SPEECH EMOTION RECOGNITION

A PROJECT REPORT

Submitted by

JOSIKA P

(2019202020)

submitted to the Faculty of

INFORMATION SCIENCE AND TECHNOLOGY

in partial fulfillment for the award of the degree

of

MASTER OF COMPUTER APPLICATIONS

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600 025

JUNE 2022

ANNA UNIVERSITY

CHENNAI - 600 025

BONA FIDE CERTIFICATE

Certified that this project report titled **"SPEECH EMOTION RECOGNITION"** is the bona fide work of **JOSIKA P (2019202020)** who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE:CHENNAI
DATE:

MS.R.L.JASMINE
TEACHING FELLOW,
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

Dr.S.SRIDHAR
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

Speech emotion recognition is a process of understanding between human emotions and acoustic features. It is a difficult process for machines to identify the emotions based on human voice as its varies according to pitch, speed, loudness. Deep learning method which is used in variety of research areas such as pattern recognition, signal processing, classification can be used for speech emotion recognition. Convolution neural network which is an artificial neural network used for image/object recognition shows remarkable results in recognition task hence it can be used, it uses different modules for speech emotion recognition and classifiers are used to differentiate emotions such as happy, sad, fear etc .

Four types of datasets namely RAVDESS, TESS, CREMA-D, SAVEE which contains audio files were used for training and testing the CNN model to predict seven emotions like happy, neutral, sad, disgust, surprise, angry, fear. The audio files are preprocessed followed by feature extraction which is done using mel frequency cepstral coefficient MFCC with LIBROSA package and model is done by 1D CNN and 2D CNN and finally testing with different live voices.

திட்டப்பணி சுருக்கம

பேச்சு உணர்ச்சி அங்கீகாரம் என்பது புரிந்து கொள்ளும் செயல்முறை மனித உணர்வுகள் மற்றும் ஒலி அம்சங்கள் இடையே உள்ள தொடர்பாகும். இயந்திரங்களுக்கு இது கடினமான செயலாகும். சுருதிக்கு ஏற்ப மாறுபடும் மனிதக் குரலின் அடிப்படையில் உள்ள உணர்வுகளை அடையாளம் காண வேகம் சத்தம் ஆகிய உணர்வுகள் தேவைப்படுகின்றன. பல்வேறு ஆராய்ச்சிகளில் பயன்படுத்தப்படும் ஆழமான கற்றல் முறை வடிவ அங்கீகாரம், சமிக்ஞை செயலாக்கம், வகைப்பாடு போன்ற பகுதிகளுக்குப் பயன்படுத்தலாம். கன்வல்யூஷன் நியூரல் நெட்வொர்க் என்பது ஒரு செயற்கை நெட்வொர்க் படம் மற்றும் பொருள் ஆகியவற்றை அங்கீகரீக்க பயன்படுத்தப்படும் நரம்பியல் நெட்வொர்க் ஆகும். இது பேச்சு வழக்கிற்கு வெவ்வேறு தொகுதிகளைப் பயன்படுத்துவது போன்ற உணர்ச்சிகளை வேறுபடுத்துவதற்கு உணர்ச்சி அங்கீகாரம் பயன்படுத்தப்படுகின்றன.

மகிழ்ச்சி, சோகம், பயம் போன்றவை. நான்கு வகையான தரவுத் தொகுப்புகள் அதாவது ஆடியோ கோப்புகளைக் கொண்ட சிஎன்என் பயிற்சி சோதனைக்கு பயன்படுத்தப்பட்டது மகிழ்ச்சி, நடுநிலை, சோகம், வெறுப்பு, ஆச்சரியம், கோபம், பயம் ஆகிய அம்சங்களை பிரித்தெடுப்பதன் மூலம் ஆடியோ கோப்புகள் முன்கூட்டியே செயலாக்கப்படுகின்றன. இது உடன் மெல் அதிர்வெண் செப்ஸ்ட்ரல் குணகம் ஐப் பயன்படுத்தி செயல்படுகிறது. பேக்கேஜ் மற்றும் மாடல் கன்வல்யூஷன் நியூரல் நெட்வொர்க் மூலம் செய்யப்படுகிறது மற்றும் இறுதியாக வெவ்வேறு நேரடி குரல்களால் சோதனை செய்யப்படுகிறது.

ACKNOWLEDGEMENT

First and foremost, I would like to express our deep sense of gratitude to our guide **MS.R.L.JASMINE**, Teaching Fellow, Department of Information Science and Technology, Anna University, for her excellent guidance, counsel, continuous support and patience. She has helped us to come up with this topic and guided us in the development of this project. She gave us the moral support to finish our mini project in a successful manner.

I express our gratitude to **Dr.S.Sridhar** Professor and Head of the Department, Department of Information Science and Technology, Anna University, for her kind support and for providing necessary facilities to carry out the work.We are thankful to the project committee members **Dr.S.Saswati Mukherjee**, Professor, **Dr.M.Vijalakshmi**, Associate Professor, **Dr.E.Uma, Assistant Professor**, **MS.P.S.Apirajitha**, Teaching Fellow, **MS.C.M.Sowmiya**, Teaching Fellow Department of Information Science and Technology, Anna University, Chennai, for their valuable guidance and technical support.

We also thank all the faculty and non-teaching staff members of the Information Science and Technology, CEG Campus, Anna University, Chennai for their valuable support throughout the course of our project work.

JOSIKA P

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AQ	Algorithm Quasi-optimal
ASR	Automatic Speech Recognition
AW	Adaptive Wavelet
CNN	Convolution Neural Network
CREMA-D	Crowd Sourced Emotional Multimodal Actors Dataset
ELM	Extreme Leaning Machines
GE	Graph Embedding
LW	Large Width
NLP	Natural Language Pre-processing
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SAVEE	Surrey Audio-Visual Expressed Emotion
SER	Speech Emotion Recognition
SL	Subspace Learning
SVM	Support Vector Machine
TESS	Toronto Emotional Speech Set
TLSL	Transfer Linear Subspace Learning

CHAPTER 1

INTRODUCTION

1.1 DEEPLEARNING TECHNIQUES

Deep learning in a single term can understand as human nervous system. Machine vision deep learning sets are made to learn over a collection of audio/image also known as training data, in order to rectify a problem. The various deep learning models trains a computer to visualize like a human. Deep learning models based on the inputs to the nodes can visualize. Hence network type is like that of a human nervous system, with every node performing under a larger network as a neuron. So, deep learning models are basically a part of artificial neural networks. Algorithms of deep learning learns in depth about the input audio/image as it passes over every neural network layer. Low-level characteristics like edges are detected by learning given to the initial layers, and successive layers collaborate characteristics from prior layers in a more philosophical representation. Images, sounds, censor data and other data are those digital forms patterns which deep learning recognizes. For prediction pre-training the data and constructing a training set and testing set . As prediction obtains an optimum node such that the predicted node provides the satisfactory output. Basis of the neurons are in different levels and created to predict at every level and the most-optimum predictions, and thereafter for the best-fit outcome use the data. It is treated as true machine intelligence. A Convolutional neural network (CNN) is a sort of feed-ahead artificial network in which the joining sequence among its nodes is motivated by presenting an animal visual-cortex. Single cortical neurons give response to the stimuli at a prohibited area of region known as the receptive areas. The receptive areas of various nodes semi-overlap so that they can match the visual area. The reply

of a single node for stimuli among its receptive area could be mathematically through the convolution operations. Convolutional network was motivated by natural procedures and are varieties of multi-layer perceptron formulated to use least quantity of pre-processing. They have broad use in image and video recognition, recommendation systems and NLP. The dimensions of the characteristics map (convolved features) is regulated by following parameters:

- Depth refers to the filter count used in the operation.
- Stride refers to the size of the filter if the size is 5*5 the the stride will be 5.
- Zero-padding is padding the input matrix with often convenient around the border in order to apply filter to 'input audio' matrix's bordering elements. Using zero padding size of the characteristics map can be governed.

The below Figure 1.1 shows the different types of layers in a CNN model. It consists of three layers namely input layer, hidden layer and output layer. The basic workflow of the three layers is given in the digram.

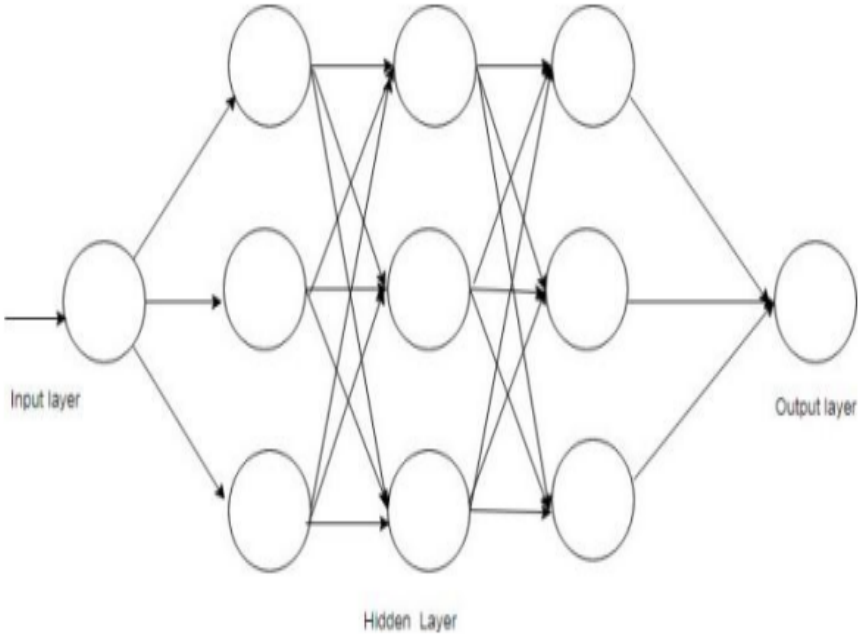

Figure 1.1: Outline of CNN with its Three Layers

1.2 PROBLEM STATEMENT

- Recognition of emotions in audio signals has been a field of study in the past. Previous work in this area included use of various classifiers like SVM, Neural networks, Bayes classifier etc.
- The number of emotions classified varied from study to study, they play an important aspect in evaluating the accuracy of the different classifiers.
- Using machine learning models to speech emotion recognition has less accuracy in order to overcome this issue have been using a model from deep learning called convolution neural network (CNN).

1.3 MOTIVATION

Speech emotion recognition (SER) is the task of recognizing the emotion from speech irrespective of the semantic contents. However, emotions are subjective and even for humans it is hard to notate them in natural speech communication regardless of the meaning. The ability to automatically conduct it is a very difficult task and still an ongoing subject of research. This project aims to help in building an emotion recognizer from speech data using a deep neural network.

1.4 OBJECTIVE

The objective is to identify emotions by using various sample audio files available. In order to achieve this, convolution neural network is used which has steps like feature learning and classification. The audio files will be converted into image files and feature extraction is done using mel frequency cepstral coefficients which is a technique to extract features from the audio files then training and testing of data will be done in order to use convolution neural network.

1.5 IMPLEMENTATION PLATFORM

- Development Platform: Jupyter Notebook
- Language: Python 3

1.6 DATASET DETAILS

1.6.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

1.6.2 Surrey Audio-Visual Expressed Emotion (SAVEE)

The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, KL), postgraduate students and researchers at the university of surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories.

1.6.3 Toronto Emotional Speech Set (TESS)

There are a set of 200 target words were spoken in the carrier phrase ”Say the word ” by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total. The dataset is organised such that each of the two female

actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

1.6.4 Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

1.7 SCOPE OF THE PROJECT

Speech emotion recognition has wide range of scope in medicine, security, entertainment, education. It helps in easy and cost effective AI method to recognize the emotions using speech. Emotional state of customers can be identified with help of speech emotion recognition. It also helps in knowing the emotions of patients and helps us in providing the timely medication.

1.8 ORGANIZATION OF THE REPORT

The thesis is organized into 6 chapters, describing each part of the project with detailed illustration and system design diagrams. The chapters are as follows:

Chapter 2: discusses the related works made on the proposed work, by analyzing issues of the existing system.

Chapter 3: discusses the functionalities of the proposed system and explains about each of the modules in detail.

Chapter 4: concentrates on the algorithm developed and used for the proposed system.

Chapter 5: illustrates the implementation results of the proposed work

Chapter 6: concludes the report by summarizing the results and proposes possible enhancements that can be extended as the future work.

The above mentioned six modules are followed up with the references which deliberately explains and list all the reference documents used during the various phases of the project, which includes the journal papers, conference papers, white papers, articles and websites referred for tutorials.

CHAPTER 2

LITERATURE SURVEY

This Chapter explains about the literature survey made on the existing system,analyzing the problem statements and issues with the observations and motivations.

2.1 SPECTRAL REGRESSION MODEL BASED ON GRAPH EMBEDDING

Here, Xinzhou Xu [1] generalized the spectral regression model joins of Extreme Leaning Machines (ELMs) and Subspace Learning (SL) was expect for overlooking the disadvantages of spectral regression based Graph Embedding (GE) and ELM. Demonstration over speech emotional corpora determined that the impact and feasibility of the techniques compared to prior methods that includes ELM and Subspace Learning (SL) techniques. The system output can be improved by exploring embedded graphs at more precise levels. Only least-square regression along with l2-norm minimization was considered in the regression stage.

2.2 HETEROGENEOUS TOKEN-USED SYSTEM TO DETECT THE SPEECH DEPRESSION

Zhaocheng Huang [2] uses a heterogeneous token-used system to detect the speech depression. Abrupt changes and acoustic areas are solely and collectively figured out in joins among different embedding methods. Detection of depression were used and probably various health problems that would affects vocal generation. Landmarks are used to pull out the information particular to

individual type of articulation at a time. This is a hybrid system. LWs and AWs hold various information. AW holds section of acoustic area into[] single token per frame, and on the contemporary the abrupt changes in speech articulation are shown by LWs. The hybrid join of the LWs and AWs permits exploitation of various details, more specifically, articulatory dysfunction into conventional acoustic characteristics are also incorporated.

