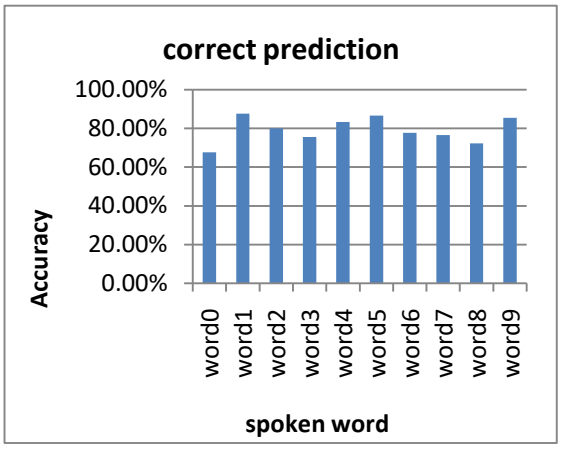

Fig 6.Correct prediction

In [7], Kavya C S et al used Local Binary Pattern (LBP) to extract features and K-nearest neighbours classifier. It gave an accuracy of 70%. In [8], Fatemeh Vakhshiteh et al used Hidden Markov Model for word recognition and produced an accuracy of 80.25%. In [1], Mei-Li Zhu et al used AutoEncoders and achieved an accuracy of 80.07%. In [9], Youda Wei and Xiadong Hu used Convolutional Neural Network(CNN) which resulted in 92.76% accuracy. In the proposed approach, using AutoEncoders and LSTM, it gave an average accuracy of 70.73% for 10 words. Fig 7 shows the comparative results.

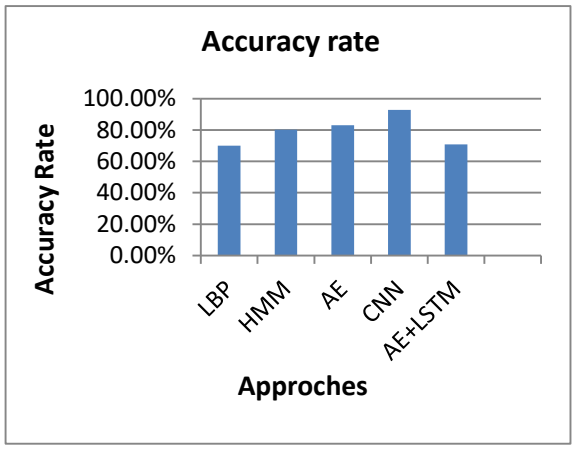

Fig 7. Comparative results.

VI. Conclusion

A new approach for building a lip reading system is proposed in this paper. Lip reading is helpful for hearing impaired in their day to day life. LSTM neural networks are useful for video processing, to understand the interdependencies between various frames. The approach is modelled upon the dataset of 19 isolated words, however, this system can further be extrapolated

March - April 2020
ISSN: 0193-4120 Page No. 2254 - 2260

for interpreting sentences being said by a speaker in a video.

VII. REFERENCES

[1] Zhu, Mei-li, Qing-qing Wang, and Jiang-lin Luo. "Lip-Reading Based on Deep Learning Model." In *Transactions on Edutainment XV*, pp. 32-43. Springer, Berlin, Heidelberg, 2019.

[2] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features," *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, Shahrekord, 2017, pp. 195-199.

[3] WenJuan, Yao & Liang, Yaling & MingHui, Du. (2010). A real-time lip localization and tacking for lip reading. 6. 10.1109/ICACTE.2010.5579830.

[4] B. T. H. Nguyen, T. V. Hieu and B. N. Dung, "Robust lip feature detection in facial images," *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Guilin, 2017, pp. 867-871.

[5] Ping-xian, Yang, Guo Rong, Guo Peng and Fang Zhaoju. "Research on lip detection based on Opencv." *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)* (2011): 1465-1468.

[6] X. Qi, T. Wang and J. Liu, "Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision," *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Harbin, 2017, pp. 151-155.

[7] Kavya, C. S., N. H. Poornima, N. Sahana, K. V. Vidyashree, and G. R. Kiranmayi. "Conversion of LIP movement to speech: An aid to physically impaired and dumb people." In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1868-1871. IEEE, 2016.

[8] Vakhshiteh, Fatemeh, Farshad Almasganj, and Ahmad Nickabadi. "Lip-reading via deep neural networks using hybrid visual features." *Image Analysis & Stereology* 37, no. 2 (2018): 159-171.

[9] Wei, Youda, and Xiaodong Hu. "Text Recognition from Silent Lip Movement Video." In *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, pp. 168-171. IEEE, 2018.

Sanika Gowaikar
Dept of Information Technology
Sinhgad College of Engineering
Pune, India
sanika.gowaikar@gmail.com

Ganesh Pathak
Dept of Computer Science & Engg.
Sathyabama Inst. Of Sci. & Tech.
Pune, India
ganeshpathak@sinhgad.edu

Shivani Devi
Dept of Information Technology
Sinhgad College of Engineering
Pune, India
shivanigdevi@gmail.com

Vaishnavi Gagare
Dept of Information Technology
Sinhgad College of Engineering
Pune, India
vaishnavi.gagare97@gmail.com

Kalyani Chavan
Dept of Information Technology
Sinhgad College of Engineering
Pune, India
chavankalyani87@gmail.com