Vision based Lip Reading System using Deep Learning

Conference Paper · July 2022

DOI: 10.1109/CCGE50943.2021.9776430

5 authors, including:

Smriti H Bhandari
Research
21 PUBLICATIONS **147** CITATIONS

SEE PROFILE

2021 International Conference on Computing, Communication and Green Engineering (CCGE)
JSPM's RSCOE, Pune, India. Sep 23-25, 2021

Vision based Lip Reading System using Deep Learning

Nikita Deshmukh
Department of Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Ashta, India
apurvamali27@gmail.com

Anamika Ahire
Department of Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Ashta, India

Smriti H Bhandari
Department of Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Ashta, India

Apurva Mali
Department of Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Ashta, India

Kalyani Warkari
Department of Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Ashta, India

2021 International Conference on Computing, Communication and Green Engineering (CCGE) | 978-1-6654-1509-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCGE50943.2021.9776430

***Abstract*— Lip reading is an approach for understanding speech by visually interpreting lip movements. Vision based lip leading system takes a video (without audio) as an input of a person speaking some word or phrase and provides the predicted word or phrase the person is speaking as output.**

This paper presents the method for Vision based Lip Reading system that uses convolutional neural network (CNN) with attention-based Long Short-Term Memory (LSTM). The dataset includes video clips pronouncing single digits. The pretrained CNN is used for extracting features from pre-processed video frames which then are processed for learning temporal characteristics by LSTM. The SoftMax layer of architecture provides the result of lip reading. In the present work experiments are performed with two pre-trained models namely VGG19 and ResNet50 and the results are compared. To further improve the performance of the system ensembled learning is also used. The system provides 85% accuracy using ResNet50 and ensemble learning.

***Keywords*— CNN; RNN; LSTM; Attention Mechanism; automatic lip reading; deep learning**

I. INTRODUCTION

In recent years, machine learning has made substantial influence on social progress, promoting the rapid growth and development in artificial intelligence technology and solving a variety of real-world problems. Many human-computer interaction and virtual reality (VR) technologies are built using automatic lip reading technology. It has the potential to be extremely useful in visual perception and human language communication. Automatic lip reading can help learning and understanding lip language and lip movements for speech recognition by reducing the time and effort required. In noisy environments where audio speech recognition may be problematic, visual lip reading plays a important role in human- computer interaction. It can also be used as a hearing aid for those who are having hearing disabilities. An automatic lip reading system can be utilized to recognize speech, making the life of hearing-impaired people easy. If audio in a video is not of good quality or audio is noisy, then speech to text recognition is difficult, so in such cases vision based lip reading system helps to get accurate result.

The feature extraction and classification are usually the two processes in traditional lip reading systems. Previously, most feature extraction algorithms used pixel values taken from the region of interest (ROI), that is, mouth as visual input in the first stage. The abstract image features are retrieved using principal component analysis (PCA), discrete cosine transform (DCT), discrete wavelet transform (DWT) etc. [1]. The second stage involves support vector machine (SVM) and hidden Markov model classifiers (HMM) for prediction based on visual features obtained from first stage [1]. This method assures that no information is lost but size of the feature vector can be very large and may also contain considerable redundancy. The transformation method concentrates the majority of the image's energy into a limited number of coefficients thus eliminating redundancy. The transform coefficients are ranked according to the importance of the data they represent [2]. The classification is mostly based on static and dynamic data. The dynamic data have the spatial and temporal variations. HMM, Artificial Neural Network (ANN), Gaussian Mixture Model (GMM) and other classifiers can be defined [3].

Due to significant growth and development in the field of computer vision along with deep learning, the research is focused on the end-to-end deep learning architectures where there is no manual intervention for extracting the features. The LSTMs were introduced about couple of decades ago. Since then, a lot of success is observed in a variety of human language technologies, such as bidirectional LSTM based acoustic models and language models in speech recognition. Oscar Koller et al. [4] demonstrated CNN-LSTM network training for recognition tasks categorizing more than 1000 classes such as action gesture, activity and sign language interpretation. Although CNNs have made tremendous progress in gesture and sign language processing, motion appears to play a significant role in these tasks and relying on the HMM sequence for capturing temporal change may not be sufficient. The goal of combining deep CNN with LSTM layers is to make it easier to train the entire network. For efficient training, the LSTM can work with a vast amount of data. They can model nonlinearity without relying on the Markov assumption, which may be an advantage over more standard models like HMM [4].

