Integration of Audio video Speech Recognition using LSTM and Feed Forward Convolutional Neural Network

Shashidhar R (shashidhar.r@sjce.ac.in)

JSS Science and Technology University,Sri Jayachamarajendra College of Engineering

https://orcid.org/0000-0002-3737-7819

Sudarshan Patil Kulkarni

JSS Science and Technology University

Research Article

Keywords: Audio visual speech recognition, lip reading, FFNN, LSTM, Deep neural network

Posted Date: March 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-173380/v1

Wireless Personal Communication (2021)
https://doi.org/10.1007/s00034-017-0623-5

Integration of Audio video Speech Recognition using LSTM and Feed Forward Convolutional Neural Network

Shashidhar R[1] · S Patilkulkarni[2]

Abstract In the current scenario, audio visual speech recognition is one of the emerging fields of research, but there is still deficiency of appropriate visual features for recognition of visual speech. Human lip-readers are increasingly being presented as useful in the gathering of forensic evidence but, like all human, suffer from unreliability in analyzing the lip movement. Here we used a custom dataset and design the system in such a way that it predicts the output for the lip reading. The problem of speaker independent lip-reading is very demanding due to unpredictable variations between people. Also due to recent developments and advances in the fields of signal processing and computer vision. The task of automating the lip reading is becoming a field of great interest. Here we use MFCC techniques for audio processing and LSTM method for visual speech recognition and finally integrate the audio and video using feed forward neural network (FFNN) and also got good accuracy. That is why the AVSR technique attract a great attention as a reliable solution for the speech detection problem. The final model was capable of taking more appropriate decision while predicting the spoken word. We were able to get a good accuracy of about 92.38% for the final model.

Keywords Audio visual speeeh recognition, lip reading, FFNN, LSTM, Deep neural network

Shashidhar R
shashidhar.r@sjce.ac.in

S Patilkulkarni
sudarshan_pk@sjce.ac.in

1 Department of Electronics and Communication Engineering, JSS Science and Technology University, Sri Jayachamarajendra college of Engineering, Mysuru, India-570006

2 Department of Electronics and Communication Engineering, JSS Science and Technology University, Sri Jayachamarajendra college of Engineering, Mysuru, India-570006

1 Introduction

Audio speech recognition is the most widely used technique today to automatically detect what a person is saying into the form of text. In modern era, it has gained a lot of popularity and we use it almost in our day to day life in the form of Google Assistant or even the Amazon Alexa. But, the common observation is that this audio speech recognition is mostly use in indoors and does not give good response in outdoors. This is due to intervention of noise. The noise adds to the audio signals and most of the necessary data are lost. That is not the case when it comes to VSR. The Visual Speech Recognition has some advantages over the Audio Speech Recognition. They are a) it is not attentive to audio noise and modification in audio environments has no effect on the data. b) Does not need the user to make a sound. In present times we have a lot of data available and even possess a high computational ability. This will also serve us with an advantage to use various machine learning and deep learning algorithms to get the best results possible.

The LRS2 database is used as most common database available [1] [7]. The feature extraction is done in the Region of Interest. The performance given by the audio speech recognition is not that good compared to the performance given by the AVSR in noisy conditions. The noise can be of different types like street, train etc. It shows that the noise independent of the type gives the same results. Lip-reading is the job of deciphering transcript from the measure of a presenter's mouth. Ahmad B A Hassanat explained different approaches of lip localization [2]. Ayaz A. Shaikh et al proposed the depth sensor camera has also been used to get the third dimension in the dataset [3]. During the creation of the dataset, the above mentioned factors have been taken care by using a headrest [3]. Themos Stafylakis et al proposed residual and LSTM techniques for LRW database and get 83% accuracy [4]. Shillingford et al has used Lipnet. In the lipnet, used two approaches to solve the problems are learning the visual features and prediction. The word error rate of this work is 89.8% and 76.8% [6].

One of the other architecture to implement Lip-reading is Long-Short Term Memory (LSTM) [7]. LSTM are used for lip-reading that determines the typical is accomplished of selectively indicating which spatiotemporal balances are important for an individual dataset. LRS2 datasets used in the model and it achieves 85.2%. G. Sterpu et al [8] looks int futuristic Deep Neural Network architectures for lip reading based on a sequence to sequence Recurrent Neural Network. This work make sure for both redeveloped and 2D/3D Convolutional Neural network visual frontends, operational monotonic consideration, and a combined connectionist Temporal Classification Sequence-to-sequence loss. This evaluated

system is done with fifty nine talkers and a terminology of over six thousand arguments on the widely accessible TCD-TIMIT dataset. Kumar et al [9] showed the set of experiments in detail for speaker dependent, out-of-vocabulary and speaker independent settings. In order to show the real time nature of audio produced in the system, the hindrance values of Lipper has been compared with other speech reading systems. The audio only accuracy is 80.25%, the annotation accuracy variance is 2.72% in audio, and Audio-visual accuracy is 81.25%, the annotation accuracy variance is 1.97% in audio-visual.

Some of common dataset used in lip reading is Grid audio-visual dataset, the work in [10] is based on the Grid audio-visual dataset. Visual dataset are recorded with a frame rate is 25 Frames per second, totally 75 frames per sample for 3 seconds. In this work LCA Net and end to end deep neural network are used. The system archives 1.3% CER and word error rate is 3.0%. Dilip kumar Murgm et. al. suggested the new-fangled SD-2D-CNN-Bidirectional Long Short term memory[11] architecture. The analysis of two different approaches like 3D-2D-Convolutional neural network-BLSTM trained with CTC loss on Characters and 3D-2D-Convolutional neural network-BLSTM trained with CTC loss on word labels for lip reading are presented. For the first approaches, word error rate is 3.2% and 15.2% for seen and unseen words respectively. Performance on grid dataset of the second approaches, word error rate is 1.3% and 8.6% for seen and unseen words respectively. The performance of the Indian English unseen dataset word error rate is 19.6% and 12.3% for the two approaches.

One of the most famous dataset used for lip reading is “Lip Reading in the wild (LRW)” [12][13][22] from BBC Tv it contains 500 targeted words. Themos Stafylakis et al used Residual networks and Bidirectional LSTMs and misclassification rate of the architecture is 11.92%. Used the same database and same method got 83% accuracy. Audio visual speech recognition is one of prospective explanation for speech recognition in noisy environment [13]. Shiliang Zhang et al used bimodal –DFNN, used 150 hours multi-condition training data and archives 12.6% phone error rate for clean test data. Word error rate is 29.98%.

Kuniaki Noda et al introduce a multi stream HMM model for integration of Audio and Visual features [14]. Word Recognition rate of MSHMM is 65% and Signal to noise ratio is 10 dB. Stavros Petridis et al Long –short Memory based end-end visual speech recognition classification [15]. The model contain of two streams which citation features straight away from the mouth. The two streams take place via bidirectional Long Short Term Memory. Databases like ouluVS2 and CUAVE used, the accuracy of the work is 9.7% and 1.5% respectively.

Fei Tao et al[16] proposed structure is likened with Conventional HMMs with observation models implemented with Gaussian mixture model(GMM) and used this channel matched word error rate is 3.70% and Channel mismatched word error rate is 11.48%. The hybrid Connectionist Temporal Classification architecture for audio-visual recognition of speech in-the wild is used in the [17].

The audio features are of many kinds. The three of them used in [18] are LPC,PLP, MFCC. The study shows that the MFCC has the highest accuracy of about 94.6% for Hindi Language in noiseless environment. Overall performance of about 86% is shown by MFCC compared to PLP which has about 83%. RNN is used for audio input prediction. For the visual features the Active Shape Models and the Active Appearance Model is used to detect the lip region [2]. It proceeds a lot of period to create and process the data to be in the format required for the

application. The objective that is defined in the work [19] can be affected by the varying light intensity, movement of the head, the distance from the camera.

