SPEECH EMOTION RECOGNITION

A Minor Project Report

Submitted To

Chhattisgarh Swami Vivekanand Technical University Bhilai, India

For

The Partial Fulfillment of Degree

Of

Bachelor of Technology

In

Computer Science & Engineering

By

Chetna Arya

Roll No.- 303302218048

&

Shweta Shrivas

Roll No.- 303302218052

Semester 7th (CSE)

Under the Guidance of

Mr. Abhishek Kumar Saw

Assistant Professor

Department of Computer Science & Engineering

S.S.I.P.M.T, Raipur

Department of Computer Science & Engineering
Shri Shankaracharya Institute of Professional Management & Technology,
Raipur (C.G.)

Session: 2021 - 2022

DECLARATION BY THECANDIDATE

We the undersigned solemnly declare that the Minor project report entitled **"*SPEECH EMOTION RECOGNITION*"** is based our own work carried out during the course of our study under the supervision of **Mr. Abhishek Kumar Saw.**

We assert that the statements made and conclusions drawn are an outcome of the project work. We further declare that to the best of our knowledge and belief that the report does not contain any part of any work which has been submitted for the award of any other degree/diploma/certificate in this University/Deemed university of India or any other country.

CHETNA ARYA
Roll No.-303302218048
Enrollment No. – BF4740
B.Tech. VII Sem (CSE)

SHWETA SHRIVAS
Roll No.-303302218052
Enrollment No.- BF4744
B.Tech. VII Sem (CSE)

CERTIFICATE BY THE SUPERVISOR

This is to certify that the Minor project report entitled **"*SPEECH EMOTION RECOGNITION*"** is a record of project work carried out under my guidance and supervision for the fulfillment of the award of degree of Bachelor of Technology in the faculty of Computer Science & Engineering of Chhattisgarh Swami Vivekananda Technical University, Bhilai (C.G.) India.

To the best of my knowledge and belief the report

i) Embodies the work of the candidate herself

ii) Has duly been completed

iii) Fulfills the partial requirement of the ordinance relating to the B.Tech degree of the University

iv) Is up to the desired standard both in respect of contents and language for being referred to the examiners.

(Signature of the Supervisor)
Mr. Abhishek Kumar Saw
Assistant Professor,
Dept of Computer Science & Engineering

Forwarded to Chhattisgarh Swami Vivekanand Technical University
Bhilai

(Signature of HOD)
Dr. J P Patra
Dept. of Computer Science &Engineering
S.S.I.P.M.T Raipur, C.G

(Signature of the Principal)
Dr. Alok Kumar Jain
S.S.I.P.M.T Raipur, C.G

CERTIFICATE BY THEEXAMINERS

The project report entitled ***“SPEECH EMOTION RECOGNITION”*** has been examined by the undersigned as a part of the examination of Bachelor of Technology in the faculty of Computer Science & Engineering of Chhattisgarh Swami Vivekanand Technical University, Bhilai.

____________________	____________________
Internal Examiner	**External Examiner**
Date:	**Date:**

ACKNOWLEDGEMENT

Working for this project has been a great experience for us. There were moments of anxiety, when we could not solve a problem for the several days. But we have enjoyed every bit of process and are thankful to all people associated with us during this period we convey our sincere thanks to our project guide **Mr. Abhishek Kumar Saw** for providing me all sorts of facilities. His support and guidance helped us to carry out the project. We owe a great dept of his gratitude for his constant advice, support, cooperation & encouragement throughout the project we would also like to express our deep gratitude to respected **Dr. J P Patra** (Head of Department) for his ever helping and support. We also pay special thanks for his helpful solution and comments enriched by his experience, which improved our ideas for betterment of the project. We would also like to express our deep gratitude to respected **Dr. Alok Kumar Jain** (Principal) and college management for providing an educational ambience. It will be our pleasure to acknowledge, utmost cooperation and valuable suggestions from time to time given by our staff members of our department, to whom we owe our entire computer knowledge and also we would like to thank all those persons who have directly or indirectly helped us by providing books and computer peripherals and other necessary amenities which helped us in the development of this project which would otherwise have not been possible.

.

(Signature of Candidate)

CHETNA ARYA

B.Tech (CSE) VII Semester

Roll No.- 303302218048

Enrollment No.- BF4740

(Signature of Candidate)

SHWETA SHRIVAS

B.Tech (CSE) VII Semester

Roll No.- 303302218052

Enrollment No.- BF4744

LIST OF ABBREVIATIONS

ML	Machine Learning
MLP	Multilayer Perceptron Classifier
ANN	Artificial Neural Network
SER	Speech Emotion Recognition
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
GB	Giga Byte
RAM	Random Access Memory
LPCC	Linear predictive cepstral coefficients
MEL	Mel scale cepstral analysis
LPA	Linear predictive analysis
RASTA	Relative spectra filtering of log domain coefficients
MFCC	Mel-Frequency Cepstrum Coefficient
MLNN	Multilayer Neural Network
SVM	Support Vector Machine
GMM	Gaussian Mixture Model
AR	Autoregressive
HMM	Hidden Markov Model
MLM	Maximum Likelihood Model
KNN	K-Nearest Neighbour
PLP	Perceptual Linear Predictive coefficients
TEO	Teager Energy Operator

LIST OF FIGURES

TABLE OF CONTENT

ABSTRACT

Speech Emotion Recognition is a system where we determine emotions from live audio. Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction. In speech emotion recognition, many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Machine Learning techniques have been recently proposed as an alternative to traditional techniques in SER. Different persons have different emotions and altogether a different way to express it. Speech emotion do have different energies, pitch variations are emphasized if considering different subjects. The speech emotion recognition is based on the Artificial Neural Network (ANN) algorithm which uses different modules for the emotion recognition and the MLP classifier is used to differentiate emotions such as happiness, anger, neutral state, sadness, etc. The dataset for the speech emotion recognition system is the speech samples and the characteristics are extracted from these speech samples using LIBROSA package. The classification performance is based on extracted characteristics. Finally we can determine the emotion of speech signal. The approach consists of three steps. First, numerical features are extracted from the sound database by using audio feature extractor. Then, feature selection method is used to select the most relevant features. Finally, a machine learning model is trained to recognize seven universal emotions: anger, sadness, happiness and neutral.

CHAPTER- 01

INTRODUCTION

1.1 Overview

As human beings' speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them.

We define a Speech Emotion Recognition system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this project we attempt to detect underlying emotions in live audio by analyzing the acoustic features of the audio data. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication.

Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them [1]. Detecting emotions is one of the most important marketing strategies in today's world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI-related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result, this type of application has much potential in the world that would benefit companies and even safety to consumers.

1.2 Application

The speech emotion recognition system is implemented as a Machine Learning (ML) model.

The steps of implementation are comparable to any other ML project, with additional

fine-tuning procedures to make the model function better.

The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data.

The second step, called feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address the several data representation and data quality issues.

The third step is often considered the core of an ML project where an algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to.

The final step is to evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms.