2.3 TRANSFER LINEAR SUBSPACE LEARNING (TLSL) FRAMEWORK FOR CROSS CORPUS RECOGNITION OF SPEECH

Peng Song [3] offers transfer linear subspace learning (TLSL) for cross corpus recognition of speech. TLSL aims to extract robust characteristics representations over corpora into the trained estimated subspace. TLSL enhances the currently used transfer learning techniques which only focuses on searching the most portable components of characteristics TLSL can reach even better results compared to the 6 baseline techniques with stats significance, and gives better outcomes , in fact all the transfer learning is more accurate than usual learning techniques. A big set back that these early transfer learning methods possess was that they concentrate on searching the portable components of characteristics that tend to ignore less informative section. The less informative parts are also significant when it comes to transfer learning results experimented that TLSL is implemented for cross-corpus recognition of speech emotion.

2.4 AUTOMATIC ENCODERS OF SPEECH EMOTION RECOGNITION

With this paper Jun Deng [4] focused on unsupervised learning with automatic encoders of speech emotion recognition. Significantly work was on joining generative and discriminative training, by partially supervised learning

algorithms designed to settings where non-labeled data was available. The process had been sequentially evaluated with 5 databases in different settings. The proposed technique enhances recognition performance by learning the prior knowledge from non-labeled data in conditions with a smaller number of libeled examples. These techniques can solve the problems in mismatched settings and incorporate the learnings from different domains into the classifiers, eventually resulting in outstanding performance. This shows that the model is having the capacity to make good use of the combination of labeled and non-labeled data for speech emotion recognition. The residual neural network displayed that intense architectures make the classifier beneficial to pull out complicated structure in image processing.

2.5 CANTONESE-SPEAKING NARRATIVE SPEECH

Ying Qin [5] presented cantonese-speaking narrative speech which is a base of completely automated assessment system. Experiments on the text characteristics driven by the proposed data could detect out the impairment of language in the aphasic speech. The AQ scores were significantly correlated with the text characteristics learned by the siamese network. The improvised representation of ASR output was leveraged as the confusion network and the robustness of text characteristics were felicitated to it. There was an immediate requirement of improving the performance of ASR on aphasic speech for speech that has more robust characteristics. It was necessary that the databases of pathological speech and other languages to apply this proposed methodology. As seen clinically the most desirable one is automatic classification of aphasia variant along with this large-scale accumulation of data is needed substantially.

CHAPTER 3

SYSTEM DESIGN AND ARCHITECTURE

This chapter consists of system design of the project with its overall architecture diagram and module diagram and brief description about the modules in the project.

3.1 SYSTEM ARCHITECTURE

The first step is data preparation which creates a list based on the audio files names from different data sets. Based on the list created emotions will be counted in order to have a idea about the actual values. In explore dataset it converts the audio file which is in wav format to waveplot which is used for feature extraction and modeling. Feature extraction is done using MFCC which uses 13 filters and extracts the MFCC spectrogram feature and stores it in an array which can be pickled and used later. Baseline model uses 1D CNN to train and test the audio samples which is then used to model with filters. Data augmentation is done by adding speed, pitch and dynamic change to the original audio file datasets and modeling it again with 1D CNN to check the accuracy. 2D CNN uses melspectrograms as the input and does the modeling which uses the image classification technique and predict the emotions using audio with a better accuracy. The below Figure 3.1 gives a detailed explanation of speech emotion recognition with different modules and executable steps in each modules.

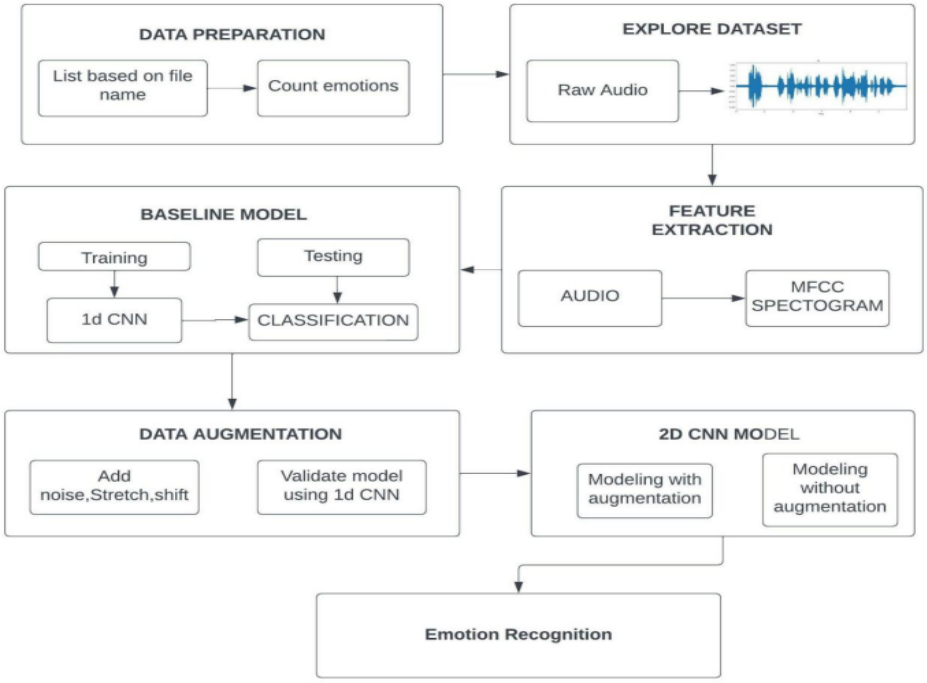

Figure 3.1: Detailed Architecture of the Project

3.2 WORKFLOW OF PROJECT

Figure 3.2 is the workflow of the project combining the steps in each modules and it shows the flow of speech emotion recognition. The audio dataset is input which undergoes data preparation, features are extracted and stored in a list to pickle for future use. Modeling is done using 1D CNN and 2D CNN with augmentation and without augmentation methods.

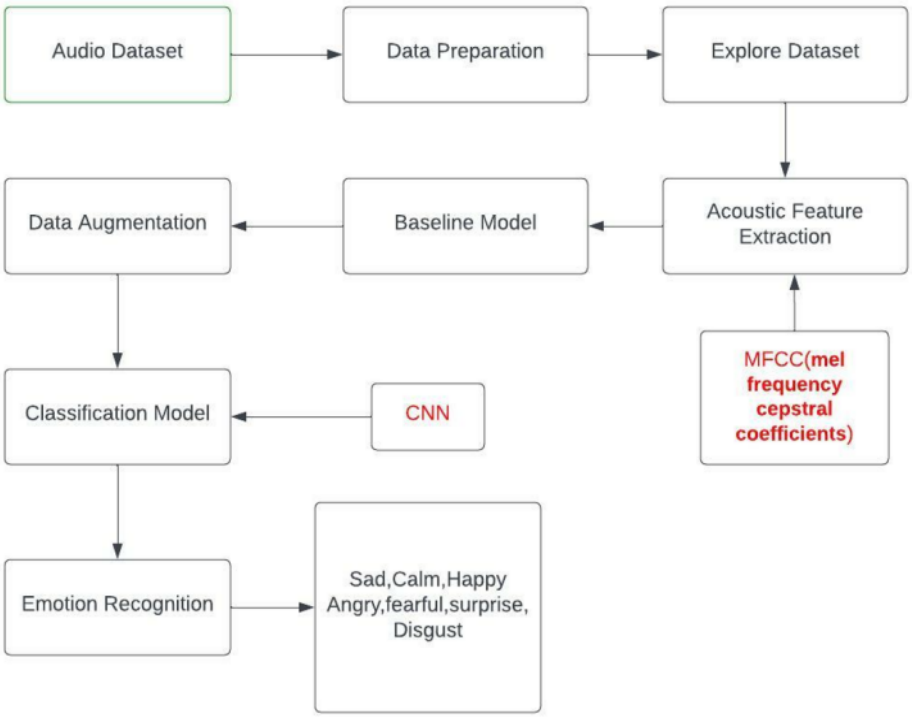

Figure 3.2: Workflow of the Project

3.3 MODULES OF THE PROJECT

The following are the modules in speech emotion recognition.

- Data Preparation
- Explore Data set
- Feature Extraction
- Baseline Model
- Data Augmentation
- 2D CNN Classification model

3.3.1 Data Preparation

Load the 4 different datasets from kaggle which contain audio files will be saving the datasets with a variable name and will group them into a list based on the file name with explains about the audio file then in order to proceed count the emotions based on gender.Figure 3.3 is the flow diagram of the module data preparation, it contains various steps takes place in this module.

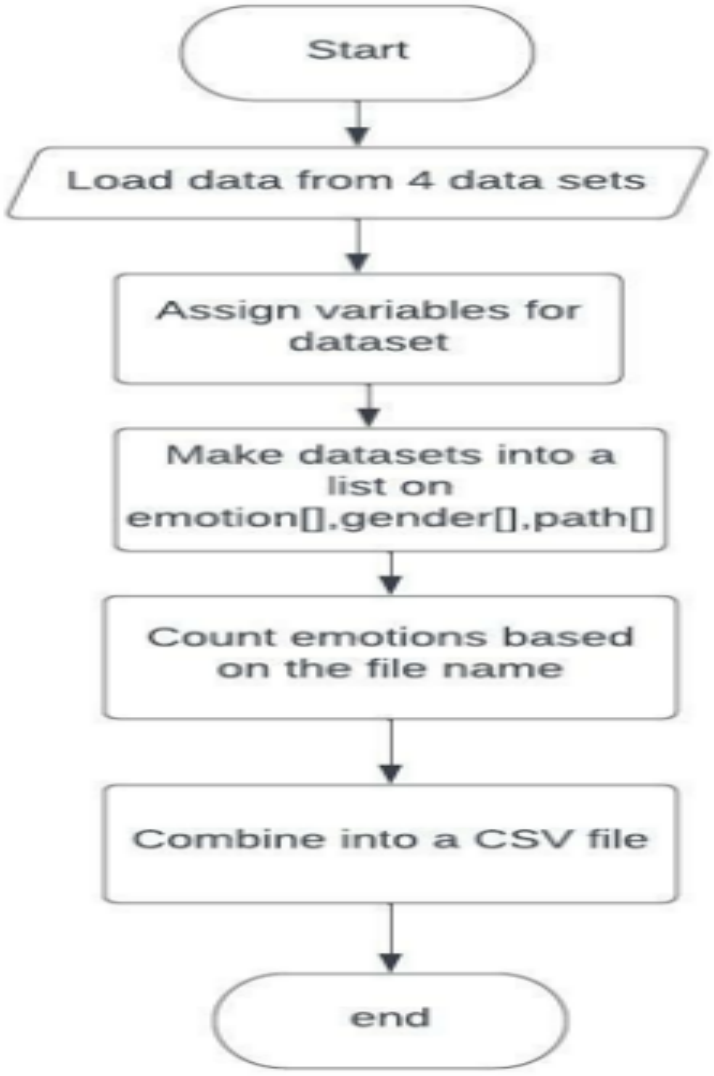

Figure 3.3: Flow Diagram of Data Preparation

3.3.2 Explore Data set

Here will explore the audio files and plot them based on time and amplitude to check the fluxations based on different emotions and different genders. Figure 3.4 is the flow diagram of the module explore data set, it contains various steps takes place in this module.

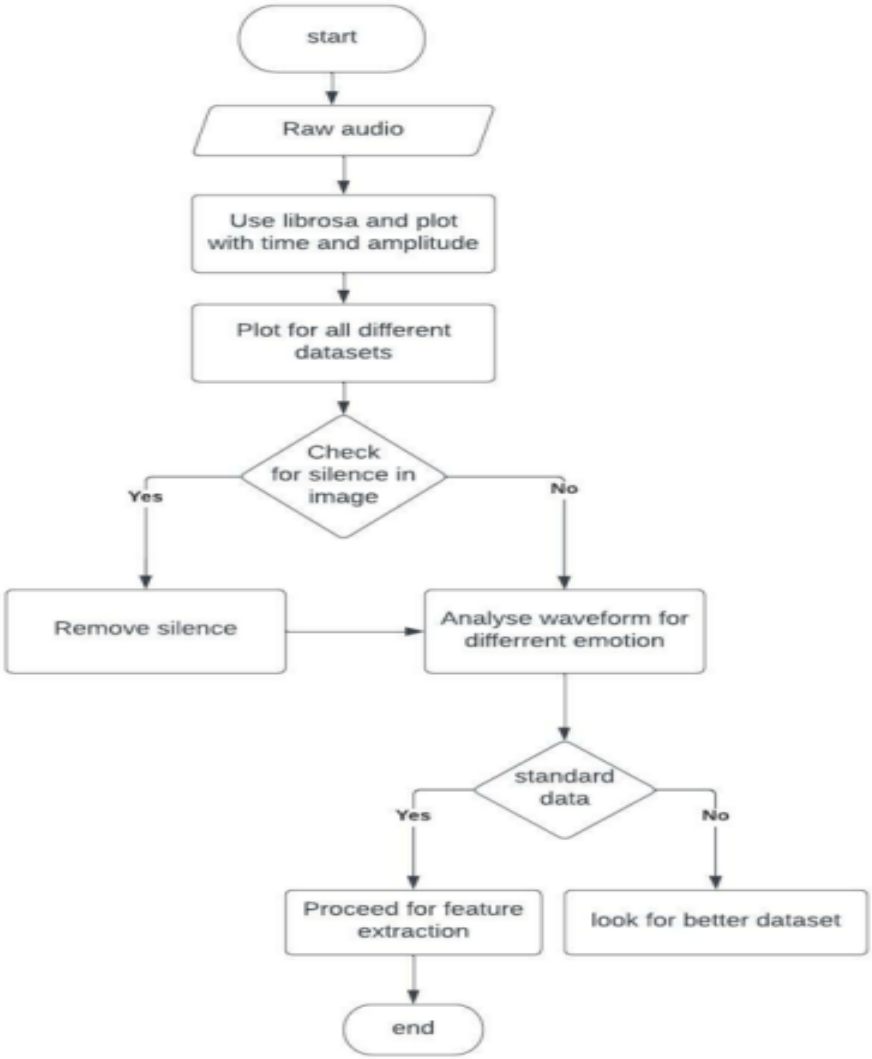

Figure 3.4: Flow Diagram of Explore Data set

3.3.3 Feature Extraction

In this feature extraction part going to use the fast fourier transform, melscale filter bank and plot the MFCC bands with respect to time and make it into a spectogram. A spectrogram is an image that displays the variation of energy at different frequencies across time. The vertical axis (ordinate) represents frequency and the horizontal axis (abscissa) represents time. The energy or intensity is encoded either by the level of darkness or by the colors. There are two general types of spectrograms: wideband spectrograms and narrow spectrograms. Wideband spectrograms has a higher time resolution than narrow spectrograms. This enables the wide-band spectrograms to show individual glottal pulses. In narrow-band have higher frequency resolution than wideband spectrograms. This feature enables the narrow-band spectrograms to resolve individual harmonics. Considering the importance of vocal fold vibration, along with the fact that glottal pulse is associated with one period of vocal fold vibration decided to convert all utterances into wide-band spectrograms. In doing so, the length of hamming windows were set to 5 ms with ms overlap. The number of DFT points was set to 512. Also, discarded the frequency information greater than 4 kHz from spectrograms since frequencies below 4000 Hz are sufficient for speech perception in many situations . In pilot studies, eliminating energy above 4000 Hz improved the performance of the algorithms. This gave 129 frequency points. All spectrogram images were, first, resized to have 129×129 pixels and, then, z-normalized to have zero mean and standard deviation close to one.

