Table of Contents

CHAPTER I

INTRODUCTION

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion play in shaping human social interaction.

Emotional displays convey considerable information about the mental state of an individual. This has opened up a new research field called automatic emotion recognition, having basic goals to understand and retrieve desired emotions.

In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions , speech, physiological signals , etc. Several inherent advantages make speech signals a good source for affective computing. For example, compared to many other biological signals (e.g., electrocardiogram), speech signals usually can be acquired more readily and economically. This is why the majority of researchers are interested in speech emotion recognition (SER).

SER aims to recognize the underlying emotional state of a speaker from his/her voice. The area has received increasing research interest all through current years. There are many applications of detecting the emotion of the persons like in the interface with robots, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, etc.

Definition of Emotion:

A definition is both important and difficult because the everyday word “emotion” is a notoriously fluid term in meaning. Emotion is one of the most difficult concepts to define in psychology. In fact, there are different definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure .

Scientific discourse has drifted to other meanings and there is no consensus on a definition. Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes.

Categorization of Emotion:

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous).

In the first approach, emotions are described with a discrete number of classes. Many theorists have conducted studies to determine which emotions are basic. A most popular example is Ekman who proposed a list of six basic emotions, which are anger, disgust, fear, happiness, sadness, and surprise. He explains that each emotion acts as a discrete category rather than an individual emotional state.

In the second approach, emotions are a combination of several psychological dimensions and identified by axes. Other researchers define emotions according to one or more dimensions. Wilhelm Max Wundt proposed in 1897 that emotions can be described by three dimensions: (1) strain versus relaxation, (2) pleasurable versus unpleasurable, and (3) arousing versus subduing.

PAD emotional state model is another three-dimensional approach by Albert Mehrabian and James Russell where PAD stands for pleasure, arousal, and dominance. Another popular dimensional model was proposed by James Russell in 1977. Unlike the earlier three-dimensional models, Russell’s model features only two dimensions which include (1) arousal (or activation) and (2) valence (or evaluation) .

MOTIVATION

A major motivation comes from the desire to improve the naturalness and efficiency of human-machine interaction. Speech emotion recognition also has an extensive application prospect on such aspects as auto supervision and control of safety system. For instance, it can be used to auto remote telephone service center for discovery of dissatisfaction of customers or remote teaching to timely recognize emotions of students for proper treatment for the purpose of improving teaching quality. It can also be used to the criminal scout for auto detection of psychological state of criminal suspects and auxiliary lie detection.

OBJECTIVES

The objective is to develop a system based on Python which can extract desired features from the speech file and using the deep learning approach can identify the emotion embedded in that particular speech .

Three key issues need to be addressed for successful SER system, namely,

(1) choice of a good emotional speech database,
(2) extracting effective features, and
(3) designing reliable classifiers using machine learning algorithms.

CHAPTER II

PROPOSED WORK

Block diagram of Speech Emotion Recognition System (SER)

SER system consists of four main steps. First is the voice sample collection. The second features vector that is formed by extracting the features. Third step is to determine which features are most relevant to differentiate each emotion. These features are introduced to machine learning classifier for recognition. This process is described in Figure 1.