Comparison results help to choose the appropriate ML algorithm most relevant to the problem. This project is completely based on machine learning and deep learning where we train the models with RAVDESS Dataset which consists of audio files which are labeled with basic emotions [2].

CHAPTER- 02

LITERATURE REVIEW

2.1 SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition is nothing but an application of the pattern recognition system in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc. In this project we have used MLP which come under ANN.

The system contains five major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in below figure. Overall, the system is based on analysis of the generation mechanism of speech signal, extracting some of features which contain information about speaker’s emotion & taking appropriate pattern recognition model to identify states of emotion.

Typically, a set of emotion having 300 emotional states [3]. Whenever, signal is passed to the feature extraction & selection process, the extracted speech features are selected in terms of emotion relevance. All over procedure revolves around the speech signal for extraction to the selection of speech features corresponding to emotions. Forward step is generation of database for training as well as testing of extracted speech features. At the end, detection of emotions has been done using classifier with the usage of pattern recognition algorithm. The Speech emotion recognition is similar to the speaker recognition system but different types of approach to detect emotions make it secure & intelligent. The evaluation of the system is depending on naturalness of the input database.

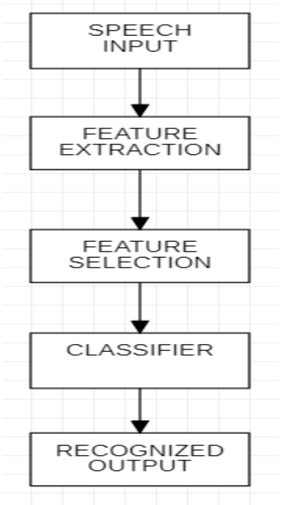


FIG 2.1 – STRUCTURE OF SE

2.2 SPEECH PARAMETERS

Some of the Time domain parameters been applied in recent researches from speech emotion recognition are briefly given.

2.2.1 Energy

In speech processing, Energy refers to the variations in amplitude of some speech signal. The voiced & unvoiced signal contain distinguish energy level so that this could easily identified with the usage of this feature. Also, initials & end of the silent part can be found out.

2.2.2 Pitch

Pitch is frequency of vibration of vocal cords of any voice signal. As during speech production process, our glottis opens and then closes with some period; such period is pitch period. Pitch range is 50-400Hz typically. It can be estimated by quantifying the period or measuring the harmonics. Fundamental frequency is a feature i.e., reciprocal of pitch.

2.2.3 Formant

In speech wave, a concentration of acoustic energy around a particular frequency is formant. The vocal tract behaves like a resonance chamber that always amplifying & attenuating certain frequencies. The spectrum of vocal tract response consists of number of resonant frequencies i.e. Formants. It ranges from 200 to 1000 Hz, 1000 to 2000 Hz and 2000-3500 Hz. This given formants carries informative information so that usually considered whereas the fourth one never considered. Each formant corresponds to a resonance in the vocal tract.

2.2.4 Mel-frequency cepstrum coefficient (MFCC):

A unique representation of spectral property of voice signals. These are the best for speaker/speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. The computation of MFCC explained in article by Mirlab [4]. An article about Spectrogram deals that Mel-frequency scale represents subjective pitch i.e., perceived pitch. Also, it takes certain properties of the human auditory system [5].

2.3 LITERATURE

A lot of research works have been performed in identification of emotion through speech processing. Performance of speech recognition systems is usually evaluated in terms of accuracy

and speed. Some of that work has already been done by many people worldwide.
Tin Lay Newel proposed a text independent method of speech emotion classification, which makes use of short time log frequency power coefficients (LFPC) to represent the speech signals and a Hidden Markov Model (HMM) is used as a classifier. A system for classification of emotional state of utterances is proposed. Six categories of speech emotions are applied in such work. A database of 60 emotional utterances, each from 12 speakers is constructed and is used to train and test the system. Average accuracy of 78% achieved & analyzed performance using LPCC and MFCC. Also goes with the fact that grouping of emotions with same characteristics enhance system performance [7].

Kamran Soltani studied the importance of the psychology and linguistics in spoken language man-machine interfaces. Along with the techniques in signal processing and analysis, it also requires psychological and linguistic analysis. The work makes use of six emotions, happiness, sadness, anger, fear, neutral and boredom. It uses pitch i.e., speech fundamental frequency, formant frequencies, energy and voicing rate as features. These speech parameters were used to train neural network classifier and the Berlin Database of Emotional Speech. Average accuracy of 77% is achieved. The work concludes that anger and neutral can be recognized easily while fear the most difficult one [8].

Jana Tuckova performed experimental analysis using parameters like fundamental frequency, formant frequency and statistical analysis was conducted for multilayer neural network (MLNN). The average accuracy obtained using this technique is 75.93% for multiword sentences while that for one-word sentences is 81.67%. Aim was to verify different knowledge from phonetics and neural network. Also classified the speech signal that are been described by musical theory. The result of research work gone mainstream to the area like prosody modelling and for analysis of disordered speech especially in children [6]. Every emotion contains different vocal parameter that exhibits diverse speech characteristics.

An MFCC-based vocal emotional recognition performed using ANN in which MFCC features were used as speech parameters and five different emotional states were considered for analysis [5][9]. Back-Propagation algorithm applied for interpretation of speaker emotion. Also, the proposed system for recognition is independent of linguistic background and achieved 60.55% of average accuracy of recognition.

An effective solution to improve human-computer interaction allowing human and computer intelligent interaction was developed [11]. It says, together with MFCC, pitch is the most frequently used parameter in recognizing speaker's gender. Other speech parameters used are formants, bandwidths, source spectral tilt, jitter and shimmer, harmonic to noise ratio. Speech features used for emotion recognition are statistical analyses of amplitude of speech, energy, pitch, formants, 12 MFCC, pitch and amplitude perturbations. The system does two experiments i.e., gender recognition and emotion recognition. Berlin Emotion Speech database is employed in this research work and support vector machine (SVM) supports as classifier. Sometimes emotions could not be correctly identified in adverse condition like in a noise corrupted telephone channel speech.

A research work investigated a filtering technique in automatic detection of emotions from telephonic speech where the MFCC, delta MFCC and delta MFCC features were incorporated with Gaussian mixture model (GMM) as classifier on Berlin database of emotional speech, while autoregressive (AR) model is employed in the proposed filtering method [3].

2.4 CLASSIFIER

Various machine learning techniques have been employed for classification purpose. Forward to the procedure of selecting features, classification is needed. Aim is to build a classification model with the help of some machine learning algorithm to predict emotional states on the basis of speech parameters [11].

Among all approaches available, mostly applied classification methods are Hidden Markov Model (HMM), Support Vector Machine (SVM), Maximum Likelihood Model (MLB), Artificial Neural Network (ANN), k-Nearest Neighbour (k-NN). Some other classifiers deserve reference here are Decision Tree, Fuzzy Classifier and many more.