The proposed method for automatic lip reading recognition has three phases. The initial process is to extract keyframes from a sample video, which are then utilized to find critical points of the lip region or mouth and locate the ROI. This ROI is computationally processed in successive frames. The VGG19 network and ResNet50 are used to extract the characteristics from the original mouth image. First step is to preprocess the input video which includes extracting keyframes and positioning of mouth. The second is an attention-based LSTM network, which uses video keyframe features that helps learning sequential information with attention weights. SoftMax layer provides the final recognition result.

This paper is ordered in five sections. Section II provides literature review focusing upon the research efforts using deep neural networks. Section III provides detailed methodology proposed for automatic lip reading. Section IV presents the results along with critical observations and comments. Section V concludes the paper providing future direction.

II. LITERATURE REVIEW

Several research papers were reviewed to understand the work done in this area of automated lip reading. The researchers have used CNN based approaches with varying architectures of pre-trained models [1][5-9] and datasets. The highest accuracy reported with single word videos is 88%. Table I provides comparison of these approaches considering Methodology, Dataset used, Performance and limitations.

III. PROPOSED WORK

The proposed system uses CNN and attention-based LSTM. The block diagram of the lip reading system is shown in Fig. 1.

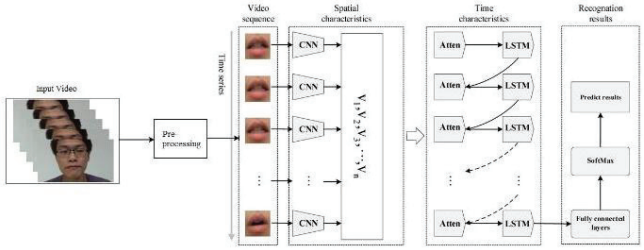

Fig. 1. Block diagram of Lip Reading System [1]

The video (without audio) is presented to the system as input. This video gets preprocessed to form the video sequence with region of interest. Then the CNN is used to extract spatial characteristics from the keyframes presented in terms of a feature vector. Further, these feature vectors will be used by attention-based LSTM to extract temporal characteristics from the sequence of frames. After LSTM the fully connected layers are added to predict the speech. The methodology followed to perform major tasks is explained in the following sub-sections.

A. Preprocessing of input video

As mentioned in Fig. 1 the input video needs to be preprocessed. To obtain the desired frames to be processed for feature extraction the steps as shown in Fig. 2 are followed. The input is the facial video of a person uttering some text. To understand some useful information from the input video there is a need to preprocess individual frames of video. Hence the first task is to extract the frames from the video.

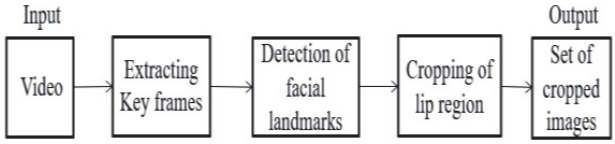

Fig. 2. Preprocessing of input video

TABLE I. LITERATURE REVIEW: COMPARISON OF SELECTED PAPERS

Paper / Year of Publication	Methodology	Limitations	Dataset	Performance
[1] /2019	"Deep convolutional neural network (VGG19) and attention – based long short-term memory"	Requires a good quality of video	Own dataset consisting of three males and three females	Accuracy of 88.2%
[5] / 2019	"Convolutional neural network along with pre-trained models (AlexNet, GoogleNet)"	Gives more accurate result specifically for alphabet level recognition.	AvLetters [10] dataset	Accuracy of 64.40%.
[6] / 2016	"VGGNet along with SVM"	Model trained from scratch did not perform well as size of dataset is small	MIRACL-VC1 [11] dataset	Accuracy of 76%.
[7] / 2017	"Deep convolutional neural network (VGG16) and attention – based long- short term memory"	Does not support larger dataset	MIRACL-VC1 dataset	Validation accuracy of 79% and test accuracy of 59%.
[8] / 2020	"Viseme concatenation and 3D Convolutional Neural Networks"	Limited training data	MIRACL-VC1 dataset	Accuracy of 76.89%.
[9] / 2017	"Combination of spatiotemporal convolutional, residual and bidirectional Long Short-Term Memory networks."	If phonetic and "visemic" content of the word pairs are similar then correct identification of the first and last viseme of a word may be difficult.	LRW [12] dataset	Accuracy of 83.00%.