- Wide-band spectrogram with 5 ms Hamming window.
- Narrow- band spectrogram with 25 ms Hamming window.

Figure 3.5 is the wide and narrow spectogram. It shows how the spectrogram is plotted over the frequency and time.

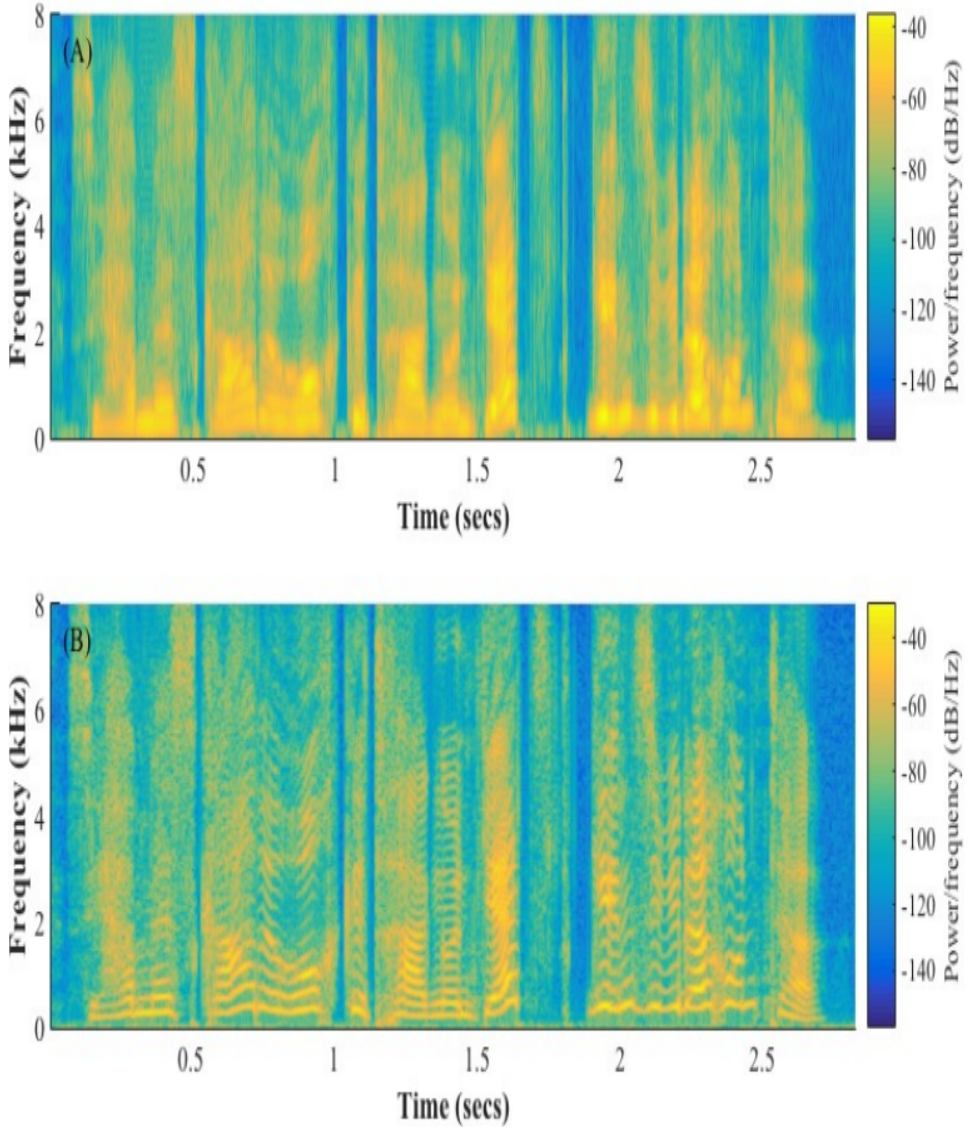

Figure 3.5: Wide Band and Narrow Band Spectrogram

Figure 3.6 is the flow diagram of MFCC algorithm which contains eight different steps to calculate the MFCC array values.

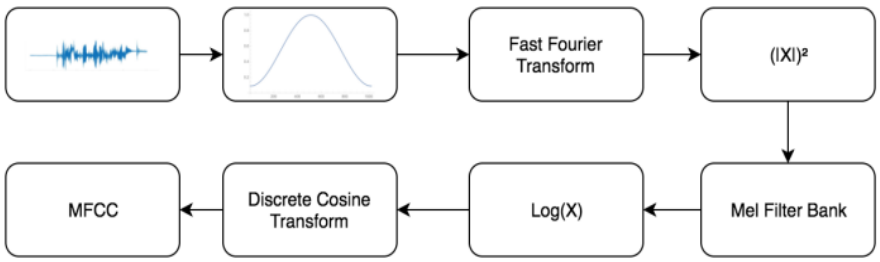

Figure 3.6: MFCC Algorithm Flow

Figure 3.7 is the flow diagram of feature extraction module, it contains various steps that are to be followed to get MFCC values for future modeling.

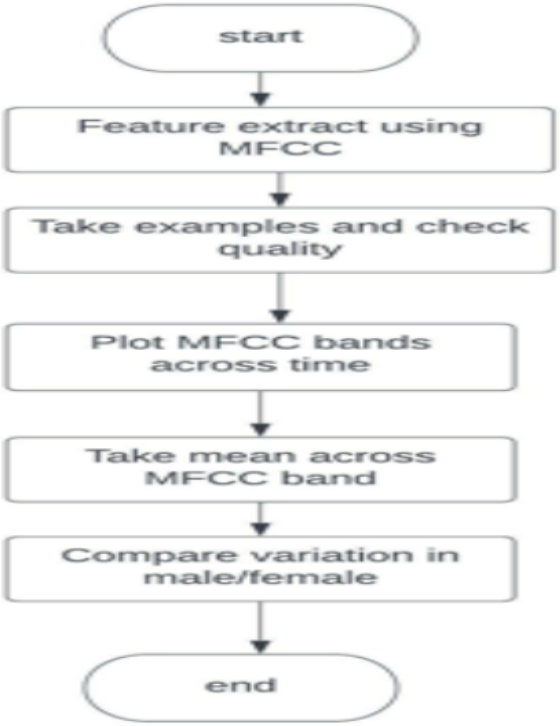

Figure 3.7: Flow Diagram of Feature Extraction

3.3.4 Baseline Model

Within this module train the model for accuracy estimations. First, import necessary modules. Then pull the data set. Will receive the sampling rate value with librosa packages and mfcc function. Thereafter this value holds other variables. Now audio files and mfcc value hold a variable consequently it will add a list. Then zip the list and hold two variables x & y. Then have represented (x, y) shape values with the use of numpy package.

Speech represented in the form of image with 3 layers. While using CNN, do consider, 1st and 2nd derivatives of speech image with time, frequency. CNN can predict, analyze the speech data, CNN can learn from speeches and identify words or utterances. Figure 3.8 is the flow diagram of Baseline model module following these steps we will get a model loss and confusion matrix which gives us information about prediction.

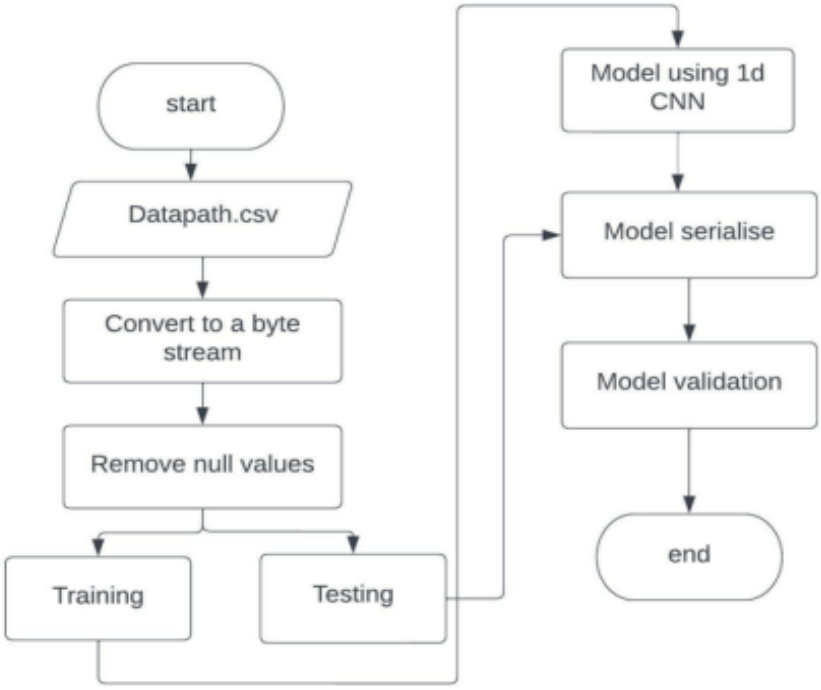

Figure 3.8: Flow Diagram of Baseline Model

3.3.5 Data Augmentation

- Static noise: This process involves addition of noise i.e. white noise to the sample. White noises are random samples distributed at regular intervals with mean of 0 and standard deviation of 1.

- Time Shifting: Shift the wave by sample rate/10 factor. This will move the wave to the right by given factor along time axis.

- Stretch: This one is one of the more dramatic augmentation methods. The method literally stretches the audio. So the duration is longer, but the audio wave gets strecthed too. Thus introducing and effect that sounds like a slow motion sound. Look at the audio wave itself,

notice that compared to the orginal audio, the strected audio seems to hit a higher frequency note. Thus creating a more diverse data for augmentation.

- Pitch: This method accentuates the high pitch notes, by normalising it sort of.

Figure 3.9 is the flow diagram of data augmentation module where we will add augmentation methods like speed, pitch and dynamic change.

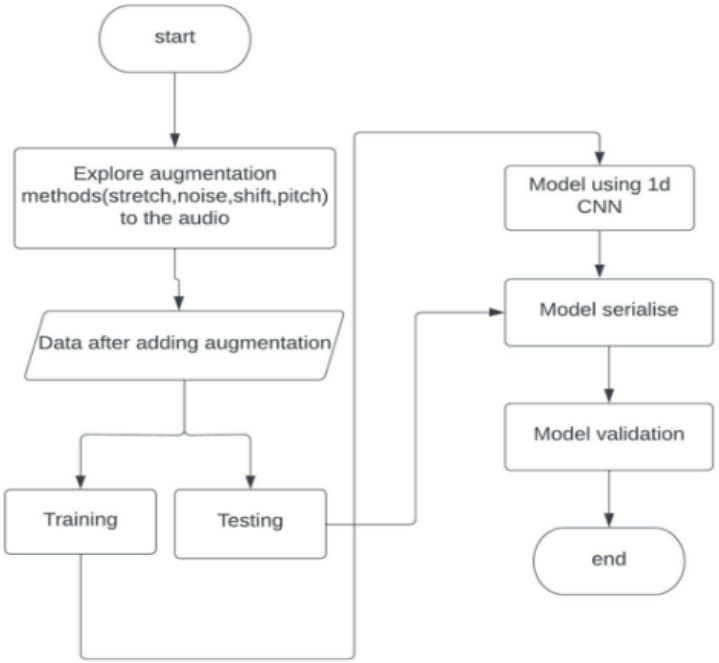

Figure 3.9: Flow Daigram of Data Augmentation

3.3.6 2D CNN Classification

Here will be adding the custom functions. 2D CNN is a array of 30 MFCC by 216 audio length as input data. Here will be finding various results using MFCC with augmentation and without augmentation and will also be using another feature extraction method to compare the accuracy rate with MFCC the method are going to use here is log-melspectrogram. Comparing all the inference with 1D CNN model and 2D CNN model with different feature extraction methods and the audio with augmentation and without augmentation to predict the correct emotions. Below explains the architecture of 2D CNN. Figure 3.10 is a digramatic representation of 2D CNN using MFCC feature vector method.

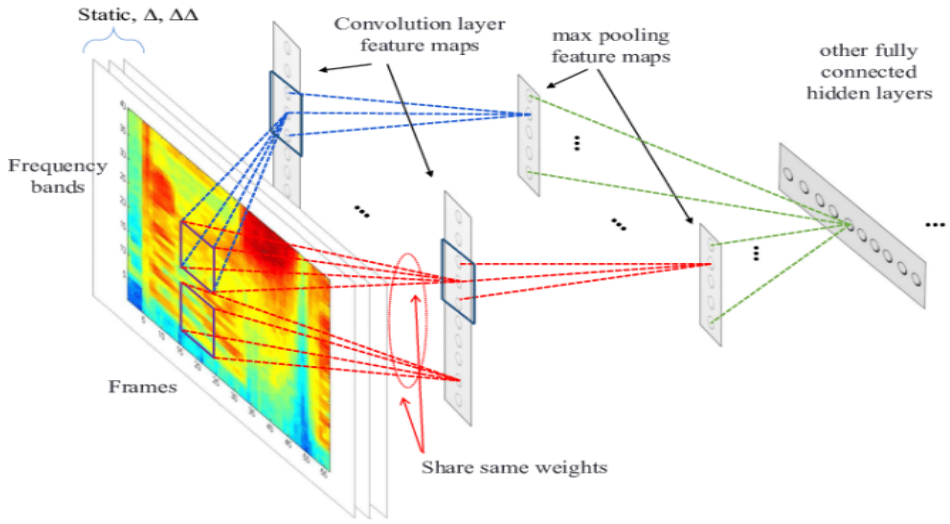

Figure 3.10: 2D CNN Using Feature Vector Method

Figure 3.11 is the flow diagram of 2D CNN module which uses MFCC feature vector method output as an input for modeling the training and testing data sets.