Ochiai et al proposed the most significant speaker clues is extracted from the dataset. This is attention based feature extraction. They have used 3 layer BLSTM with 512 units each. They further suggest to use datasets in which the video is taken from different angles. Doing so can improve the accuracy as it is common for humans to move their heads while talking [20].

Joon Son Chung et al proposed a new set of database called LRS it contains 100,000 normal sentences from BBC television [21].

Jha. V. P. Namboodiri et al used Charlie Chaplin videos, The word spotting technique achieves 35% upper despicable typical accuracy over recognition-based technique on extensive LRW dataset. Determine the request of the technique by word recognizing in a standard speech videos are "The great dictator" by Charlie Chaplin[22].

Z. Thabet et al applied a machine learning approaches to identify lip-reading and used nine dissimilar classifiers has remained applied and verified, reporting their misperception mediums among dissimilar clusters of arguments. The classification procedure went on more than one classifier but these three classifiers got the best outcomes which are Gradient Boosting, Support Vector Machine and logistic regression with results 64.7%, 63.5% and 59.4% respectively[23].

Yaman kumar et al proposed a speech reading or lip-reading is the method of empathetic and receiving phonetic topographies from a presenter's visual features such as movement of mouths, face, teeth and tongue. It has an extensive range of hypermedia applications such as in investigation, Internet telephony and as an aid to a person with hearing impairments [24].

Y. Lu et al proposed a visual lip reading is a technology which associations machine vision and language perception.

In the lip reading techniques, first detection of the face region from the input image then extract the mouth region of the speakers and determine the pronunciation of these features by appreciation model, in this manner they recognizing the speech contents [25].

Lip-reading in general or in particular can be used to enhance the thoughtful of what a people says and also it is greatly beneficial for hearing impaired [26, 27] people. Thus a main usage would be hearing impaired people who would mostly benefited with accurate text of what the speaker is talking. Many experiments and research studies have shown that people with [26, 27] and without [28] hearing impairment practice visual signals for understanding and enhance to understand the words of the speaker.

The rest of this paper is scheduled as follows: In section 2 explain in details of methodology. Section 3 explains the database creation and challenges faced while create the database. Section 4 explains the result and application of the work and section 5 conclude the proposed work.

2 Methodology

Different Studies mostly on human machine interfaces have showed that utilizing not only the available audio information but also the speech information present in

the visual data can boost the accuracy of speech recognition. The addition of visual information to automatic speech recognition is found to improve accuracy especially in acoustically noisy conditions where audio data is corrupted. The proposed work used English dataset for recognition of audio visual speech. The custom English dataset contains two forty five videos from five speakers pronouncing seven English words like 'ABOUT','BOTTLE', 'DOG', 'ENGLISH', 'GOOD', 'PEOPLE', and 'TODAY'. The model expects a video of speaker uttering one of the seven pre-trained words and displays the result in the form of text. However, this work does not take account of predicting speakers of different accents or it doesn't deal with predicting the sentence or individual phonemes.

Methods involved in developing Audio visual speech Recognition system. It consists of six steps explain one by one given below.

Data extraction: The video is a custom data. It contains of audio data too. We extracted the both audio and video in separate files. In our proposed work we have taken two forty five videos consisting of seven words. The audio has to be extracted from the given MP4 file. We took custom dataset video and in the first step extract the frames from the custom video. The output will be in color image and it has to be converting into gray scale to decrease the computational overhead.

Audio Feature Extraction: To accomplish audio speech recognition audio dataset holding essential features has to be shaped. An element called moviepy derived in nearby to achieve this task. Here we used custom dataset it in the form of MP4 format, this element agrees us to extract audio from MP4 files. To extract the audio part we used open source model called 'Librosa'. The Audio preprocessing and recognition here we used five audio methods the methods, are MFCCS, CHROMA, MEL, CONTRAST, TONNETZ

Video Feature Extraction:

In order to extract video features, first we have grouped the videos of same words into a single folder. Videos are taken individually from the folders and using the "shapepredictor68facelandmarks.dat" file, which is available on internet to locate the landmarks as you can see in the picture. In Figure 1 used the shape predictor to predict the human face. For our work required only lip part so we have neglected all other features and considered only the features around the mouth region. Figure 7 shows how video is divided into number frames. The coordinate values of each position is occupied and kept in an array. Then the mean deviance of the coordinates of that position is taken as the feature. This affords us with two coordinates multiplied by twenty places, in total forty video features for a video which are stored in "Videofeatures.csv

Figure 1: Preprocessing Step in single Frame.

Training models and saving weights:

Audio features are extracted using simple deep neural network in the keras model and to train the visual speech we use long short term memory recurrent neural network. For the integration of audio and visual is trained by estimate made by audio and video based models. The weights of the models are kept using Model Check Point process existing in Keras.

Loading the models:

The models with corresponding architecture are loaded with weights which are saved in previous step with the extraction of hdf5 and the 3 models are ready to predict the output.

Performance Evaluation:

The testing and evaluate the audio and visual speech can be predicted using the trained models. First we evaluate and get the confession matrix of audio part, after that we get the confession matrix of the video part and then finally combine the audio and video after getting the predictions and we get the confession matric of the combine model.

Block diagram:

The above Figure 3 is the block diagram depicts the working of AVSR. Here we use input data as video. In the first step extract the audio from video and do the processing using different techniques. Second we take video as an input but in this video, audio is absent. After that do the face detection and lip localization. Extract the lip coordinates and train the data using deep neural network called LSTM and train the audio data using simple DNN and finally combined using feed forward neural network.

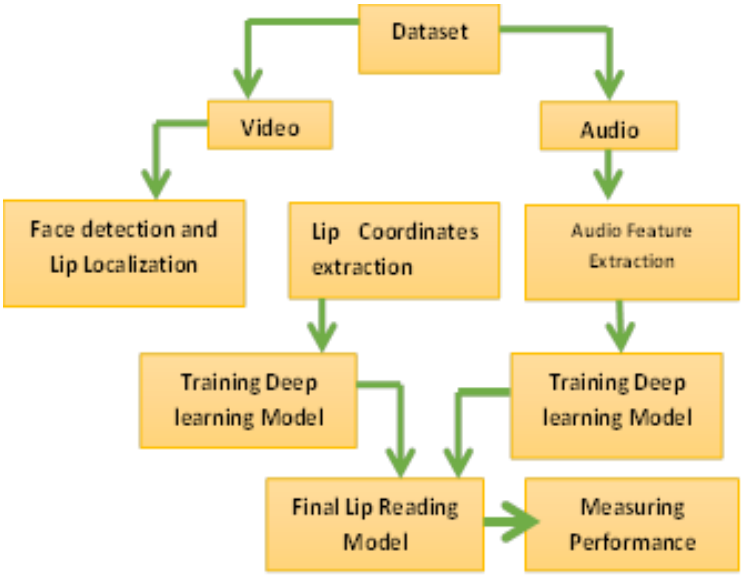
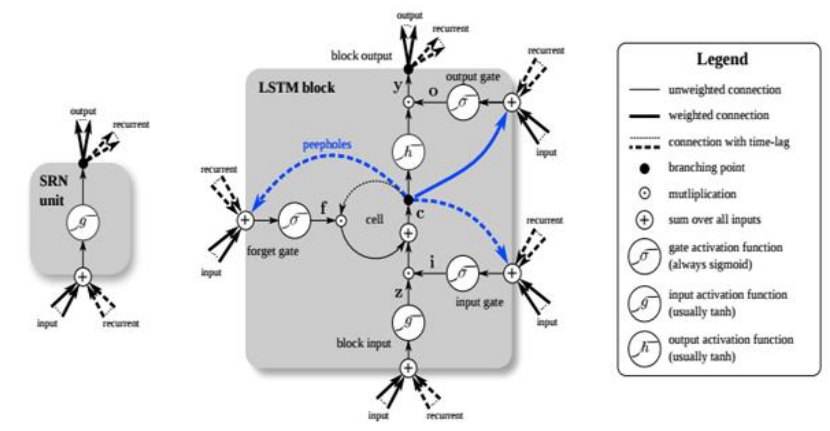


Figure.3 Block Diagram of AVSR

2.1 Mathematical Model of LSTM:



(a) (b)

Figure. 4 Complete Diagram of the Simple Recurrent Network unit (a) and a Long Short-Term Memory block (b) as used in the hidden layers of a recurrent neural network

Forward Pass

Let x^4 be the input vector at period t, N be the total number of LSTM blocks and M is input number. Then we get the following weights for an LSTM layer.