These entire classification models have their own advantages & drawbacks according to their application area. Classifier gives a decision based on the patterns of test speech sample and patterns of trained database. GMM is more suitably applied for global features that has been extracted from training utterances and achieves two levels of accuracy. Classifiers like ANN have their own strength in identifying nonlinear boundaries separating the emotional states as well. Out of many, Feed-forward & Multilayer perceptron layer neural network are frequently used model. ANN exploit concept of acoustical phonetic & pattern recognition. In speech recognition, HMM is used as classifier for classifying sequential data as it represents a set of various states. Also represents probabilities of making a transition from a state to another. HMM has achieved success in modelling of temporal information in speech spectrums [7].

CHAPTER- 03

PROJECT REQUIREMENT ANALYSIS

3.1 PROJECT OBJECTIVE IN DETAIL

The speech emotion recognition system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the mode outperform previous state-of-the-art methods in assigning data to a minimum of one among 4 emotion categories (i.e., angry, happy, sad and neutral).

Choosing to follow the lexical features would require a transcript of the speech which might further require a further step of text extraction from speech if one wants to predict emotions from real-time audio. Similarly, going forward with analyzing visual features would require the surplus to the video of the conversations which could not be feasible in every case while the analysis on the acoustic features are often wiped out real-time while the conversation is happening as we'd just need the audio data for accomplishing our task. Hence, we elect to analyze the acoustic features during this work. The field of study is termed as Speech Processing and consists of three components: Speaker Identification, Speech Recognition, Speech Emotion Detection.

An emotion one out of a delegated set of emotions is identified with each unit of language (word or phrase or utterance) that was spoken, with the precise start of every such unit determined within the continual acoustic signal. A striking nature unique to humans is that the ability to change conversations supported the spirit of the speaker and also the listener.

3.2 SOFTWARE REQUIREMENT

Python programming language is used. Python is an interpreter, high-level, and general- purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects [10].

- Python3
- Pycharm
- Librosa
- Soundfile
- Pickle
- Pyaudio
- Sklearn
- NumPy
- Speech recognition
- Os
- Glob2

3.3 HARDWARE REQUIREMENT

- i5 8th gen processor
- 2 GB Graphic Card
- 8GB RAM

CHAPTER- 04

PROBLEM IDENTIFICATION & DESIGN

4.1 FLOWCHART

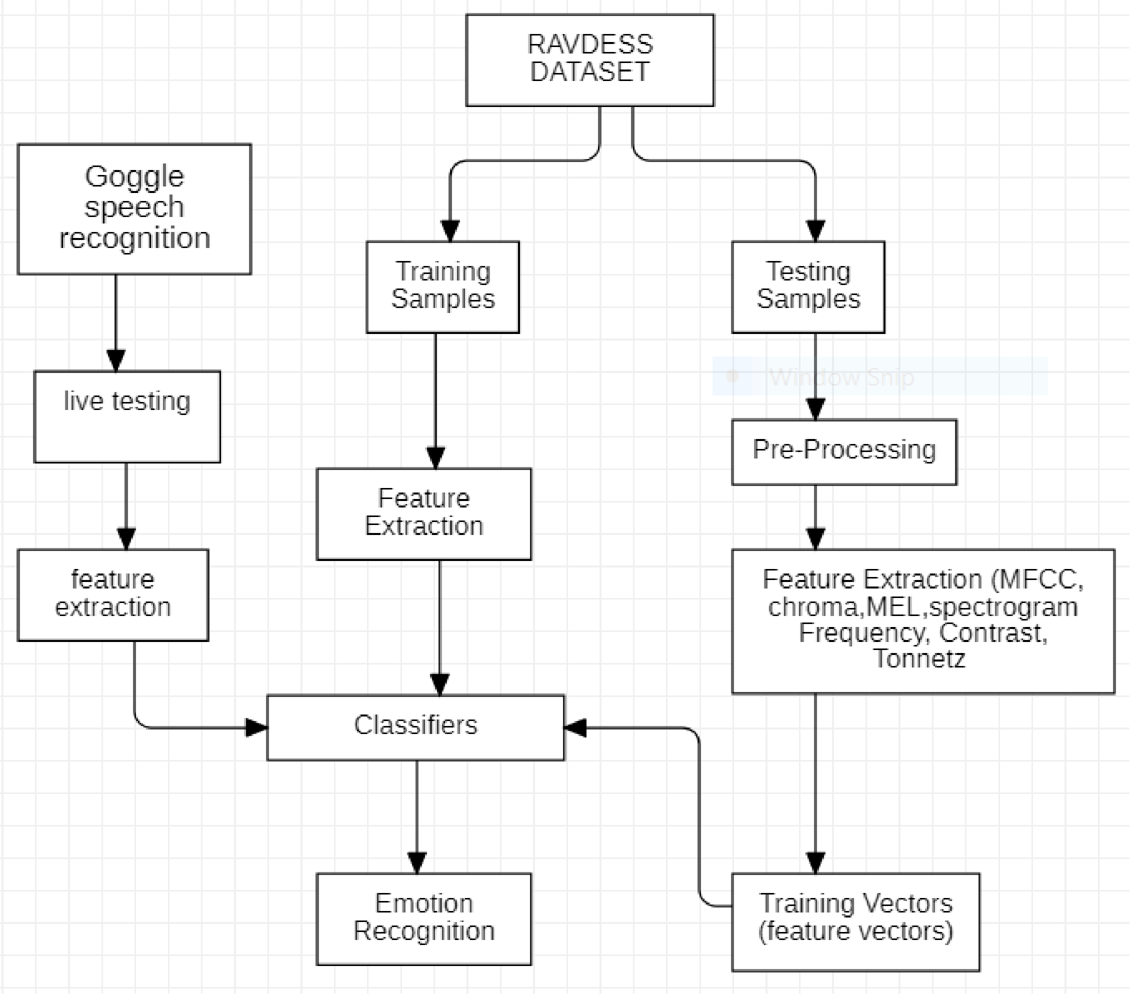


FIG 4.1: FLOWCHART OF SER

The above flowchart shows the work flow of the project. RAVDESS dataset is used for training and testing of the project which is divided in 75% and 25% respectively. In testing, the dataset goes through pre-processing and then features extraction is performed like MFCC, MEL, chroma, tonnetz. Then classifier MLP work on this extracted feature and give appropriate output. In training, the features are extracted and then classifier uses it to determine the underlying emotion. The project works in such a way that we use live testing using Google speech recognition and then features are extracted. MLP classifier uses the features to determine emotion.