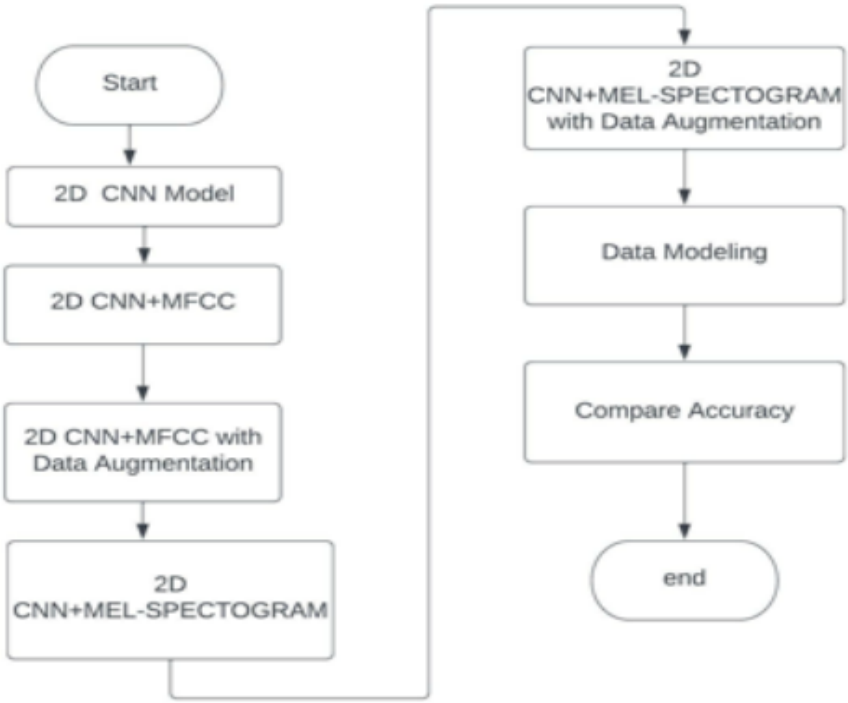

Figure 3.11: Flow Diagram of 2D CNN

CHAPTER 4

ALGORITHM IMPLEMENTATION

This section explains in detail the various modules in the system. Each module includes the input for the module, process flow for the module and output for the module in detail.

4.1 DATASET GATHERING

For speech emotion recognition data is gathered from RAVDESS, TESS, SAVEE, CREMA-D which consists of 12,162 files. These datasets were combined and will go through feature extraction process followed by modelling.

4.2 DATA PREPARATION

Here gathered dataset set will be processed. Variables will be assigned and list will be created based on emotion, gender and will be combined into a csv file.

Algorithm 4.2 Data Preparation

```
1: Input: Dataset containing audio files
2: Output: Count of emotions
3: Load dataset from kaggle
4: Assign variables for dataset
5: for each file of dataset do
6:     Save emotions as list
7:     Save path as list
8:     Save gender as list
9: end for
10: Count Emotion based on the list
11: Combine all into a csv file
```

4.3 EXPLORE DATA

Here librosa library is used to find various analysis to the raw audio in order to ensure that our raw audio file is valid for further analysis.

Algorithm 4.3 Explore Data

1: Input: Get raw audio files
2: Output: Analysis from raw audio
3: Read raw audio
4: Using librosa plot time amplitude graph
5: **for** each *audio* of *dataset* **do**
6: Plot time, amplitude graph
7: **end for**
8: Check for silence in the image
9: Analyse waveform for different emotion
10: Confirm for valid data
11: Proceed for feature extraction
12: Count Emotion based on the list
13: Combine all into a csv file

4.4 FEATURE EXTRACTION

Feature extraction is an important step to make machines learn through feature categorization. Here feature extraction has been done through MFCC.

Algorithm 4.4 Feature Extraction

1: Input: Audio data
2: Output: MFCC feature extracted array values
3: Read dataset
4: Use MFCC algorith for feature extraction
5: **for** each *data* of *savedcsv* **do**
6: extract features in array
7: **end for**
8: Find the statistical features
9: Compare the variations between male and female voice
10: Count Emotion based on the list
11: Combine all into a csv file

4.4.1 MFCC-Mel Frequency Cepstral Coefficeint

MFCC, short for Mel-Frequency Cepstral Coefficient. MFCC is a sentence, is an ”image” of the vocal tract that delivers the sound. The initial phase in any programmed is speech acknowledgment framework is to remove the valuable component that recognize the pieces of the sound sign that are useful for distinguishing the etymological substance and disposing of the various stuff which conveys data like foundation commotion, feeling and so on. Mel Frequency Cepstral Coefficients (MFCCs) are an element broadly utilized in programmed discourse and speaker acknowledgment.

Algorithm 4.4 Feature Extraction

1: Pre emphasis
2: Framing
3: Windowing
4: Fast Fourier Transform
5: Mel filter Bank

4.5 MODEL DETAILS

CNN in speech processing is a successful end product of deep learning methodology. Research prove that CNN has been useful in extracting raw signals in appliications like image recognition and speech recognition. Since speech is a input spectograms and MFCCs play a vital role as they are commonly represent speech features along with CNN for emotion detection.

Algorithm 4.5 CNN model

```
Input: Testing audio
Output: Predict speech emotion
Feed the pre-trained model
Given a input video frames on Train and Test
Split the audio into frames
for each image of load do
    Read frames store in a folder
    Load the image size(64,64)
    Img= currentframe/255
    Predict current audio using trained model.
    Val =Model.predict(audiova)
end for
The audio classified
```

4.5.1 1D CNN Model with MFCC

Mel Frequency Cepstrum (MFC), is a representation of the short-term power spectrum of sound. It is based on a linear cosine transform of a log power spectrum on a non-linear mel-scale of frequency. As MFCC is a popular speech feature widely used in various speech processing applications, it is used for speech detection. The hyper parameters and the python package (librosa) used for MFCC generation are similar to the ones described for spectrogram generation. The only difference is that 40 MFCCs per window are generated compared to the earlier mentioned 128 spectrogram coefficients per window. This model also consists of 4 sets of parallel convolutional layers, followed by maxpooling layers and two more FC layers, similar to the one described in the previous section. As the input size is different to that of model 2A and 2B experimented with kernels of different sizes.

4.5.2 2D CNN Model with Spectrograms

A spectrogram is a representation of speech over time and frequency. 2D convolution filters help capture 2D feature maps in any given input. Such

rich features cannot be extracted and applied when speech is converted to text and or phonemes. Spectrograms, which contain extra information not available in just text, gives further capabilities in attempts to improve emotion recognition. The following model uses mel-frequency Spectrogram as input to a 2D CNN. Spectrograms are generated when Short Term Fourier Transform (STFT) is applied on windowed audio or speech signal. The audio is sampled at 22050Hz. Windowing is then carried out on each audio frame using a "hann" window of length 2048. Fast Fourier Transform (FFT) windows of length 2048 are then applied on the said windowed audio samples with an STFT hop-length equal to 512. The obtained spectrogram magnitudes are then mapped to the mel-scale to get mel-spectrograms. 128 spectrogram coefficients per window are used in this model. The Mel-frequency scale puts emphasis on the lower end of the frequency spectrum over the higher ones, thus imitating the perceptual hearing capabilities of humans. Librosa python package, along with the above mentioned parameters, to compute the mel-spectrograms. A sample spectrogram corresponding to audio "I cannot... you are not here by choice. Nobody would ride this bus by choice." is shown below in Figure 4.1.

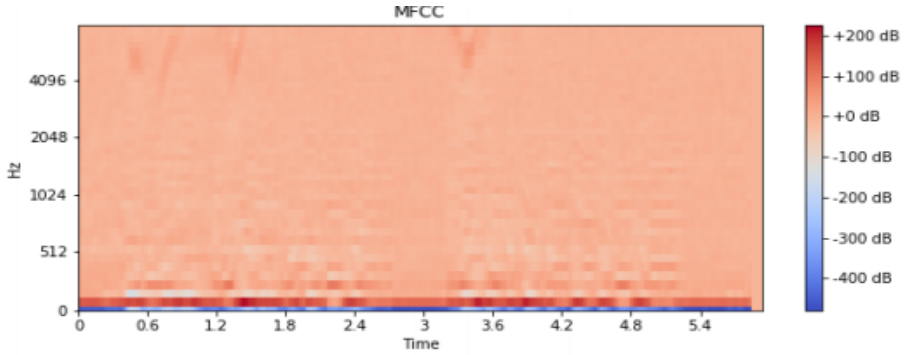

Figure 4.1: MFCC of a Audio Sample

In CNN model, take Spectrogram input with a maximum image width of 256 (number of windows). Since dataset has audios of varying lengths, trim long duration audio files to a fixed duration (4 seconds), which covers 75 percentile of all audio data samples of the dataset. This decision was made under

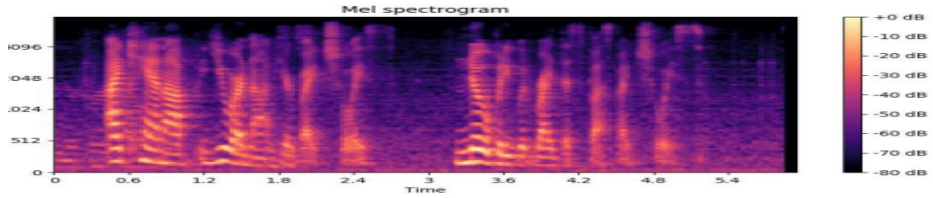

Figure 4.2: Spectrogram of a Audio Sample

the assumption that 6 the frequency variations that characterize the emotionality of the speech data will be present throughout the dialogue and hence will not be lost by this reduction in length. The above Figure 4.2 details the 2D CNN architecture used to detect emotion using spectrograms. A set of 4 parallel 2D convolutions are applied on the spectrogram to extract its features. The input shape of the spectrogram image is 128 x 256 (number of Mels x number of windows). 200 2D-kernels are used for each of the parallel convolution steps. Figuring out the optimal kernel size is a difficult and time taking task, which may depend on several factors all which cannot be clearly defined. To prevent choosing one single kernel size that could possibly be sub-optimal decided to use kernels of different sizes, each of which is fixed for a single parallel path, to take advantage of the different patterns picked up by each kernel. The sizes of each of the kernels in their respective parallel CNN paths are 12 x 16, 18 x 24, 24 x 32, and 30 x 40. The features generated in the said convolution layers are then fed to their respective max-pool layers, which extracts features from each filter as the pool size is exactly half along the width and height of the convolution output. The extracted features are fed to the Fully Connected (FC) layer. This model makes use of two FC layers of sizes 400 and 200. Batch normalization is applied to both the FC layers. The activation function used in the convolutional layers and the first FC layer is the Rectified Linear Unit (ReLU). The output of the last FC layer is then fed to a softmax layer, which classifies the input speech signal among different emotion classes.

CHAPTER 5

IMPLEMENTATION AND RESULTS

In experimental results consists of the details about the hardware and software requirements to the project and the experiments that have been performed along with their outcomes. The detailed result of the project is also portrayed in this chapter.

5.1 HARDWARE REQUIREMENTS

In this project, a computer with sufficient processing power is needed. This project requires too much processing power, due to the image and video

1. Operating System - Windows
2. RAM - Minimum 8 GB
3. Hard Disk - Minimum 100 GB
4. Graphic card - NVIDIA GeForce GTM Titan

5.2 SOFTWARE REQUIREMENTS

1. Operating System – Windows 8+
2. Programming Language – Python 3.8

3. PyTorch 1.11.0 21
4. Framework – Streamlit
5. Libraries – Librosa, keras

5.3 1D CNN RESULTS AND ANALYSIS

After 1D CNN is done with around 100 epochs for the training and test dataset we can see the model loss graph for training and testing datasets. Figure 5.1 is the loss that starts to plateau at around 50 epochs.

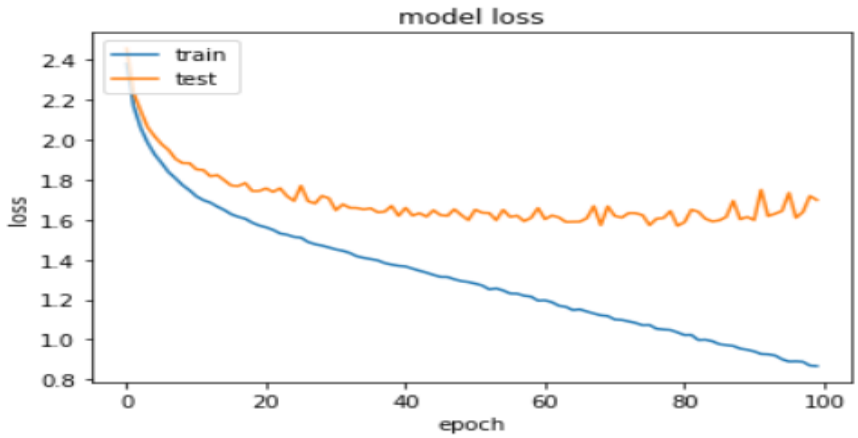

Figure 5.1: Model Loss for Training and Testing Data Set

The below Figure 5.2 is the Actual and predicted labels for our datsets with an accuracy of 42.45% using simple MFCC feature extraction method. Gender prediction is higher compared to the emotions. A random ten samples have been displayed to check the accuraccy.