Usually, videos are acquired with a frame rate of 30 frames per second. To build the model, it becomes essential to extract the features and to get the hidden relationship of the sequence of frames. The way the word gets pronounced and the length of each utterance is different concerning the subject uttering the word. Also, it may happen that while pronouncing a specific word there is a series of redundant information regarding the lip movements. Therefore, the redundant information must be removed from all the extracted original frames. It will also help to balance with training speed and recognition results. Thus, the proposed method does not use all the frames from the video sequence, instead only the keyframes are being extracted and the lip regions are segmented for further processing.

Extraction of keyframes

The video or the time of the utterance or the total number of frames is split into 10 equal intervals, and a random frame is picked from each interval as a keyframe. Thus, video gets converted into 10 keyframes providing uniform input length. For example, if a video comprises of 40 frames, then it is partitioned into 10 parts with equal intervals. That is, 10 partitions with four frames per interval. Further after randomly picking 1 frame per partition, the entire video now gets converted into a sequence of 10 frames.

Detection of facial landmarks

The keyframes extracted from the video are further processed to detect the Facial landmarks. Detecting facial landmarks can be seen as a subset of the problem of shape prediction. Given the input typically ROI indicating object of attention, the shape predictor localizes interest points or the key points around the shape. In the context of facial landmarks, these methods attempt to identify key facial structures on the identified face.

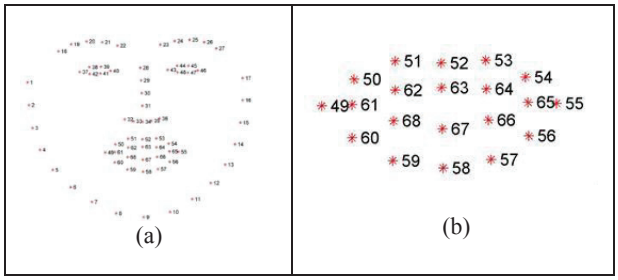

Fig. 3. Detection of facial landmarks and cropping of lip region (a) Detection of facial landmarks (b) Localization of mouth /lips for cropping

It involves a two-step process: i) Locating the face in the image. ii) Identifying the most important facial structures in the ROI of face. The dlib library is used for detecting landmarks. The facial landmark detector of the dlib library is based on the research paper by Kazemi and Sullivan [13]. This algorithm provides 68 facial landmarks. The indexes of these landmarks can be viewed as in the image shown in Fig. 3 (a).

Localization of mouth /lip region

After detection of facial landmarks, referring to the points representing the lips in the image, the lip region is cropped. The points that were considered for cropping the lip area are 49, 55, 52, and 58 as shown in Fig. 3 (b) which are the extreme points depicting the lips. Each keyframe is cropped as per the coordinates of the lips and resized to the standard size of 224 x 224.The output of the preprocessing step is the set of 10 cropped lip images per video.

B. Implementation of CNN and LSTM

The set of cropped images are given as input to the spatial characteristics block consisting of CNN as shown in Fig. 4 The VGG19 and ResNet50 CNN architectures are used. VGG19 and ResNet50 give the set of spatial features (Feature vector) for each input frame. VGG19 architecture is as shown in Fig. 4. and Fig. 5 describes the architecture of ResNet50 pre-trained network.

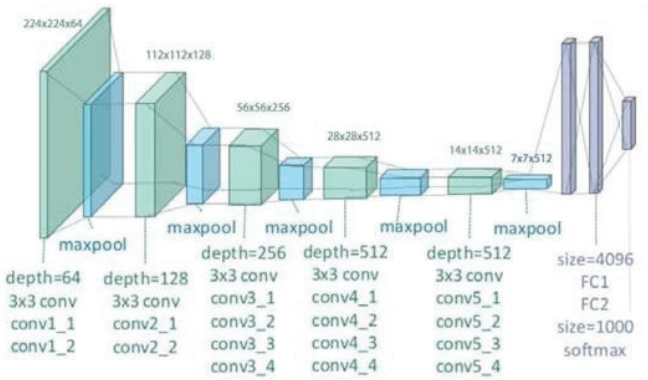

Fig. 4. Architecture of CNN VGG19

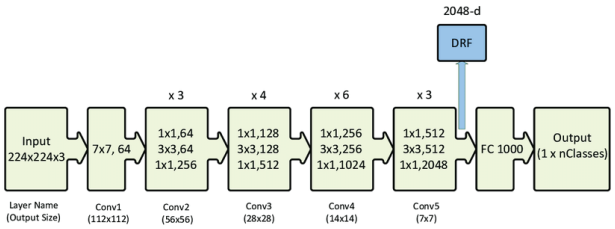

Fig. 5. Architecture of ResNet50 Pre-trained Network [14]

