JSS Mahavidyapeetha

JSS SCIENCE AND TECHNOLOGY UNIVERSITY
MYSURU-570006

FINAL YEAR B.E PROJECT REPORT
2020-2021

SPEECH RECOGNITION
An approach for Multi-language

Submitted by

Name	USN	e-mail	Phone-no
Anoopa S	01JST17EC016	anoopakedila@gmail.com	+91 9148522162
M Snehith Reddy	01JST17EC052	m.snehithreddy000@gmail.com	+91 7674027890
Mantena Vinay Krishna	01JST17EC055	mantenavinay@gmail.com	+91 9849337989
Anita Bhagashetti	01JST17EC014	anitabhagashetti7377@gmail.com	+91 8884746761

Submitted in partial fulfilment of the requirement of academic event in BE

Under the Guidance of

Prof. Shashidhar R
Assistant Professor
Dept. Of E&C

JSS SCIENCE AND TECHNOLOGY UNIVERSITY
MYSURU-570006

JSS Mahavidyapeetha

JSS SCIENCE AND TECHNOLOGY UNIVERSITY MYSURU-570006

CERTIFICATE

Certified that the project work entitled "**SPEECH RECOGNITION An approach for Multi-language**" carried out by **Anoopa S, M Snehith Reddy, M Vinay Krishna and Anita Bhagashetti**, bonafide students of Sri Jayachamarajendra College of Engineering, Mysuru in partial fulfillment for the award of Bachelor of Engineering in ELECTRONICS & COMMNUNICATION ENGINEERING of the JSS Science And Technology University, Mysuru during the year 2020-21. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the final report.The project report has been approved as it satisfies the requirements in respect of Project work prescribed for the degree.

Dr. M.N. Jayaram
Professor and Head,
Dept. of E & C
SJCE, Mysuru

Prof. Shashidhar R.
Assistant Professor,
Dept. of E & C
SJCE, Mysuru

Dr. S B Kivade
Principal, SJCE, Mysuru

External Viva

Name of Examiners	Signature with Date
1.	.....
2.	.....
3.	.....

JSS Mahavidyapeetha

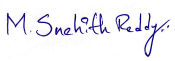
JSS SCIENCE AND TECHNOLOGY UNIVERSITY MYSURU-570006

DECLARATION

We **Anoopa S, M Snehith Reddy, Mantena Vinay Krishna and Anita Bhagashetti** bearing the USN : 01JST17EC016, 01JST17EC052, 01JST17EC055 and 01JST17EC014 respectively students of B.E. in Electronics and Communication Engineering, JSS Science and Technology University, SJCE, Mysuru, do hereby declare that this project entitled "**SPEECH RECOGNITION An approach for Multi-language**" has been independently carried out by us, under the guidance of **Prof. Shashidhar R** and has not been submitted elsewhere for the award of degree.

Date: 30-07-2021
Place: Mysuru

Anoopa S
(USN: 01JST17EC016)



M Snehith Reddy
(USN: 01JST17EC052)

Mantena Vinay Krishna
(USN: 01JST17EC055)

Anita Bhagashetti
(USN: 01JST17EC014)

Acknowledgements

This project would never have been before your eyes, if it were not for them, their inspiration, patience and support.

We express our sincere thanks to Dr. S B Kivade, Principal & Dean, JSS Science and Technology University, SJCE, Mysuru for his encouragement and constant support during the entire course.

We owe special thanks to Dr. M N Jayaram, Professor and Head, Department of Electronics and Communication Engineering, JSS Science and Technology University, SJCE, Mysuru, a wonderful person who supported, encouraged and motivated all the students of our batch.

We are grateful to our Guide Prof. Shashidhar R, Assistant Professor, Department of Electronics and Communication Engineering, JSS Science and Technology University, SJCE, Mysuru for his valuable guidance, constant encouragement and inspiration towards completion of this project.

We also thank our panel members, Prof. B A Sujathakumari, Associate Professor, Dept of E & C, Prof. Thygaraja Murthy, Associate Professor, Dept of E & C and Prof. M S Praveen Kumar, Assistant Professor, Dept of E & C, for their support and encouragement towards the completion of the project.

Our hearty thanks to all the other Teaching and Non-Teaching staff for their advice and support throughout the academic study at the department.

Last but not the least, we owe a lot to our parents for their constant wishes and encouragement which served as a "Beacon Light" and crowned out efforts with success.

Project members

Anoopa S

M Snehith Reddy

Mantena Vinay Krishna

Anita Bhagashetti

Abstarct

Communication is all about expressing one's thoughts to another person through speech and other facial expressions. It is an essential part of one's lives as it enables people to interact with the world. But for people with hearing impairment, it is difficult to communicate without any assistance like hand signs, facial expressions, lip-reading etc. Hearing-impaired people find it difficult to interpret the lip movements unless they are trained and thereby it is tough to detect the spoken words for them. Audio Visual speech recognition (AVSR) systems for lip reading makes them to understand speech and maintain social activities without depending on the auditory perception. Thus, one can visualize AVSR system as a lifeline for people with hearing impairment which helps them in providing a way to understand the words that are being tried to conveyed to them through speech.

In this work, an AVSR system is proposed, which is the combination of both audio and visual processing, that implies integration of both visual and audio recognition processes working separately. Visual speech processing consists of face detection, lip localization, feature extraction and recognition. Whereas, the audio processing consists of audio feature extraction and recognition. The proposed architecture includes 1D CNN model for audio and LSTM model for visual and feed forward network for integration uses the custom dataset and achieved train accuracy of 93.33% and test accuracy of 92.26% for Kannada dataset and train accuracy of 94.67% and test accuracy of 91.75% for English dataset. The use of this Hybrid model clearly out performs all conventional methodologies that have been proposed in the past.

Contents

List of Figures

List of Tables

Acronyms

3DMM 3D Morphable Model

AAM active appearance model

AliNN Alignment neural network

ANN Artificial neural network

ASHVF appearance and shape-based hybrid visual features

ASR Automatic Speech Recognition

AVSR Audio Visual Speech Recognition

BF-LSTM Bifocal Long Sshort Term Memory

BGRU Bidirectional Gated Recurrent Unit

CFI Concatenated Frames of Image

CNN Convolutional Neural Network

CTC Connectionist Temporal Classification

DAS Dynamic Audio Sensor

DCT Discrete Cosine Transform

DNN Deep neural networks

DTW Dynamic Time Warping

DTW Dynamic Time warping

DVS Dynamic Visual Sensors

DWT Discrete Wavelet Transform

EM Expectation-Maximization

FFMPEG Fast Forward Moving Picture Expert Group

GMM Gaussian mixture model

GRU Gated recurrent units

HMM Hidden Markov Model

LSTM Long Short Term Memory

MFCC Mel-Frequency Cepstral Coefficients

NB Naive Bayes

PCA Principal component analysis

RNN-T Recurrent Neural Network Transducer

RTF Rich Text Format

SNR Signal to Noise Ratio

SVM Support Vector Machine

TCN Temporal Convolutional Networks

TLPT Time based Label-Preserving Transform

Chapter 1

Introduction

This chapter gives a brief overview of AVSR and its history. The discrete steps involved in automatic lip reading is also discussed followed by the motivating factors that lead the team to take up this work. The problem is then defined and objectives are set that defines the course of action of the proposed work.

1.1 Overview

In recent trends pattern recognition has become important issue which allows computers to imitate human brain to interpret things.When compared to other recognition systems such as fingerprint, gesture or facial recognition, audio visual speech recognition is more beneficial and robust which makes it important building block of Human Computer interface. Pattern recognition, computer vision and image processing are important steps in lip reading.

Nowadays lip reading is becoming very important technique implemented in recognition systems where several lip reading techniques may be used to improve performance of recognition models. Looking at history of lip reading one has to go back to 1954 when the first work related to lip reading was proposed. Later the system was given by Petajan from university of Illinois which was a popular method in 1980s.

After that there has been a number of researches in the field of lip reading. Since audio signal is sensitive to noisy environment, pixel-based method and artificial neural network (ANN) were used for lip reading to build a recognition model in 1989. In 1993, Hidden Markov Models (HMMs) was used in his lip reading systems to achieve sentence recognition rate of 25%. A lip reading system using color motion video which combined snake model, principal component analysis (PCA) and HMM was proposed [4] to achieve accuracy of 94% for 10 isolated words. To improve performance of multi modal continuous lip reading recognition. Further this work was focussed on context-dependent deep neural networks (DNN) system [5] which realized deeper layers of the network for visual parameters to achieve absolute word accuracy of 84.7% with a massive 33% increase on the baseline HMM. Many companies and institutions are investing on researches in the field of lip reading.

Haar feature and Adaboost cascade classifiers were used to track the speaker's facial expressions and lip movements in an open source system invented by Intel. This system has got the ability to enhance word recognition accuracy and processing speed. British scientists have designed a computer for lip reading to differentiate between various languages such as German, Arabic, Italian, Polish etc with great accuracy. Google and Oxford universities have discovered an outstanding lip reading AI software where it may be known to find out speaker's lip movements on BBC TV shows. It turned out to be great with 46.8% accuracy when compared to trained lip specialist which was only 12.8% in same test.

Audio visual speech recognition (AVSR) is a method that uses image processing capabilities in lip reading to aid speech recognition systems in recognizing undeterministic phone among probability decisions. Thus the focus is on building audio visual speech recognition model to interpret both audio as well as visual data. Performance and limitations of hybrid model used for audio-visual speech recognition are to be assessed and specified later so that it may help for further research in the field.

1.1.1 Face Detection

Face detection may be visualized as a computer vision issue which involves locating faces in images. Face detection is the primary step in face bio-metrics, and its efficiency has a great impact on the performance of additional operations. It is a challenging task for humans to solve, so feature-based techniques such as the cascade classifier are used to solve it. At present deep learning methods have achieved good results on standard face detection data-sets. Face detection is generally considered to be the introductory step towards a number of face-oriented technologies like face identification or recognition. Face detection has many beneficial applications. Face detection may be termed as a particular event of object-class recognition. In object-class recognition or detection, the main job is to determine the positions and sizes of all entities in an image that are part of a given class. Face detection algorithms concentrate on the discovery of human faces.

Face detection is similar to image recognition in which the image under observation is compared bit by bit. Any face related feature changes in the database will not give a valid comparison. Face recognition has an important role in any sort of face related image-processing applications. Nowadays, a lot of works related to face detection or recognition have been proposed to make it more progressed and efficient. Lip-area extraction is the most significant part of the process to get good recognition rate. Many innovations have been made for extracting facial image from the face.The active appearance model (AAM) is a kind of model in which the shape as well as grey-level appearance (the one which shows only shades of grey colour with no other colours) may be determined. It is hard to directly identify or recognize lip regions because various other parts like moustache, eyes, nose, eyebrows, and body are observed in target image.

1.1.2 Lip Localization

Lip-area extraction is the most significant part of the process to get good recognition rate. Many innovations have been made for extracting facial image from the face.The active appearance model (AAM) is a kind of model in which the shape as well as grey-level appearance (the one which shows only shades of grey colour with no other colours) may be determined. It is hard to directly identify or recognize lip regions because various other parts like moustache, eyes, nose, eyebrows, and body are observed in target image.

1.1.3 Feature Extraction and Recognition Models

Snakes or active-contour models are usually used for shape analysis and object detection by making use of de-formable templates. There are different pixel based and model based methods that may be used for feature extraction.Hybrid models are the models which are a combination of two or more methods in order to interpret and analyze the data. These models give more accurate results with high accuracy rate. Hidden Markov model (HMM) is an important statistical method for continuous sequence categorization like speech recognition, dynamic hand-gesture identification and face related data (facial expressions) recognition.

Considering recent researches in the field, there are different methods of lip reading like Dynamic Time Warping (DTW), template matching, Hidden Markov model (HMM) and Artificial Neural Networks (ANN).HMM is the appropriate model for correctly depicting the information related to movement of the lips. The training and testing phases of HMM system includes extraction of features using DWT or DCT from mouth part, which may then be given as inputs to figure out the variables of the system, after which the word may be identified during testing phase.

1.1.4 AVSR

Audio visual speech recognition (AVSR) is an approach that uses image processing techniques in lip reading in assisting speech recognition systems. Audio visual speech recognition (AVSR) is a technique that uses image processing capabilities in lip reading to aid speech recognition systems in recognizing undeterministic phones or giving preponderance among near probability decisions. The fundamentals of AVSR system which emphasizes on combining audio and video processing of speech signal.

AVSR primarily consists of two main part; The Audio recognition and Visual recognition (Lip reading). While, the audio recognition consists of feature extraction and recognition processes, the video recognition consists of face detection, lip localization, feature extraction and recognition. Combined, these two sources of speech information result in better automatic recognition rates than were obtained from either source alone. We chose to map the visual signal into an acoustic representation closely related to the vocal tract's transfer function. Given such a mapping, the visual signal could be converted and then integrated with the acoustic signal prior to any symbolic encoding.

1.2 Motivation

Communication is nothing but expressing one's thoughts to another person through speech. Lip reading is the process of depicting spoken words by detecting lip movement. For example, Hearing -impaired people use lipreading extensively in their daily conversations to understand one another in chattering environment and in situations where the audio speech signal is not readily understandable. Hearing-impaired people find difficult to interpret the lip movement unless they are trained and thereby it is tough to detect the spoken words for them. Audio video speech recognition for lip reading makes them to understand speech and maintain social activities without depending on the auditory perception. Thus the source of motivation behind lip reading and audio visual speech recognition is to assist people with hearing impairment and to provide a way to understand what is being said to them. It helps the deaf people to figure out spoken words so that they can take active part in conversations.

1.3 Problem Definition

Primarily there are two points of interest with the design of an AVSR application

1. Gaining of an appropriate representation of the visual speech modality.
2. The effective integration of the acoustic and visual speech modalities in the presence of a variety of degradations.

1.4 Objectives

The following objectives were planned to be achieved during the course of this project.

1. Develop a database for multi language model (English and Kannada)
2. Develop an algorithm for Audio Speech Processing.
3. Develop an algorithm for Visual Speech Recognition.
4. Integration of both audio and video.
5. Validation of results

1.5 Organization of the report

This report is organized in the following way. Chapter 2 describes the existing work on AVSR and the methods they have used and its obtained accuracy, Chapter 3 explains the system architecture of the proposed methodology and Chapter 4 list outs the software and hardware components used and chapter 5 gives the results of the proposed method with screenshots of result and chapter 6 speaks about the conclusion of the work.

Chapter 2

Literature Survey

An extensive literature survey has been conducted prior to the beginning of the proposed work and have been well documented for further reference and also to figure out the drawbacks in the existing systems. A summary of the literature survey is then explained briefly.

2.1 Previous Research

2.1.1 Audio Recognition Systems

The method proposed is for enhancing speech recognition dataset with a pre-trained model and script. [1] have used Korean news videos and scripts to obtain dataset. Comparing the piece of data obtained from the pre-trained model with the ground truth script, they produced the pair of audio and script. In each pair, the audio clip has the word exactly fitted in it and the script is perfect as it is written by them manually. In the experiments on news videos and scripts, it is showed that their method extracts automatic speech recognition dataset in more efficient and obtained 67% accuracy.

Xutai Ma et al. here have presented a method of enhancing speech recognition dataset with a pre-trained model and script. Comparing the piece of data obtained from the pre-trained model with the ground truth script, they produced the pair of audio and script. In each pair, the audio clip has the word exactly fitted in it and the script is perfect as it is written by them manually. [2]In the experiments on news videos and scripts, it is showed that their method extracts automatic speech recognition dataset in more efficient and obtained more accuracy.

Jon Macoskey et al. have presented a Bifocal RNN-T, a new variant of the Recurrent Neural Network Transducer (RNN-T) architecture designed for improved inference time latency on automatic speech recognition. The model enables a dynamic pivot for its runtime compute pathway, namely taking advantage of keyword spotting to select which component of the network to execute for a given audio frame. [3]To accomplish this, we leverage a recurrent cell we call the Bifocal LSTM (BF-LSTM), which we detail in the paper. The architecture is compatible with other optimization strategies such as

quantization, sparsification, and applying time reduction layers, making it especially applicable for deployed, real-time speech recognition settings. We present the architecture and report comparative experimental results on voice assistant speech recognition tasks. Specifically, we show our proposed Bifocal RNN-T can improve inference cost by 29.1% with matching word error rates and only a minor increase in memory size.

An automatic speech recognition system based on an adaptive Gaussian mixture technique for audio signal modality. After feature extraction stage, for robust density estimation an adaptive mixture estimation method is used based on optimal minimization of the integral square distance between the true density that represents the speech features and the approximated mixture. This estimation is relatively complex because of its complex representation of the density and the issues with Expectation-Maximization (EM) algorithm. [4]The technique proposed in this work not only shows its performance through the experimental results of the paper but also provides a natural and efficient way of including both audio and video into the robust automatic speech recognition program of study.