4.2 DATA FLOW DIAGRAM/MECHANISM

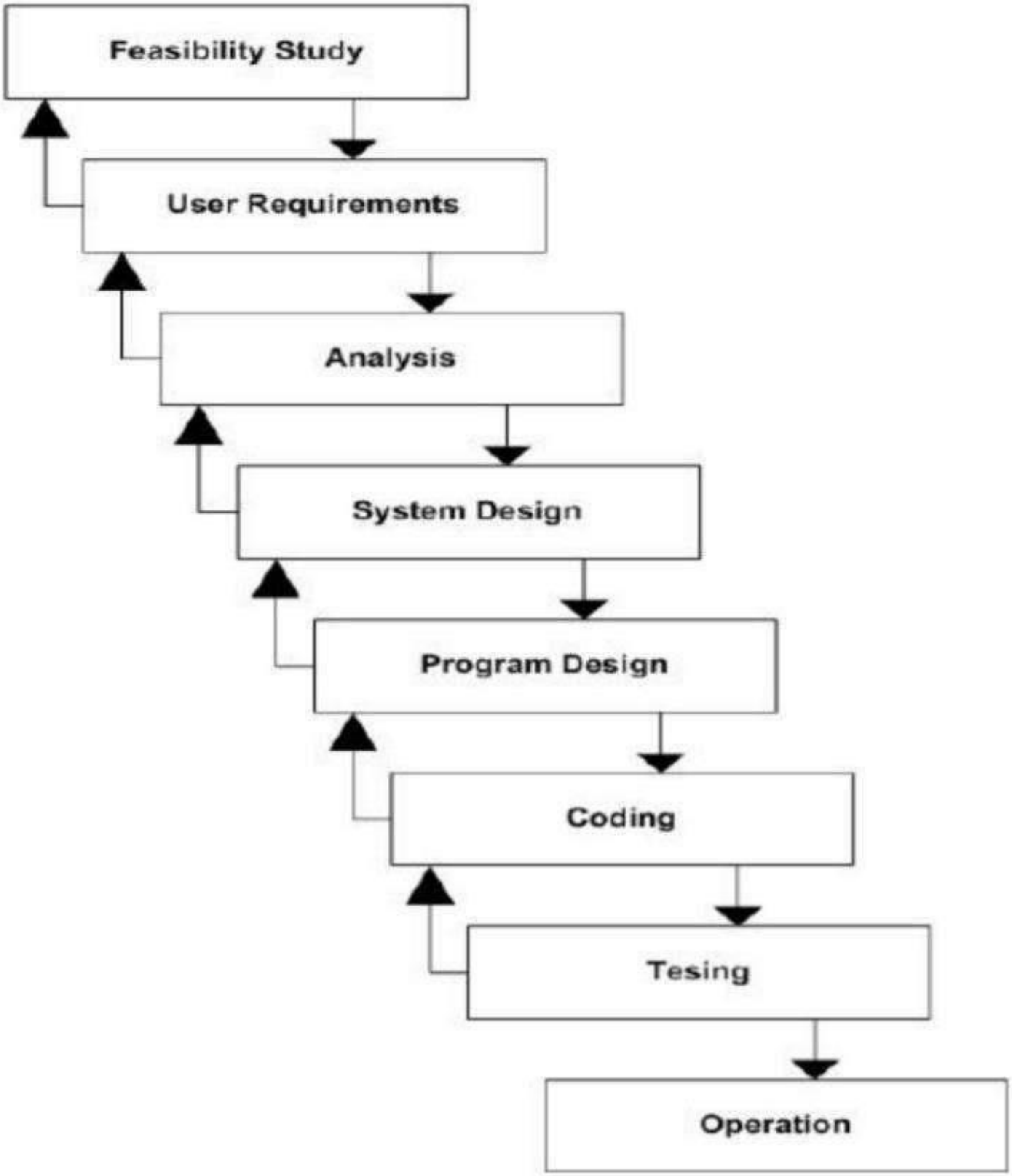


FIG. 4.2 WATERFALL MODEL

Waterfall approach was first SDLC Model to be used widely in Software Engineering to ensure success of the project. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In this Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially.

The following illustration is a representation of the different phases of the Waterfall Model. The sequential phases in Waterfall model are −

- **Requirement Gathering and analysis** − All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
- **System Design** − The requirement specifications from first phase are studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.
- **Implementation** − With inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
- **Integration and Testing** − All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
- **Deployment of system** − Once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
- **Maintenance** − There are some issues which come up in the client environment. To fix those issues, patches are released. Also, to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

All these phases are cascaded to each other in which progress is seen as flowing steadily downwards (like a waterfall) through the phases. The next phase is started only after the defined set of goals are achieved for previous phase and it is signed off, so the name "Waterfall Model". In this model, phases do not overlap [13].

4.3 ALGORITHM

Multi-layer perceptron (MLP) is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer, as shown in Fig. below. The input layer receives the input signal to be processed. The required task such as prediction and classification is performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed forward network in a MLP the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases of MLP are pattern classification, recognition, prediction and approximation [12].

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptron's (with threshold activation). Multilayer perceptron's are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

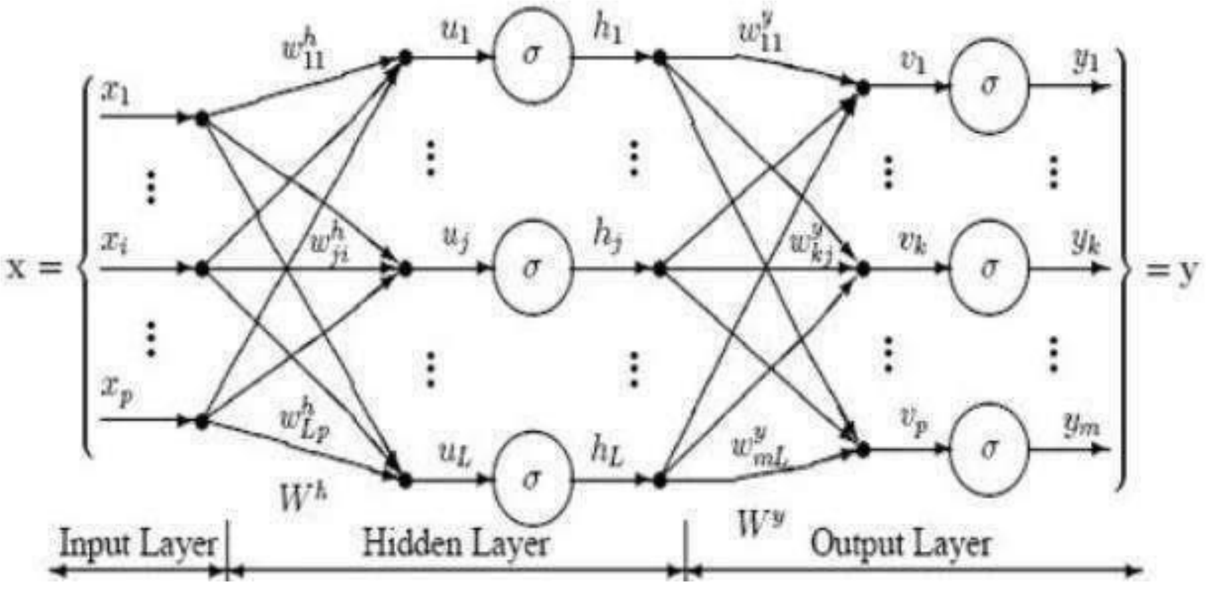


FIG:4.3 MLP MODEL

The multilayer perceptron is applied for supervised learning problems. The multi-layer perceptron is also issued for the purpose of classification. The MLP is made to train on the given dataset. The training phase enables the MLP to learn the correlation between the set of inputs and outputs. During training, the MLP adjusts model parameters such as weights and biases in order to minimize the error. The MLP uses Back propagation, to make weight and bias adjustments relative to the error. The model parameter has been set with hidden layer 300, iteration 500, which have found to be best by grid search.