The pre-processed lip images of size 224 x 224 are given as input to VGG19 and ResNet50 networks. For the purpose of feature extraction, VGG19 architecture up to the first fully connected layer is used. For ResNet50 also the layers up to FC1000 are used. Thus, both the architectures provide feature maps for each frame of individual video. The size of feature vector are 4096 and 2048 dimensions for VGG19 and ResNet50 respectively. The set of feature vectors then are provided as an input to the Time characteristics block of Fig. 1 consisting of attention-based LSTM. LSTM is a special type of RNN which will learn long-term dependency information. LSTM with an attention mechanism is used so that model pays greater attention to the effective areas of the entire video. The feature vector for every frame is thus weighted and used as an input to the LSTM. The attention-based LSTM layers are introduced to add temporal information of lip sequences. LSTM consists of different memory blocks called cells. There are three gates, input, output, and forget gate. The first step in LSTM will be of deciding to drop useless information from the cell state by using cell state and the new information will be stored in the cell state. Each time the cell state will be updated as per input given for learning sequence information (sequence relation) as per time interval. The sequence information is provided to the fully connected layer which then flattens the sequence information into a single vector, which is given to the SoftMax layer. The SoftMax layer converts the output in terms of probabilistic result and based on the probability, results are predicted.

C. Performance analysis

Performance analysis is done by using a confusion matrix. The dataset gets divided into training and testing datasets to do the performance analysis of the algorithm/system. The overall accuracy gives the rate of correct prediction. Additionally, to get more insights about the results, class-wise and subject-wise accuracies are also obtained.

Further to improve the performance of system, ensemble learning is experimented.

Ensemble Learning

Ensemble methods in statistics and machine learning combine multiple learning algorithms to provide higher predictive performance than any of the individual learning algorithms. Thus, for implementation of ensemble method, seven different models are used. Experiments are performed with CNN + LSTM and also with CNN + attention-based LSTM using VGG19 and ResNet50. Each model votes for a particular class. The class with maximum voting will provide the final result of the method.

IV. RESULTS AND DISCUSSION

Dataset - The dataset [15] has total 540 videos of 6 persons pronouncing digits from zero to nine. Six people are facing the camera and speak digit 0 to 9. All the videos are recorded into the full-frontal pose. Each speaker speaks each digit nine times.

After implementing the proposed methodology, the prediction results are obtained with following architectures and models:

A. Using VGG19 as pre-trained model:

1) *VGG19 + Simple LSTM (single model)*
2) *VGG19 + attention-based LSTM (single model)*
3) *VGG19 + Ensemble with LSLM (7 models of LSTM)*
4) *VGG19 + Ensemble with attention-based LSLM (7 models of attention-based LSTM)*

B. Using ResNet50 as pre-trained model:

5) *ResNet50 + Simple LSTM (single model)*
6) *ResNet50 + attention-based LSTM (single model)*
7) *ResNet50 + Ensemble with LSLM (7 models of LSTM)*
8) *ResNet50 + Ensemble with attention-based LSLM (7 models of attention-based LSTM)*

The class-wise, subject-wise and overall performance is analyzed to gain more insights about the proposed methodology.

C. Class-wise accuracy

Fig. 6 shows comparison of class-wise accuracies of each digit for four different models of VGG19.

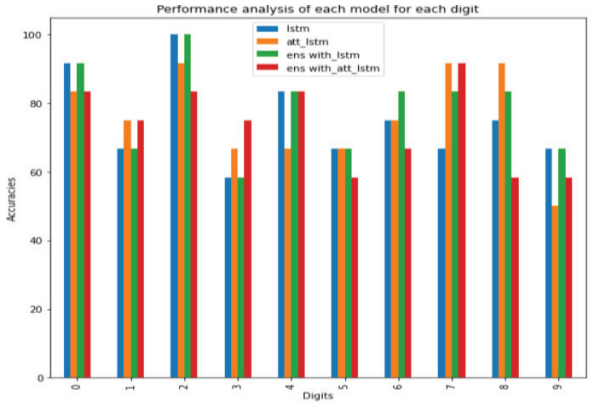

Fig. 6. Class-wise accuracies (VGG19)

Fig. 7 shows comparison of class-wise accuracies of each digit for four different models of ResNet50. It is observed that, Digit 2 has the highest and digit 3 has the lowest accuracy as compared to other digits.