Swapna Agarwal et al. have proposed a meta-learning based few-shot approach for generating personalized 2D talking heads where the lip animation is driven by a given audio. The idea is 2 that the model is meta-trained with a dataset consisting of a large variety of subjects' ethnicity and vocabulary. They showed that meta-trained model is capable of generating realistic animation for previously unseen face and unseen audio when finetuned with only a few-shot examples for a very short time (180 seconds). [5]Considering the fact that facial expressions driven by audio are mainly expressed through motion around lips, they have used lip only for the animation. They have used GRID and TCD-TIMIT datasets and their own custom dataset of Asian people. Both qualitative and quantitative analysis show that animations generated by such meta-learned model surpasses the state-of-the-art methods both in terms of realism and time taken.

The referred paper analyses the function of voice translation software on mobile platform based on the existing technology. To improve the quality of translation, a phrase based translation statistical model is proposed and the training process is analysed in detail. Then the architecture of the oral translation system is designed by using the speech recognition function of Android framework and the strategy of multi translation engine coordination. For the problem of data sparseness, they have proposed a modelling method which combines the language model based on words and the language model with parts of speech. [6]The log linear model is used to analyse the semantic of corpus annotation. The parameters obtained from training reflect the structural information of sentences. Finally, the actual test on the mobile client shows that the real-time speech translation system can not only effectively speed up the speed of data check and audit, but also appropriately improve the accuracy of speech recognition and ensure the quality of translation work.

To boost the performance of Mask-CTC, Yosuke Higuch et al. have proposed the encoder network architecture by employing a recent architecture called Conformer.[7] Next, they have done the training and decoding methods by introducing auxiliary objective

to predict the length of a partial target sequence, which allows the model to delete or insert tokens during inference. Experimental results on different ASR tasks show that the proposed approaches improve Mask CTC significantly, outperforming a standard CTC model (15.5% → 9.1% WER on WSJ). Moreover, Mask-CTC now achieves competitive results to AR models with no degradation of inference speed (¡ 0.1 RTF using CPU). They have also showed a potential application of Mask CTC to end-to-end speech translation

2.1.2 Visual Recognition Systems

A combination of spatiotemporal convolutional, residual and bidirectional Long Short-Term Memory networks is used for audio visual speech recognition [8]. They evaluated it on the Lipreading In-The Wild benchmark, a database from BBC TV broadcasts. The proposed network attains word accuracy upto 83.0%, resulting in 6.8% absolute improvement over the current state-of-the-art, without using information about word boundaries during testing or training

The fusion of audiovisual features with an alignment neural network (AliNN), based on recurrent neural network (RNN) with attention model. The proposed front-end model can automatically learn the alignment from the data and resulting aligned features are concatenated and fed to back-end ASR systems [9]. They have evaluated with matched and mismatch channel with various noise condition. According to the results, it yields 24.9% absolute improvement over the baseline with Gaussian mixture model with hidden Markov model (GMM-HMM) back-end and 2.4% with DNN-HMM back-end.

A new approach to the field of weakly supervised learning in the video domain. They have used sign language data for their experimentation. The approach exploits sequence constraints within each independent stream and combines them by explicitly imposing synchronization points to make use of parallelism that all sub-problems share. They have done this with multi-stream HMMs while adding intermediate synchronization constraints among the streams[10], they embedded powerful CNN-LSTM models in each HMM stream following the hybrid approach. They clearly outperformed the state-of-the-art on all data sets and observe significantly faster convergence using the parallel alignment approach.

A new architecture called Hahn Convolutional Neural Network is proposed. The complexity level is reduced enormously by minimizing number of layers and parameters. [11] They have used OuluVS2, AVLetters and BBC LRW datasets for their experiments. The classification results are 93.72% on OuluVS2, 59.23% on AVLetters and 46,6% on BBC LRW.

A deep learning model for using feature visualization. The explanation is based on the Grid dataset which includes inputs from both females and males [12]. There is in depth explanation for the hidden internal layers of the DNN. The conclusion for the paper lies on CNN as a self-learning networks which express a high accuracy for feature extraction and feature visualization. These features and explanation can bring improvements in the use of lip reading and DNN.

The method developed for lipreading using Convolution Neural Network(CNN). The

method uses a set of frames obtained from a video which gives better result for lip-reading as reported by Chung and Zisserman et al. using CNN. CNN like AlexNet, VGG and GoogleNet can work well for lip-reading using the CFI (Concatenated Frames of Images)[13]. The method is proposed to apply CNN to two types of dataset, 1. Full-lip images and 2. Patches around tracked lips obtained by face-alignment. In this method for data preprocessing we used TLPT (Time based Label-Preserving Transform) which converts the videos of the dataset into the CFIs (Concatenated Frames of Image) which include both types of dataset i.e. full-lip image and its lip landmarks. As CNN requires a large dataset to train itself TLPT is very helpful. After preprocessing the model is trained using CNN algorithm where the CFIs are introduced as classified for prediction. The Accuracy of trained model d1 and d2 was an average of 87.0 and 89 respectively. Using both the types of dataset the performance of the method was better than using any single dataset alone.

The Algadhy R et al. proposed a method on lip reading using neural network. The dataset set used were GRID corpus which is an audio-video benchmarked dataset. The dataset with single and dual modality were used. Either audio or video were also used as single modality for training and testing the dataset [14]. Dynamic Audio Sensor(DAS) and Dynamic Visual Sensors(DVS) were used to detect the spiking in the audio and video signals. In the next step, Pre-processing of GRID corpus dataset consisting of audio and video recording of 1000 sentences spoken by 34 speakers (18 male and 16 female) speaking a sentence of 6 words each was processed. The facial area was detected using the OpenCV face detector to extract the lip movement. In the Feature extraction process, the model was trained with RNN for audio features and CNN for video features. For this work, 90% of dataset is used for training and 10% of dataset is used for testing. For single modality, accuracy was 83.83% while for dual modality using DAS and DVS the accuracy was around 79.66%.

The importance of lip-motion for speech recognition especially for foreign language learners is mentioned. The proposed technique for visual speech analysis uses lip-tracing in 2d-view of speaker. The author used 2D videos to train 3D Morphable Model(3DMM). [15] This technique is used to train 3DMM from the images and videos. 3DMM is trained using a software called FaceGen. The steps followed in this method are 1. To build 3DMM and 2. Mapping 2d video and audio to 3d synthetic head. FaceGen is used to construct synthetic head poses. FaceGen and Principle Component Analysis(PCA) is used to create head poses and locate vertex correspondence. In Mapping 2d video to 3DMM is used by camera matrix method by Huber. Also, Gold Standard Algorithms used for 2D video to maps it to 3DMM. Two datasets were used for the testing and training of the method containing front and the side- view video of the speaker. Face Analyzer was used to track facial feature of 2d video of dataset to map it to 3DMM. The Experiments showed that using both front and side- view of the dataset will improve the accuracy of the result.

The Wei J Yang et al. have used audio-less video for prediction. Some of the challenges faced are intensity, pronounciation, speaking speed, same lip sequences of different characters and variation in length of the sequence of images etc.[16] Database

used for the experiment is MIRACLE-VCL consist of the sequential image of a person speaking a phrase or a word. The model used 12 layer of CNN with two layer of batch normalization to train and extract the visual features. The model includes two steps. 1] creation of concentrated image from image from the image sequence, 2] encoding with training of image. In the first step, hair Cascade Facial Landmark detector is used to extract only lip portion from images. In next step two level of batch normalization is performed which is done to reduce unwanted variations. In this model, every next layer is dependent on previous layer. Using this method an accuracy of 96% and a validation accuracy of 52.9% have been obtained.

To gain more accurate lip-reading analysis using 3D feature extraction is used. The issues with the dataset as noticed by the authors are the distance between camera and speaker, variation in the illumination,geometric features such as height, width and perimeter of the lip and the varying shape of mouth depending on speakers head position. The dataset is created by taking images of four males and four females. [17] The recording is done using a data collection system supported by Kinect Face Tracking Software Development Kit so as to get the real time input RGB images. In preprocessing data augmentation techniques are applied with translation, rotation, mirror reverse and color change. It also describes the muscle details which are to be captured while recording such as elevator angular orris, buccinator, zygomatic etc. Two different methods are proposed, first is model-based method, and second is image-based method. They have used technique used the densely connected convolutional network (Dense Nets). The accuracy of 98.75% is achieved based on this techniques.

2.1.3 Audio visual Recognition Systems

The key contributions of [18] are: (1) comparing two models for lip reading, one using a CTC loss, and the other using a sequence-to-sequence loss. (2) investigating to what extent lip reading is complementary to audio speech recognition, especially when the audio signal is noisy. Datasets used for audio-visual speech recognition: LRS2-BBC, and LRS3-TED. The Connectionist Temporal Classification (CTC) is the version of HMM model which predicts frame-wise labels and then looks for the optimal alignment between the frame-wise predictions and the output sequence.The main weakness of CTC is that, it assumes each unit is independent. The second type is sequence-to-sequence models (seq2seq) that first read all of the input sequence before predicting the output sentence. The results demonstrate that the mouth movements provide important cues in speech recognition when the audio signal is noisy; and give an improvement in performance even when the audio signal is clean. The gains when using the audio-visual TM-seq2seq compared to the audio-only model are similar.

Petridis et al. have used hybrid CTC/attention architecture for audio-visual recognition of speech in-the-wild. Dataset used: the LRS2 database [19]. The proposed audiovisual model leads to an 1.3% absolute decrease in word error rate compared to audio only model. This audio-visual model significantly outperforms the audio-based model (up to 32.9% absolute improvement in word error rate) for several different types of noise as the signal-to-noise ratio decreases.

A Recurrent Neural Network (RNN) based AVSR is proposed in this research. Here the audio features mechanism is modelled by Mel-frequency Cepstrum Coefficient (MFCC) and further processed by RNN system, whereas the visual features mechanism is modelled by Haar-Cascade Detection with OpenCV and again, it is further processed by RNN syste [20]. Then, both of these extracted features were integrated by multimodal RNN-based features-integration mechanism. The performance in terms of the speech recognition rate and the robustness of the proposed AVSR system were evaluated using speech under clean environment and Signal to Noise Ratio (SNR) levels ranging from -20 dB to 20 dB with 5 dB interval.

The proposed model consists of two streams, one per modality, which extract features directly from the raw images and waveforms, respectively. Each stream consists of a ResNet which extracts features from the raw inputs. This is followed by a 2-layer BGRU network which models the temporal dynamics in each stream. Finally, the information of the different streams/modalities is fused via another 2-layer BGRU which models the joint temporal dynamics. Dataset used: 500 words from the LRW database[21]. The proposed system results in an absolute increase of 0.3% in classification accuracy over the end-to-end audio-only model and an MFCC-based system. The end to end audio-visual fusion model also significantly outperforms (up to 14.1% absolute improvement) the audio-only models under high levels of noise.

K Tan et al. address joint speech separation and dereverberation, which aims to separate target speech from background noise, interfering speech and room reverberation.The proposed novel multimodal network that exploits both audio and visual signals. The proposed network architecture adopts a two-stage strategy, where a separation module is employed to attenuate background noise and interfering speech in the first stage and a dereverberation module 6 to suppress room reverberation in the second stage[22]. This network achieves a 21.10improvement in ESTOI and a 0.79 improvement in PESQ over the unprocessed mixtures.

Jadczyk et al. describes an audio-visual speech recognition system for the Polish language as well as a set of performance tests under various acoustic conditions[23]. This paper present the overall structure of AVASR systems with three main areas: audio feature extraction, visual feature extraction, and (subsequently) audio-visual speech integration. The recordings contain only faces (frontal view) on bright background with rather invariant lighting conditions, and feature full HD quality with 25 frames per second. Each speaker has been recorded for about 10 minutes, totalling about 4 hours. With the Active Appearance Model (AAM) and multi-stream Hidden Markov Model (HMM), this can improve system accuracy by reducing the word error rate by more than 30%.

H. Meutzner et al. proposed a state-based integration scheme that uses dynamic stream weights in DNN-based audio-visual speech recognition. DNN-based systems are able to outperform the GMM-based system for each SNR, where the largest relative improvements are seen for very low SNR conditions. [24]The video data was also taken from the GRID corpus, which contains clean facial video recordings for each utterance (contains recordings from 34 speakers). On an average this method has an accuracy of

87.20%.

Debnath et al. proposes appearance and shape-based hybrid visual features (ASHVF) to embed all the information into compact features set for better visual speech recognition. Three classifiers, ANN, SVM, and NB are used for hybrid classification. The performances of all individual classifiers are evaluated along with the hybrid classifier. Audio–visual English digit database VISW is used here for AV-ASR. The dataset consists of 10 speakers, 6 males and 4 females. Each speaker uttered each word 10 times.[25] So, one word is uttered 100 times by the different speaker. This proposed Hybrid classifier gives a accuracy of over 78.45% over the other methods.

Martienez et al. address the limitations of this Bidirectional Gated Recurrent Unit (BGRU) model and propose changes which further improve its performance. The BGRU layers are replaced with Temporal Convolutional Networks (TCN). Greatly simplify the training procedure, which allows us to train the model in one single stage. They use the Lip Reading in the Wild (LRW) and LRW1000 databases which are the largest publicly available lip reading datasets in English and Mandarin, respectively, in the wild. There are 1000 word classes and a total of 718,018 samples with total duration of approximately 57 hours. [26]This proposed model results in an absolute improvement of 3.2%, in these datasets which is the new state-of-the-art performance.

M Hao et al. proposes a simpler architecture of 3D-2D-CNN-BLSTM network with a bottleneck layer. It also present analysis of two different approaches for lip reading on this architecture. In the first approach, 3D-2D-CNNBLSTM network is trained with CTC loss on characters (ch CTC). Then BLSTM-HMM model is trained on bottleneck lip features (extracted from 3D 2DCNN-BLSTM ch-CTC network) in a traditional ASR training pipeline. In the second approach, same 3D-2D-CNN-BLSTM network is trained with CTC loss on word labels (w CTC). The first approach shows that bottleneck features perform better compared to DCT features. [27]The proposed 3D-2D-CNN-BLSTM w-CTC has given state-of-the art results with relative improvement of 55% and 24.5% on Grid seen and unseen test sets with 1.3% WER and 8.6% WER respectively.

In [28], F Tao et al. , used deep neural networks (DNN) have emerged as powerful alternatives for feature space modelling, which have made end-to-end frameworks feasible. The visual features are directly extracted from the pixels with convolutional neural networks (CNNs). The proposed multi-task learning (MTL) structure has separate subnetworks for acoustic and visual features, which are later combined with recurrent neural network (RNN). Accuracy of this paper is 95% 7 in ideal channel and in challenging channel in noisy area accuracy is 91%. And in clean area 94%.

W Feng et al. used CNN to extract features from video and passes it to LSTM 1 layer. Audio is directly passed onto another LSTM layer.[29] Then these two layers are combined and connected and passed onto softmax layer to determine the target words. Model achieved the accuracy of 84.4%±1.7%(no noise environment) and - 81.4%±1.8%(SNR 20db) on AVLetters.

J. Yu et al. addresses the three issues associated with the construction of audio-visual speech recognition (AVSR) systems. First, the basic architecture designs i.e. end-to-end and hybrid of AVSR systems are investigated. Second, purposefully designed modality

fusion gates are used to robustly integrate the audio and visual features. [30]Third, in contrast to a traditional pipelined architecture containing explicit speech separation and recognition components, a streamlined and integrated AVSR system optimized consistently using the lattice-free MMI(LF-MMI) discriminative criterion is also proposed. Hybrid LF-MMITDNN system is used. The model achieved word error rate of 5.93.

Y Yuan et al. proposes an auxiliary loss multimodal GRU (alm-GRU) model including three parts: feature extraction, data augmentation and fusion & recognition. First, alm-GRU captures the temporal information with the correlation between frames; second, alm-GRU equipped with a novel loss is an end-to-end network combining both fusion and recognition, which can consider the specific information of categories. during feature extraction, in the path of video, the region of interest (ROI) is obtained firstly, then the data augmentation strategies including generative component are employed to increase the number of images. [31] Finally, the CNN is used to extract the image features of ROI. In data augmentation step effects like jitter, small angle rotation etc. are used for video, Gaussian noise to audio and also GANs are used. The DNN architecture has two streams: video and audio. Video and audio are both extracted features after feature extraction, the features are sent to the processing streams at the same time. Here, in order to acquire more balanced representation, this paper introduced auxiliary loss connection. The auxiliary loss connection lies after video GRU and audio GRU, then data are mapped to the target space using merge and mapping block. Model achieved an accuracy of 89.34% for AVLetters 85.53% for AVDigits.

P Zhou et al. proposes a novel multi-modality attention method to integrate information from audio and video for audio-visual speech recognition using Seq2Seq model. The attention based Seq2seq architecture consists of a sequence encoder, a sequence decoder and an attender. Because the encoding process tends to be lossy for long input sequences, an attention mechanism is introduced to automatically select the most relevant information from encoder memory to help the decoder to predict accurate unit at each decoding step.[32] This research proposes to use modality attention in a WLAS end-to-end system for the sake of eliminating the same feature length . constraint. Model achieved 36% relative improvement comparing to LAS in 0dB SNR.