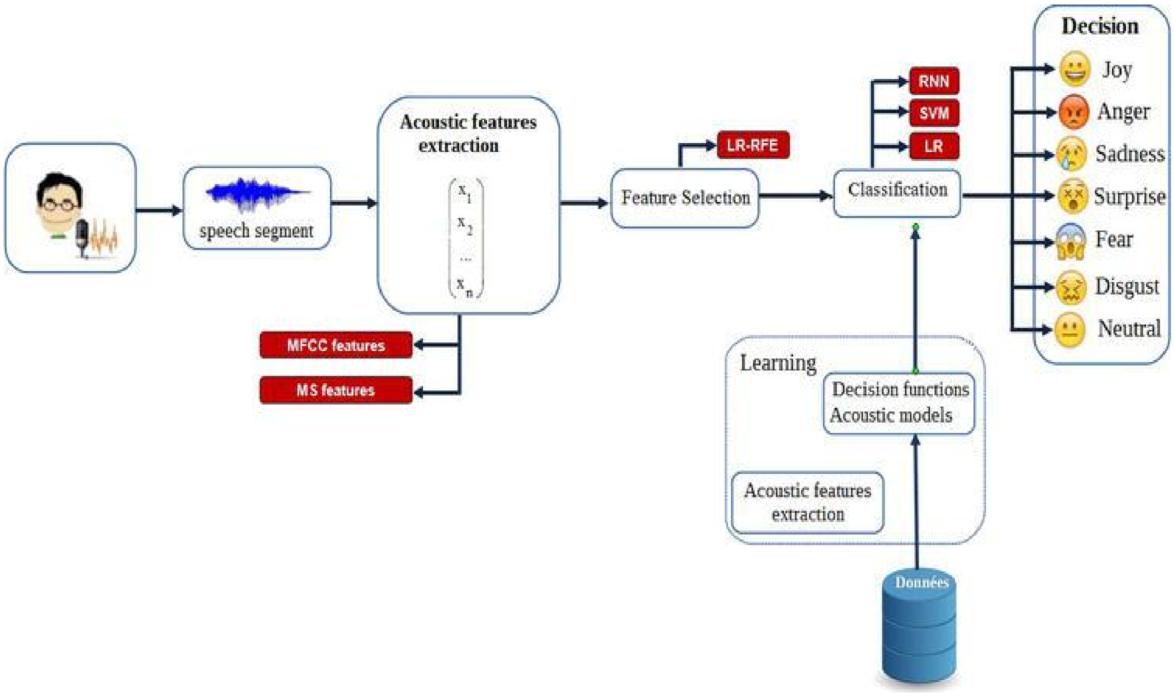


Figure 1.Block diagram of the proposed system.

The whole pipeline is as follows

- Preparing the Dataset: : download and convert the dataset to be suited for extraction.
- Loading the Dataset: This process is about loading the dataset in Python which involves extracting audio features, such as obtaining different features such as MFCC from the speech signal.
- Training the Model:The extracted features have to be mapped to respective emotion labels.
- Testing the Model: Model is then can be tested with different input and efficiency, accuracy of the developed model can be measured.

Dataset for training the model:

One of the popular dataset source for speech emotion recognition is RAVDESS.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.

Feature extraction

The emotional feature extraction is a main issue in the SER system. Many researchers have proposed important speech features which contain emotion information,such as energy, pitch, formant frequency, Linear Prediction Cepstrum Coefficients(LPCC), Mel-frequency cepstrum coefficients (MFCC), and modulation spectral features (MSFs) . Thus, most researchers prefer to use combining feature set that is composed of many kinds of features containing more emotional information .However, using a combining feature set may give rise to high dimension and redundancy of speech features; thereby, it makes the learning process complicated for most machine learning algorithms and increases the likelihood of overfitting.Therefore, feature selection is indispensable to reduce the dimensions redundancy of features.
The speech signal contains a large number of parameters that reflect the emotional characteristics.

■ MFCC:

Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration.The method of extracting MFCC as shown in figure 2.

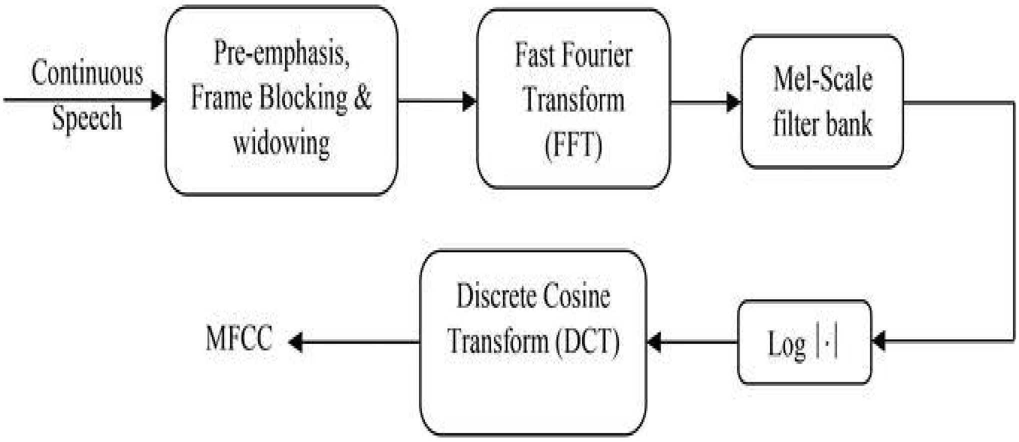

Continuous Speech
Pre-emphasis, Frame Blocking & widowing
Fast Fourier Transform (FFT)
Mel-Scale filter bank
Log |·|
Discrete Cosine Transform (DCT)
MFCC

Figure 2 .Scheme of MFCC extraction

Any sound generated by humans is determined by the shape of their vocal tract (including tongue, teeth, etc). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC (which is nothing but the coefficients that make up the *Mel-frequency cepstrum*) accurately represents this envelope.