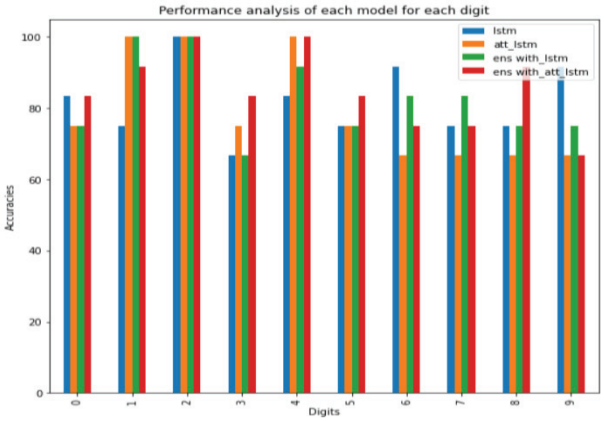

Fig. 7. Class-wise accuracies (ResNet50)

B. Subject-wise accuracy

Fig. 8 shows comparison of subject-wise accuracies of each speaker for four different models of VGG19.

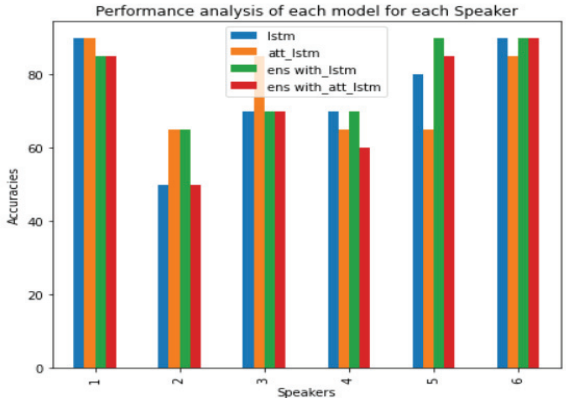

Fig. 8. Subject-wise accuracies (VGG19)

Fig. 9 shows comparison of subject-wise accuracies of each speaker for four different models of ResNet50.

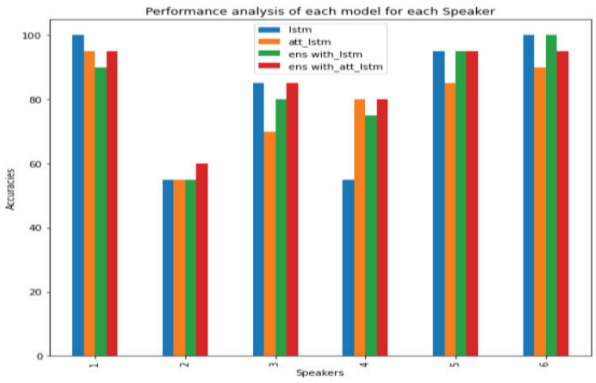

Fig. 9. Subject wise accuracies (ResNet50)

It is observed that, speaker 6 has the highest accuracy and speaker 2 has the lowest accuracy in all the four models as compared to other speakers in both VGG19 and ResNet50 architectures. In order to find the reason, the input frames are visualized for these speakers.

(a) Input video sequence of Speaker 2 for digit 0

(b) Input video sequence of Speaker 6 for digit 0

Fig. 10. Input video sequences for digit 0

Fig. 10 (a) shows the input video sequence of speaker 2 for digit 0 whereas Fig. 10 (b) shows the input video sequence of speaker 6 for the digit 0. The lip movement of speaker 2 is much similar in all the frames; so, may be the features or information distinguishing the pronunciation of various digits cannot be captured well in case of speaker 2. That may be the reason for getting less accuracy for speaker 2.