Inputs weights: W_z, W_i, W_f, $W_n \in R^{NxM}$

Recurrent Weights: R_z, R_i, R_f, $R_n \in R^{NxN}$

Peephole Weights: P_i,P_f, $P_o \in R^N$

Bias Weights : b_z, b_i, b_f, $b_o \in R^N$

Then the vector formulas for a vanilla LSTM layer forward pass can be written as:

$Z^t = W_z x^t + R_z y^{t-1} + b_z$

$z^t = g(z^t)$ block input

$i^t = W_i x^t + R_i y^{t-1} + P_i \Theta C^{t-1} + b_i$

$i^t = \sigma(I^t)$ Input gate

$f^t = W_f x^t + R_f y^{t-1} + P_f \Theta C^{t-1} + b_f$

$f^t = \sigma(f^t)$ Forget gate

$C^t = Z^t \Theta i^t + C^{t-1} + f^t$ cell

$\Delta^t = W_o X^t + Ro Y^{t-1} + Po \Theta C^t + bo$

$O^t = \sigma(\Delta^t)$ output gate

$Y^t = h(c^t) \Theta O^t$ block output

Where σ, g and h are point wise nonlinear initiation functions. The logistic sigmoid ($\sigma(x)=\frac{1}{1+e-x}$) is used as postern activation function and the hyperbolic tangent (g(x) = h(x)=tanh(x)) is usually used as the block input and output activation function. Point wise development of two vectors is denoted by Θ.

Backpropagation through Time:

The deltas inside the LSTM block are then calculated as:

$\partial y^t = \Delta^t + Rz^T \partial z^{t+1} R_1^T \partial I^{t+1} + Ro^T \partial O^{t+1}$

$\partial O^t = \partial y^t \Theta h(c^t) \Theta \sigma^t(\partial^t)$

$\partial e^t = \partial y^t \Theta O^t \Theta h^t(e^t) + Pa \Theta \partial O^t + P_i \Theta \partial f^{t+1} + P_f \Theta \partial f^{t+1} + \partial e^{t+1} \Theta f^{t+1}$

$\partial f^t = \partial e^t \Theta e^{t-1} \Theta \sigma^1(f^t)$

$\partial f = \partial c^t \Theta z^t \Theta \sigma^1(f^t)$

$\partial z^t = \partial C^t \Theta I^t \Theta g^1(z^t)$

Δ^t = vector of details passed down from the layer above.

E = loss function is formally corresponds to $\frac{\partial E}{\partial y}$ but including the recurrent dependencies.

$\partial x^t = W_z^T \partial Z^t + W_i^T \partial I^t + W_f^T \partial f^t + Wa^T \partial O^t$

3. Database

3.1 Dataset Creation

Data-set is created for both English and Kannada Words using an extensive setup which includes an electronic gimbal for stable video and a Smartphone with sufficient storage space. In table 1 mention the parameter of the dataset features.

Table 1 Dataset Features

Parameter	Value
Resolution	80x1920P
Frame/Second	FPS
Storage Duration of Video	1.20PS
Storage Size of video	Mb

The dataset is embraced of interrelated audio and lip movement data in various videos of multiple topics construing the identical words. The formation of the dataset was finished to enable the progress and proof of procedures charity to train and test the method that contains of lip-motion. The data set is a gathering of videos of agrees declaiming a fixed screenplay that is planned to be used to train software to recognize lip-motion patterns.

The recordings were collected in a controlled, noise-free, indoor setting with a smart-phone capable of recording at 4K resolution. This data-set consists of around 240 video samples per person. 11 male and 13 female subjects, with ages ranging from 18 to 30, volunteered for the data-set creation process. This data-set can be used for speech recognition, lip reading applications. Around 240 video samples were collected per subject.

3.2 Challenges while Creating Dataset

There were various challenges that were encountered during the data-set creation process which are explained below

- Interference of external noise may cause disruption for audio feature extraction. Noise free environment is an important requirement of data-set creation.

- Lip movement of an individual should be in conjunction with each other in order to extract the lip feature, random movement of lip leads to error.
- Each person who is ready to give database has to spare around 30 to 45 minutes reciting the words, which can be tedious.
- Recording a video of person with moustache or beard leads to difficulty in detecting the lip movement.
- Selection of English and Kannada words to prepare database was difficult as some of the words have similar pronunciation.

Creating database for both Kannada and English words is tedious and time consuming as it is needed to take a number of samples for the same word from different people, and there is need to regularly upload the videos to hard disk or system in order to clear space in mobile so that new videos may be captured.

4. Result and Discussion

The metrics used to measure the concert of the model are accuracy in classification and confusion matrix.

Accuracy rate: It is defined as the number of exact predictions by model divided by total number of predictions.

$$\textbf{Accuracy} = \frac{\textbf{Number of Correct Predictions by the model}}{\textbf{Total Number of Predicions made}}$$

Confusion Matrix:

The confusion matrix defined the number predictions of each word. Predictions are correct and incorrect by the classical model are with the equivalent classes and total values.

The Figure5 shows the feed forward neural architecture of the audio only model and in Figure 6 shows the confusion matrix of the audio model and here shows the correct and incorrect predictions of the words and the audio only model accuracy is 91.42%. Figure 7 shows the confusion matrix of the video model and the accuracy is 80%. In Figure 12 shows the confusion matrix of combine model and the accuracy of combined model is 92.38%.

In Figure 8 shows the accuracy of the audio and video combined model and Figure 9 shows the loss curve of the combined model. Figure 10 shows the accuracy curve of the LSTM model, this algorithm used for visual speech recognition. Figure 11 shows the loss details of the LSTM model.