- Chroma Features

The term ***chroma feature*** or ***chromagram*** closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing speech or music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale. One main property of chroma features is that they capture harmonic and melodic characteristics of speech/music, while being robust to changes in timbre and instrumentation. Two of the main chroma features are
(a) Chroma vector:
A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
(b) Chroma deviation*:*
The standard deviation of the 12 chroma coefficients.

- Mel Spectogram

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

It partitions the Hz scale into bins, and transforms each bin into a corresponding bin in the Mel Scale, using a overlapping triangular filters.Mel Spectrogram, is a Spectrogram with the Mel Scale as its *y* axis..An example for melspectogram is as shown in fihure 3.

The Mel Scale is used to provide greater resolution for more informative (lower) frequencies

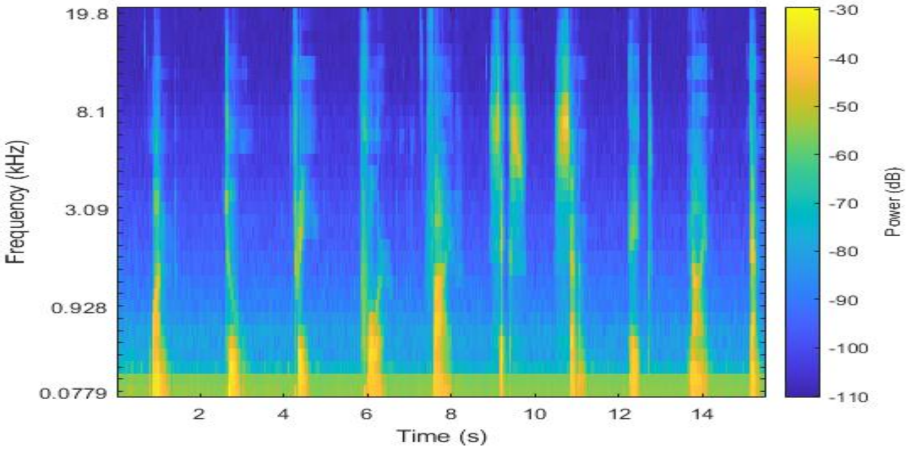


Figure 3: MEL SPECTOGRAM

- Spectral Contrast

Spectral Contrast considers the spectral peak, spectral valley and their difference in each sub-band. For most audio, the strong spectral peaks roughly correspond with harmonic components; while non-harmonic components, or noises, often appear at spectral valleys. Thus, Spectral Contrast feature could roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum

- Tonal Centroid features

Estimates tonal centroids as coordinates in a six-dimensional interval space.It allows spatial representations of tonal distance and tonal relationships.

Model/Classifier Development:

Many machine learning / Neural Network algorithms have been used for discrete emotion classification.The goal of these algorithms is to learn from the training samples and then use this learning to classify new observation. In fact, there is no definitive answer to the choice of the learning algorithm; every technique has its own advantages and limitations.

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Support Vector Machines (SVM) are also can be used. It is an optimal margin classifier in machine learning. It is also used extensively in many studies that related to audio emotion recognition .It can have a very good classification performance compared to other classifiers especially for limited training data.

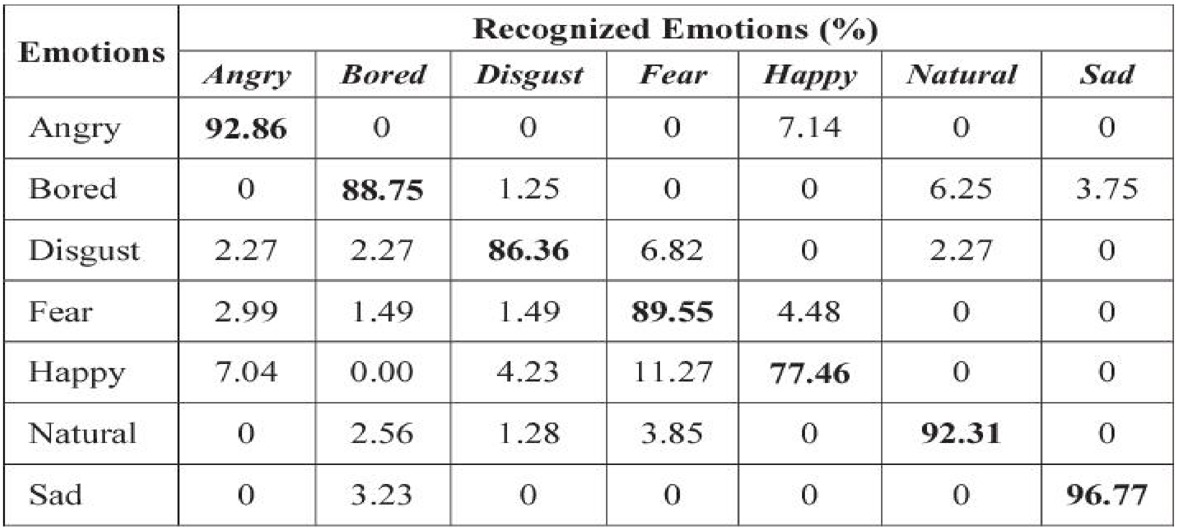
The developed Classifier can be to used to trained with the extracted features and their corresponding labels so that it can predict the emotion present in future input.