J Wu et al. inspired by the TasNet structure which contains three parts, an audio encoder/ decoder and a separation network. Video is extracted by passing it to 3d-conv layer followed by ResNet -18 and followed by 1d-conv layer. The audio is encoded by performed 1d conv and then apply relu activation function. The fusion process is performed through a simple concatenation operation over the convolution channel dimensions, followed by a position-wise projection P to reduce the feature dimension. Model achieved an accuracy of 14.02dB and 9.92dB Si-SNR on two and three-speaker test sets on LRS2.

This study proposes to implement a BRNN-based AV-SAD system with advanced LSTMs (ALSTMs), which includes multiple connections to frames in the past. The model contains 3 networks. First network through which acoustic features are passed on to two fully connected layers then to A-LSTM and LSTM layer. Second network also contains the same layers through which video features are passed onto.[34] The last

network which fuses both the network contains 8 two LSTM layers, a fully connected layer and a softmax layer in order. Achieved an F1 score of 89.7% in clean and HD recordings on their custom dataset.

M Wand et al. , used audio data to pre-processed with OpenSMILE. For video data, raw pixel features are used. The mouth ROI is extracted from the full-face videos as follows: First, facial landmarks are detected with the DLib facial landmark detector. Finally, all videos are converted into grayscale. In the fusion experiments, the video stream is up sampled by a factor of 4 to obtain 100 frames/second as for the audio. This study deals with two models one is single modality and the other is fusion model. [35]In single modality either video or audio is passed on to fully connected layers and after applying dropout then passed onto LSTM where as in fusion both audio and video are given as input to the same network (without the last LSTM layer) and then fusing them with the same network used in single modality model. Achieved an accuracy of 78.2%±5.5% with audio only model in -5db noise 90.3%±3.5%(when training on all noise levels) with fusion model on GRID audiovisual corpus.

2.2 Summary

From literature survey it is clear that AVSR systems comprise of both audio and visual data processing, and the important steps involved in the process of lip reading are face detection, lip localization or extraction followed by feature extraction and recognition. Here, the data needs to be trained in neural network system and tested in order to interpret the lip movements as text output. Different methods of lip reading and performance of hybrid models used for AVSR systems were explored. Lip reading database is the basis of lip reading recognition system and can influence the accuracy of recognition directly. Face detection and lip region localization being the primary step in lip reading systems may be performed either by using active appearance model (AAM) or by using Viola-Jones Algorithm. Hybrid models are the models which are a combination of two or more methods in order to interpret and analyse the data. Using such models it is possible to get accurate results with high accuracy rates for AVSR systems.

MFCC is considered to be better and popular audio feature extraction method when compared to RASTA-PLP or LPCC because it gives consistent results and is robust to noise. Along with that DTW based on template matching may be used for recognition. HMM is suitable for describing lip movement information perfectly. So it is very important in feature extraction. The HMM system consists of two phases: training phase and testing phase. In the training phase, features are extracted using DCT/DWT from the mouth regions and they are given as inputs to estimate the parameters of HMM. Then, the word is recognized in the testing phase.But it is observed that HMM with DWT based features performs better than HMM with DCT based features. Different hybrid models such as ANN-HMM, DBN-HMM, CNN-HMM, CNN LSTM etc may be used for feature extraction and recognition processes to get high recognition rates. Good recognition rates in AVSR systems increases its efficiency and makes it perform better than other systems thereby serves as a valuable resource to assist deaf people.

Chapter 3

System Architecture and Methodology

This chapter deals with the pipeline and the methodologies used in implementing Audio-Visual Speech Recogniton. To implement Audio-Visual Speech Recognition, two datasets were created with English and Kannada languages. English dataset contains a total of 737 videos of 9 different words 'About','Bad','Bottle', 'Come','Cow', 'Good','Pencil', 'Read', 'Where' with each word having 80 odd videos. Kannada dataset contains a total of 673 videos of 8 different words 'Avanu','Bagge','Bari', 'Howdu','Illa', 'Janarige', 'Kathe', 'Nale' with each word having 80 odd videos. There are a total of 16 speakers where they utter the words in Kannada and English datasets on 5 or 6 instances each.

3.1 Block diagram

The research is fulfilled by conducting four steps. The block diagram of Audio-Visual Speech Recognision is shown in Figure 3.1.

- Dataset Creation
- Audio Model
- Visual Model
- Fusion Model

First a dataset is created and video features and audio features are extracted and then using Deep Learning algorithms models are created for the classification using only audio, using only video and using both audio and video.

3.1.1 Dataset Creation

As the first step dataset is created with the above specification. Each video in the dataset is of 1920 x 1080 resolution. Then these videos are clipped using Microsoft video editor

with each video containing the duration of 1 sec at which the word is uttered and the video rate has been adjusted to 30 fps. Then the dataset is divided in two (Train and Validation) in such a way that the validation set has 21 videos which is nearly equal to 25% of the data and the train contains the remaining 75% of the data.

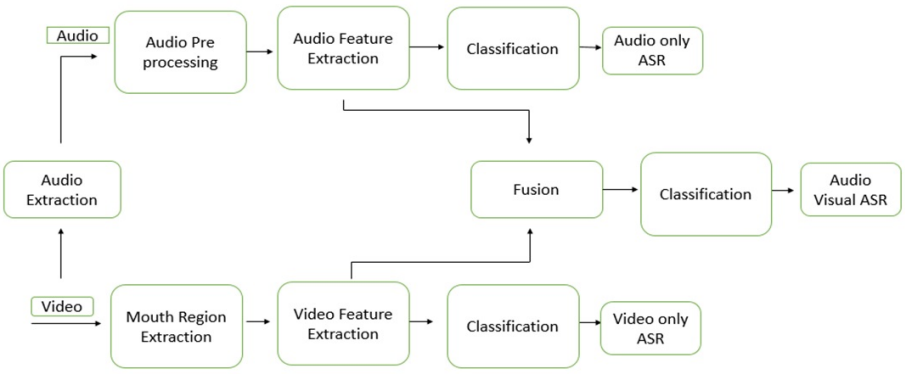


Figure 3.1: **Block Diagram of Audio-Visual Speech Recognition**

3.1.2 Audio Model

First audio files are created from the video dataset and saved in .wav formats using ffmpeg.Then the features are extracted from the audio using librosa which is an open source module that is available in python. In the intention to create a more classified feature vector, a total of five features from the audio are extracted. They are

- MFCCS
- CHROMA
- MEL
- CONTRAST
- TONNETZ

MFCCS

Mel Frequency Cepstral Coefficients of a signal are small set of features usually about 10-20 values which concisely describe overall shape of the spectral envelope.The envelope of the time power spectrum of a speech signal is a representative of the vocal tract and MFCC accurately represents this envelope.

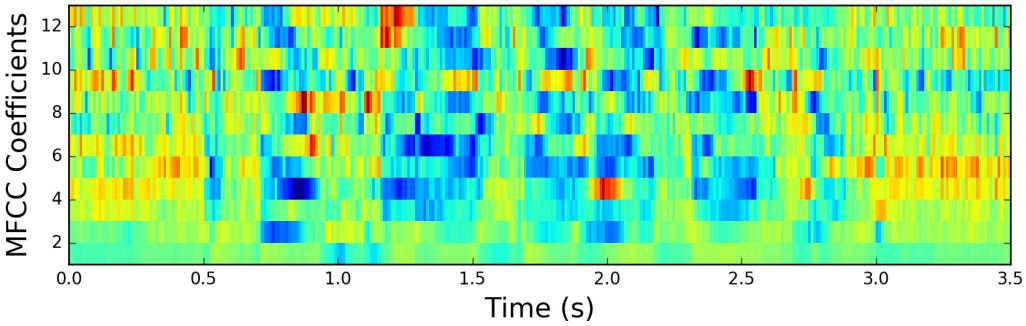


Figure 3.2: **Spectogram of MFCC**

CHROMA

Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency can give useful musical information about the audio and may even reveal perceived musical similarity that is not apparent in the original spectra.

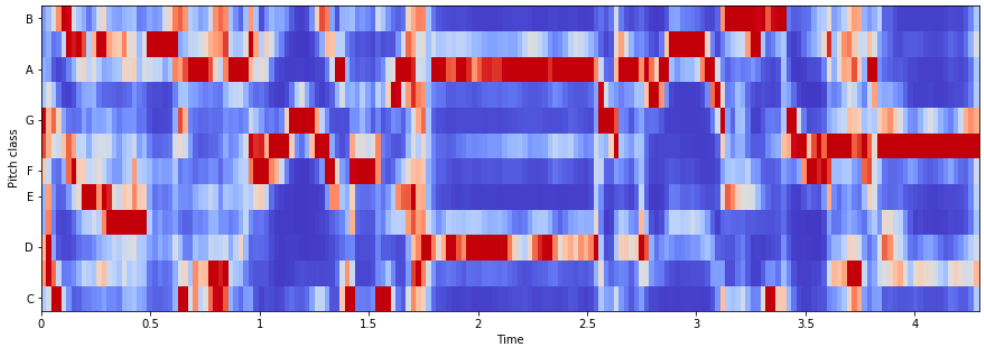


Figure 3.3: **Chroma Features**

MEL

The Mel Scale is the result of non-linear transformation of the frequency scale. This Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also “sound” to humans as they are equal in distance from one another.In contrast to Hz scale, where the difference between 500 and 1000 Hz is obvious, whereas the difference between 7500 and 8000 Hz is barely noticeable.

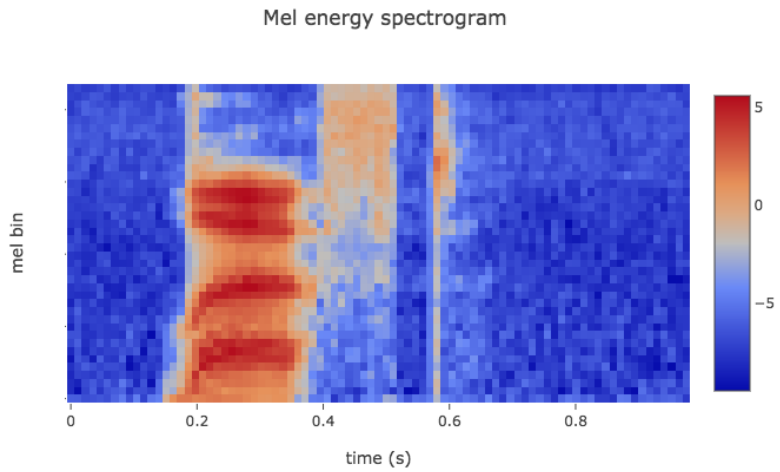


Figure 3.4: **Mel Features**

TONNETZ

A Tonnetz, or "tone-network" in German, is a two-dimensional representation of the relationships among pitches.Individual pitches are laid out along multiple axes: each intersection is displayed as a cell with a specific pitch.

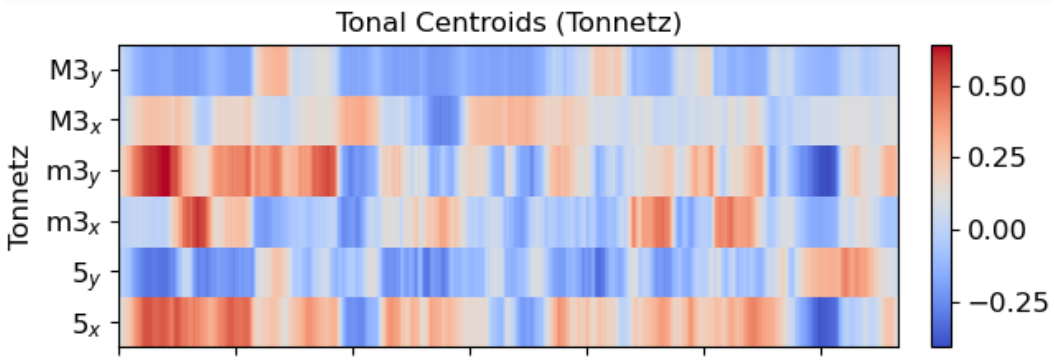


Figure 3.5: **Tonal Centroids**

CONTRAST

Contrast considers the spectral peak, the spectral valley, and their difference in each frequency sub-band.

All these features are combined to get generate the feature verctor of size 193 * 1. Next a Convolutional Neural Network using 1 Conv1D layers, followed by MaxPooling1D layer,

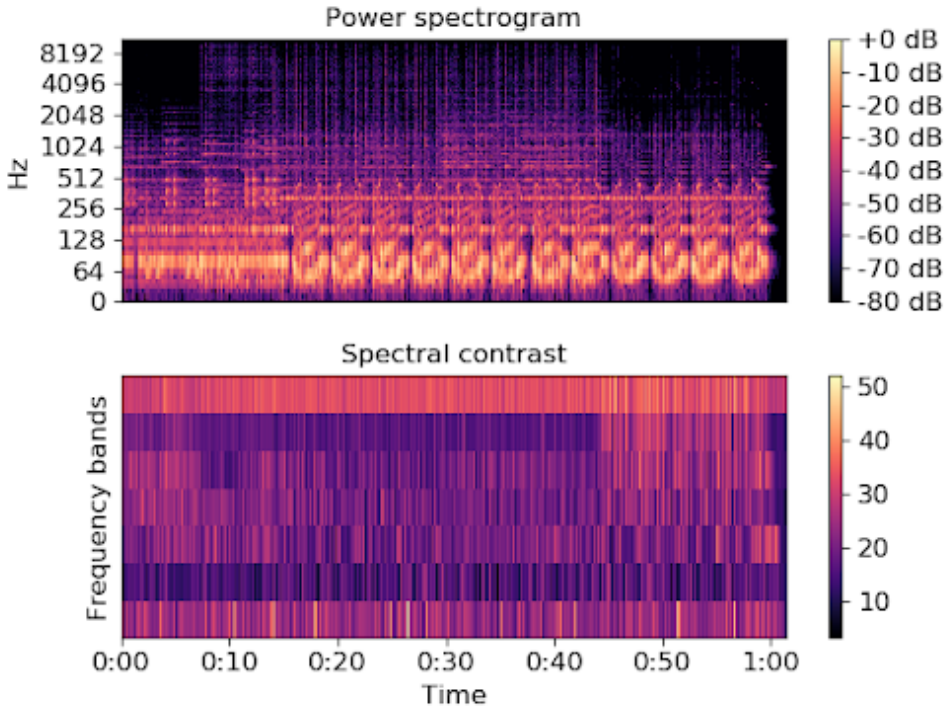


Figure 3.6: **Spectral Contrast**

Batchnormalization layer, dropout and then followed by two Dense Layers is created. One-dimensional CNNs work with sequence in one dimension, and tends to be useful in various signal analysis over the fixed length signals. They work well for analysis of audio signals, for instance. The output from the corresponding layers of audio model will be as follows

$$a^1 = b^1 + \sum_{i=1}^{193} conv1D(w_i, X\_train\_audio[i] \tag{3.1}$$

$$y^1 = R(a^1) \tag{3.2}$$

$$y^2 = R(w^2 y^1 + b^2) \tag{3.3}$$

$$y^3 = R(w^3 y^2 + b^3) \tag{3.4}$$

$$y^4 = R(w^4 y^3 + b^4) \tag{3.5}$$

where y^i is the output vector of layer i, R is the ReLu activation function, w^i is the weights of layer i, b^i is the bias of layer i.

Finally a softmax layer was attached for to this network for the classification. The loss function which is used to train the model is cross entropy which is given by

$$L^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} log(\hat{y}^{<t>}) - (1 - y^{<t>}) log(1 - \hat{y}^{<t>}) \tag{3.6}$$

3.1.3 Visual Model

First mouth region is extracted from the video using dlib library which is available in python 3 as shown in Figure 3.7 in each frame of the video.Then the regions are converted into grey scale to reduce the complexity of the model as shown in Figure 3.8. Then the position of the outer lip coordinates are extracted and saved in the feature vector.