Out[108]:

	actualvalues	predictedvalues
170	male_sad	female_disgust
171	female_neutral	female_neutral
172	male_angry	female_angry
173	female_disgust	female_disgust
174	male_angry	male_angry
175	female_fear	female_angry
176	male_neutral	male_neutral
177	female_fear	female_fear
178	female_happy	female_happy
179	female_neutral	female_sad

Figure 5.2: Actual vs Predicted Using 1D CNN

The confusion matrix is plotted for gender based emotions based on 1D CNN for training and testing datasets with an accuracy of 42.45 %. Further evaluation is done by checking gender based and emotion based in order to improve the model. Figure 5.3 is the confusion matrix plotted against true label and predicted label for gender based speech emotoins with 42.45 % accuracy.

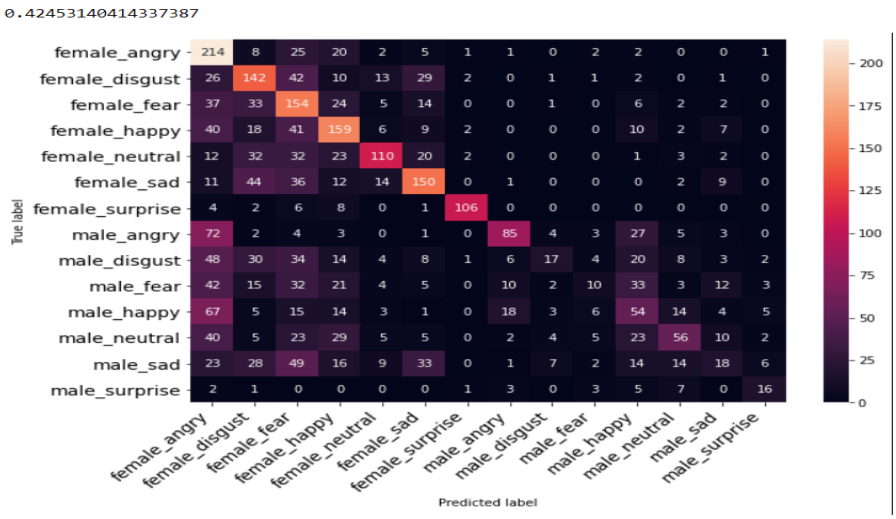

Figure 5.3: Confusion Matrix for Gender Based Emotions Using 1D CNN

With just gender we get a 74 % accuracy. The model is especially precise in capturing female voices. However, male voices tends to be harder and it does make higher mistakes thinking its female.Figure 5.4 gives the confusion matrix using 1D CNN based only on gender.

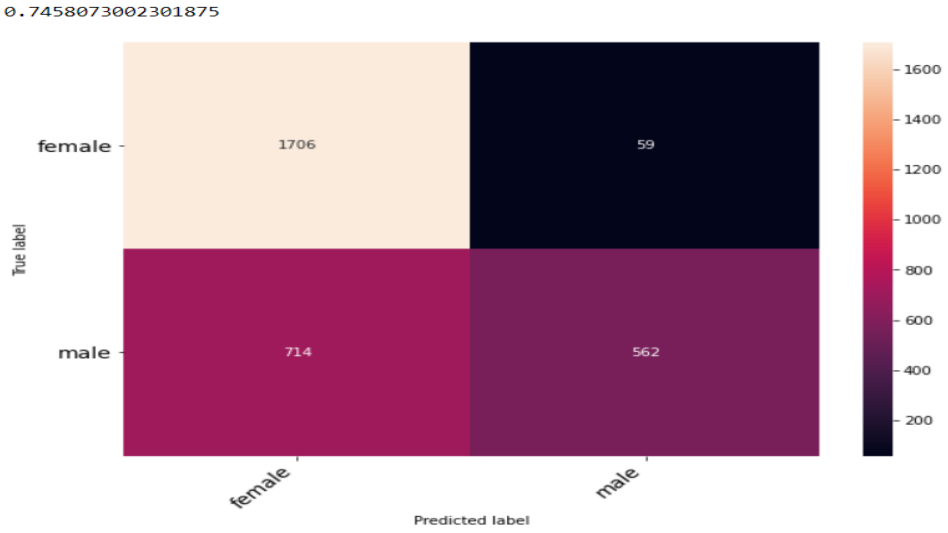

Figure 5.4: Confusion Matrix for Gender Using 1D CNN

Ignoring gender and considering only core emotions got an accuracy of around 49.39 % and the precision and recall score for anger and surprise is higher compared to other emotions.Thus next process improves this accuracy. Figure 5.5 gives the confusion matrix using 1D CNN based only on emotion.

5.4 1D CNN WITH AUGMENTATION

The number of epochs increased to 150 . In 1D CNN model without augmentation, set to 100 epochs and the log loss plot shows that it has reached full potential at a log loss of about 1.6 after about 50 epochs, and further epochs doesn't really make it more accurate. With data augmentation however, it seems to indicate that it hasn't quite plateau yet and could still get better.

Hence the number of epochs has been increased so it can achieve its

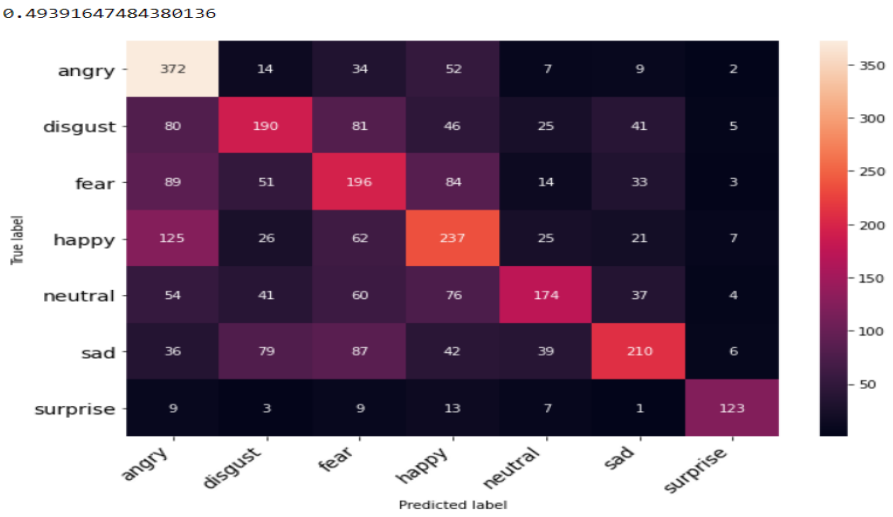

Figure 5.5: Confusion Matrix for Emotion Using 1D CNN

full potential. Originally when I set to 100 epochs, the plot indicated it hasn't plateau. Rerun at 150it shows that around 100 epochs diminishing returns sets in. So 150 epochs considered instead of 100 epochs. Figure 5.6 is the loss for 150 epochs after adding data augmentation methods .

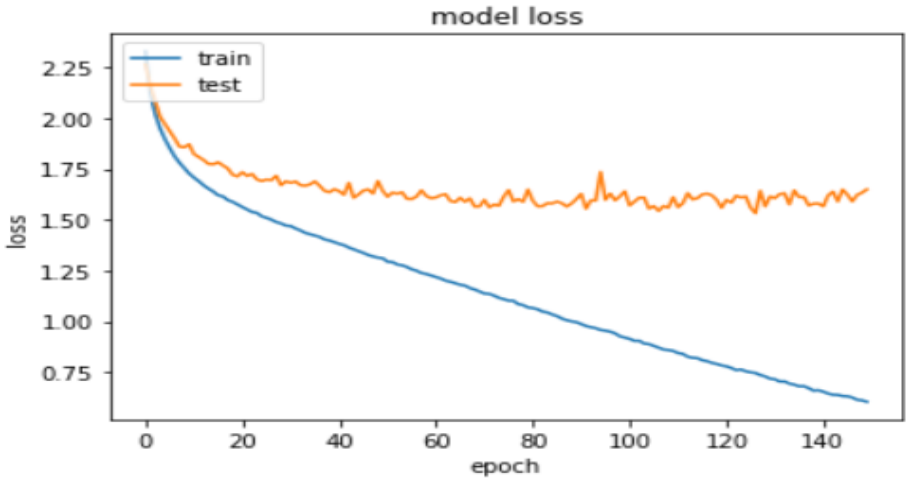

Figure 5.6: Model Loss for Training and Testing Data set with Augmentation

Figure 5.7 is the Actual and predicted after adding data augmentation methods like speed, pitch and dynamic change for our datsets with an accuracy of 49.39 %. Here the predicted labels run better than after adding augmentation methods.

Out[74]:

	actualvalues	predictedvalues
170	female_surprise	female_neutral
171	female_fear	female_fear
172	female_neutral	female_neutral
173	female_neutral	female_neutral
174	male_surprise	male_surprise
175	male_disgust	female_fear
176	female_fear	female_fear
177	female_disgust	female_disgust
178	male_neutral	male_happy
179	male_fear	female_neutral

Figure 5.7: Actual vs Predicted Values After Augmenation Using 1D CNN

Thus after data augmentation methods like speed, pitch and dynamic change our accuracy has been increased to 47.41 %. Figure 5.8 is confusion matrix drawn comparing the predicted and actual labels here both gender and emotion is taken consideration and it shows that the accuracy is 47.42 % which is higher compared to simple 1D CNN model.

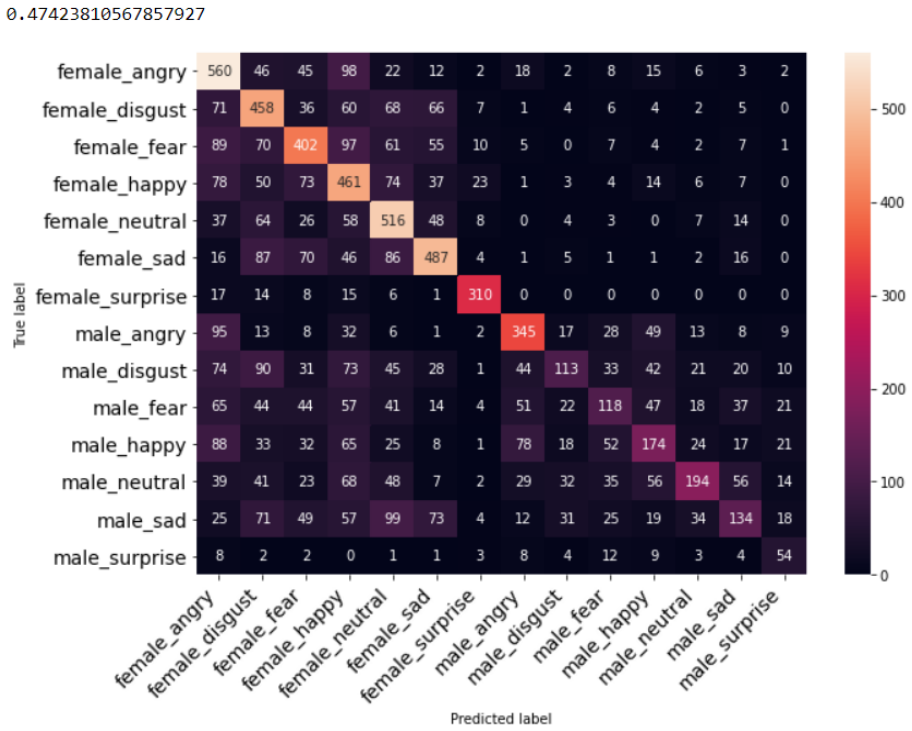

Figure 5.8: Confusion Matrix for Gender Based Emotions with Data Augmentation Using 1D CNN

Figure 5.9 is the confusion matrix based only on gender for training and testing samples and the accuracy of finding gender has been improved after data augmentation methods to 79.89 %.

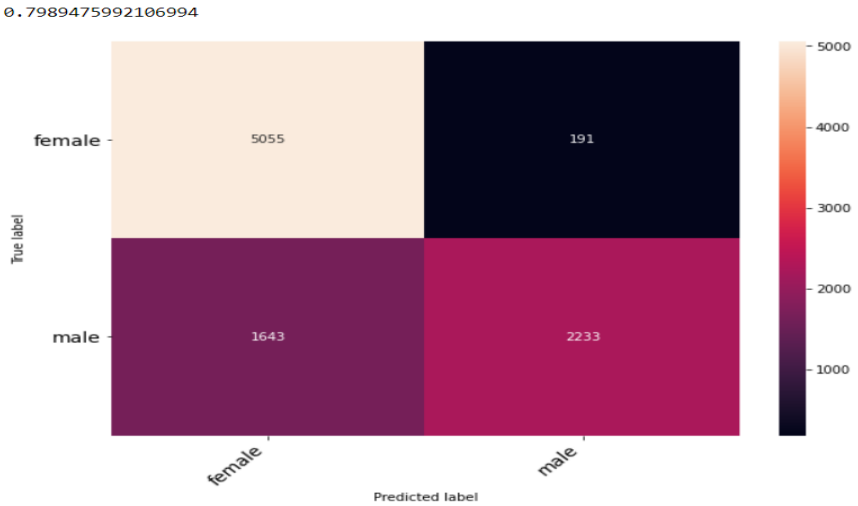

Figure 5.9: Confusion Matrix for Gender with Data Augmentation Using 1D CNN

Figure 5.10 is the confusion matrix based on core emotions for training and testing samples and the accuracy of finding emotions has been improved after data augmentation methods to 52.72 %.This information can be used to improve the model further.