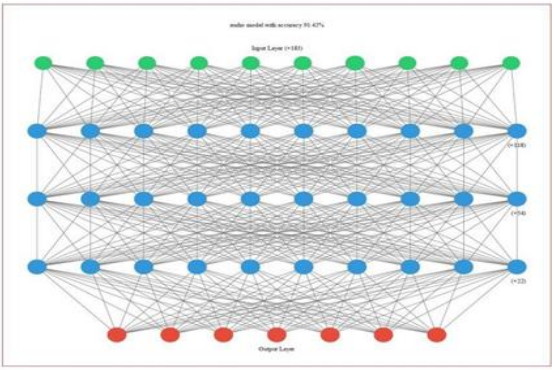


Figure 5: Feed Forward Neural Architecture of Audio Model

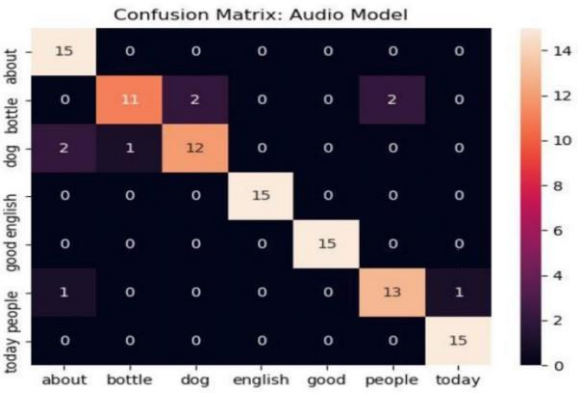


Figure 6: Confusion matrixes for audio model

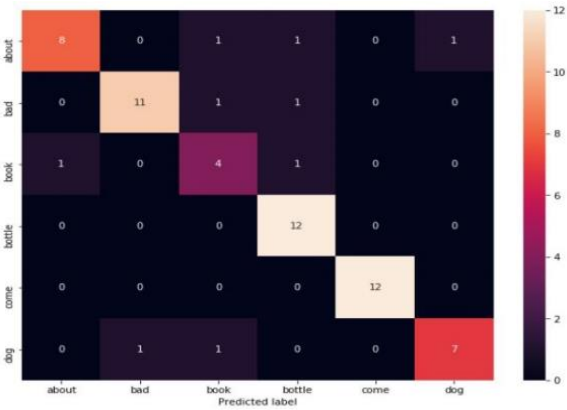


Figure 7: Confusion matrixes for video model

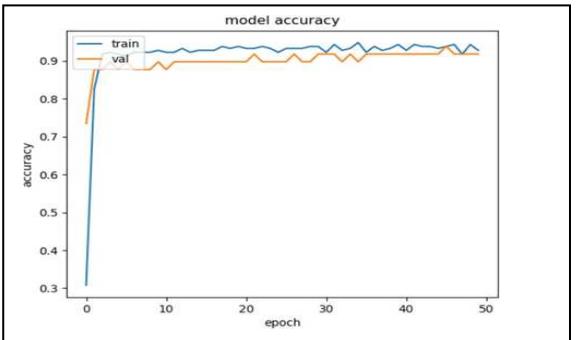


Figure 8: Accuracy curve of combined Model

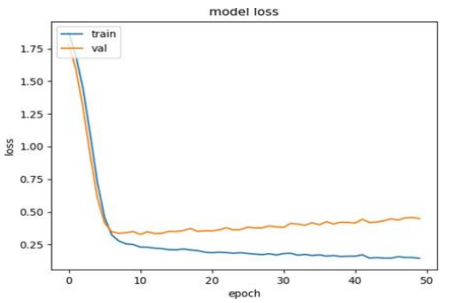


Figure 9: Loss curve of Combined Model

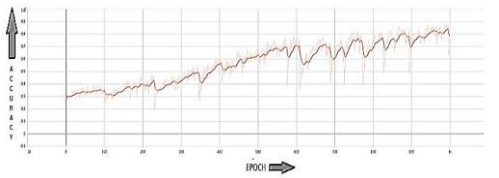


Figure 10: Accuracy curve for LSTM model

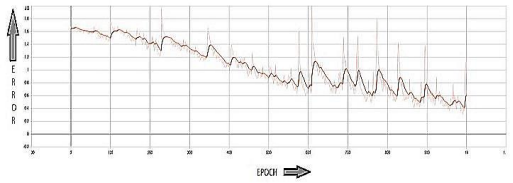


Figure 11: Loss curve for LSTM model

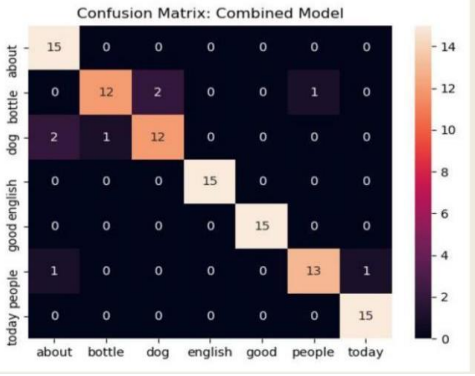


Figure 12: Confusion matrix of combined model

:

Output: For testing the developed model, an user friendly user interface has been developed using Python and are shown in following Figures.

Table 2 Comparison with existing work

Dataset	Source	WER/Accuracy
LRS2-BBC	BBC	8.2% WER
LRS3-TED	TED & TEDx	5 to 6% WER
Custom Dataset	Created own dataset	92%word accuracy.

Applications

- AVSR technique can be used in forensics so that the crime branch people can understand what have been spoken just with the help of lip reading video.
- Hearing-impaired individuals can read lips and lip reading claims may benefit them to improve their lip imitation skills.
- AVSR is also used in Human Computer Interaction related applications to improve user experience.

5. Conclusions

In this work we develop audio visual speech recognition for custom dataset and the dataset contains English words. First we extract the audio from the video using five different techniques and got 91.42% accuracy and recognition of visual speech using LSTM technique and got 80% accuracy. Eventually, combined the LSTM and feed forward neural network to get a better accuracy in AVSR model. The combined audio and video involving feed forward neural network and LSTM recurrent neural network and got 92.38% accuracy.

Acknowledgements
The authors would like to acknowledge dataset created by ourselves with the help of faculty members and UG and PG students. Based on email request we will share the database to the researchers for the purpose of research.

1. Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman. Deep Audio-visual Speech Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018 pp 1,2,6.
2. Ahmad B. A. Hassanat. Speech and Language Technologies Edition 1, Chapter 14, Visual Speech Recognition, Publisher: InTech, Editors: Prof. Ivo Ipsic, pp 285-289.
3. Ayaz A. Shaikh and Dinesh K. Kumar. Visual Speech Recognition Using Optical Flow and Support Vector Machines International Journal of Computational Intelligence and Applications Volume 10, 2011, pp 171.
4. Themos Stafylakis, Georgios Tzimiropoulos, Combining Residual Networks with LSTMs for Lipreading. Interspeech 2017, pp 3653.
5. Yao WenJuan, Liang YaLing, and Du MingHui. A real-time lip localization and tacking for lip reading. 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE). 20 - 22 August 2010, pp 364.
6. Shillingford, Brendan & Assael, Yannis & Hoffman, Matthew & Paine, Thomas & Hughes, C´ıan & Prabhu, Utsav & Liao, Hank & Sak, Hasim & Rao, Kanishka & Bennett, Lorrayne & Mulville, Marie &

Coppin, Ben & Laurie, Ben & Senior, Andrew & Freitas, Nando. (2018). Large-Scale Visual Speech Recognition. Interspeech 2018.
7. Courtney, Logan & Sreenivas, Ramavarapu. Learning from Videos with Deep Convolutional LSTM Networks. arXiv preprint. arXiv:1904.04817 (2019)
8. G. Sterpu, C. Saam and N. Harte, "Can DNNs Learn to Lipread Full Sentences?," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 16-20.
9. Kumar, Yaman & Jain, Rohit & Salik, Khwaja & Shah, Rajiv Ratn & Yin, Yifang & Zimmermann, Roger. (2019). Lipper: Synthesizing Thy Speech Using Multi-View Lipreading. Proceedings of the AAAI Conference on Artificial Intelligence. 33. pp. 2588-2595.
10. K. Xu, D. Li, N. Cassimatis and X. Wang, "LCANet: End-to-End Lipreading with Cascaded Attention-CTC," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, 2018, pp. 548-5558.
11. Margam, Dilip & Aralikatti, Rohith & Sharma, Tanay & Thanda, Abhinav & K, Pujitha & Roy, Sharad & Venkatesan, Shankar. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. Published in ArXiv 2019
12. Stafylakis, Themos & Khan, Muhammad Haris & Tzimiropoulos, Georgios. (2018). Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs.
13. S. Zhang, M. Lei, B. Ma and L. Xie, "Robust Audio-visual Speech Recognition Using Bimodal Dfsmn with Multi-condition Training and Dropout Regularization," ICASSP 2019 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6570-6574
14. Noda, Kuniaki & Yamaguchi, Yuki & Nakadai, Kazuhiro & Okuno, Hiroshi & Ogata, Tetsuya. (2014). Audio-visual speech recognition using deep learning. Applied Intelligence. 42.pp. 722–737 .
15. S. Petridis, Z. Li and M. Pantic, "End-to-end visual speech recognition with LSTMS," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2592-2596, doi: 10.1109/ICASSP.2017.7952625.
16. F. Tao and C. Busso, "Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1290-1302, July 2018.
17. S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos and M. Pantic, "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 513-520.
18. Y. Goh, K. Lau and Y. Lee, "Audio-Visual Speech Recognition System Using Recurrent Neural Network," 2019 4th International Conference on Information Technology (InCIT), Bangkok, Thailand, 2019, pp. 38-43.
19. J. Wang, L. Wang, J. Zhang, J. Wei, M. Yu and R. Yu, "A Large-Scale Depth-Based Multimodal Audio-Visual Corpus in Mandarin," 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, United Kingdom, 2018, pp. 881-885.
20. Ochiai, Tsubasa & Delcroix, Marc & Kinoshita, Keisuke & Ogawa, Atsunori & Nakatani, Tomohiro. (2019). Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues.Interspeech 2019.pp. 2718-2722.