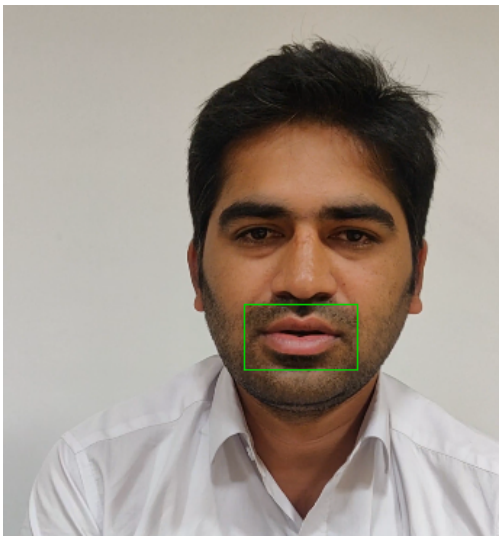


Figure 3.7: **Mouth ROI extraction**

Figure 3.8: **Conversion to GreyScale**

Then a model with network of LSTM's and dense layers(Deep LSTM Network) is created. Long Short-Term Memory (LSTM) networks is a type of RNN(recurrent neural network) which is can learn order dependence in a sequence prediction problem. It contains three gates:

1. Forget Gate

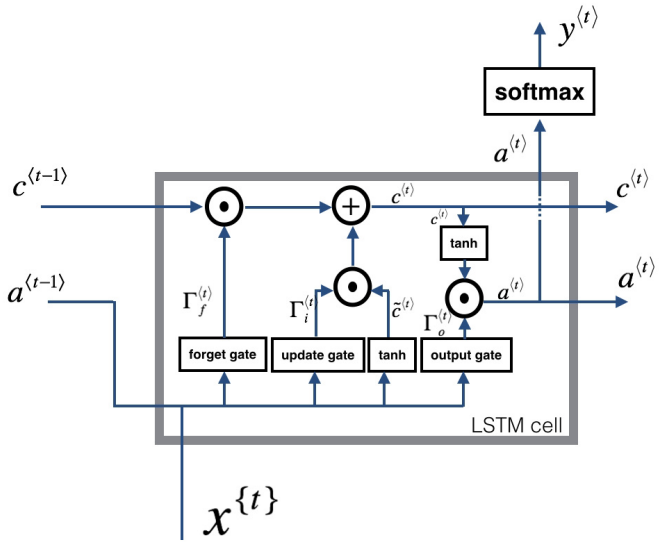


Figure 3.9: **Structure of an LSTM cell**

2. Input Gate
3. Output Gate

Forget Gate decides whether to keep or forget the info from the previous timestamps and Input Gate quantifies the importance of that data coming as an input and the Output Gate figures the most relevant output that it has to generate.

The model which is created contains an LSTM layer with 128 hidden units and 8 time stamps as the first layer.The operaions inside LSTM cell are as follows

$$\tilde{c}^{<t>} = tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c) \tag{3.7}$$

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u) \tag{3.8}$$

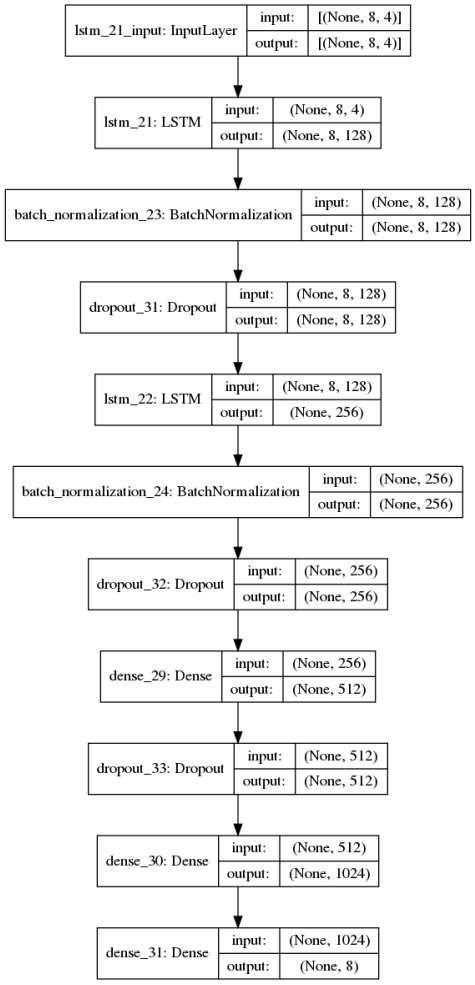
$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f) \tag{3.9}$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o) \tag{3.10}$$

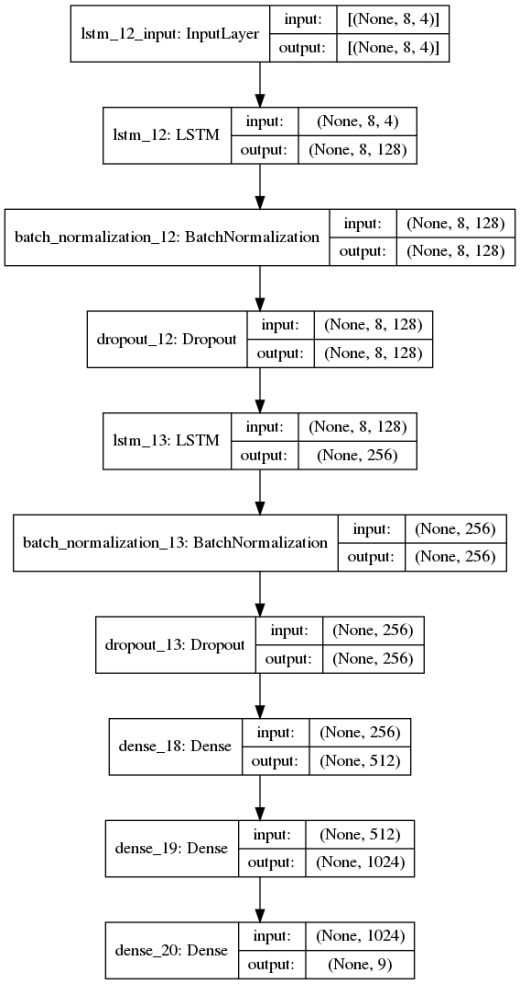
$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \tag{3.11}$$

$$a^{<t>} = \Gamma_o * tanh(c^{<t>}) \tag{3.12}$$

where c represents the memory cell, t represent time stamp, Γ_u represents the update gate. Γ_f represents the forget gate, Γ_o represents the output gate, σ represents the sigmoid function, W_u represents the weights of update gate, W_f represents the weights of forget gate, W_o represents the weights of output gate, b represents bias, $\tilde{c}^{<t>}$ represents



(a) Model used to train Kannada dataset



(b) Model used to train English dataset

Figure 3.10: Overview of Visual model

candidate cell variables.

And next one more LSTM layer is introduced. So the resulting output from the second layer will be

$$\tilde{c1}^{<t>} = tanh(W1_c[a1^{<t-1>}, a^{<t>}] + b1_c) \tag{3.13}$$

$$\Gamma 1_u = \sigma(W1_u[a1^{<t-1>}, x1^{<t>}] + b1_u) \tag{3.14}$$

$$\Gamma 1_f = \sigma(W1_f[a1^{<t-1>}, x1^{<t>}] + b1_f) \tag{3.15}$$

$$\Gamma 1_o = \sigma(W1_o[a1^{<t-1>}, x1^{<t>}] + b1_o) \tag{3.16}$$

$$c1^{<t>} = \Gamma 1_u * \tilde{c1}^{<t>} + \Gamma 1_f * c1^{<t-1>} \tag{3.17}$$

$$a1^{<t>} = \Gamma 1_o * tanh(c1^{<t>}) \tag{3.18}$$

This is followed by three dense layers. So again the output equations from the these three dense layers will be

$$y^2 = R(W2 * a1 + b2) \tag{3.19}$$

$$y^3 = R(W3 * y^2 + b3) \tag{3.20}$$

$$y^4 = R(W4 * y^3 + b4) \tag{3.21}$$

and finally a softmax layer was attached for to this network for the classification. The loss function which is used to train the model is cross entropy which is given by

$$L^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} log(\hat{y}^{<t>}) - (1 - y^{<t>}) log(1 - \hat{y}^{<t>}) \tag{3.22}$$

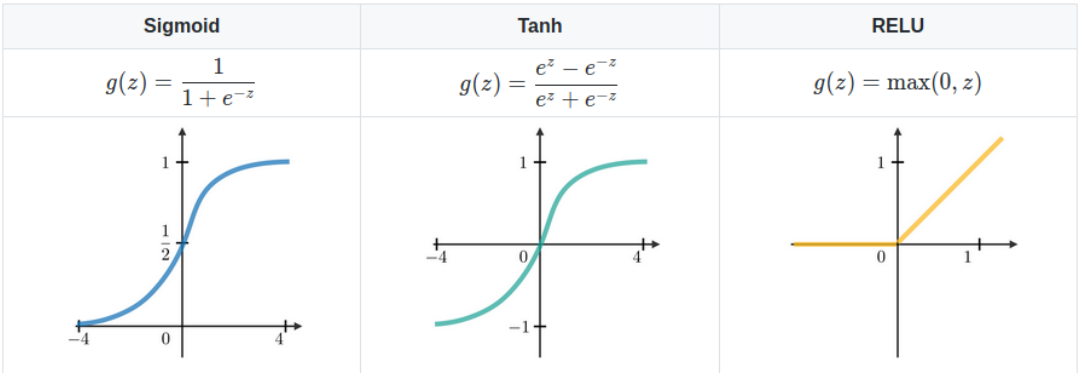


Figure 3.11: **Activation functions used**

Video model overview is as shown in Figure 3.10 and the activation functions used in the models are shown in Figure 3.11.

3.1.4 Fusion Model(AVSR)

Fusion model contains a total of four parts.

1. Audio only part
2. Visual only part
3. Combination of Audio only and Visual only parts
4. Combination of all the above 3 parts

In audio-only part features are extracted in the same way as the Audio model. Then a Deep Convolutional Neural Network is created. The model which is created is the replica of the Audio model except in place of a softmax layer there is an additional dense layer.

So, the equations of this part are as follows

$$a^1 = b^1 + \sum_{i=1}^{193} conv1D(w_i, X\_train\_audio[i] \tag{3.23}$$

$$y^1 = R(a^1) \tag{3.24}$$

$$y^2 = R(w^2 y^1 + b^2) \tag{3.25}$$

$$y^3 = R(w^3 y^2 + b^3) \tag{3.26}$$

$$y^4 = R(w^4 y^3 + b^4) \tag{3.27}$$

$$y^5 = R(w^5 y^4 + b^5) \tag{3.28}$$

In Video-only model, Video features as extracted the same way as the Video model. Then a deep LSTM network is created.The first layer which is LSTM layer contains 128 hidden units with 8 time stamps.

$$\tilde{c}^{<t>} = tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c) \tag{3.29}$$

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u) \tag{3.30}$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f) \tag{3.31}$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o) \tag{3.32}$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \tag{3.33}$$

$$a^{<t>} = \Gamma_o * tanh(c^{<t>}) \tag{3.34}$$

It is followed by Batch Normalization layer and dropout layer and next one more LSTM layer with 8 timestamps amd 128 hidden units is introduced.

$$\tilde{c1}^{<t>} = tanh(W1_c[a1^{<t-1>}, a^{<t>}] + b1_c) \tag{3.35}$$

$$\Gamma 1_u = \sigma(W1_u[a1^{<t-1>}, x1^{<t>}] + b1_u) \tag{3.36}$$

$$\Gamma 1_f = \sigma(W1_f[a1^{<t-1>}, x1^{<t>}] + b1_f) \tag{3.37}$$

$$\Gamma 1_o = \sigma(W1_o[a1^{<t-1>}, x1^{<t>}] + b1_o) \tag{3.38}$$

$$c1^{<t>} = \Gamma 1_u * \tilde{c1}^{<t>} + \Gamma 1_f * c1^{<t-1>} \tag{3.39}$$

$$a1^{<t>} = \Gamma 1_o * tanh(c1^{<t>}) \tag{3.40}$$

This LSTM layer is followed by dropout and a dense layer. So, the resulting equation from this dense layer will be

$$y_v = R(W2 * a1^{<t>} + b2) \tag{3.41}$$

Overview of this part of the model is shown in Figure 3.13 and the architecture of LSTM network is shown in Figure 3.14

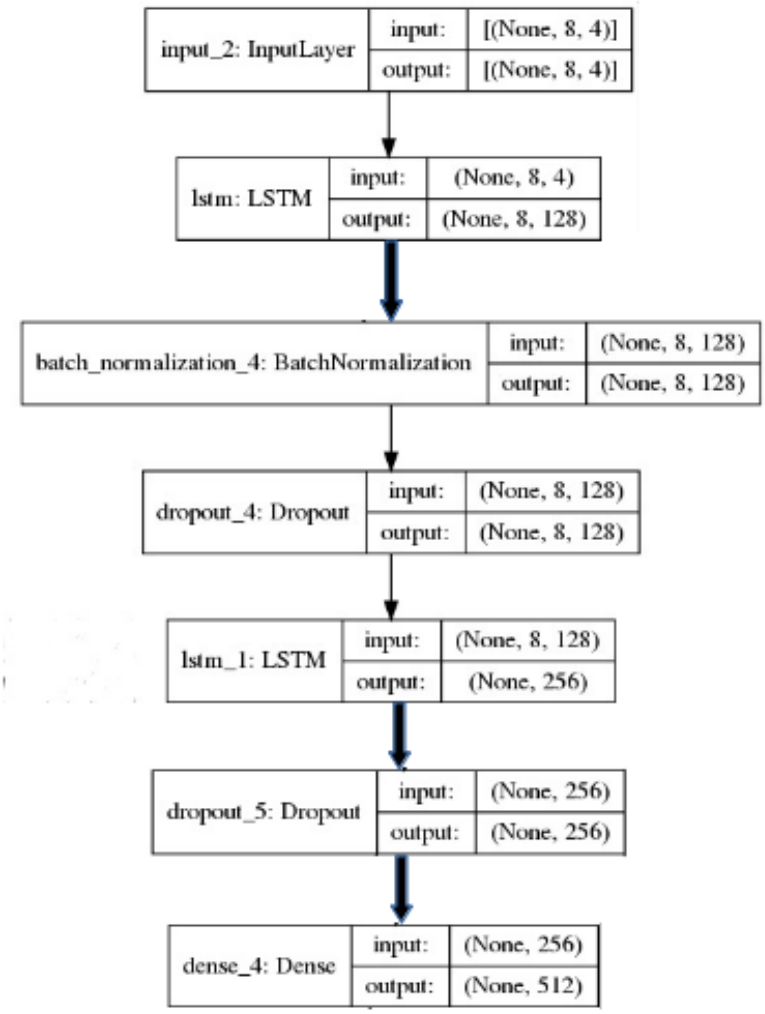


Figure 3.12: **Overview of Visual-only part in Fusion model**

In "Combination of Audio-only and Visual-only parts" the feature map from the first dense layer from Audio only part is concatenated with the feature map from first LSTM layer from Visual only part. then from equations 3.3 and 3.12

$$a_c = [a1^{<t>}, y^2] \tag{3.42}$$

and the resulting feature map is passed as an input to Deep Feedforward Neural Network which contains three dense layers while the first two dense layers are followed by a Batch normalization layer and a dropout layer respectively. Feedforward Neural network is the basic feedforward neural network which does not form any loops. The architecture of Feedforward Neural network is shown in Figure 3.15 and the overview of the model is