CHAPTER- 05

METHODOLOGY

5.1 SOFTWARE USED

5.1.1 PYTHON

Python is a high-level, interpreted scripting language developed in the late 1980s by Guido van Rossum at the National Research Institute for Mathematics and Computer Science in the Netherlands. The initial version was published at in 1991, and version 1.0 was released in 1994. The Latest version of python is 3.9. Python has huge number of modules for covering every aspect of programming [10]. These modules are easily available for use hence making Python an extensible language. Python is a scalable programming language because it provides an improved structure for supporting large programs than shell-scripts.

5.1.2 PYCHARM

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains (formerly known as IntelliJ).[5] It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSs), and supports web development with Django as well as data science with Anaconda.

5.1.3 LIBROSA

Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation (using LSTM's), Automatic Speech Recognition. It provides the building blocks necessary to create the music information retrieval systems.

5.1.4 SKLEARN

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

5.1.5 NUMPY

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. We used NumPy as it provides 50x faster an array object (called Nd array) for our sound file.

5.1.6 PYAUDIO

Pyaudio provides Python bindings for PortAudio, the cross-platform audio I/O library. With Pyaudio, you can easily use Python to play and record audio on a variety of platforms". We are using Pyaudio to get the audio from the user.

5.1.7 SOUNDFILE

SoundFile can read and write sound files. File reading/writing is supported through libsndfile, which is a free, cross-platform, open-source (LGPL) library for reading and writing many different sampled sound file formats that runs on many platforms including Windows, OS X, and Unix**.** SoundFile can read and write sound files. This library helps us to open the sound file.

5.1.8 PICKLE

Pickle in Python is primarily used in serializing and deserializing a Python object structure. In other words, it's the process of converting a Python object into a byte stream to store it in a file/database, maintain program state across sessions, or transport data over the network. The pickled byte stream can be used to re-create the original object hierarchy by unpickling the stream. It is used for the saving model**.**

5.1.9 OS

The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system dependent functionality. The *os* and *os.path* modules include many functions to interact with the file system.

5.2 FEATURE USED

In the project, audio features which are supported and used to analyse audio data are as follows: -

5.2.1 Mel Frequency Cepstral Co-efficients (MFCC)

MFCC features represent phonemes (distinct units of sound) as the shape of the vocal tract (which is responsible for sound generation) is manifest in them [4]. The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phones in the speech. MFCC is calculated using this equation: -

$$\hat{\mathrm{C}}_n = \sum_{n=1}^{k} \left(\log\hat{\mathrm{S}}_k\right) cos \left[n\left(k-\frac{1}{2}\right)\frac{\pi}{k}\right]$$

FIG 5.1 – MFCC EQUATION

5.2.2 Chroma

It is used for harmonic and melodic characteristics of music, meaningfully characterized pitches of music in 12 different categories. Chroma-based features, which are also referred to as “pitch class profiles”, are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale.

5.2.3 MEL Spectrogram Frequency (mel)

It deals with human perception of frequency; it is a scale of pitches judged by listeners to be equal distance from each other. It scales the frequency in order to match more closely what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies). A frequency measured in Hertz (f) can be converted to the Mel scale using the following formula: -

$$mel\,(f) = 2595\ x\ \log_{10}\,(1 + f/700)$$

FIG 5.2 – MEL EQUATION

5.2.4 Spectral Contrast

In an audio signal, the spectral contrast is the measure of the energy of frequency at each timestamp. Since most of the audio files contain the frequency, whose energy is changing with time. It becomes difficult to measure the level of energy. Spectral contrast is a way to measure that energy variation. High contrast values generally correspond to clear, narrow-band signals, while low contrast values correspond to broad-band noise. Here the energy contrast is measured by comparing the mean energy in the peak energy frame to that of the bottom or valley energy frame.

5.2.5 Tonal Centroid Features (TONNETZ)

This phenomenon considers that the audio file is of 6 pitch classes by merging some of the classes together so the drawn chroma feature graph from this method classifies the pitches only in 6 classes which we call Tonnetz.

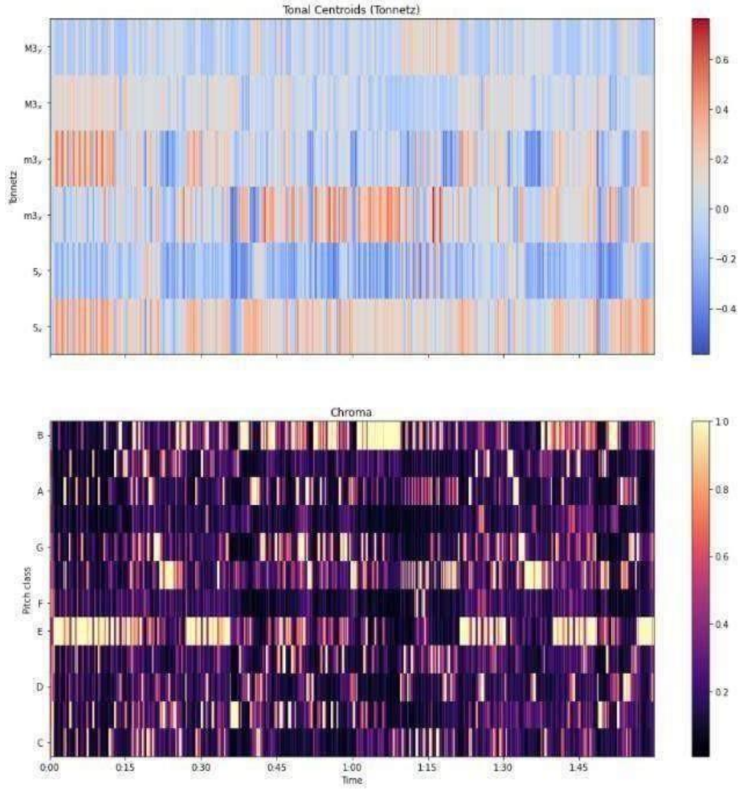


FIG 5.3 – THE ABOVE FIGURE SHOWS

TONNETS VS CHROMA FOR SAME AUDIO

5.3 DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) has been used. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB).

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions.

Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). The utterances of the speech are kept constant by speaking only 2 statements of equal lengths. Each of the RAVDESS files has a unique filename.

The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics [2].

FILENAME IDENTIFIERS

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the ‘neutral’ emotion.
- Statement (01 = “Kids are talking by the door”, 02 = “Dogs are sitting by the door”).
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

FILENAME EXAMPLE: 02-01-06-01-02-01-12.mp4

- Video-only (02)
- Speech (01)
- Fearful (06)
- Normal intensity (01)
- Statement “dogs” (02)
- 1st Repetition (01)
- 12th Actor (12)
- Female, as the actor ID number is even.