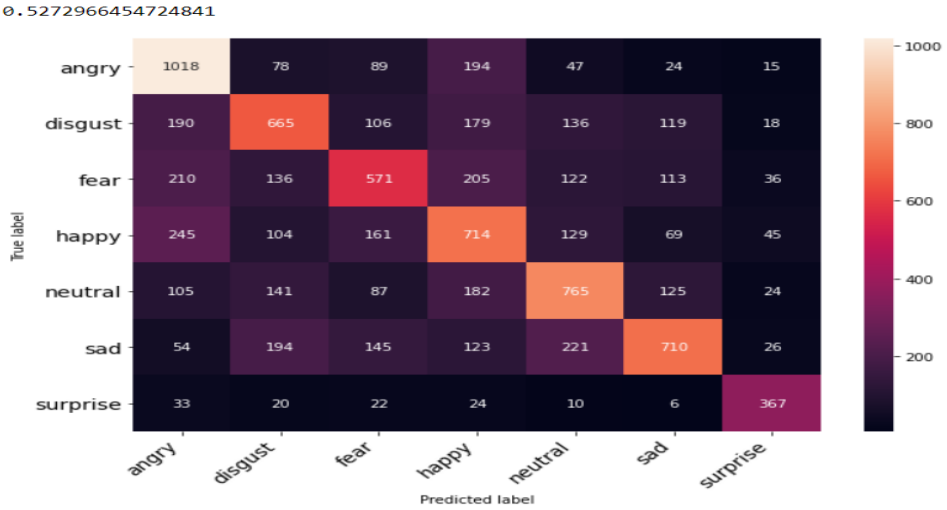

Figure 5.10: Confusion Matrix for Gender with Data Augmentation Using 1D CNN

5.5 2D CNN MODEL

The 2D CNN takes in a 2D array of 30 MFCC bands by 216 audio length as input data imagine it as a 30 x 216 pixel image and just like in images, we could include a 3 Dimension. It got 4 convolution blocks of batch normalisation, max pooling and a dropout node.

5.5.1 MFCC without Augmentation

2D CNN using training and testing samples with MFCC feature vector and without augmentation gives accuracy of 65.04 %. Figure 5.11 is a graph plotted against actual and predicted using mfcc without augmentation method with 2D CNN model.

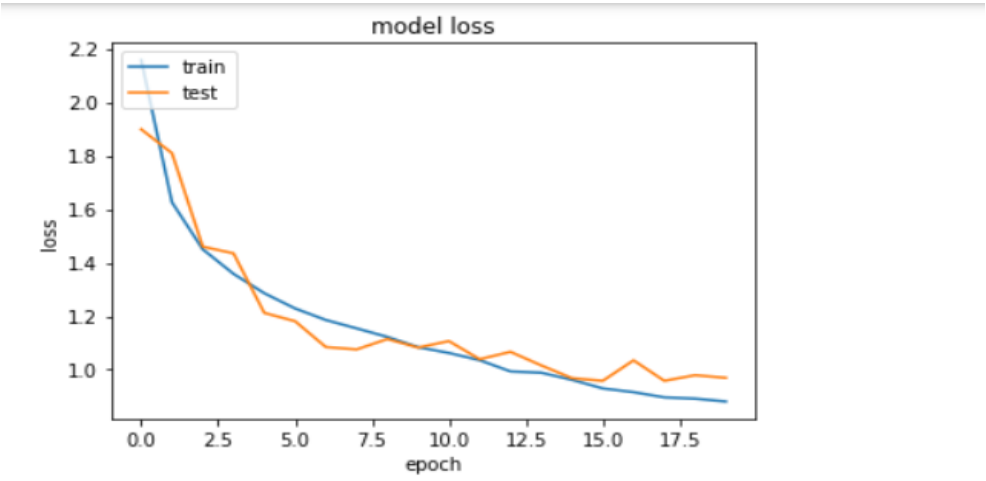

Figure 5.11: Model Loss Graph with 2D CNN without Augmentation

Figure 5.12 is a confusion matrix plotted against actual and predicted values using mfcc without augmentation method with 2D CNN model.

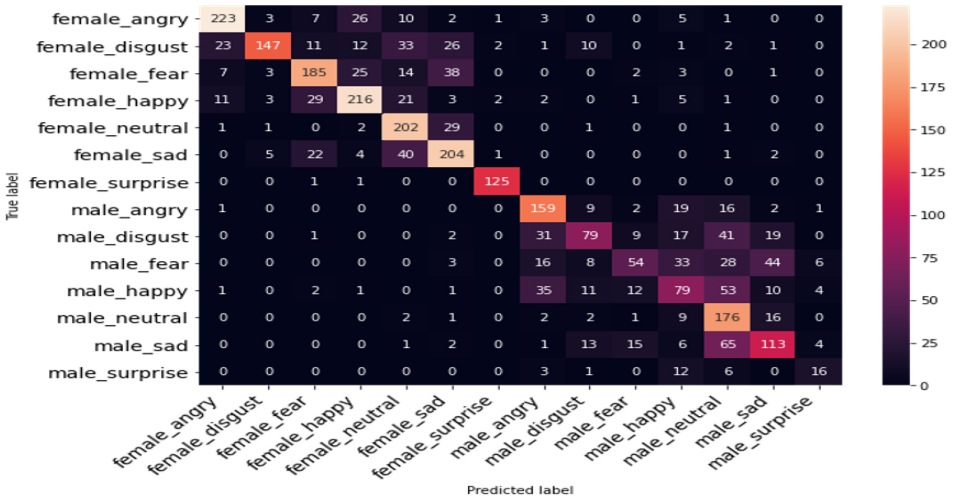

Figure 5.12: Confusion Matrix for Gender Emotion Using 2D CNN without Augmentation

5.5.2 MFCC with Augmentation

2D CNN using training and testing samples with MFCC feature vector and with augmentation gives accuracy of 62.68% which is low compared to MFCC without augmentation method. Figure 5.13 is a graph and confusion matrix plotted against actual and predicted values using mfcc with augmentation method with 2D CNN model.

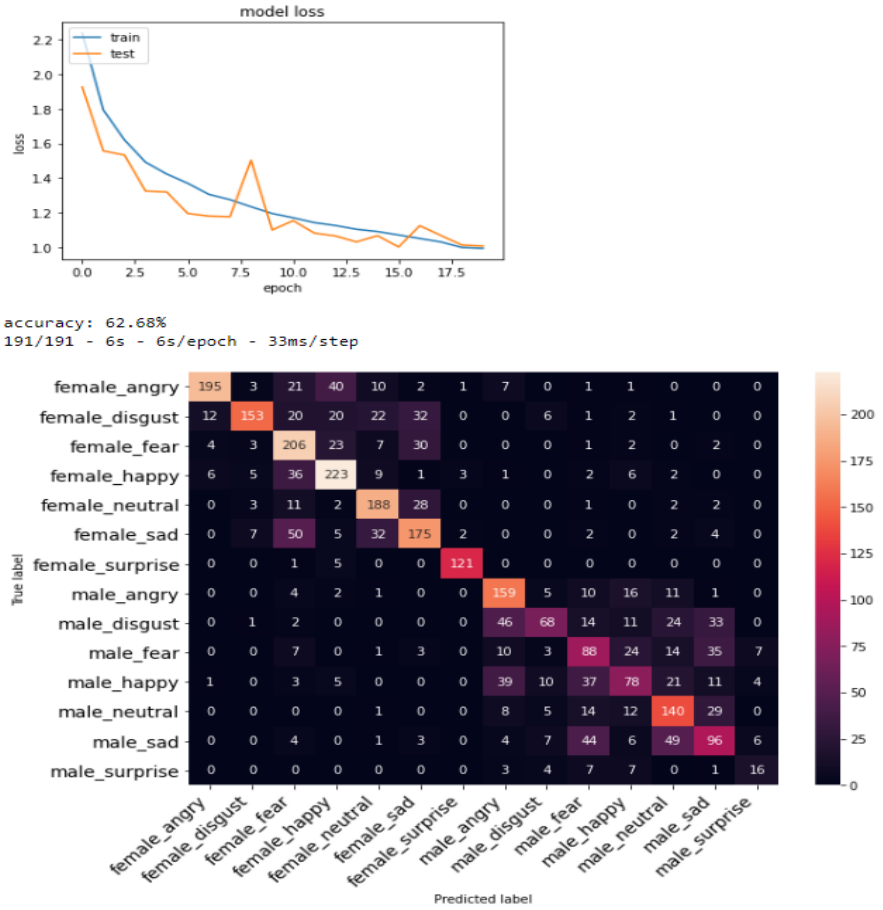

Figure 5.13: Confusion Matrix for Gender with Data Augmentation Using 1D CNN

5.5.3 Log-Melspectogram without Augmentation

2D CNN using training and testing samples with log-melspectogram feature vector and without augmentation gives accuracy of 62.22% which is low compared to MFCC with augmentation method. Figure 5.14 is a graph and confusion matrix plotted against actual and predicted values using Log-melspectogram withoout augmentation method with 2D CNN model.

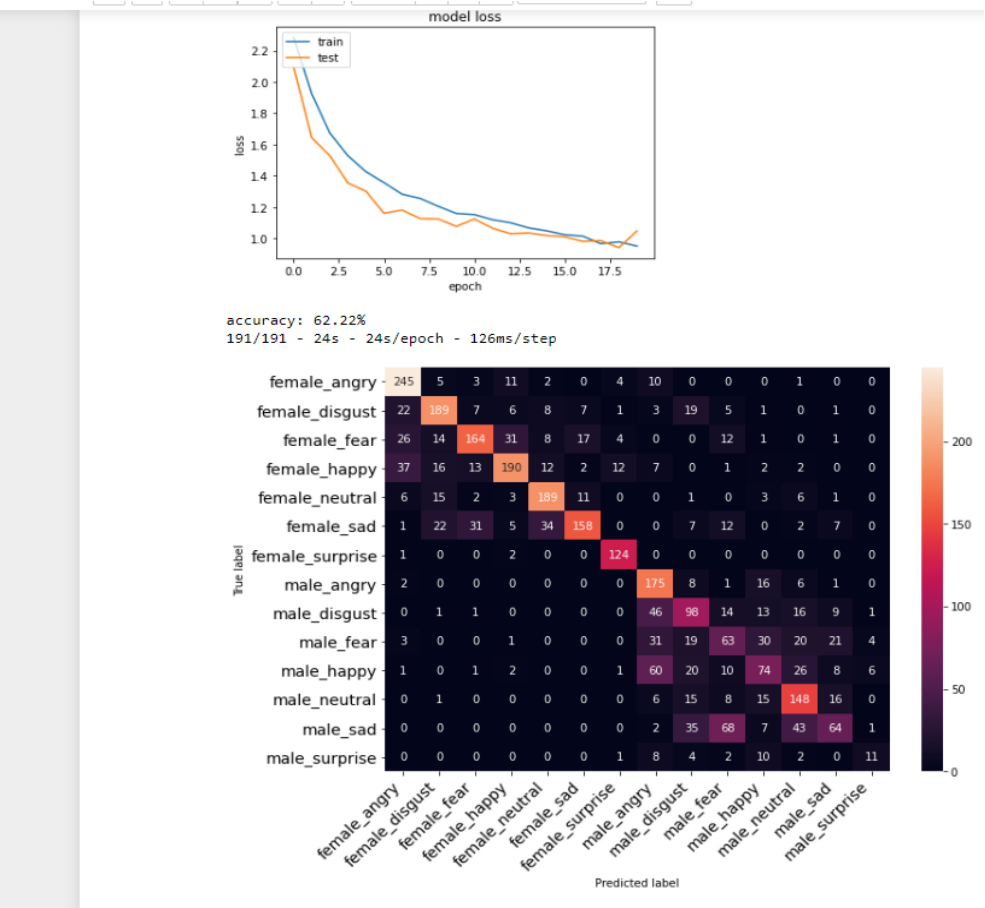

Figure 5.14: Confusion Matrix for Gender with Data Augmentation Using 1D CNN

5.6 PERFORMANCE ANALYSIS

I have checked the performance of overall system. This system give accuracy level Predicted speech emotions. Hence 2D CNN with MFCC along with data augmentation has performed well in speech emotion recognition compared to other methods. Table 5.1 reference in comparison for existing system and proposed system.

Table 5.1: Loss and Accuracy

MODEL	LOSS	ACCURACY
1D CNN with MFCC	57.55%	42.45%
1D CNN with MFCC with Augmentation	50.61%	49.39%
2D CNN with MFCC without Augmentation	34.96%	65.04%
2D CNN with MFCC with Augmentation	37.32%	62.68%
2D CNN with Log-melspectrogram without Augmentation	37.78%	62.68%

5.7 TEST CASES

To check the accuracy of the system some sample test cases are used. Following are the test cases:

1.Only audio files with be allowed to choose. The figure 5.15 illustrates that only audio files can be choosed for our prediction. Selecting some other files with some other format is not supported.

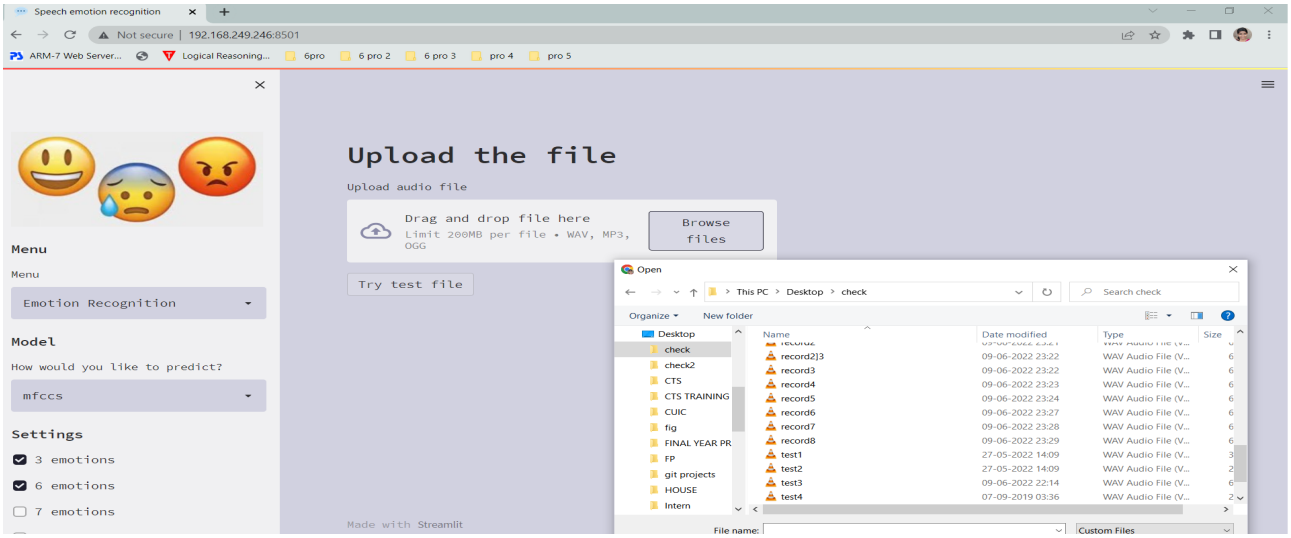

Figure 5.15: Audio File Selection Validation

- Output : Only audio files are allowed
- Result : Passed

2.Try test file button tests the default audio. The figure 5.16 illustrates that try test button predicts the graph and emotion for the test audio file which we have included in the folder.