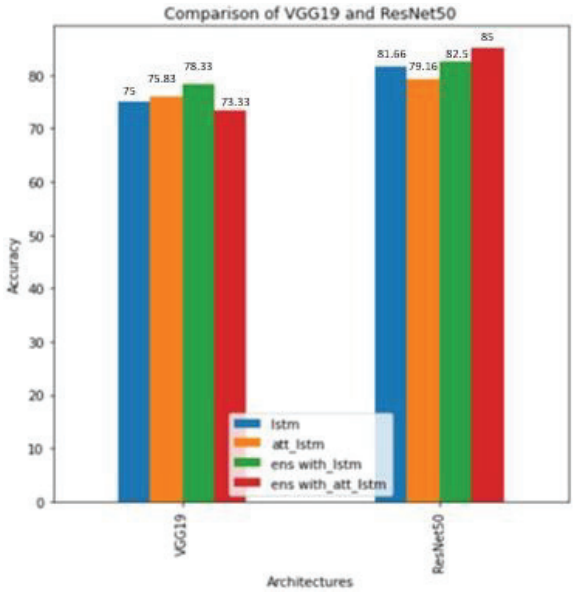

Fig. 11. Comparison between accuracies of VGG19 and ResNet50

Fig. 11 shows the comparison between the overall accuracies obtained from VGG19 and ResNet50 architectures.

ResNet50 gives more accuracy as compared to VGG19 in all the four models. The highest accuracy obtained is with ResNet50 with ensemble attention-based LSTM model which is 85%.

V. CONCLUSION

The proposed methodology for Vision based lip reading system is a combination of CNN along with LSTM. The system works with the videos (with no audio) having single word uttered by the speaker. Initially preprocessing of videos is done to get the keyframes with localization and cropping of mouth or the lip region. CNN is used for feature extraction and LSTM is used for learning the sequence information. Two pre- trained CNN architectures namely VGG19 and ResNet50 are used. The final result is predicted by two fully connected layers followed by the SoftMax layer. The dataset used for experimentation includes video clips pronouncing single word (digit). To improve accuracy of the system the ensemble learning is used. It is found that ResNet50 provides improved results than VGG19 and also ResNet50 is faster than VGG19. The overall accuracy of the proposed method using ResNet50 is 85%.

There is a scope to improve performance of developed system. The implemented system predicts the digits i.e. one word at a time. Also, presently the experiments are performed with only one dataset. To check the robustness of the system, it should be tested with few more datasets. In future it may be extended to predict more number of words from a given input sequence at a time.

REFERENCES

[1] Lu, Y.; Li, H. “Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory,” Appl. Sci. 2019, 9, 1599. https://doi.org/10.3390/app9081599

[2] Patricia Scanlon , Richard Reilly and Philip de Chazal, “Visual Feature Analysis for Automatic Speech reading,” International Conference on Audio-Visual Speech Processing, September 2003

[3] Priyanka P. Kapkar and S. D. Bharkad, “Lip Feature Extraction and Movement Recognition Methods”, International Journal of Scientific & Technology Research, vol.8, August 2019.

[4] Oscar Koller, Sepehr Zargaran and Hermann Ney, “Re-Sign: Re-Aligned End-to-End Sequence Modeling with Deep Recurrent CNN-HMM,” Human Language Technology & Pattern Recognition Group RWTH Aachen University, Germany, 2017

[5] Tayyip Ozcan and Alper Basturk, "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models," Balkan Journal of Electrical and Computer Engineering, Vol. 7, No. 2, April 2019.

[6] Amit Garg, Jonathan Noyola, Sameep Bagadia, “Lip reading using CNN and LSTM,” 2016

[7] Abiel Gutierrez and Zoe-Alanah Robert, "Lip Reading Word Classification", Stanford University, 2017

[8] Pooventhiran G, Sandeep A, Manthiravalli K, Harish D, Karthika and Renuka D, “Speaker-Independent Speech Recognition using Visual Features,” International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 2020

[9] Themos Stafylakis and Georgios Tzimiropoulos, “Combining Residual Networks with LSTMs for Lip reading,” Computer Vision Laboratory University of Nottingham, UK, arXiv:1703.04105v4 [cs.CV]

[10] http://www2.cmp.uea.ac.uk/~bjt/avletters/

[11] https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1

[12] https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html

[13] Vahid Kazemi and Josephine Sullivan, “One Millisecond Face Alignment with an Ensemble of Regression Trees,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[14] Mahmood, Ammar et al. “Automatic Hierarchical Classification of Kelps Using Deep Residual Features.” Sensors (Basel, Switzerland) vol. 20,2 447. 13 Jan. 2020, doi:10.3390/s20020447

[15] https://drive.google.com/u/1/uc?id=1ftS9GHYkOyQ-hQdFvamTblYyIWQCTWud&export=download