5.4 PRE-PROCESSING MODULE

This is the Pre-Processing module, once after getting the input from user, the input speech is pre-processed.

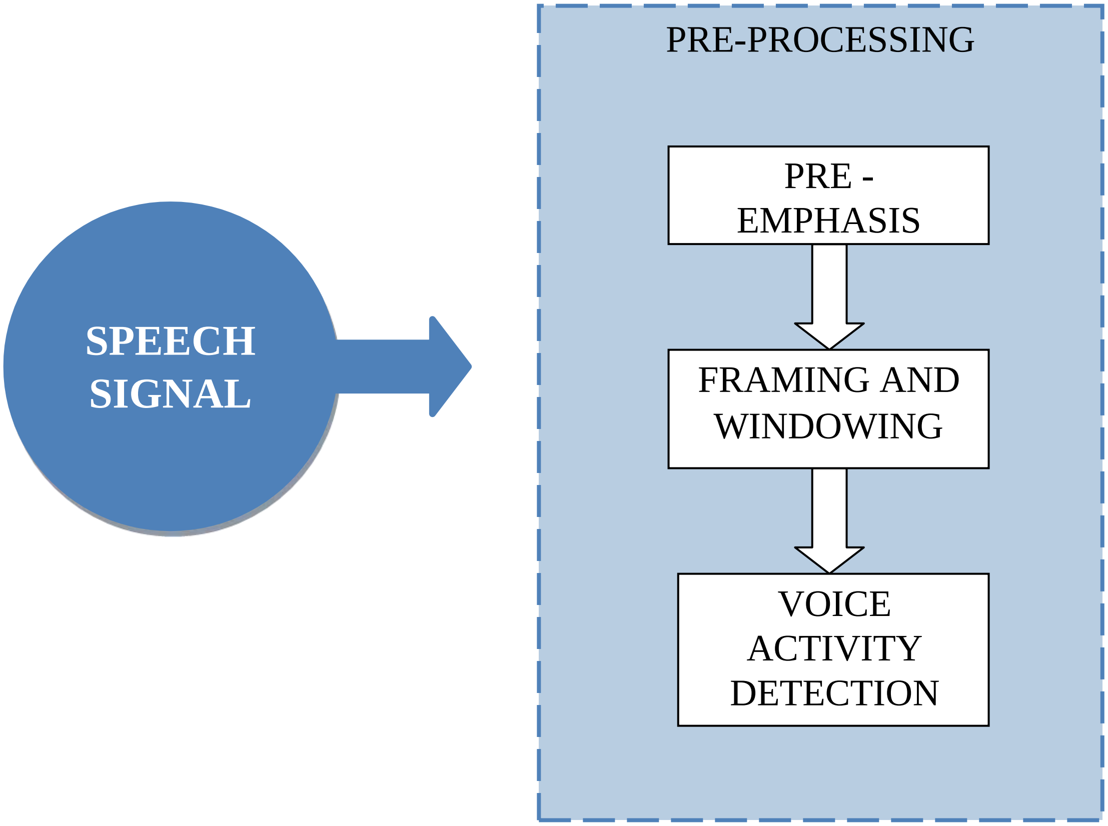


FIG 5.4 – PRE-PROCESSING MODULE

CHAPTER- 06

TESTING AND SNAPSHOTS

6.1 TESTING

After the feature extraction, we need to make the system knows about the feature for instance, we are using only “angry”, “sad”, “neutral”, “happy” emotions in our system. Then the step is Defining the file which is what emotion then training the machine using each feature extracted to the known emotion and testing, where we are using 75 % of data for training and 25 % for the testing by splitting.

At first loading the data from a folder which can be done using python library glob and getting base name using os library as we know RAVDEES dataset is made such a way that emotion on 2nd base so declaring X for feature and y for emotion. X is obtained from “extract feature” and y obtained using “base name” after splitting the base name and we know the “base name” refers to what emotion.

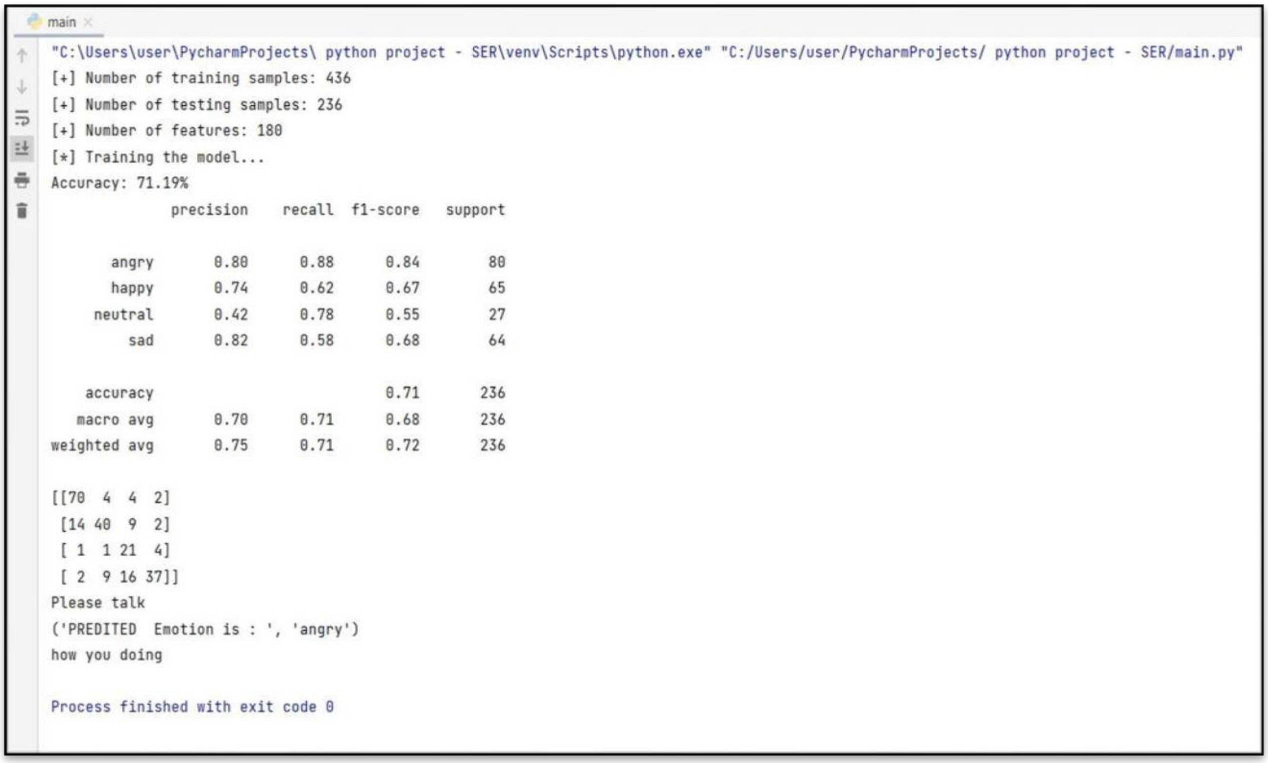
For testing the live voice, “Live testing .py” has been created where we are using “pyaudio” module to take voice and we add some noise to make the feature extraction better and it is then passed to extract sound feature and where mlp classifier predicts the emotion and We have added the extended feature that give translation of voice to text using “speech Recognition” module. which need to installed and imported.