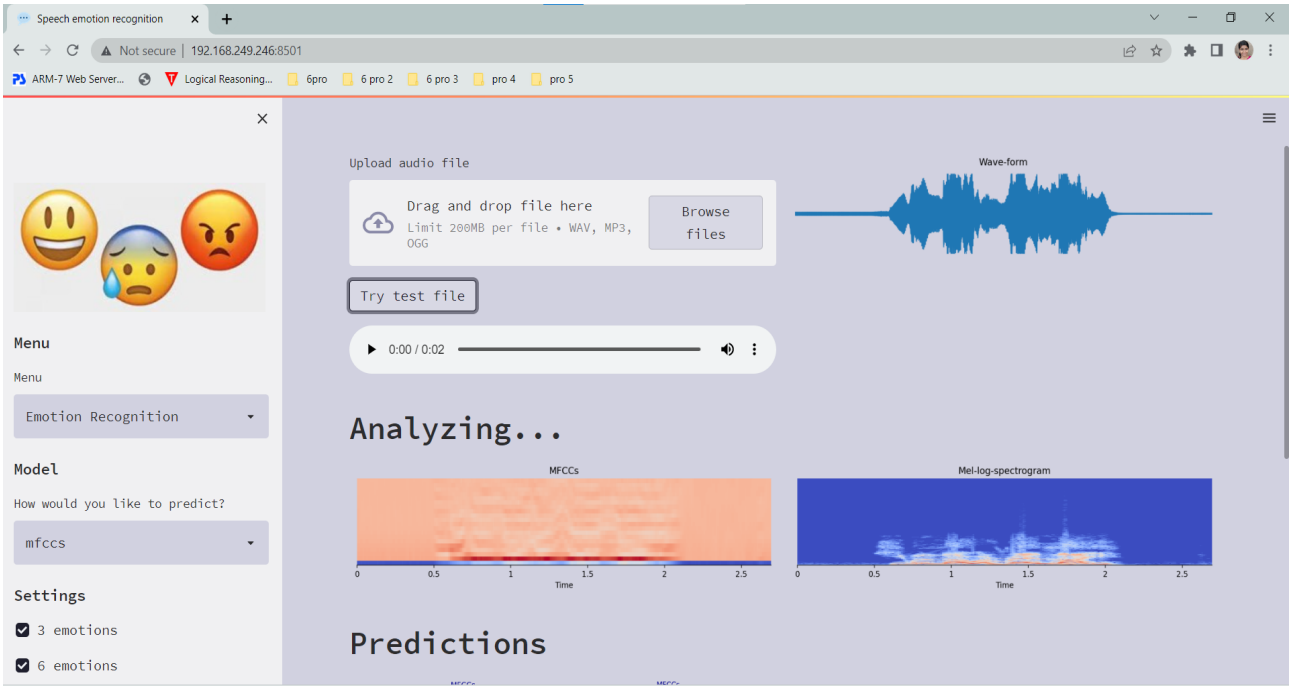

Figure 5.16: Test File Analysis

- Output : Tests default audio and plots for that audio
- Result : Passed

3.Selecting files without speech. Figure 5.17 illustrates that plain audio /music without speech will not be considered for prediction as there is no speech to conclude the emotions.

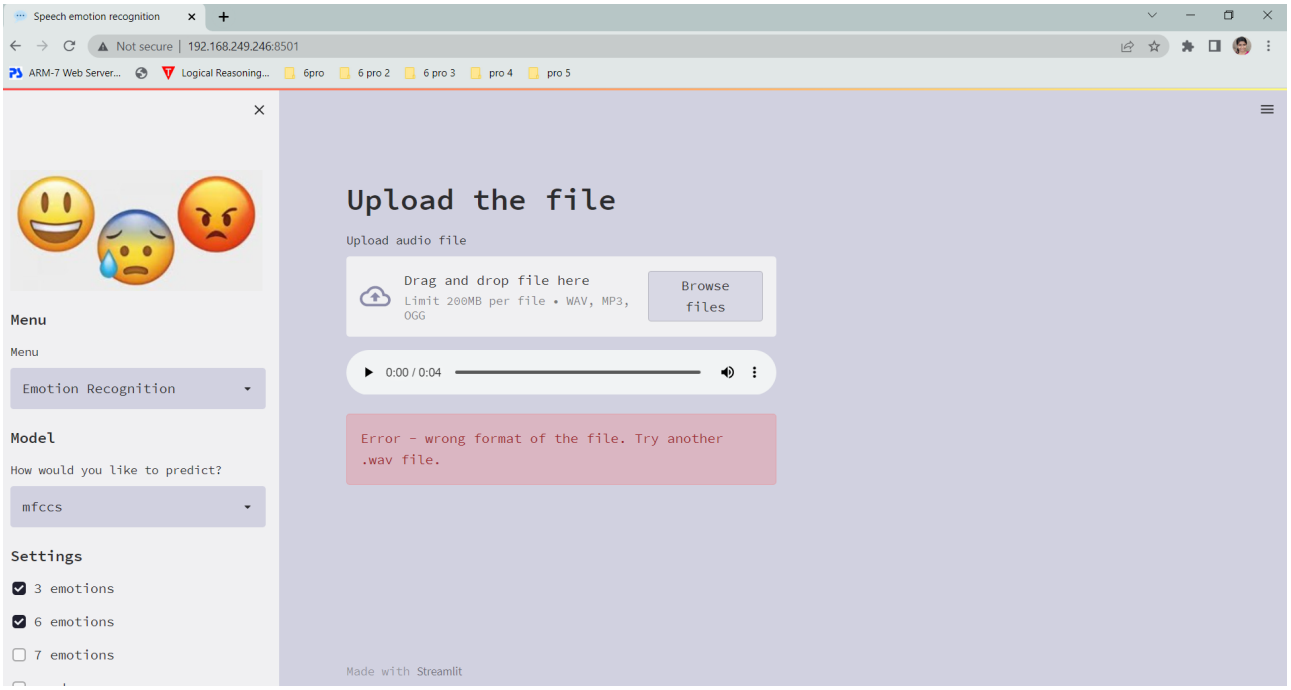

Figure 5.17: Audio without Speech

- Output : Shows error when audio is without speech
- Result : Passed

4.Selecting empty audio. Figure 5.18 illustrates that there will be no predictions made for an empty audio as there is no frequency change in it.

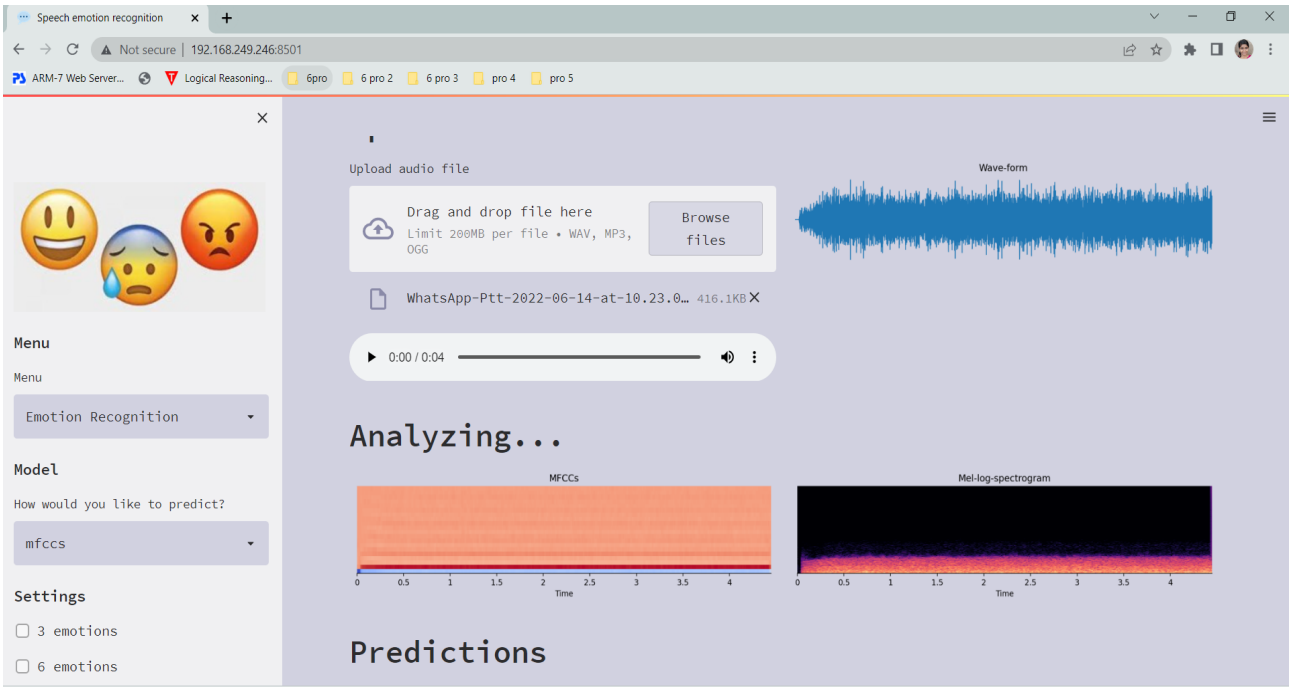

Figure 5.18: Predictions for Empty Audio

- Output : No prediction for empty audio
- Result : Passed

5.Selecting only 3 emotions. The figure 5.19 illustrates that on clicking 3 emotion filter button it will map our audio to only 3 emotions such as positive, negative, neutral.

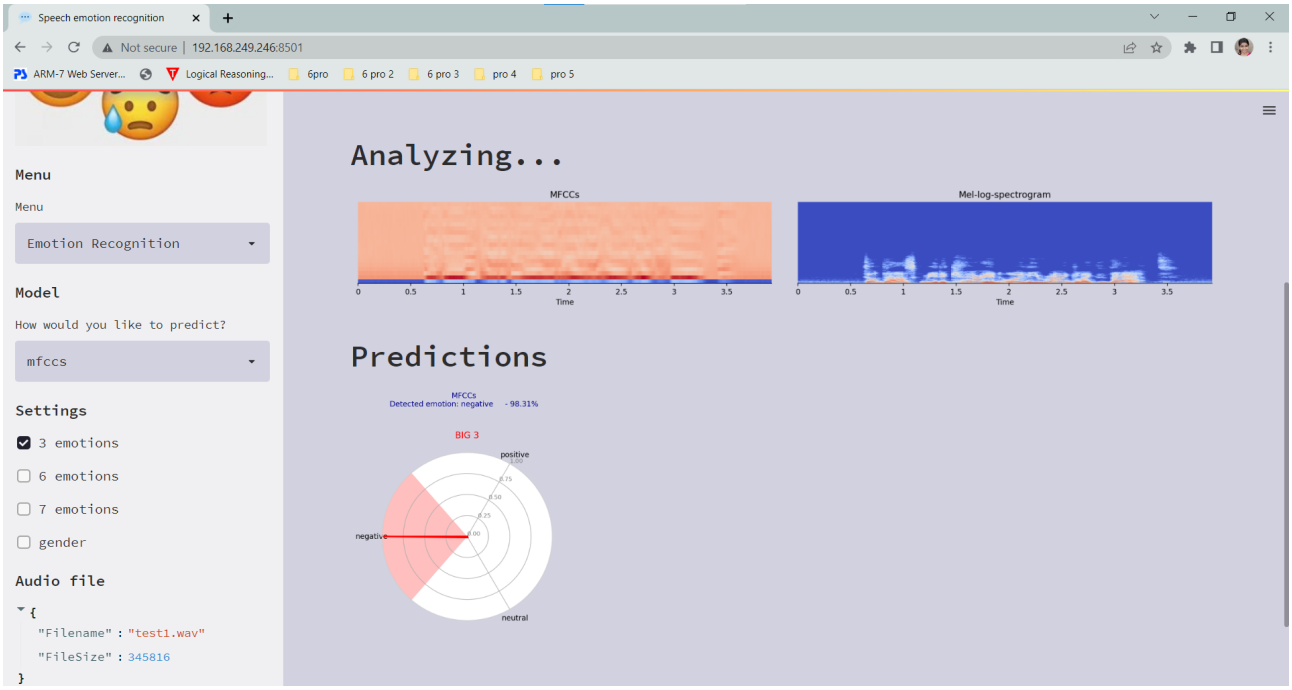

Figure 5.19: Prediction of Only Three Emotions

- Output : 3 emotions output (positive,negative,neutral)
- Result : Passed

6.Selecting only 6 emotions. The Figure 5.20 illustrates that on clicking 6 emotion filter button it will map our audio to only 6 emotions such as fear, angry, neutral, happy, sad, surprise.