shown in Figure 3.16

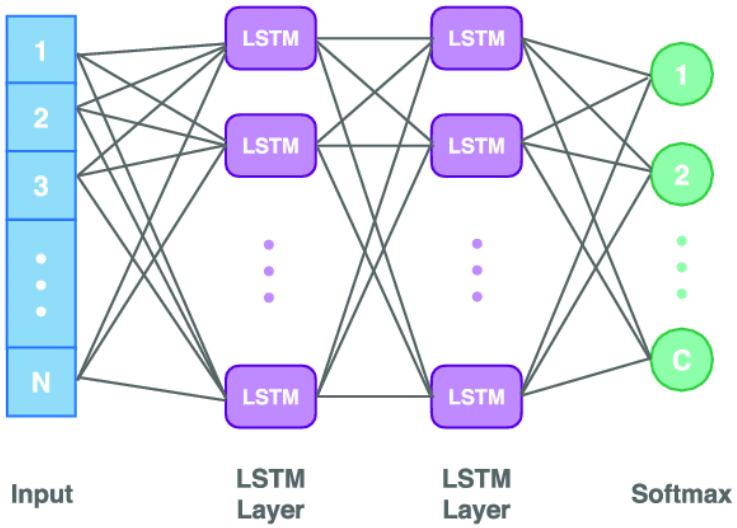


Figure 3.13: **Architecture of LSTM Network**

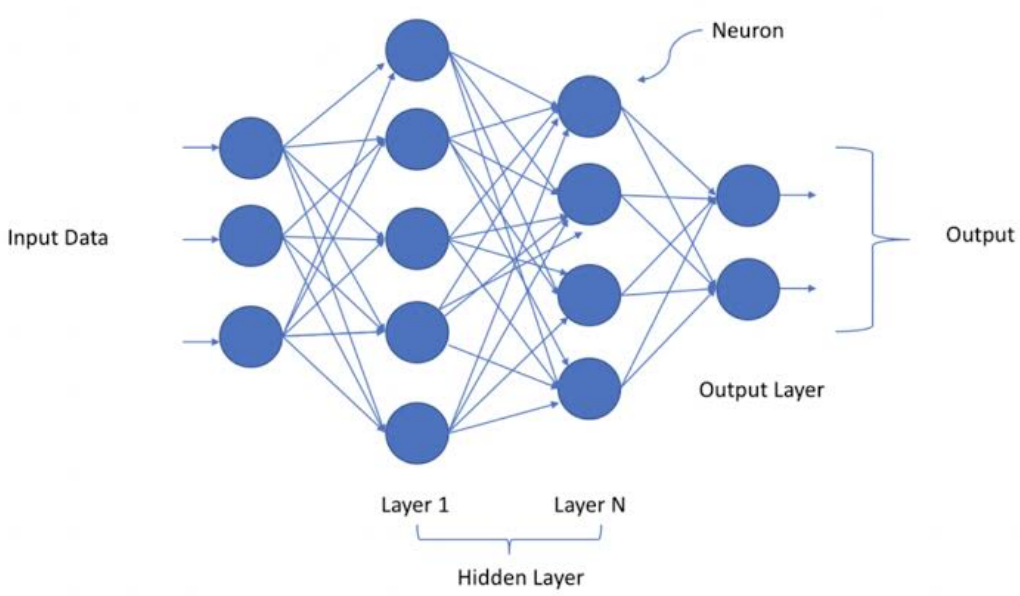


Figure 3.14: **Architecture of Feedforward Neural Network**

$$y_{d1} = R(W_{d1} * a_c + b_{d1}) \tag{3.43}$$

$$y_{d2} = R(W_{d2} * y_{d1} + b_{d2}) \tag{3.44}$$

$$y_{d3} = R(W_{d3} * y_{d2} + b_{d3}) \tag{3.45}$$

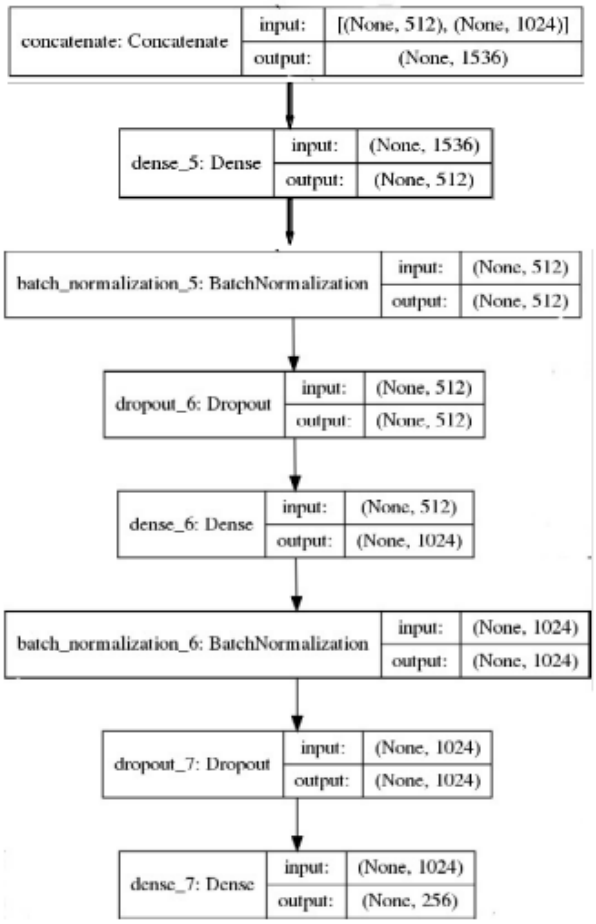


Figure 3.15: **Overview of Combining Audio-only and Visual-only parts in Fusion model**

And as the final step all above three parts are combined so the vector formed by this will be a combination of output vector of all the above three parts. Therefore from equations 3.28, 3.41, 3.45

$$a_{c2} = [y^5, y_v, y_{d3}] \tag{3.46}$$

Then this is passed as an input to Deep Feedforward Neural Network which contains three dense layers followed b a Batch Normalization lyer and a dropout layer.

$$y_{c1} = R(W_{c1} * a_{c2} + b_{c1}) \tag{3.47}$$

$$y_{c2} = R(W_{c2} * y_{c1} + b_{c2}) \tag{3.48}$$

$$y_{c3} = R(W_{c3} * y_{c2} + b_{c3}) \tag{3.49}$$

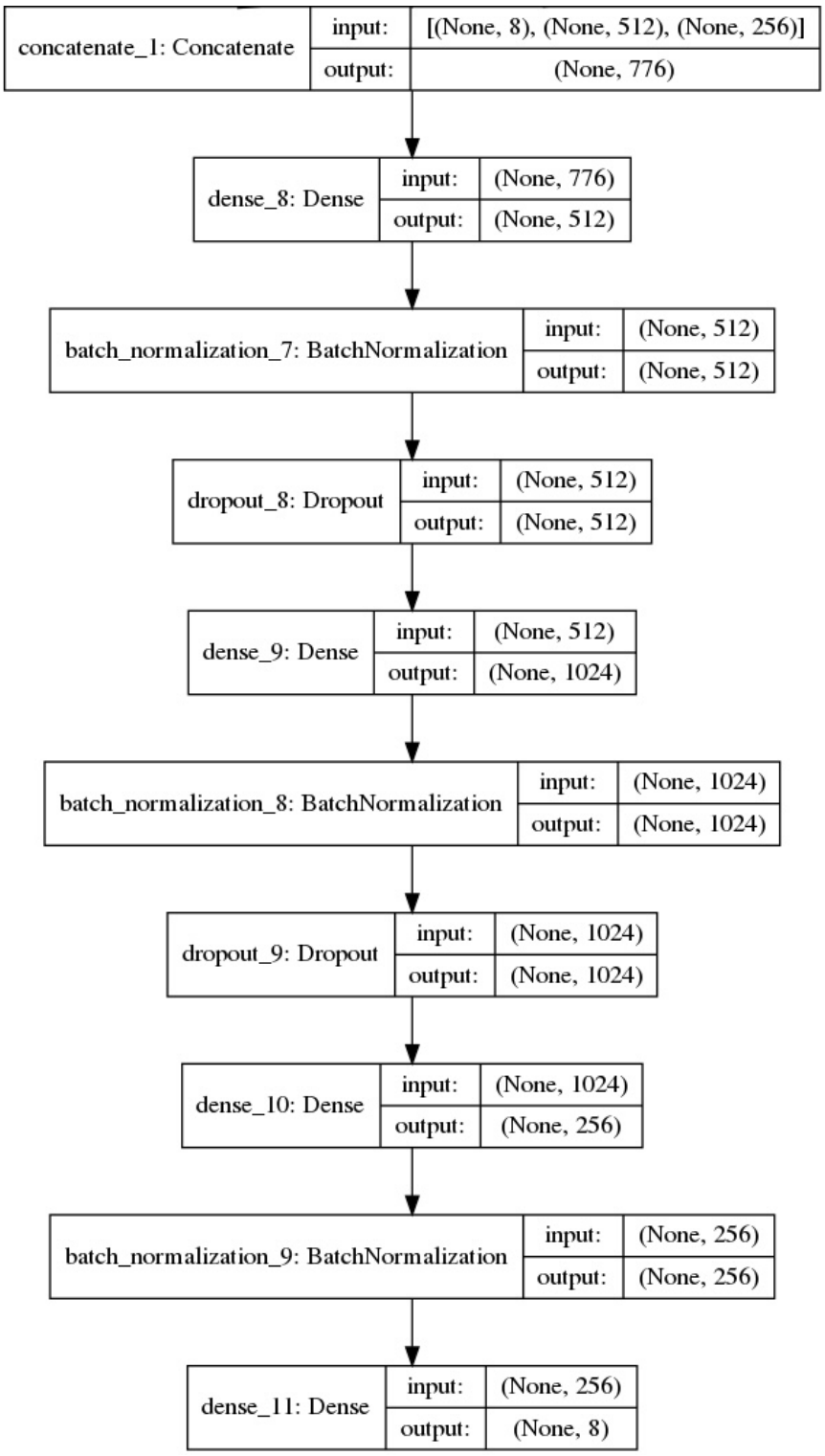


Figure 3.16: **Overview of Combining all three previous parts in Fusion model**

Chapter 4

Hardware and Software Components

4.1 Hardware requirements

4.1.1 Camera

In day-to-day life Camera is used a good amount of time for taking photos or record videos. In this paper, the database videos have been recorded with a DSLR Camera. A video quality of 1080p with 30 frames per second will be sufficient irrespective of the source used for recording say either a Camera or a Mobile Phone.

4.1.2 Laptop

A portable computer. The following are the specifications of the laptop for a smooth and efficient run of programs:-
Processor: Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
RAM: 16GB
Operating System: Windows or Linux

4.2 Software Components

4.2.1 Windows Video Editor

Windows Video Editor editor is the default video editor that is present in Windows for free. Windows video editor is a simple tool and has many advantages. The main advantage of using the windows video editor is say if the video has a irregular frames per second or has a frame per second more 30 frames per second then, when the video is cut to specific seconds or minutes the windows video editor makes sure that the frame per second in the edited video will equal to 30 frames per second.

4.2.2 Python

Python has always been a interpreted and high-level, general-purpose programming language. The design of Python's philosophy is emphasized by code readability which has its notable use with significant amount of indentation. Its a language that constructs as well as it even supports object-oriented approach which help in aiming to help the programmers write logical, clear code in case of small-scale and large-scale projects.

Python was a dynamically typed with garbage collection added to it. It even supports for some of the multiple programming paradigms, which includes structured (procedural, particularly), functional programming and object-oriented. Python was quite often described as the "Batteries Included" programming language because of it's comprehensive standard library.

Guido van Rossum has started working on Python in the year 1980s, which comes as a successor to the ABC programming language, and the initial release which was in 1991 was named as Python 0.9.0.Then, Python 2.0 has been released in early 2000 and introduced a lot of new features, like list comprehensions, garbage collection system. Python 3.0 has been released in the year 2008 and it was actually a huge major revision of the prevision language version that is not at all completely backward-compatible. Now, Python has been consistently ranking as one of the most popular programming languages in the word.

4.2.3 FFmpeg

FFMPEG is one of the popular python modules that helps in processing the data which is of either a photo or a video. It is a complete, cross-platform solution which is used for recording, converting and streaming of audio as well as the video. There are quite wide range of options which helps in processing the data for the ease of grtting the required solution.

FFmpeg has always been a leading multimedia framework, which is capable in decoding, encoding, transcoding, mux, demux, stream, filter and play pretty much anything that the humans and computers or machines have ever created. It supports the most unknown ancient or old formats up to the cutting edge. It does not matter if they were actually designed by some of the community, the standards committee or a any of the corporations out there. FFmpeg is very highly portable which means FFmpeg compiles, runs, and passes any of the testing infrastructure FATE across different Operating systems like Linux, Mac OS X, Microsoft Windows, the BSDs, Solaris, etc. and under a huge wide range in variety of build environments, machine architectures, and configurations.

The FFmpeg project module helps in trying to provide the very best technically possible answer for developers of app(application) and the users. In order, To achieve this FFmpeg combine the best available free software options. FFmpeg has a slight favor on thier particular own code to keep the dependencies on other libs low and also to maximize the code cross-sharing between several parts of FFmpeg.

4.2.4 Keras

Keras is the deep learning Application programming interface written in Python, running on top of the ML(Machine Learning) platform with TensorFlow. It has been developed with a sheer focus in a enabling tremendous and fast experimentation. Keras has been able to go from idea to the expected result as fast as possible.

Keras is Simple but definitely not the simplistic. It reduces developer a good amount of cognitive load which free the user to focus on the parts of the problem which has more importance and actually matter. Keras is Flexible. It has the ability to adopt the principle of progressive disclosure of complexity, like simple workflows are made quick and easy, while arbitrarily advanced workflows are made possible with the help of a clear path that builds upon what the user has already learned. Keras is Powerful. It provides the required industry strength, scalability and performance. it is used by prestigious companies and organizations including , Waymo, NASA, and YouTube.

4.2.5 Virtual Environment

The Virtual Environment is a tool which helps in keep dependencies required for different projects and are separated by creating isolated python virtual environments for these. This has been one of the most and quite important tools where most of the python developers have been using.

The virtual environment must be used whenever user work in any of the python based projects. It is generally best practise to have one new virtual environment for each and every python based project that user work on. So that the dependencies of each and every project can isolated from the system and as well as from each other.

Chapter 5

Implementation and Testing

5.1 Result Analysis

Audio Model

- Achieved training accuracy of 93.86% and test accuracy of 91.07% for Kannada Audio-dataset trained for 70 epochs as shown in Figure 5.1
- Achieved train accuracy of 93.67% and test accuracy of 91.53% for English Audio-dataset trained for 60 epochs as shown in Figure 5.2

```
val_accuracy: 0.8869
Epoch 195/200
16/16 [==============================] - 0s 15ms/step - loss: 0.1657 - accuracy: 0.9473 - val_loss: 0.4070 -
val_accuracy: 0.8869
Epoch 196/200
16/16 [==============================] - 0s 14ms/step - loss: 0.2451 - accuracy: 0.9289 - val_loss: 0.3733 -
val_accuracy: 0.8988
Epoch 197/200
16/16 [==============================] - 0s 14ms/step - loss: 0.2303 - accuracy: 0.9225 - val_loss: 0.3204 -
val_accuracy: 0.9167
Epoch 198/200
16/16 [==============================] - 0s 22ms/step - loss: 0.1930 - accuracy: 0.9325 - val_loss: 0.3248 -
val_accuracy: 0.9226
Epoch 199/200
16/16 [==============================] - 0s 14ms/step - loss: 0.1578 - accuracy: 0.9423 - val_loss: 0.3979 -
val_accuracy: 0.9048
Epoch 200/200
16/16 [==============================] - 0s 14ms/step - loss: 0.2224 - accuracy: 0.9386 - val_loss: 0.3906 -
val accuracy: 0.9107
```

Figure 5.1: **Training Kannada dataset with Audio Model**

```
val_accuracy: 0.9153
Epoch 195/200
18/18 [==============================] - 0s 18ms/step - loss: 0.2683 - accuracy: 0.9140 - val_loss: 0.2569 -
val_accuracy: 0.9101
Epoch 196/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2551 - accuracy: 0.9052 - val_loss: 0.2800 -
val_accuracy: 0.8942
Epoch 197/200
18/18 [==============================] - 0s 16ms/step - loss: 0.2698 - accuracy: 0.9117 - val_loss: 0.2455 -
val_accuracy: 0.8995
Epoch 198/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2771 - accuracy: 0.8985 - val_loss: 0.2019 -
val_accuracy: 0.9206
Epoch 199/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2879 - accuracy: 0.9038 - val_loss: 0.2432 -
val_accuracy: 0.9206
Epoch 200/200
18/18 [==============================] - 0s 17ms/step - loss: 0.1807 - accuracy: 0.9367 - val_loss: 0.2052 -
val_accuracy: 0.9153
```

Figure 5.2: **Training English dataset with Audio Model**

- Model accuracy and Model loss plots of the Kannada dataset are as shown in Figure 5.3 and Figure 5.4 respectively.

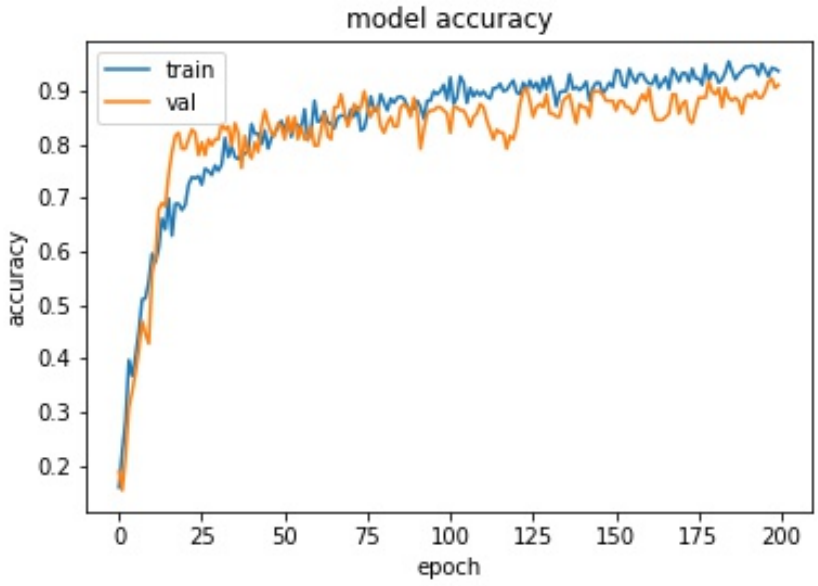


Figure 5.3: **Accuracy plot of Kannada dataset with Audio Model**

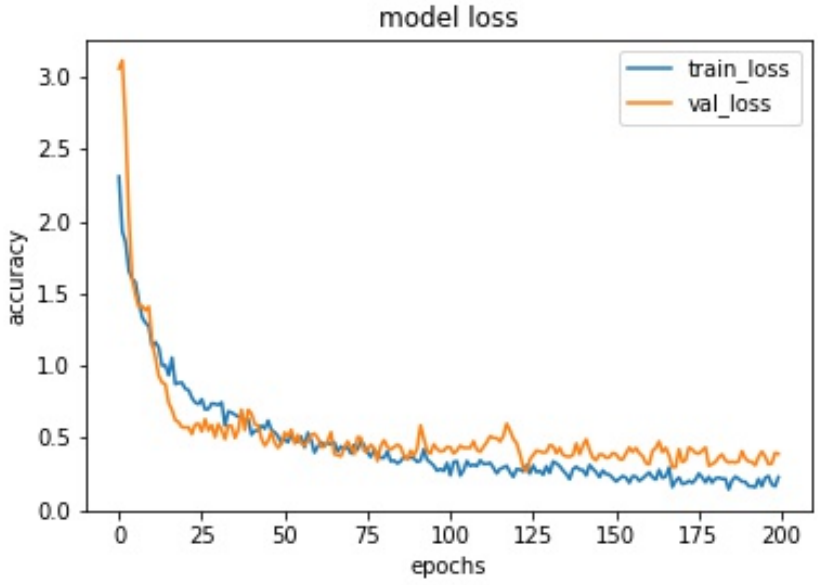


Figure 5.4: **Loss plot of Kannada dataset with Audio Model**

- Model accuracy and Model loss plots of the English dataset are as shown in Figure 5.5 and Figure 5.6 respectively.