After testing a number of live audios, the compiler gives warning of maximum iteration overhead. With this information, we increase the number of iterations. As well as during testing you can change the splitting model of training and testing. As the number of iteration increase, the time for running this program also increases exponentially.

The testing of live audio can be emphasized by varying pitch and energy. For testing, the internet of the device should be available. As the system uses Google voice recognizer, it will show result when the system has internet access. The neurons in ML uses random weights, therefore there is possibility of accuracy of audio to vary.

6.2 SNAPSHOTS

OUTPUT 1: -

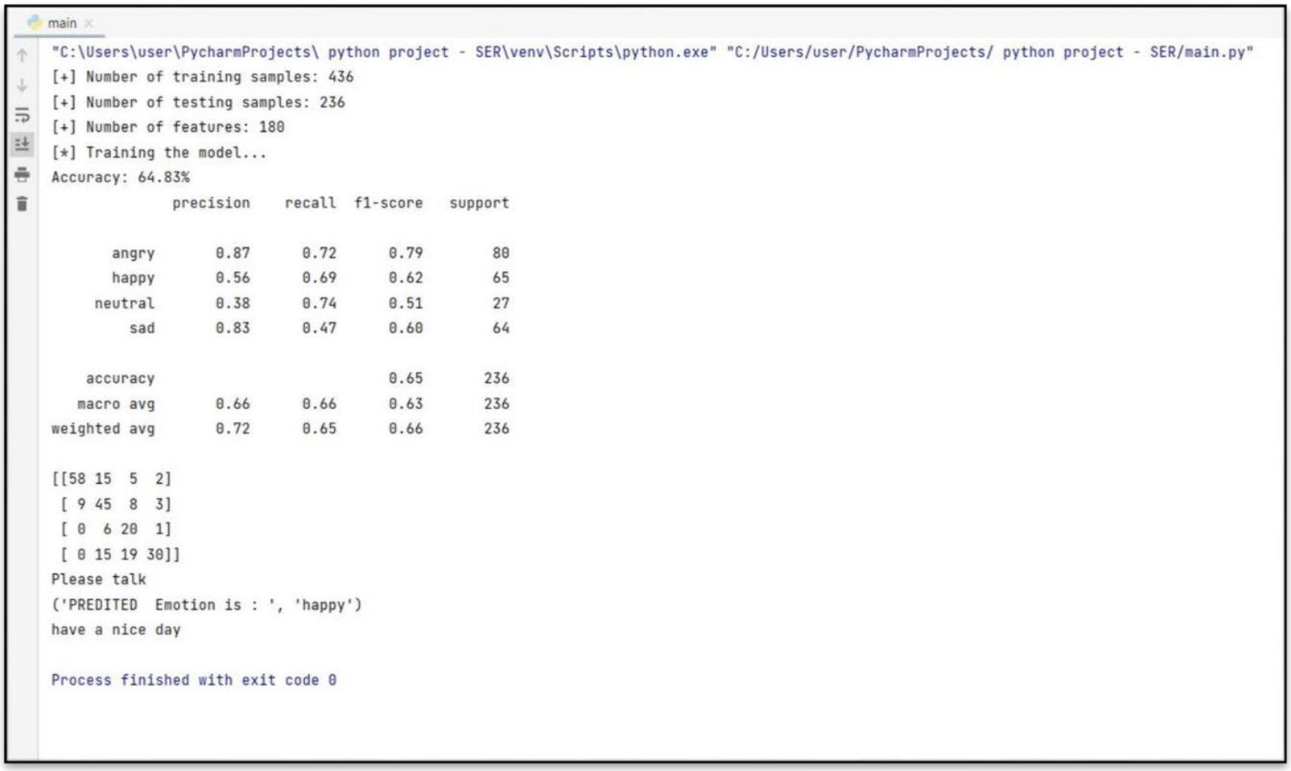


```
"C:\Users\user\PycharmProjects\ python project - SER\venv\Scripts\python.exe" "C:/Users/user/PycharmProjects/ python project - SER/main.py"
[+] Number of training samples: 436
[+] Number of testing samples: 236
[+] Number of features: 180
[*] Training the model...
Accuracy: 71.19%
              precision    recall  f1-score   support

       angry       0.80      0.88      0.84        80
       happy       0.74      0.62      0.67        65
     neutral       0.42      0.78      0.55        27
         sad       0.82      0.58      0.68        64

    accuracy                           0.71       236
   macro avg       0.70      0.71      0.68       236
weighted avg       0.75      0.71      0.72       236

[[70  4  4  2]
 [14 40  9  2]
 [ 1  1 21  4]
 [ 2  9 16 37]]
Please talk
('PREDITED  Emotion is : ', 'angry')
how you doing

Process finished with exit code 0
```

FIG 6.1: THE FIGURE SHOWS THE PREDICTED EMOTION ANGRY FOR LIVE AUDIO TEST 1.

We can see that the system uses confusion matrix and has accuracy 71.19%. The predicted emotion is angry for the phrase “how you doing”.

OUTPUT 2: -



```
"C:\Users\user\PycharmProjects\ python project - SER\venv\Scripts\python.exe" "C:/Users/user/PycharmProjects/ python project - SER/main.py"
[+] Number of training samples: 436
[+] Number of testing samples: 236
[+] Number of features: 180
[*] Training the model...
Accuracy: 64.83%
              precision    recall  f1-score   support

       angry       0.87      0.72      0.79        80
       happy       0.56      0.69      0.62        65
     neutral       0.38      0.74      0.51        27
         sad       0.83      0.47      0.60        64

    accuracy                           0.65       236
   macro avg       0.66      0.66      0.63       236
weighted avg       0.72      0.65      0.66       236

[[58 15  5  2]
 [ 9 45  8  3]
 [ 0  6 20  1]
 [ 0 15 19 30]]
Please talk
('PREDITED  Emotion is : ', 'happy')
have a nice day

Process finished with exit code 0
```

FIG 6.2: THE FIGURE SHOWS THE PREDICTED EMOTION HAPPYFOR LIVE AUDIO TEST 2

We can see that the system uses confusion matrix and has accuracy 64.83%. The predicted emotion is happy for the phrase “have a nice day”.

CHAPTER- 07

RESULT AND DISCUSSION

7.1 RESULT

The evaluation of the speech emotion recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. The project speech emotion recognition was able to understand and detect the underlying emotion of speaker. The emotion is categorized in four types – anger, happy, neutral and sad. We used 75 % data for training and 25% for testing. The result had frequently showed anger, happy emotion than sad and neutral emotions. We had used grid confusion matrix with MLP classifier. After adding more data set of sad and neutral audios, the system was able to detect the emotion to a satisfactory level.