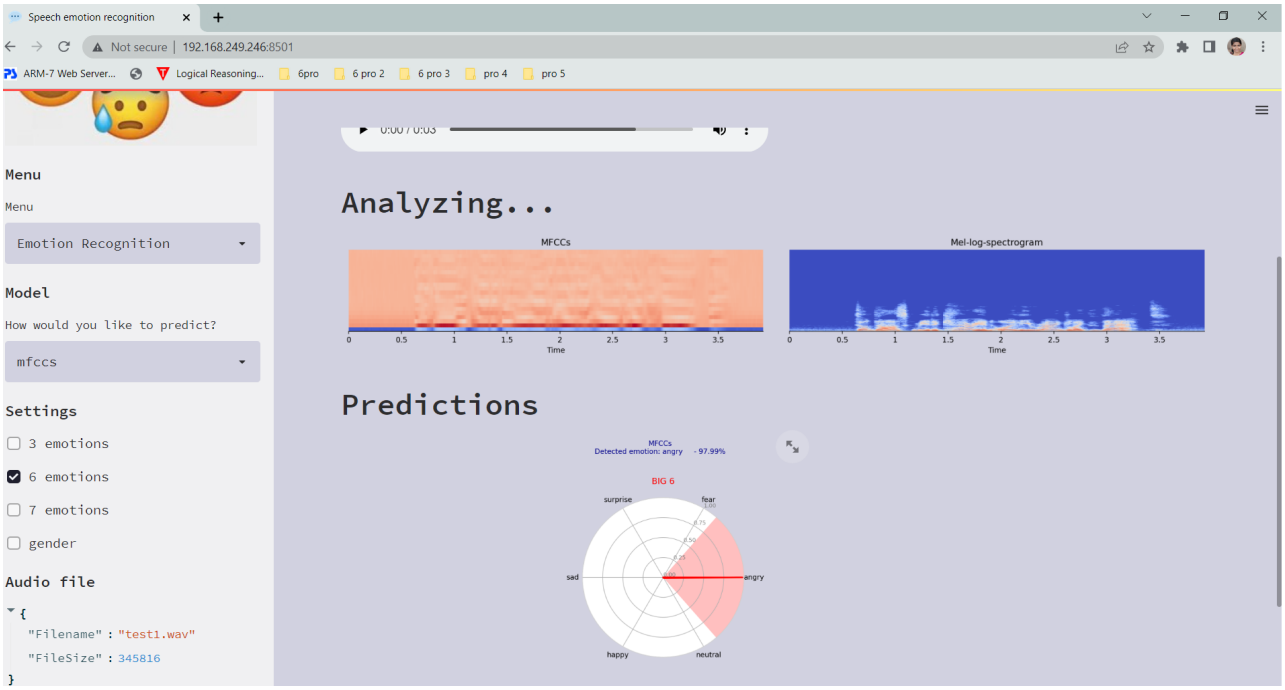

Figure 5.20: Predction of Six Emotions

- Output : 6 emotions output (’fear’, ’angry’, ’neutral’, ’happy’, ’sad’, ’surprise’)
- Result : Passed

7.Selecting 7 emotions. The Figure 5.21 illustrates that on clicking 6 emotion filter button it will map our audio to only 6 emotions such as fear, angry, neutral, happy, sad, surprise, disgust.

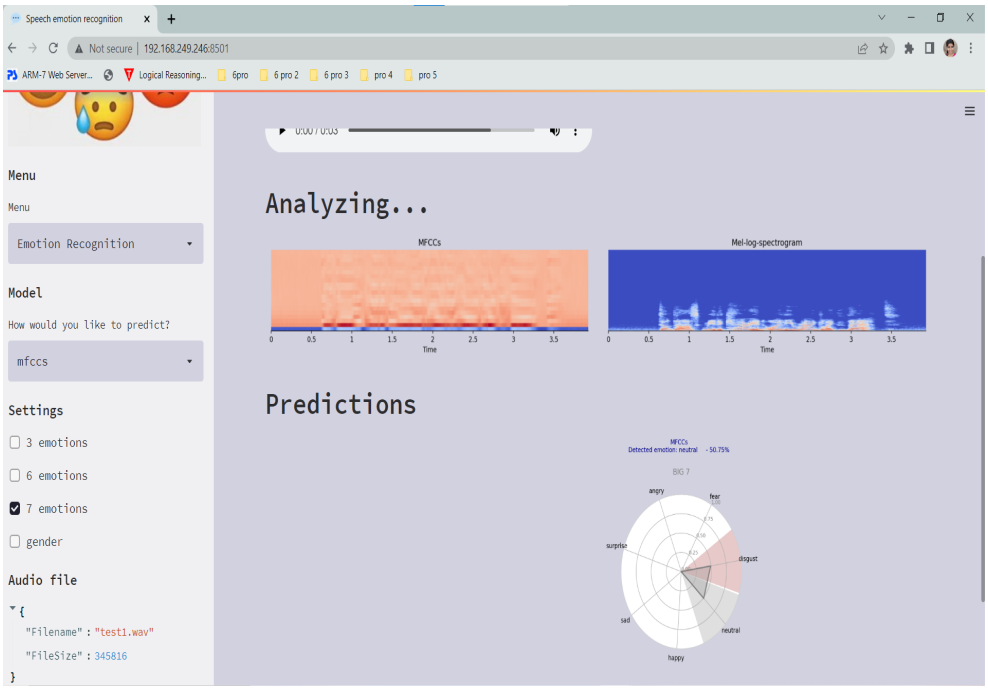

Figure 5.21: Prediction of Seven Core Emotions

- Output : 7 emotions output (’fear’, ’disgust’, ’neutral’, ’happy’, ’sad’, ’surprise’, ’angry’)
- Result : Passed

8.Selecting Gender. The figure 5.22 illustrates that on clicking gender filter it will give the gender of the speaker in the audio.

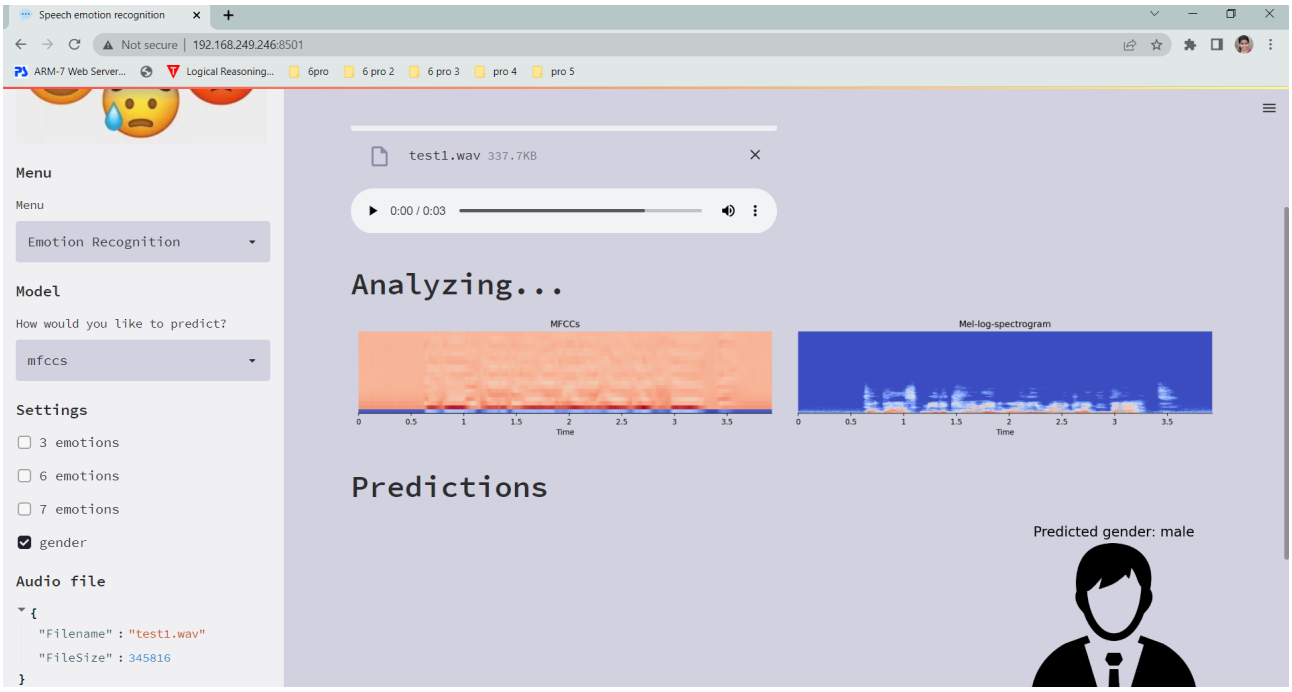

Figure 5.22: Gender Prediction

- Output : Should show male or female
- Result : Passed

9.Selecting all filter tabs. Figure 5.23 gives information on selecting all filters it shows respective values in the specified space.

Figure 5.23: Selecting All Filters

- Output : All filters are chosed
- Result : Passed

10.Selecting Dataset details from drop down. Figure 5.24 gives the detail information of the datsets used for the prediction.

Figure 5.24: Dataset Details

- Output : Dataset details shown
- Result : Passed

11.The figure 5.25 gives the output when a file of greater size with speech is uploaded to find the emotion.

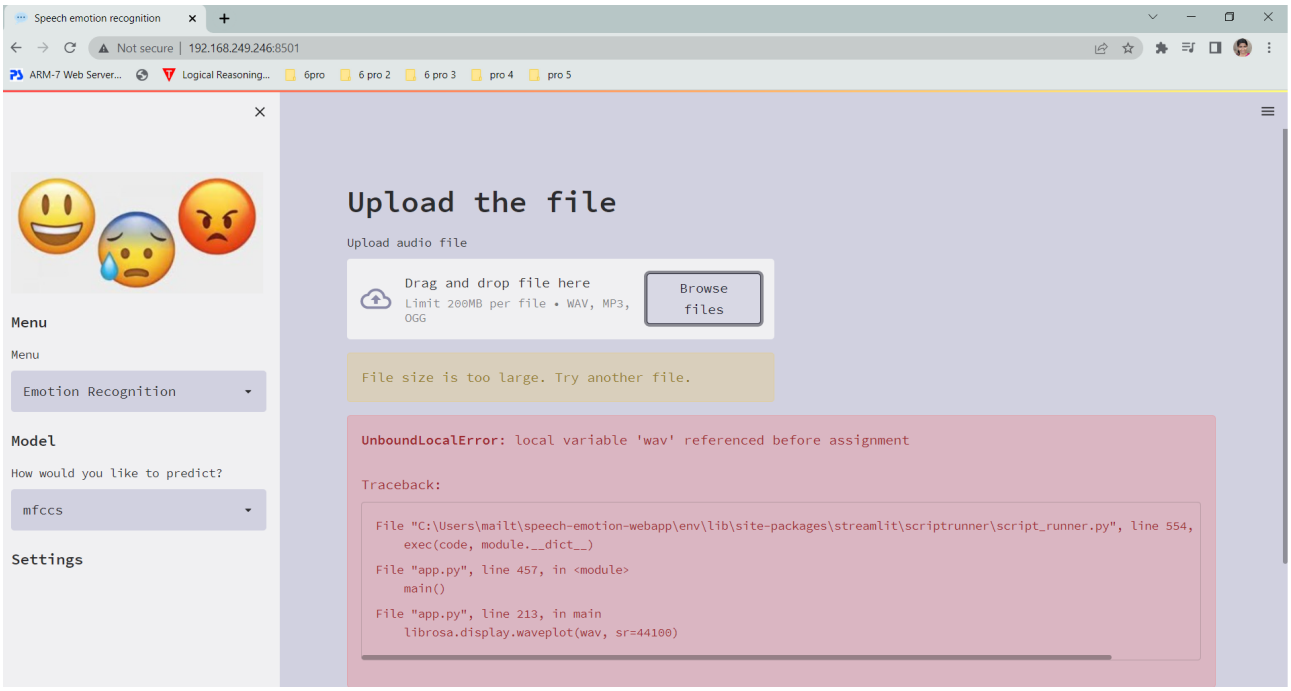

Figure 5.25: Selecting File Greater than 200 MB

- Output : shows error when large size audio is chosen
- Result : Passed

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Speech emotion recognition here is done with multiple CNN models and with different feature extraction methods. 2D CNN without data augmentation produce an output with 65.04 % accuracy while other methods and models produced less accuracy. Mel frequency cepstral coefficient performed well in feature extraction method compared to log-melspectogram feature extraction method. Data augmentation produced less accuracy with both MFCC and log-melspectogram. Hence MFCC along with 2D CNN can be used for speech emotion recognition.

6.2 FUTURE WORK

In future real time applications based on above work can be used to identify the emotions of customers. Patient diagnosis can also be done effectively using this model. More feature extraction methods can be used to compare with current model. The proposed models can be used for emotion-related applications such as conversational chatbots, social robots, etc. where identifying emotion and sentiment hidden in speech may play a role in the better conversation.

REFERENCES

[1] Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn W Schuller. Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition. *IEEE Transactions on Multimedia*, 21(3):795–808, 2018.

[2] Zhaocheng Huang, Julien Epps, Dale Joachim, and Vidhyasaharan Sethu. Natural Language Processing Methods for Acoustic and Landmark Event-Based Features in Speech-Based Depression Detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):435–448, 2019.

[3] Weijian Zhang and Peng Song. Transfer Sparse Discriminant Subspace Learning for Cross-Corpus Speech Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:307–318, 2019.

[4] Peng Song, Shifeng Ou, Xinran Zhang, Yun Jin, Wenming Zheng, Jinglei Liu, and Yanwei Yu. Transfer Semi-Supervised Non-Negative Matrix Factorization for Speech Emotion Recognition. *IEICE TRANSACTIONS on Information and Systems*, 99(10):2647–2650, 2016.

[5] Ying Qin, Tan Lee, and Anthony Pak Hin Kong. Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia. *IEEE journal of selected topics in signal processing*, 14(2):331–345, 2019.

[6] Bingxin Liu, Qiang Zhang, Ying Li, Wen Chang, and Manrui Zhou. Spatial-Spectral Jointed Stacked Auto-Encoder-Based Deep Learning for Oil Slick Extraction from Hyperspectral Images. *Journal of the Indian Society of Remote Sensing*, 47(12):1989–1997, 2019.

[7] Harini Murugan Apoorv Singh, Kshitij Kumar Srivastava. Speech Emotion Recognition Using Convolutional Neural Network (CNN). *International Journal of Psychosocial Rehabilitation*, 24(8):1–20, 2020.

[8] Matthew D Zeiler, M Ranzato, Rajat Monga, Min Mao, Kun Yang, Quoc Viet Le, Patrick Nguyen, Alan Senior, Vincent Vanhoucke, Jeffrey Dean, et al. On Rectified Linear Units for Speech Processing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3517–3521. IEEE, 2013.