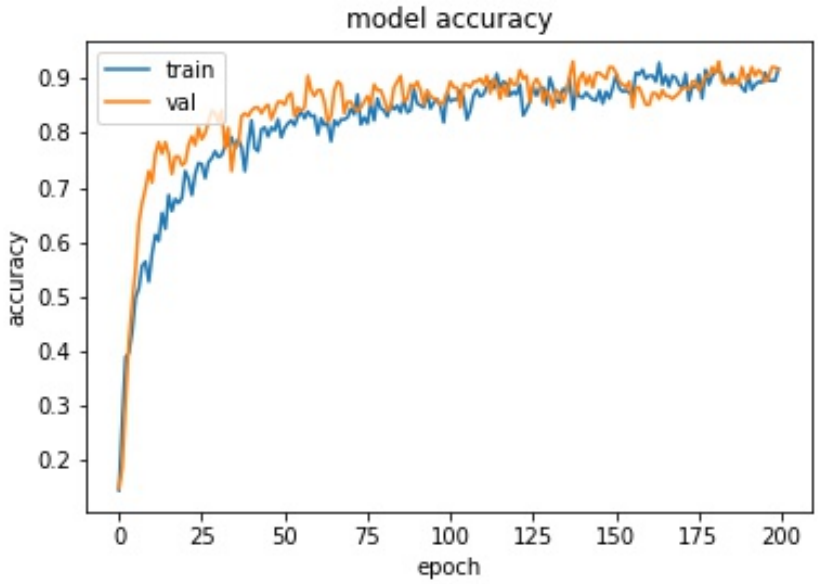


Figure 5.5: **Accuracy plot of English dataset with Audio Model**

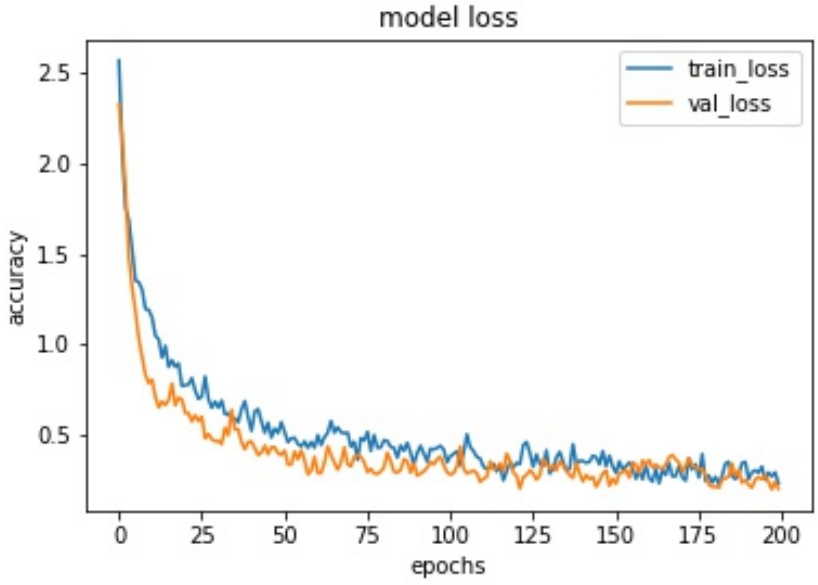


Figure 5.6: **Loss plot of English dataset with Audio Model**

Confusion matrix indicates the number of words that are being classified as various classes that are available.

The Confusion Matrix of the Audio models which are used to train Kannada dataset and English dataset are shown in Figure 5.7 and Figure 5.8 respectively. From Figure 5.7 it can be observed that percentage of "Avanu videos" in the test dataset classified as Avanu are 95.23%, percentage of "Bagge videos" in the test dataset classified as Bagge are 80.95%, percentage of Bari videos" in the test dataset classified as Bari are 90.47%, percentage of "Howdu videos" in the test dataset classified as Howdu are 85.71%, percentage of "Illa videos" in the test dataset classified as Illa are 85.71%, percentage of "Janarige videos"in the test dataset classified as Janarige are 100%, percentage of "Kathe videos" in the test dataset classified as Kathe are 100%, percentage of "Nale videos" in the testdataset classified as Nale are 90.47%.

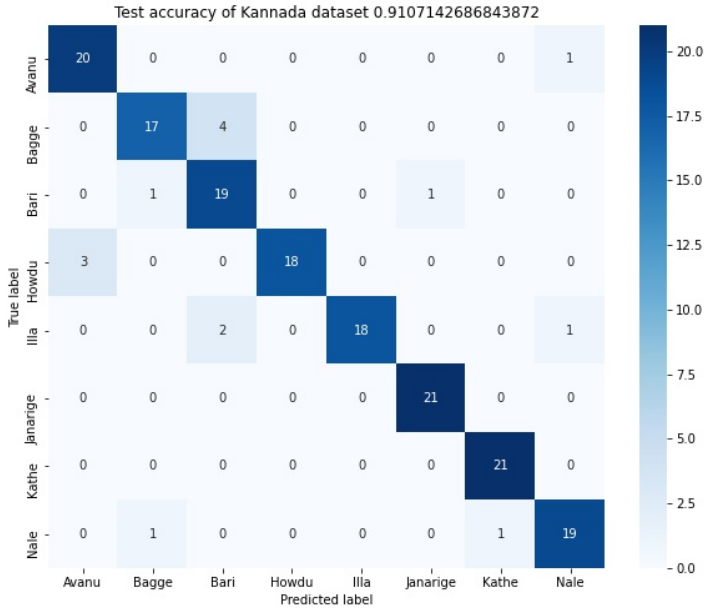


Figure 5.7: **Confusion Matrix of Kannada dataset with Audio Model**

From Figure 5.8 it can be observed that the percentage of "About videos" in the test dataset classified as About are 80.95%, percentage of "Bad videos" in the test dataset classified as Bad are 100%, percentage of "Bottle videos" in the test dataset classified as Bottle are 85.71%, percentage of "Come videos" in the test dataset classified as Come are 85.71%, percentage of "Cow videos" in the test dataset classified as Cow are 85.71%, percentage of "Good videos" in the test dataset classified as Good are 95.23%,, percentage of "Pencil videos" in the test dataset classified as Pencil are 100%, percentage of "Read videos" in the test dataset classified as Read are 90.47%, percentage of "Where

videos" in the test dataset classified as Where are 100%

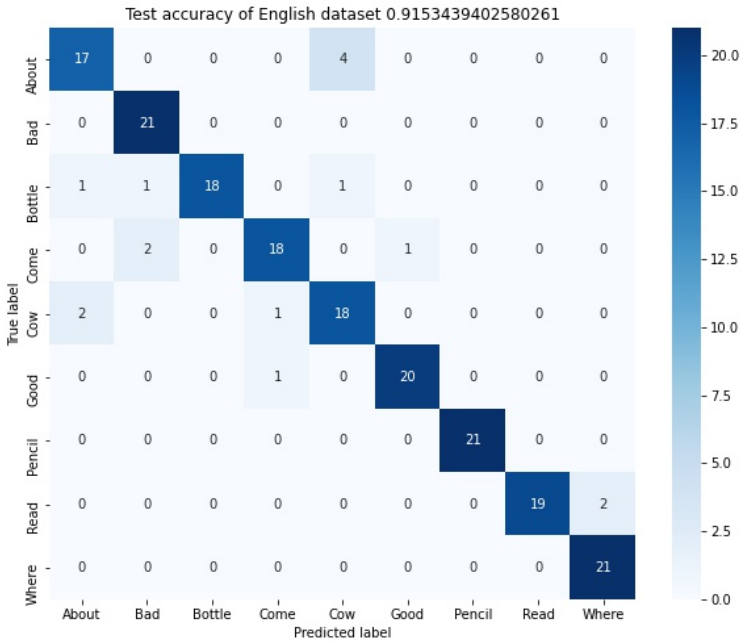


Figure 5.8: **Confusion Matrix of English dataset with Audio Model**

Classification report gives Precision, Recall and F1 score for every classification where

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.2}$$

$$F1score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{5.3}$$

The classification report of kannada dataset and english dataset are as shown in Figure 5.9 and Figure 5.10 respectively.

	precision	recall	f1-score	support
Avanu	0.87	0.95	0.91	21
Bagge	0.89	0.81	0.85	21
Bari	0.76	0.90	0.83	21
Howdu	1.00	0.86	0.92	21
Illa	1.00	0.86	0.92	21
Janarige	0.95	1.00	0.98	21
Kathe	0.95	1.00	0.98	21
Nale	0.90	0.90	0.90	21
accuracy			0.91	168
macro avg	0.92	0.91	0.91	168
weighted avg	0.92	0.91	0.91	168

Figure 5.9: **Classification Report of Kannada dataset with Audio Model**

	precision	recall	f1-score	support
About	0.85	0.81	0.83	21
Bad	0.88	1.00	0.93	21
Bottle	1.00	0.86	0.92	21
Come	0.90	0.86	0.88	21
Cow	0.78	0.86	0.82	21
Good	0.95	0.95	0.95	21
Pencil	1.00	1.00	1.00	21
Read	1.00	0.90	0.95	21
Where	0.91	1.00	0.95	21
accuracy			0.92	189
macro avg	0.92	0.92	0.92	189
weighted avg	0.92	0.92	0.92	189

Figure 5.10: **Classification Report of English dataset with Audio Model**

Video Model

- Kannada dataset is trained for 70 epochs.
- English dataset is trained of 60 epochs.
- Achieved train accuracy of 77.57% and test accuracy of 75.00% for Kannada Video-dataset as shown in Figure 5.11.

```
accuracy: 0.6548
Epoch 65/70
16/16 [==============================] - 1s 42ms/step - loss: 0.7296 - accuracy: 0.7496 - val_loss: 1.3356 - val_
accuracy: 0.7440
Epoch 66/70
16/16 [==============================] - 1s 41ms/step - loss: 0.7387 - accuracy: 0.7509 - val_loss: 1.2507 - val_
accuracy: 0.6964
Epoch 67/70
16/16 [==============================] - 1s 41ms/step - loss: 0.7092 - accuracy: 0.7489 - val_loss: 1.3634 - val_
accuracy: 0.7202
Epoch 68/70
16/16 [==============================] - 1s 41ms/step - loss: 0.7492 - accuracy: 0.7312 - val_loss: 1.4295 - val_
accuracy: 0.6786
Epoch 69/70
16/16 [==============================] - 1s 41ms/step - loss: 0.6151 - accuracy: 0.7850 - val_loss: 1.3445 - val_
accuracy: 0.7619
Epoch 70/70
16/16 [==============================] - 1s 42ms/step - loss: 0.6286 - accuracy: 0.7757 - val_loss: 1.1903 - val_
accuracy: 0.7500
```

Figure 5.11: **Training Kannada dataset with Video Model**

- Achieved train accuracy of 77.48% and test accuracy of 76.19% for English Video-dataset as shown in Figure 5.12.

```
Epoch 55/60
18/18 [==============================] - 1s 42ms/step - loss: 0.7268 - accuracy: 0.7414 - val_loss: 1.1511 - val_
accuracy: 0.6667
Epoch 56/60
18/18 [==============================] - 1s 41ms/step - loss: 0.7306 - accuracy: 0.7416 - val_loss: 1.0958 - val_
accuracy: 0.6984
Epoch 57/60
18/18 [==============================] - 1s 41ms/step - loss: 0.6274 - accuracy: 0.7793 - val_loss: 1.2342 - val_
accuracy: 0.7143
Epoch 58/60
18/18 [==============================] - 1s 39ms/step - loss: 0.6977 - accuracy: 0.7549 - val_loss: 1.3743 - val_
accuracy: 0.6720

Epoch 59/60
18/18 [==============================] - 1s 38ms/step - loss: 0.7350 - accuracy: 0.7513 - val_loss: 1.1761 - val_
accuracy: 0.6720
Epoch 60/60
18/18 [==============================] - 1s 38ms/step - loss: 0.6593 - accuracy: 0.7748 - val_loss: 1.0564 - val_
accuracy: 0.7619
```

Figure 5.12: **Training English dataset with Video Model**

- Model accuracy and Model loss plots of the Kannada dataset are as shown in Figure 5.13 and Figure 5.14 respectively.

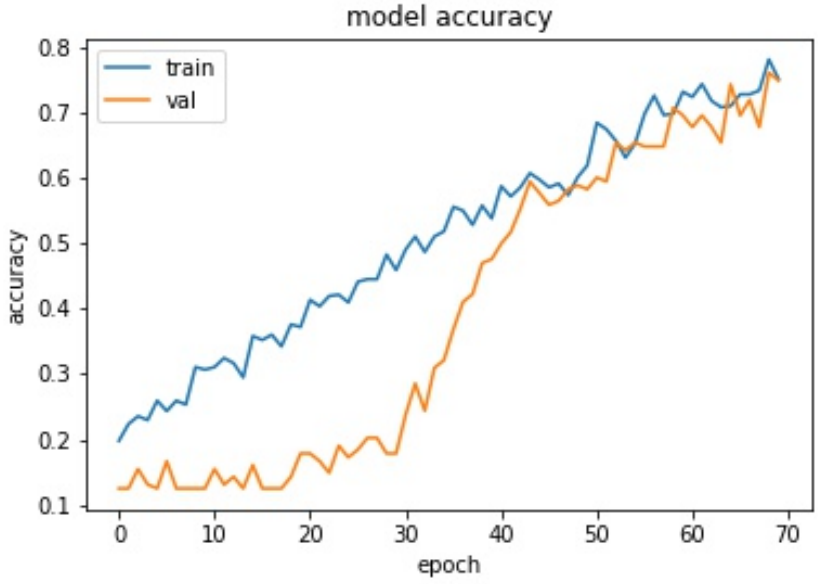


Figure 5.13: **Accuracy plot of Kannada dataset with Video Model**

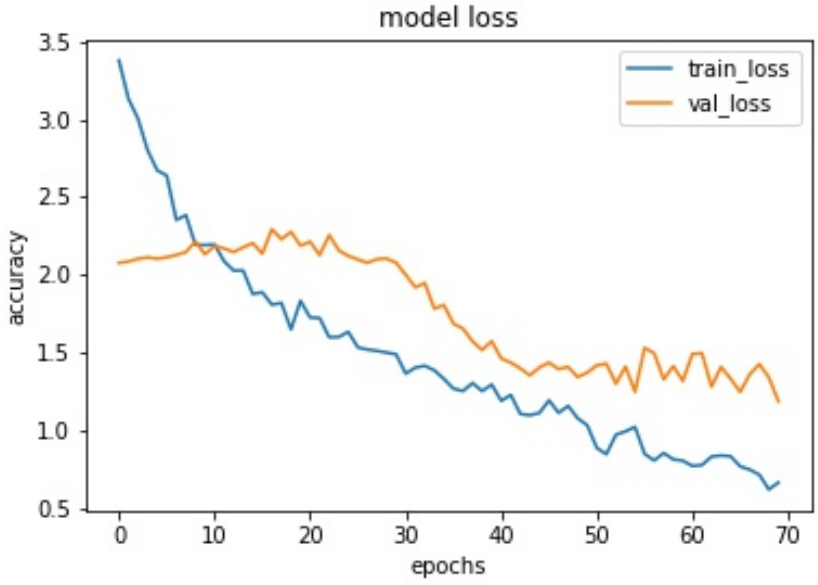


Figure 5.14: **Loss plot of Kannada dataset with Video Model**

- Model accuracy and model loss plots of the English dataset are as shown in Figure 5.15 and Figure 5.16 respectively.