21. J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Lip Reading Sentences in the Wild," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3444-3453.
22. Jha, V. P. Namboodiri and C. V. Jawahar, "Word Spotting in Silent Lip Videos," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018, pp. 150-159, doi: 10.1109/WACV.2018.00023.
23. Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba and M. Elshehaly, "Lipreading using a comparative machine learning approach," *2018 First International Workshop on Deep and Representation Learning (IWDRL)*, Cairo, 2018, pp. 19-25, doi: 10.1109/IWDRL.2018.8358210.
24. Yaman Kumar, Mayank Aggarwal, Pratham Nawal, Shin'ichi Satoh, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Harnessing AI for Speech Re- construction using Multi-view Silent Video Feed. In 2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic ofKorea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3241911
25. Y. Lu and Q. Liu, "Lip segmentation using automatic selected initial contours based on localized active contour model," *Eurasip J. Image Video Process.*, vol. 2018, no. 1, 2018, doi: 10.1186/s13640-017-0243-9.

26. Lynne E Bernstein, Marilyn E Demorest, and Paula E Tucker. 1998. What makes a good speechreader? First you have to find one. Hearing by eye II: Advances in the psychology ofspeechreading and auditory-visual speech (1998), 211–227
27. Marc Marschark, Dominique LePoutre, and Linda Bement. 1998. Mouth move- ment and signed communication. Hove, United Kingdom: Psychology Press Ltd. Publishers.
28. Quentin Summerfield. 1992. Lipreading and audio-visual speech perception. Phil. Trans. R. Soc. Lond. B 335, 1273 (1992), 71–78

Figures

Figure 1

Preprocessing Step in single Frame.

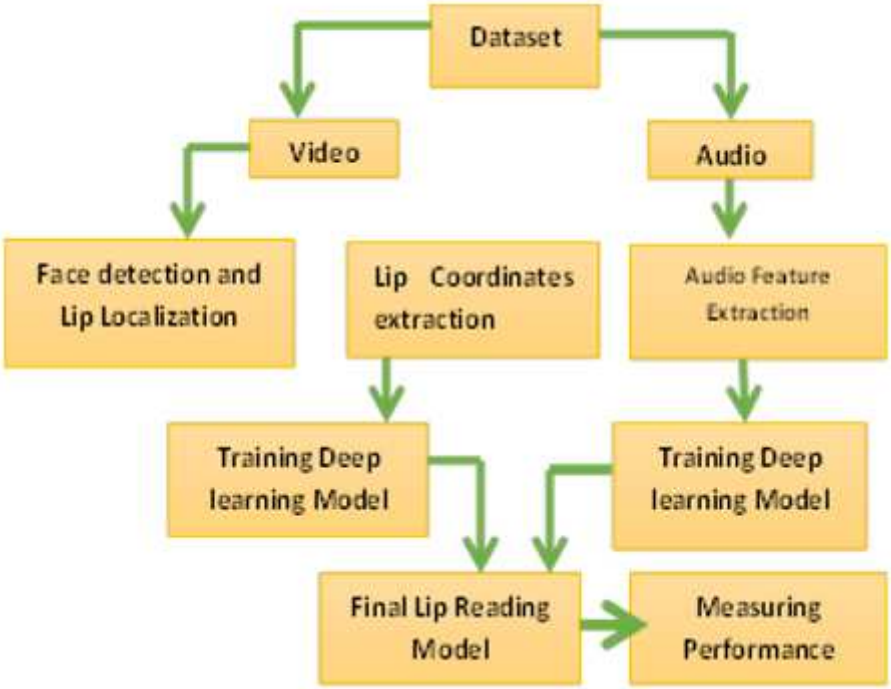


Figure 2

Block Diagram of AVSR

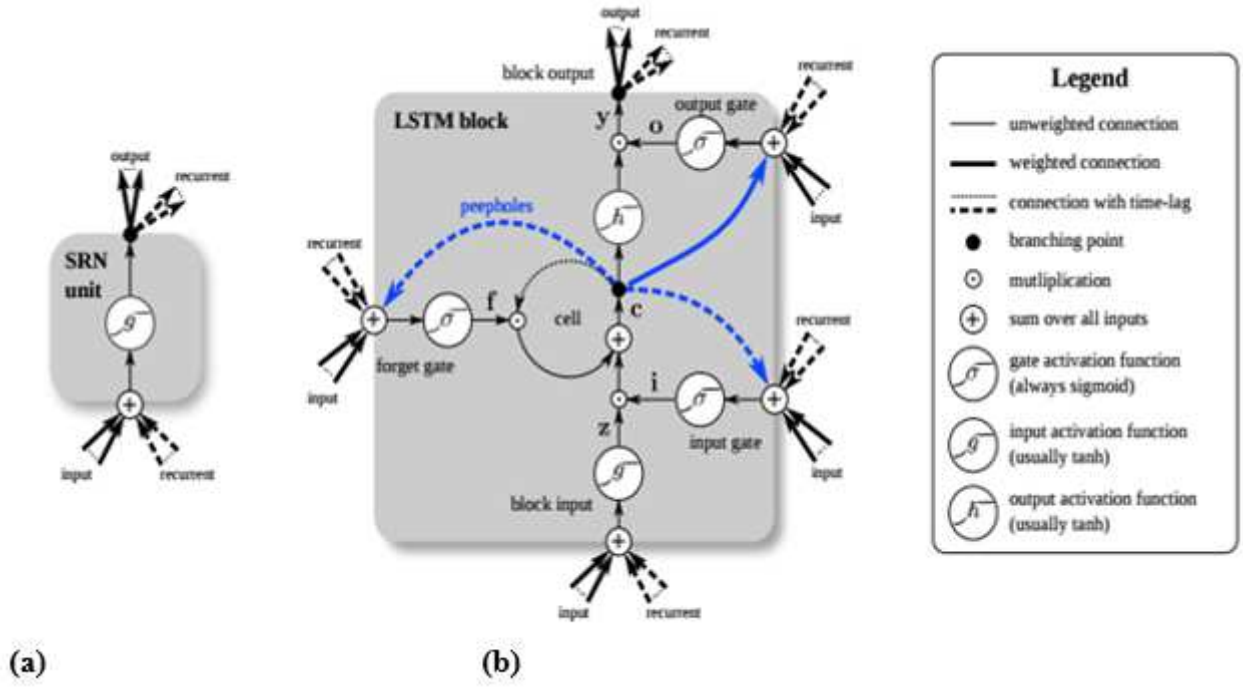


Figure 3

Complete Diagram of the Simple Recurrent Network unit (a) and a Long Short-Term Memory block (b) as used in the hidden layers of a recurrent neural network