7.2 DISCUSSION

Some of the important research issues in speech emotion recognition are discussed below in brief.

- Majority of the research results produced on emotional speech recognition have used databases with limited number of speakers. While developing emotion recognition systems using limited speaker databases; speaker specific information may play considerable role, if speech utterances of the same speakers are used for training and testing the models. On the other hand, developed models may produce poor results, due to lack of generality, if speech utterances of different speakers are used for training and testing the models. Therefore, there is a need of larger emotional speech database with reasonably large number of speakers and text prompts. Emotion recognition studies have to be conducted on large databases in view of speaker, text and session variabilities.

- Most research on emotional speech mainly focuses on characterizing the emotions from classification point of view. Hence, the main task carried out was deriving the emotion specific information from speech, and using it for classifying the emotions. On the other hand, emotion synthesis through speech is also an important task. Here, emotion specific information may be predicted from the text, and then it has to be incorporated during synthesis. For predicting the emotion specific information, appropriate models have to be developed using sufficiently large emotional speech corpus. In emotion synthesis, the major issues are the design of accurate prediction models and preparation of appropriate emotional speech corpus.

- Expression of emotions is a universal phenomenon, which may be independent of speaker, gender and language. Cross lingual emotion recognition study may be another interesting work for further research. The emotion recognition models developed using the utterances of a particular language should yield appreciably good recognition performance for any test utterance of the other language. By using cross lingual emotion analysis, one can group the languages based on their emotional similarity.

- Majority of the work done and results produced are on recognizing speech emotions using simulated databases. Real challenge is to recognize speech emotions from natural emotions. The features and techniques discussed in the literature may be applied to the natural speech corpora, to analyze emotion recognition. Realization of this, needs the collection of good natural emotional speech corpus, covering wide range of emotions, which is another challenge.

- In real time applications such as call analysis in the emergency services like ambulance and fire brigade, verification of emotions to analyze genuineness of requests is important. In this context, under the framework of emotion verification appropriate features and models can be explored.

- Most of the today's emotion recognition systems experience high influence of speaker specific information during emotion classification. An efficient technique may be developed to remove speaker specific information from the speech utterance.

- The effect of emotion expression also depends upon the linguistic contents of the speech. Identification of emotion salient words from emotional speech, and the features extracted from these words along with other conventional features may enhance emotion recognition performance.

- More often emotion classification task is performed using single model (i.e., GMM, ANN, or SVM). Hybrid models can be explored for studying their performance in the case of emotion recognition. The basic idea behind using the hybrid models is that, they derive the evidence from different perspectives, and hence, the combination of evidence may enhance the performance, if the evidence is complementary in nature.

- The trend of emotion recognition is not clearly known in the case of many other languages. It would be helpful to evaluate the established features on different Indian languages for emotion recognition. This helps to comment on whether the methods and features used in literature are language independent? This analysis is also helpful to group languages based on their emotion characteristics, which in turn would improve the performance of language identification systems.

- The study on discrimination of emotions may be extended to the emotion dimensions (arousal, valance and power), that are derived from the psychology of production and perception of emotions. Deriving the appropriate speech features related to the emotion dimensions can be explored for further improving the recognition performance.

- Expression of emotions is a multi-modal activity. Therefore, other modalities like facial expression, bio-signals may be used as the supportive evidence along with the speech signal for developing the robust emotion recognition systems.

CHAPTER- 08

CONCLUSION & SCOPE OF FURTHER WORK

8.1 CONCLUSION

For emotion recognition system different databases are used. On the basis of ability, they have to recognize a speech recognition system can be separated in different classes are isolated, connected, spontaneous and continuous words. Relevant emotional features extraction from the speech is the second important step in emotions recognition. To classify features there is no unique way but preferably acoustic and linguistic features taxonomy is considered separately.

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc. [14].

A few possible steps that can be implemented to make the models more robust and accurate are the following: -

- An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
- Figuring out a way to clear random silence from the audio clip.
- Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.
- Following lexical features-based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
- Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

8.2 SCOPE OF FUTURE WORKS

An alternate approach that could be explored for this project is splitting the classifying task into two distinct problems. A separate model could be used to classify gender and then separate models for each gender to classify emotion could be utilized. This could possibly lead to a performance improvement by segregating the task of emotion classification by gender. Some possible features to explore concerning speech would be MFCC Filter banks or features extracted using the perceptual linear predictive (PLP) technique. These features could affect the performance of models in the emotion classification task.

The global emotion detection and recognition market size is projected to grow from USD 21.6 billion in 2019 to USD 56.0 billion by 2024, at a Compound Annual Growth Rate (CAGR) of 21.0% during the forecast period [15]. Factors such as the rising need for socially intelligent artificial agents, increasing demand for speech-based biometric systems to enable multifactor authentication, technological advancements across the globe, and growing need for high operational excellence are expected to work in favor of the market in the near future.

8.3 REFERNCES

1. Umair Ayub. Speech emotion recognition using machine learning < https://medium.com/analytics-vidhya/speech-emotion-recognition-using-machine-learning-df31f6fa8404>.
2. RAVDESS DATASET (Ryerson Audio-Visual Database of Emotional Speech and Song) < https://paperswithcode.com/dataset/ravdess>
3. Jouni Pohjalainen and Paavo Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. International Conference on Acoustic, Speech and Signal Processing, 980-984, 2014
4. Mirlab Audio Signal Processing tutorials, “Speech feature MFCC Calculation guide” <https:// mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp>
5. Mandar Gilke , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues and Madhavi Pednekar. MFCCbased vocal emotion recognition using ANN, International Conference on Electronics Engineering and Informatics, 150-154, 2012.
6. Jana Tuckova and Martin Sramka. Emotional speech analysis using Artificial Neural Networks. Proceedings of the International Multi conference on Computer Science and Information Technology, 141-147, 2010.
7. Tin Lay New, Say Wei Foo and Liyanage C. De Silva. Speech emotion recognition using Hidden Markov Models. Speech Communications 41, 603-623, 2003.
8. Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on Signal Processing and its Applications, IEEE, 2007.
9. Sirko Molau, Michael Pitz, Ralf Schl¨uter, and Hermann Ney. Computing mel-frequency cepstral coefficients on the power spectrum. IEEE Transaction, 2011.
10. Website: < https://www.python.org>
11. Ian Mcloughlin, Applied Speech and Audio Processing with MATLAB Examples. Cambridge University Press 2009. ISBN-13 978-0-521-51954-0/2.
12. https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron
13. Website: < https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model>
14. Al-Talabani, A., Sellahewa, H., & Jassim, S. A. (2015). Emotion recognition from speech: tools and challenges. Mobile Multimedia/Image Processing, Security, and Applications 2015, 9497(May 2020), 94970N.

15. Website - <https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market>