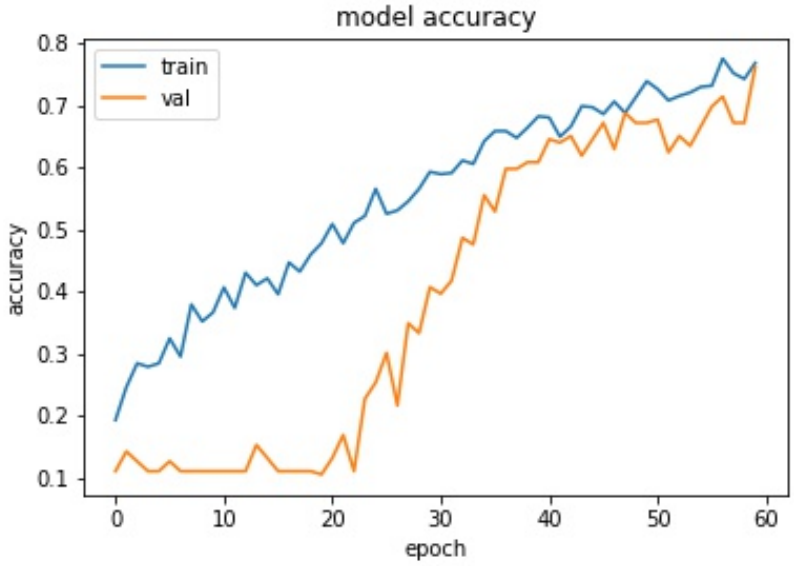


Figure 5.15: **Accuracy plot of English dataset with Video Model**

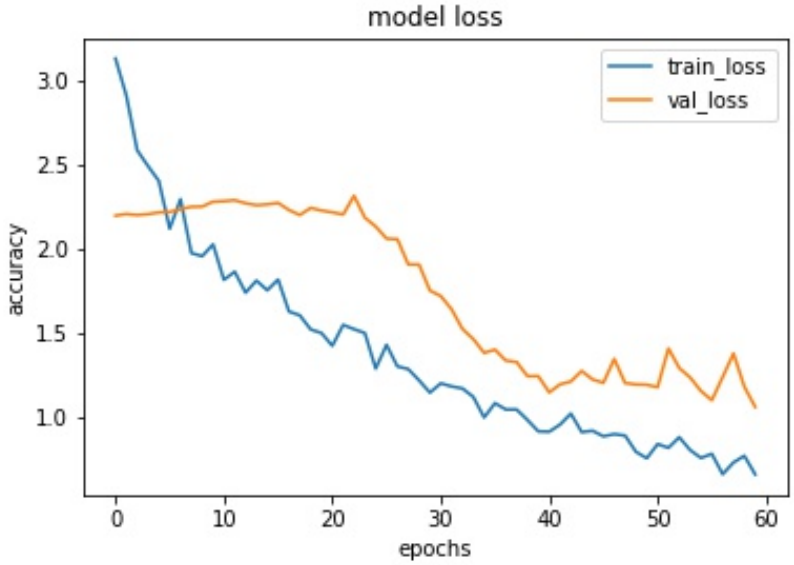


Figure 5.16: **Loss plot of English dataset with Video Model**

The Confusion Matrix of the Video models which are used to train Kannada dataset and English dataset are shown in Figure 5.17 and Figure 5.18 respectively. From Figure 5.17 one can observe that percentage of "Avanu videos" in the test dataset classified as Avanu are 57.14%, percentage of "Bagge videos" in the test dataset classified as Bagge are 80.95%, percentage of "Bari videos" in the test dataset classified as Bari are 85.71%, percentage of "Howdu videos" in the test dataset classified as Howdu are 85.71%, percentage of "Illa videos" in the test dataset classified as Illa are 76.19%, percentage of "Janarige videos" in the test dataset classified as Janarige are 71.42%, percentage of "Kathe videos" in the test dataset classified as Kathe are 61.90%, percentage of "Nale videos" in the test dataset classified as Nale are 80.95%

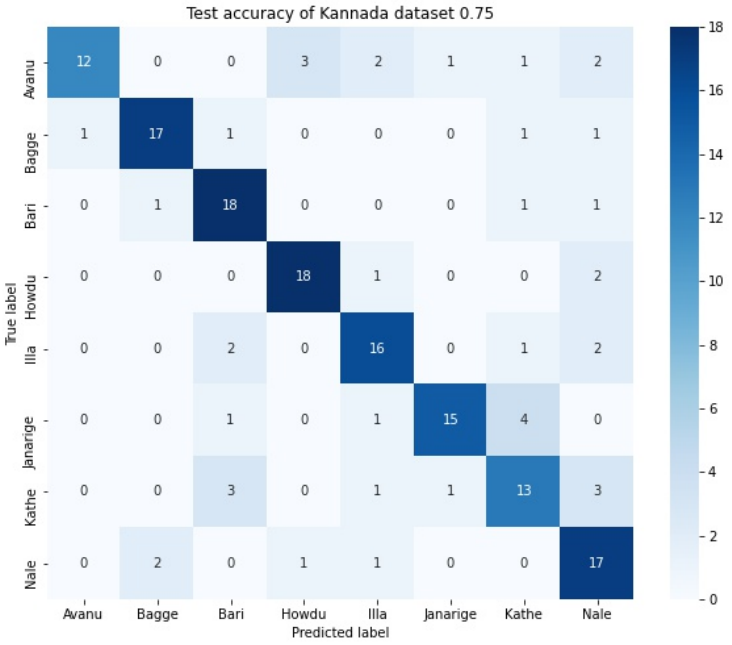


Figure 5.17: **Confusion Matrix of Kannada dataset with Video Model**

From Figure 5.18 one can observe that percentage of "About videos" in the test dataset classified as About are 90.47%, percentage of "Bad videos" in the test dataset classified as Bad are 71.42%, percentage of "Bottle videos" in the test dataset classified as Bottle are 76.19%, percentage of "Come videos" in the test dataset classified as Come are 66.66%, percentage of "Cow videos" in the test dataset classified as Cow are 80.95%, percentage of "Good videos" in the test dataset classified as Good are 71.42%,, percentage of "Pencil videos" in the test dataset classified as Pencil are 80.95%, percentage of "Read videos" in the test dataset classified as Read are 85.71%, percentage of "Where videos" in the test dataset classified as Where are 61.90%

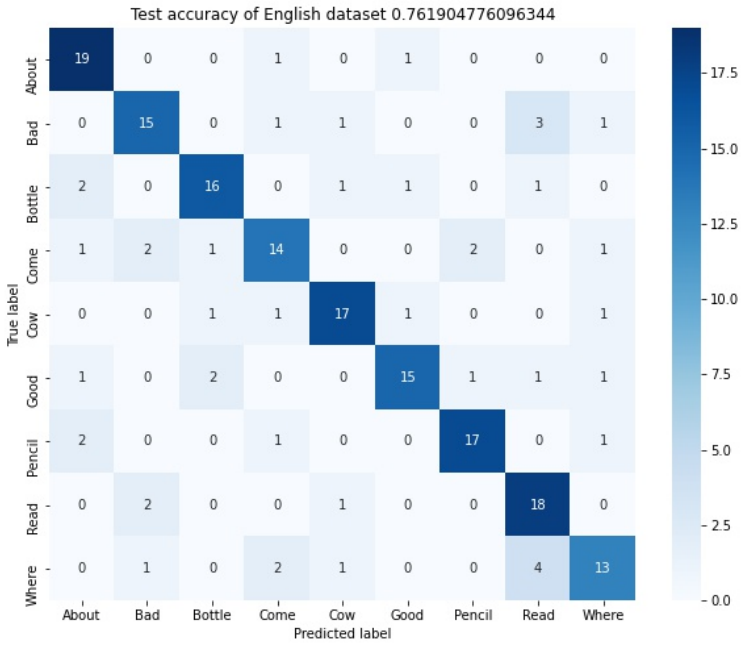


Figure 5.18: **Confusion Matrix of English dataset with Video Model**

The classification report of kannada dataset and english dataset are as shown in Figure 5.19 and Figure 5.20 respectively.

	precision	recall	f1-score	support
Avanu	0.92	0.57	0.71	21
Bagge	0.85	0.81	0.83	21
Bari	0.72	0.86	0.78	21
Howdu	0.82	0.86	0.84	21
Illa	0.73	0.76	0.74	21
Janarige	0.88	0.71	0.79	21
Kathe	0.62	0.62	0.62	21
Nale	0.61	0.81	0.69	21
accuracy			0.75	168
macro avg	0.77	0.75	0.75	168
weighted avg	0.77	0.75	0.75	168

Figure 5.19: **Classification Report of Kannada dataset with Video Model**

	precision	recall	f1-score	support
About	0.76	0.90	0.83	21
Bad	0.75	0.71	0.73	21
Bottle	0.80	0.76	0.78	21
Come	0.70	0.67	0.68	21
Cow	0.81	0.81	0.81	21
Good	0.83	0.71	0.77	21
Pencil	0.85	0.81	0.83	21
Read	0.67	0.86	0.75	21
Where	0.72	0.62	0.67	21
accuracy			0.76	189
macro avg	0.77	0.76	0.76	189
weighted avg	0.77	0.76	0.76	189

Figure 5.20: **Classification Report of English dataset with Video Model**

The propsed method accuracy is compared with existing methods and the comparison result is shown in Table 5.1

Table 5.1: Results of proposed method is compared with existing method for visual speech Recognition

Method	DWT\LDA[34]	PCA\LDA [35]	LSTM(Kannada)	LSTM (English)
Accuracy	68.04%	75%	75%	76.19%
Dataset	Custom AAVC	RML	Custom	Custom

Fusion Model

- Kannada dataset is trained for 100 epochs.
- Even English dataset is trained of 100 epochs.
- Achieved train accuracy of 93.33% and test accuracy of 92.26% for Kannada dataset as shown in Figure 5.21.

```
accuracy: 0.8988
Epoch 95/100
16/16 [==============================] - 1s 61ms/step - loss: 0.3318 - accuracy: 0.9008 - val_loss: 0.3727 - val_
accuracy: 0.8810
Epoch 96/100
16/16 [==============================] - 1s 62ms/step - loss: 0.2652 - accuracy: 0.8956 - val_loss: 0.3521 - val_
accuracy: 0.9226
Epoch 97/100
16/16 [==============================] - 1s 61ms/step - loss: 0.2829 - accuracy: 0.9125 - val_loss: 0.5693 - val_
accuracy: 0.8571
Epoch 98/100
16/16 [==============================] - 1s 62ms/step - loss: 0.2261 - accuracy: 0.9245 - val_loss: 0.3755 - val_
accuracy: 0.8988
Epoch 99/100
16/16 [==============================] - 1s 67ms/step - loss: 0.2684 - accuracy: 0.9068 - val_loss: 0.3492 - val_
accuracy: 0.9107
Epoch 100/100
16/16 [==============================] - 1s 64ms/step - loss: 0.2372 - accuracy: 0.9333 - val_loss: 0.2952 - val_
accuracy: 0.9226
```

Figure 5.21: **Training Kannada dataset with Fusion Model**

- Achieved train accuracy of 94.67% and test accuracy of 91.75% for English dataset as shown in Figure 5.22.

```
accuracy: 0.8981
Epoch 95/100
17/17 [==============================] - 1s 67ms/step - loss: 0.1211 - accuracy: 0.9520 - val_loss: 0.4226 - val_
accuracy: 0.9029
Epoch 96/100
17/17 [==============================] - 1s 88ms/step - loss: 0.1787 - accuracy: 0.9386 - val_loss: 0.3460 - val_
accuracy: 0.9126
Epoch 97/100
17/17 [==============================] - 2s 95ms/step - loss: 0.1553 - accuracy: 0.9438 - val_loss: 0.3529 - val_
accuracy: 0.9029
Epoch 98/100
17/17 [==============================] - 2s 92ms/step - loss: 0.1661 - accuracy: 0.9470 - val_loss: 0.3456 - val_
accuracy: 0.9126
Epoch 99/100
17/17 [==============================] - 2s 91ms/step - loss: 0.1960 - accuracy: 0.9422 - val_loss: 0.3751 - val_
accuracy: 0.9126
Epoch 100/100
17/17 [==============================] - 2s 93ms/step - loss: 0.2047 - accuracy: 0.9467 - val_loss: 0.3180 - val_
accuracy: 0.9175
```

Figure 5.22: **Training English dataset with Fusion Model**

- Model accuracy and Model loss plots of the Kannada dataset are as shown in Figure 5.23 and Figure 5.24 respectively.

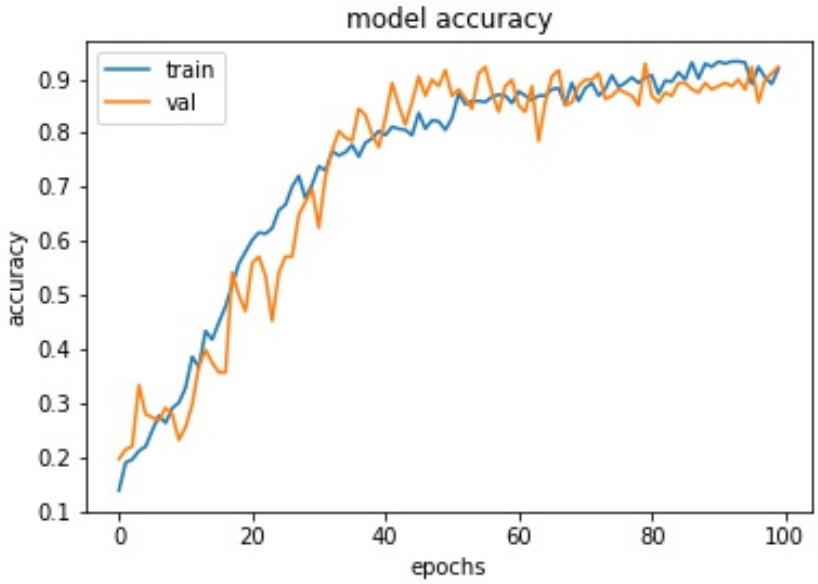


Figure 5.23: **Accuracy plot of Kannada dataset with Fusion Model**

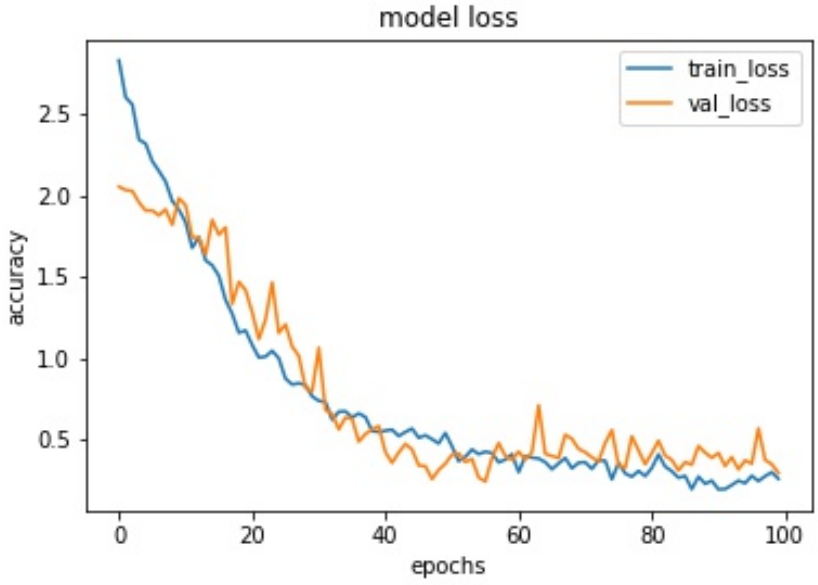


Figure 5.24: **Loss plot of Kannada dataset with Fusion Model**

- Model accuracy and Model loss plots of the English dataset are as shown in Figure 5.25 and Figure 5.26 respectively.