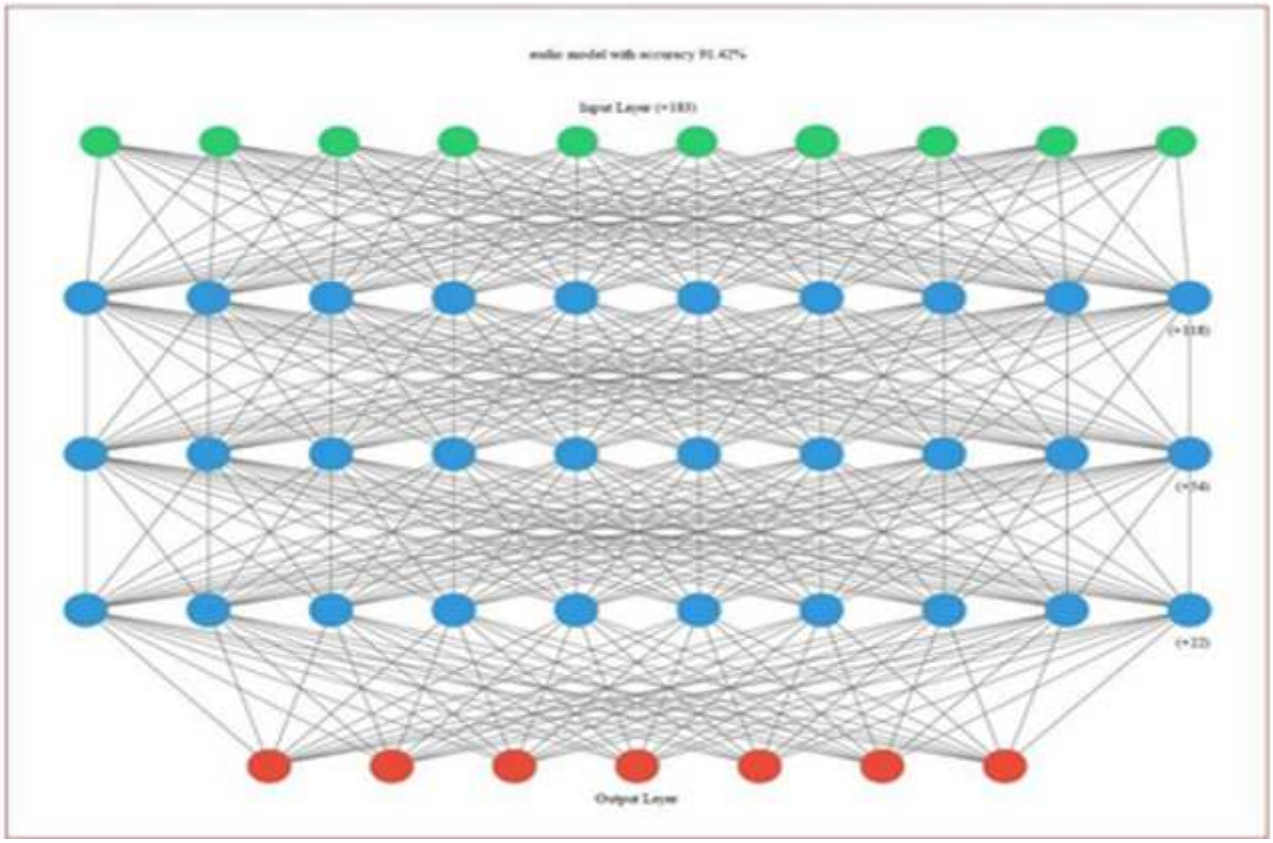


Figure 4

Feed Forward Neural Architecture of Audio Model

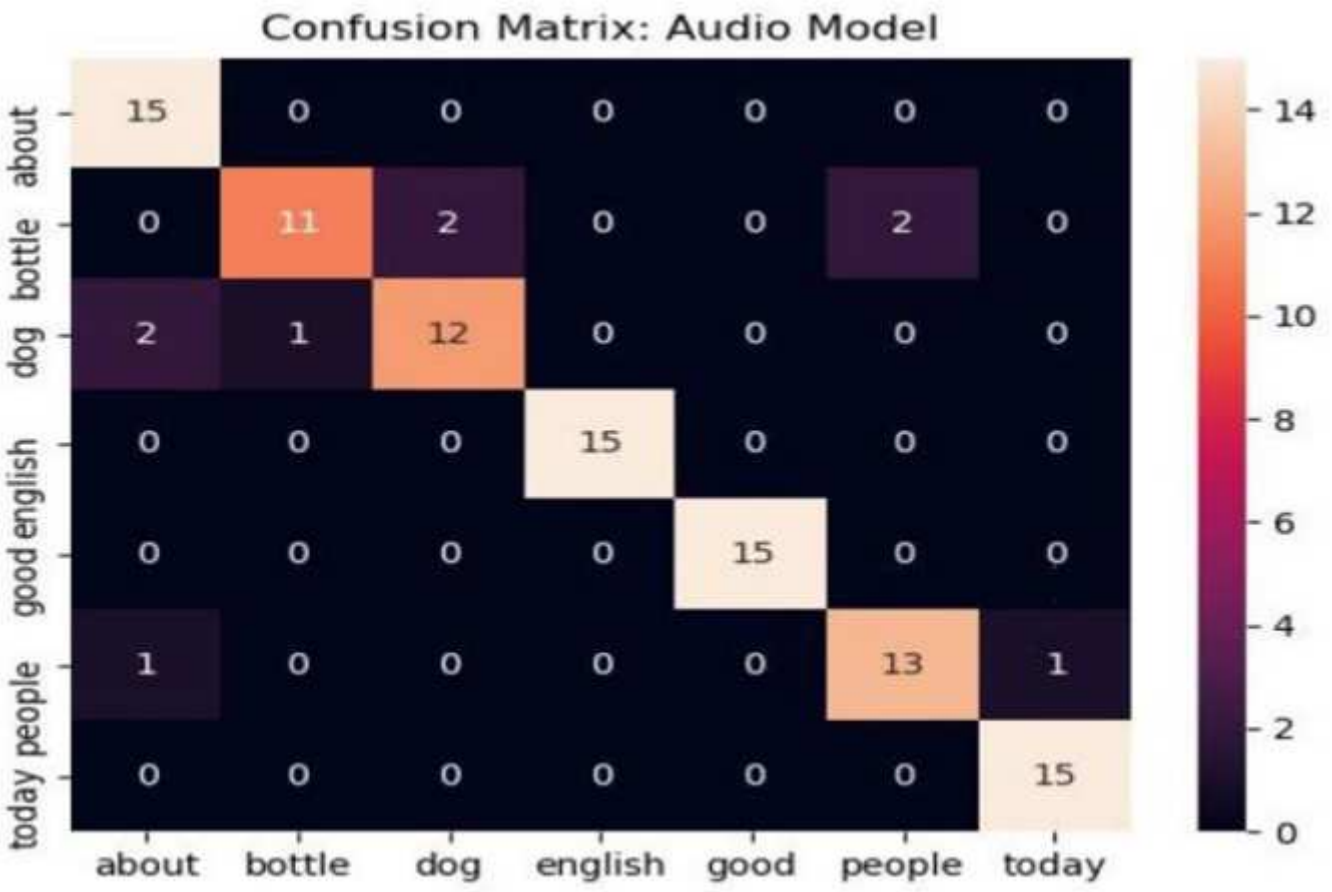


Figure 5

Confusion matrixes for audio model

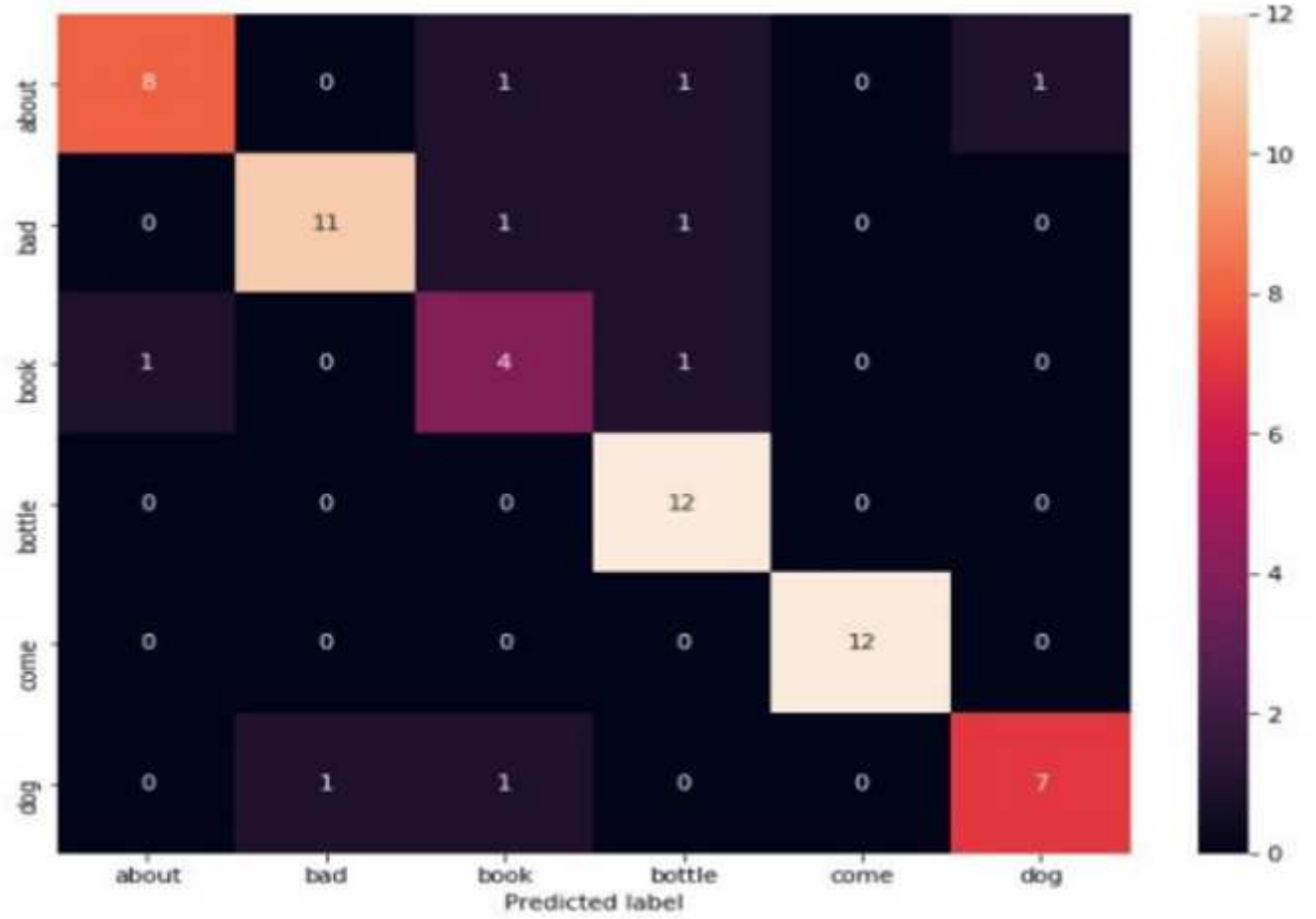


Figure 6

Confusion matrixes for video model

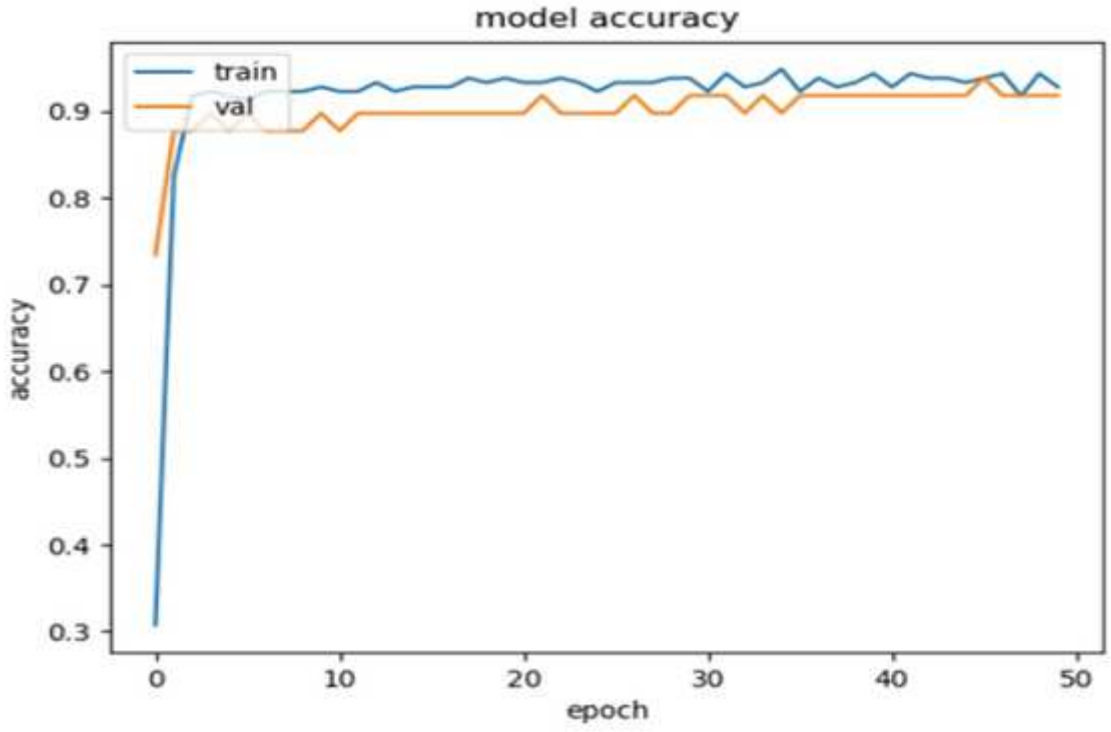


Figure 7

Accuracy curve of combined Model

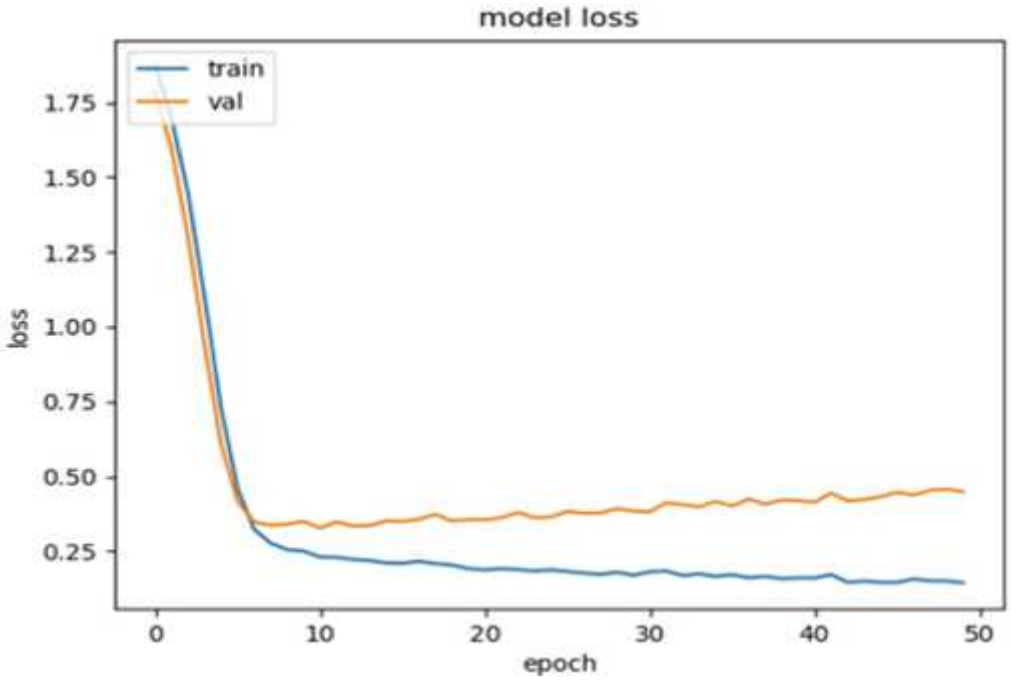


Figure 8

Loss curve of Combined Model