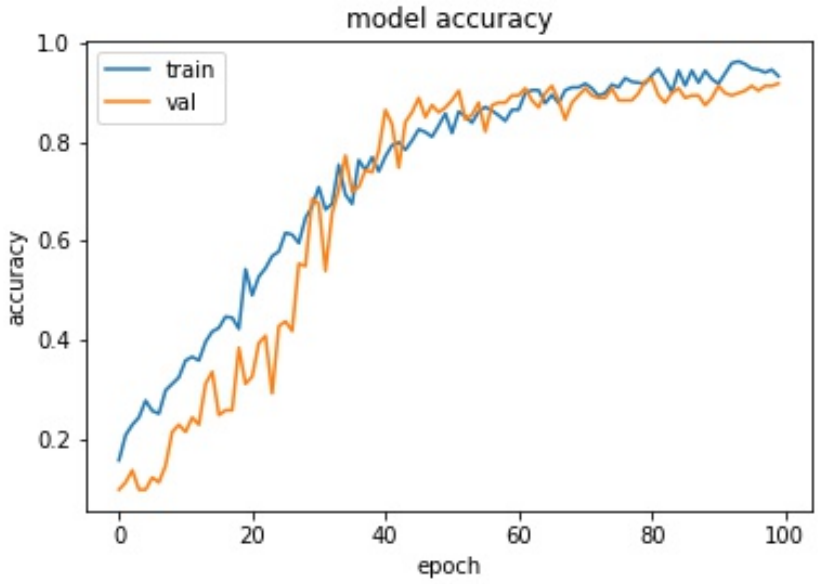


Figure 5.25: **Accuracy plot of English dataset with Fusion Model**

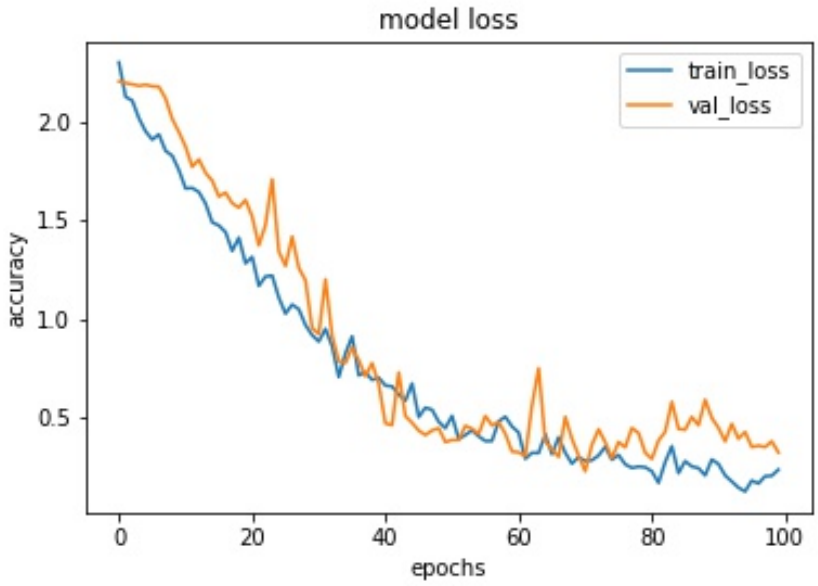


Figure 5.26: **Loss plot of English dataset with Fusion Model**

The Confusion Matrix of the Fusion models which are used to train Kannada dataset and English dataset are shown in Figure 5.27 and Figure 5.28 respectively. From Figure 5.27 one can observe that percentage of "Avanu videos" in the test dataset classified as Avanu are 95.23%, percentage of "Bagge videos" in the test dataset classified as Bagge are 76.19%, percentage of "Bari videos" in the test dataset classified as Bari are 80.95%, percentage of "Howdu videos" in the test dataset classified as Howdu are 100%, percentage of "Illa videos" in the test dataset classified as Illa are 95.23%, percentage of "Janarige videos" in the test dataset classified as Janarige are 100%, percentage of "Kathe videos" in the test dataset classified as Kathe are 100%, percentage of "Nale videos" in the test dataset classified as Nale are 90.47%

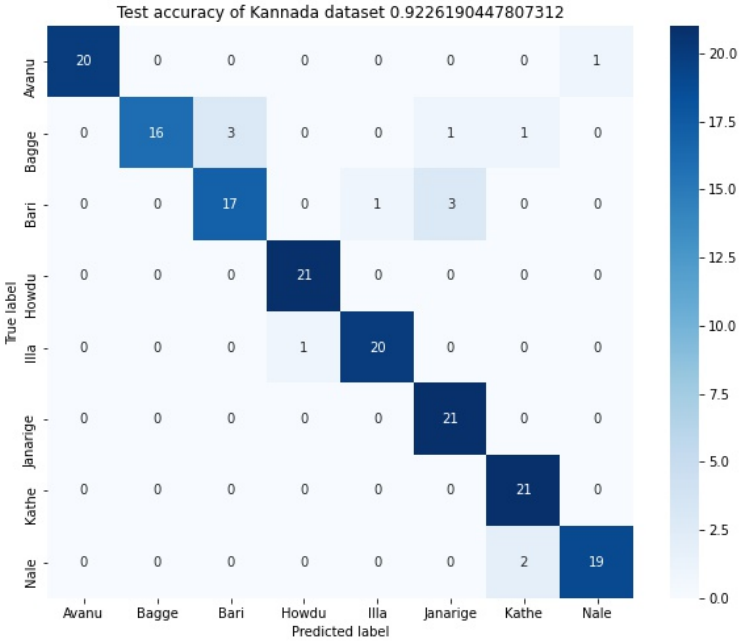


Figure 5.27: **Confusion Matrix of Kannada dataset with Fusion Model**

From Figure 5.28 one can observe that percentage of "About videos" in the test dataset classified as About are 90.47%, percentage of "Bad videos" in the test dataset classified as Bad are 100%, percentage of "Bottle videos" in the test dataset classified as Bottle are 85.71%, percentage of "Come videos" in the test dataset classified as Come are 95.23%, percentage of "Cow videos" in the test dataset classified as Cow are 90.47%, percentage of "Good videos" in the test dataset classified as Good are 100%,, percentage of "Pencil videos" in the test dataset classified as Pencil are 90.47%, percentage of "Read videos" in the test dataset classified as Read are 95.23%, percentage of "Where videos" in the test dataset classified as Where are 90.27%

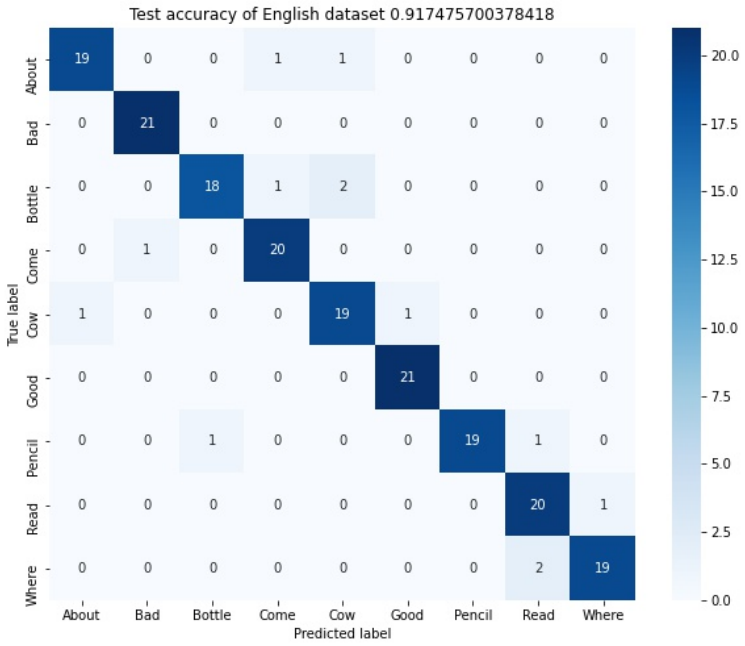


Figure 5.28: **Confusion Matrix of English dataset with Fusion Model**

The classification report of kannada dataset and english dataset are as shown in Figure 5.29 and Figure 5.30 respectively.

	precision	recall	f1-score	support
Avanu	1.00	0.95	0.98	21
Bagge	1.00	0.76	0.86	21
Bari	0.85	0.81	0.83	21
Howdu	0.95	1.00	0.98	21
Illa	0.95	0.95	0.95	21
Janarige	0.84	1.00	0.91	21
Kathe	0.88	1.00	0.93	21
Nale	0.95	0.90	0.93	21
accuracy			0.92	168
macro avg	0.93	0.92	0.92	168
weighted avg	0.93	0.92	0.92	168

Figure 5.29: **Clssification Report of Kannada dataset with Fusion Model**

	precision	recall	f1-score	support
About	0.86	0.90	0.88	20
Bad	0.91	1.00	0.95	20
Bottle	0.95	0.88	0.91	24
Come	0.92	0.92	0.92	24
Cow	0.88	0.85	0.86	26
Good	0.96	1.00	0.98	23
Pencil	1.00	0.90	0.95	21
Read	0.88	0.92	0.90	25
Where	0.91	0.91	0.91	23
accuracy			0.92	206
macro avg	0.92	0.92	0.92	206
weighted avg	0.92	0.92	0.92	206

Figure 5.30: **Classification Report of English dataset with Fusion Model**

The propsed method accuracy is compared with existing methods and the comparison result is shown in Table 5.2

Table 5.2: Results of proposed method is compared with existing method for audio visual speech Recognition

Method	Decision Fusion [34]	PCA\LDA [35]	LSTM(Kannada)	LSTM (English)
Accuracy	76.79%	82.5%	92.26%	91.74%
Dataset	Custom AAVC	RML	Custom	Custom

Chapter 6

Conclusion

An audio visual speech recognition model to interpret both audio as well as visual data is tried to be built. All the objectives that were formulated were approached systematically and completed to the fullest. Custom data-set for Kannada and English language words is created. Audio and video recognition for English and Kannada language words are performed separately and the corresponding performances are evaluated and compared with the performance of previous methodologies and implementations. The proposed architecture includes 1D CNN model for audio and LSTM model for visual and feed forward network for integration. We used the custom dataset and achieved train accuracy of 93.33% and test accuracy of 92.26% for Kannada dataset and train accuracy of 94.67% and test accuracy of 91.75% for English dataset. In conclusion, it can be seen that the proposed methodology does outperform other existing methodologies and hence a hybrid model is always the best possible way to achieve high performance measures in Audi-Visual Speech recognition.

6.1 Advantages and Limitations

The following are some of the advantages of the AVSR model:

- Every module, library used in the AVSR model, in helping to recognize the word spoken by the speaker, is open source. It is available to everyone.
- The AVSR model which is designed in this paper works for native English language and also for native Kannada Language.
- The Integration of video model along with the audio model helps in recognizing the word a lot better.

The following are some of the Limitations of the AVSR model:

- The AVSR model proposed can only recognize a single word.
- This model can not recognize sentences.

- This is not a end-to-end model.
- The camera angle used while creating the data-set was straight to the face of the speaker.

6.2 Future Work

There is huge amount of research part in order to improve this model. Some of the future work that can be implemented to improve the proposed model are

- To make the video in different angles other than straight to the face to the speaker.
- To make it able to recognize sentences.
- To make it a real-time model.

References

[1] Minsu Kwon, Ho-Jin Choi, “Automatic Speech Recognition Dataset Augmentation with Pre Trained Model and Script ”, 2019 IEEE International Conference on Big Data and Smart Computing (BigComp)

[2] Xutai Ma, Yongqiang Wang, Mohammad Javad Dousti, Philipp Koehn, Juan Pino “Streaming Simultaneous Speech Translation with Augmented Memory Transformer” 30 Oct 2020

[3] Jon Macoskey Grant P. Strimel Ariya Rastrow “Bifocal Neural ASR: Exploiting Keyword Spotting for Inference Optimization” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[4] Abdennour, Kemouche, Nabil. Aouf, “Automatic speech recognition enhancement using adaptive GMM estimation” IEEE Xplore: 12 November 2018

[5] Swapna Agarwa India Dipanjan Das Brojeshwar Bhowmick “Realistic Lip Animation from Speech for Unseen Subjects using Few-shot Cross-modal Learning” 2020 28th European Signal Processing Conference (EUSIPCO)

[6] Lv Ping “Mobile Platform Speech Intelligent Recognition and Translation APP” 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)

[7] Yosuke Higuch, Shinji Watanabe; Tetsuji Ogawa; Tetsunori Kobayash.” Improved Mask-CTC for Non-Autoregressive End-to-End” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[8] Themos Stafylakis, Georgios Tzimiropoulos, ”Combining Residual Networks with LSTMs for Lipreading” 2017, March(IEEE)9

[9] Fei Tao, Carlos Busso,”Aligning Audiovisual Features for Audiovisual Speech Recognition”, 2018 IEEE International Conference on Multimedia and Expo (ICME)

[10] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, Richard Bowden ”Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos”,IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 42, Issue: 9, Sept. 1 2020)

[11] Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi,"Lip reading with hahn convolutional neural networks",Image Vis. Comput., vol. 88, pp. 76–83, Aug. 2019 IEEE.

[12] cSantos, T. I., &; Abel, A. (2019, March). Using Feature Visualisation for Explaining Deep Learning Models in Visual Speech. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 231-235). IEEE.

[13] Jang, D. W., Kim, H. I., Je, C., Park, R. H., &; Park, H. M. (2019). Lip Reading Using Committee Networks With Two Different Types of Concatenated Frame Images. IEEE Access, 7, 90125-90131.

[14] Li, X., Neil, D., Delbruck, T., &; Liu, S. C. (2019, May). Lip Reading Deep Network Exploiting Multi-Modal Spiking Visual and Auditory Sensors. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE

[15] Algadhy, R., Gotoh, Y., & Maddock, S. (2019, May). 3D Visual Speech Animation Using 2D Videos. In ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2367-2371). IEEE.

[16] NadeemHashmi, Gupta, Mittal, Kumar, Nanda, Gupta (2018, August). A Lip Reading Model Using CNN with Batch Normalization. In 2018 Eleventh International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

[17] Wei, J., Yang, F., Zhang, J., Yu, R., Yu, M., &; Wang, J. (2018, November). Three Dimensional Joint Geometric- Physiologic Feature for Lip-Reading. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1007-1012). IEEE

[18] T. Afouras, J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Deep Audio-visual Speech Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2018.2889052.

[19] Petridis, Stavros, et al. "Audio-Visual Speech Recognition With a Hybrid CTC/Attention Architecture." 2018 IEEE Spoken Language Technology Workshop (SLT), Spoken Language Technology Workshop (SLT), 2018 IEEE 2018: 513.

[20] 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, doi: 10.1109/INCIT.2019.8912049.10

[21]]S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos and M. Pantic, "End-to-End Audiovisual Speech Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 6548-6552, doi: 10.1109/ICASSP.2018.8461326.

[22] K. Tan, Y. Xu, S. Zhang, M. Yu and D. Yu, "Audio-Visual Speech Separation and Dereverberation With a Two-Stage Multimodal Network," in IEEE

Journal of Selected Topics in Signal Processing, vol. 14, no. 3, pp. 542-553, March 2020, doi: 10.1109/JSTSP.2020.2987209.

[23] Jadczyk, Tomasz. "Audio-visual speech processing system for Polish applicable to human computer interaction." Computer Science [Online], 19.1 (2018): 41. Web. 28 Dec. 2020.

[24] H. Meutzner, N. Ma, R. Nickel, C. Schymura and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5320-5324, doi: 10.1109/ICASSP.2017.7953172.

[25] Debnath, S., Roy, P. Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition. SIViP (2020). https://doi.org/10.1007/s11760-020- 01717-0.

[26] Martinez, Brais & Ma, Pingchuan & Petridis, Stavros & Pantic, Maja. (2020). Lipreading using Temporal Convolutional Network.

[27] M. Hao, M. Mamut, N. Yadikar, A. Aysa and K. Ubul, "A Survey of Research on Lipreading Technology," in IEEE Access, vol. 8, pp. 204518-204544, 2020, doi: 10.1109/ACCESS.2020.3036865.

[28] F. Tao and C. Busso, "End-to-End Audiovisual Speech Recognition System With Multitask Learning," in IEEE Transactions on Multimedia, vol. 23, pp. 1-11, 2021, doi: 10.1109/TMM.2020.2975922.

[29] W. Feng, N. Guan, Y. Li, X. Zhang and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 681-688, doi: 10.1109/IJCNN.2017.7965918.

[30] J. Yu et al., "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6984-6988, doi: 10.1109/ICASSP40776.2020.9054127.

[31] Y. Yuan, C. Tian and X. Lu, "Auxiliary Loss Multimodal GRU Model in Audio-Visual Speech Recognition," in IEEE Access, vol. 6, pp. 5573-5583, 2018, doi: 10.1109/ACCESS.2018.2796118.

[32] P. Zhou, W. Yang, W. Chen, Y. Wang and J. Jia, "Modality Attention for End-to-end Audio visual Speech Recognition," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6565-6569, doi: 10.1109/ICASSP.2019.8683733.

[33] J. Wu et al., "Time Domain Audio Visual Speech Separation," 2019 IEEEAutomatic Speech Recognition and Understanding Workshop (ASRU), SG,Singapore, 2019, pp. 667-673, doi: 10.1109/ASRU46091.2019.9003983.11.

[34] Mr. Befkadu Belete Frew, 2019, Audio-Visual Speech Recognition using LIP Movement for Amharic Language, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 08, Issue 08 (August 2019),

[35] J. Y. R. Cornejo and H. Pedrini, "Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2019, pp. 3396-3402, doi: 10.1109/SMC.2019.8914193.

[36] Tao, Fei Busso, Carlos. (2018). Audiovisual Speech Activity Detection with-Advanced Long Short-Term Memory. 1244-1248. 10.21437/Interspeech.2018-2490.[35] M. Wand, J. Schmidhuber and N. T. Vu, "Investigations on End-to-End Audiovisual Fusion," 2018 IEEE International Conference on Acoustics, Speechand Signal Processing (ICASSP), Calgary, AB, 2018, pp. 3041-3045, doi:10.1109/ICASSP.2018.8461900

[37] M. Wand, J. Schmidhuber and N. T. Vu, "Investigations on End- to-End Audiovisual Fusion," 2018 IEEE International Conference on Acoustics, Speechand Signal Processing (ICASSP), Calgary, AB, 2018, pp. 3041-3045, doi:10.1109/ICASSP.2018.8461900