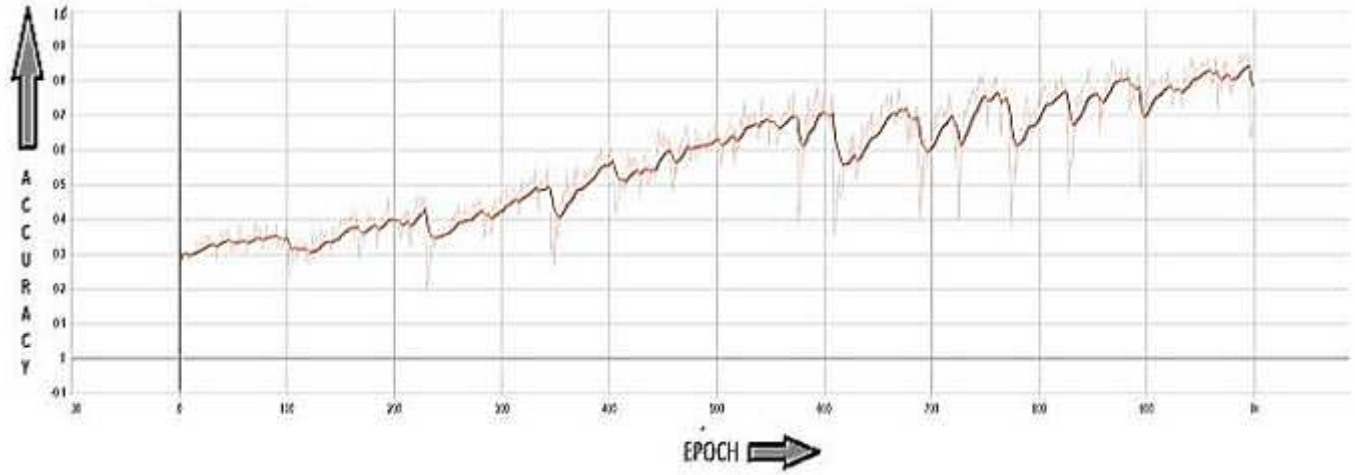


Figure 9

Accuracy curve for LSTM model

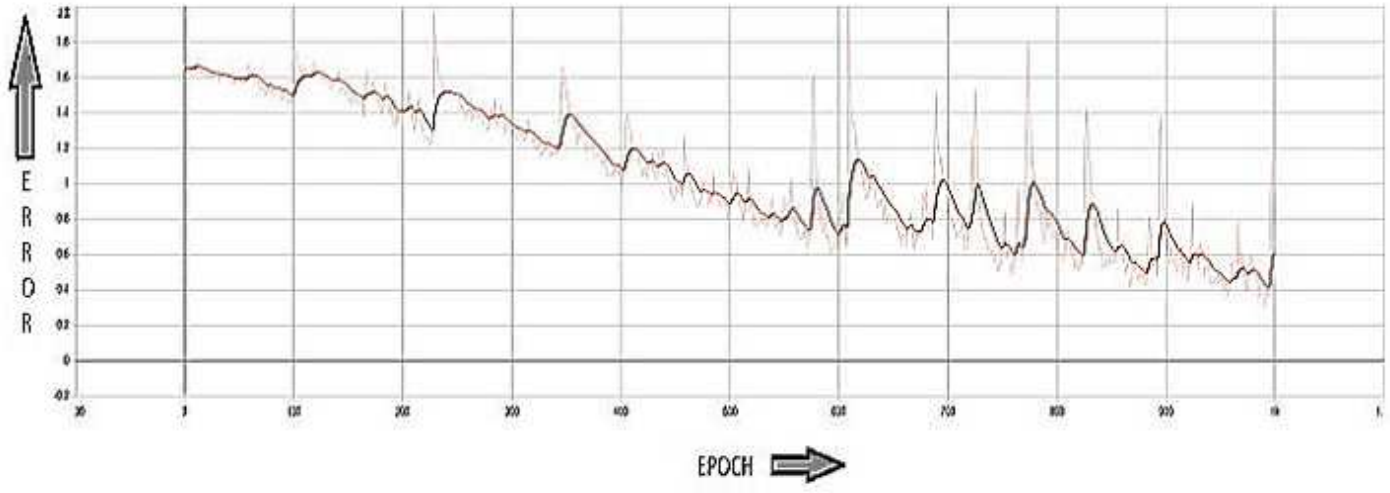


Figure 10

Loss curve for LSTM model

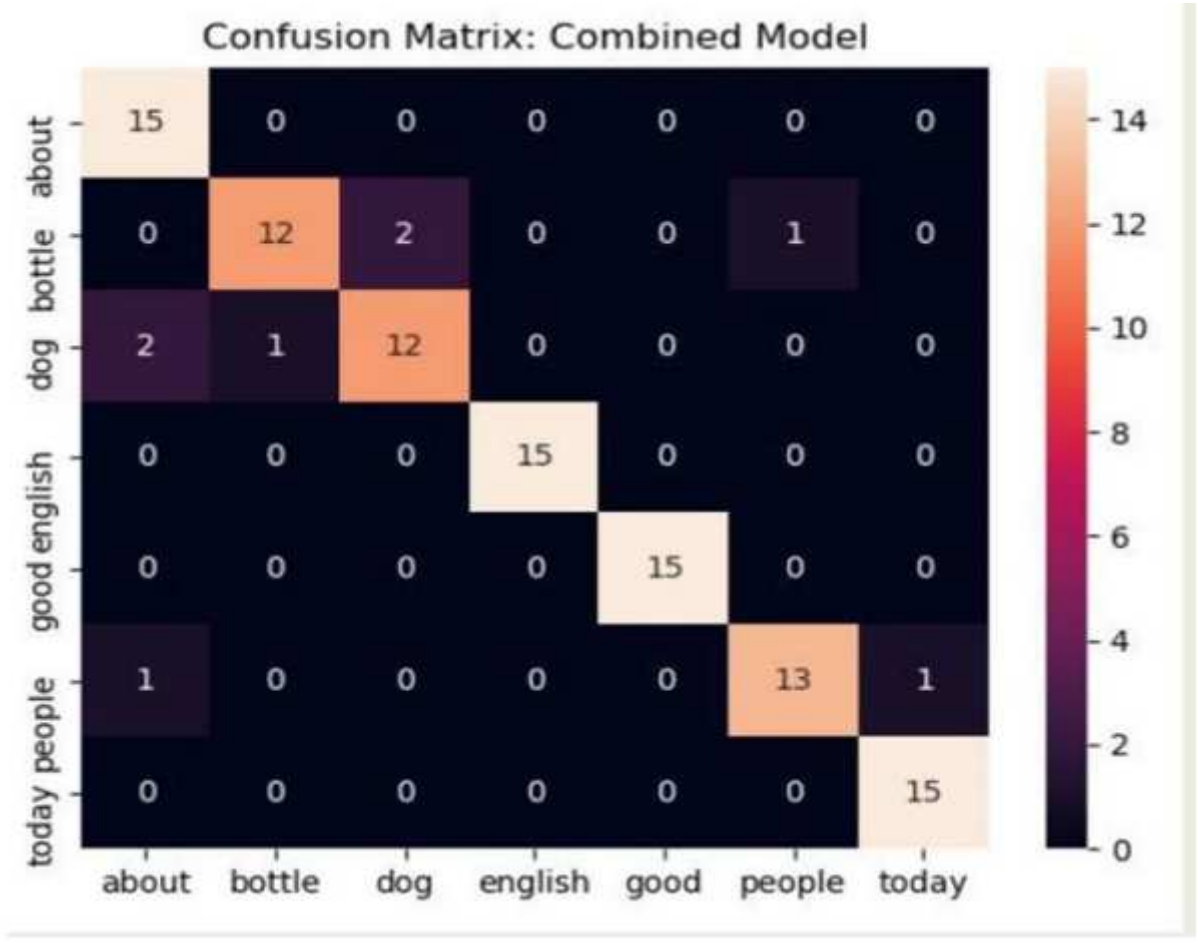


Figure 11

Confusion matrix of combined model