PROJECT OUTCOME:

SER system based on neural network / machine learning classifier analyses the given input speech file and extracts the specified features. Then it makes a prediction based on its prior training with the dataset containing huge number of audio samples and current extracted features . It chooses one of the labels from the emotion set:

"neutral", "calm", "happy", "sad","angry","fearful", "disgust","surprised"

The confusion matrix from the results of SER developed by :Thapanee S et.al from Khon Kaen University of Thailand is as shown in Figure 4.



Emotions	**Recognized Emotions (%)**						
	Angry	***Bored***	***Disgust***	***Fear***	***Happy***	***Natural***	***Sad***
Angry	**92.86**	0	0	0	7.14	0	0
Bored	0	**88.75**	1.25	0	0	6.25	3.75
Disgust	2.27	2.27	**86.36**	6.82	0	2.27	0
Fear	2.99	1.49	1.49	**89.55**	4.48	0	0
Happy	7.04	0.00	4.23	11.27	**77.46**	0	0
Natural	0	2.56	1.28	3.85	0	**92.31**	0
Sad	0	3.23	0	0	0	0	**96.77**

Figure 4. Confusion Matrix :SVM classifier on BERLIN dataset

APPLICATIONS:

- Speech emotion recognition has an extensive application prospect on such aspects as man-machine interaction and auto supervision and control of safety system.
- it can be used to auto remote telephone service center for discovery of dissatisfaction of customers or remote teaching to timely recognize emotions of students for proper treatment for the purpose of improving teaching quality.
- It can also be used to the criminal scout for auto detection of psychological state of criminal suspects and auxiliary lie detection
- The medical field: In the world of telemedicine where patients are evaluated over mobile platforms, the ability for a medical professional to discern what the patient is actually feeling can be useful in the healing process.

CHAPTER III

LITERATURE SURVEY:

Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems.

Cao et al. [2] proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification. This ranking method, instruct SVM algorithms for particular emotions, treating data from every utterer as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM achieves two advantage, first, for training and testing steps in speaker- independent it obtains speaker specific data. Second, it considers the intuition that each speaker may express mixed of emotion to recognize the dominant emotion. Ranking approaches achieves substantial gain in terms of accuracy compare to conventional SVM in two public datasets of acted emotional speech, Berlin and LDC.

Lee et al. [3] represent a hierarchical computational structure to identify emotions. This method via following layers of binary classifications, maps input speech signal in one of the corresponding emotion classes. The main concept of different level in tree is to solve the classification task in easiest way to diminish error propagation. AIBO and USC IEMOCAP databases are employed to evaluate the classification method. Over the baseline SVM, the absolute result improve the accuracy archives an absolute improvement of 72.44%- 89.58%. The consequence proves the reported hierarchical method is efficient for classifying emotional speech in various databases [3].

Albornoz et al. [4] investigate a new spectral feature in order to determine emotions and to characterize groups. In this study based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75% [4].

Dai et al. [1] proposed a computational approach for recognition of emotion and analysis the specifications of emotion in voiced social media such as Wechat. This approach approximate the mixed emotion and dynamic fluctuations in position- arousal-dominance (PAD) by extracting 25 acoustic features of speech signals and employing trained least squares-support vector regression (LV-SVR) model as well. The experimental results demonstrates the recognition rate for different emotion are different and the average rate of recognition achieves 82.43%, which is the best existing result by similar examination [1].

Thus the literature survey points out that the accuracy achieved in the detection of the speech emotion hugely depends on the features extracted from the speech for analysis. The number of features chosen for training the model must be minimum but also those features must contain most of the speech information related to emotion aspect.

The recognition accuracy can be improved by testing with dataset with variations,and experimenting with different features used for recognition by the model.Enhancement of the robustness of emotion recognition system is possible by combining databases and by fusion of classifiers. The effect of training multiple emotion detectors can be investigated by fusing these into a single detection system

CHAPTER IV

REQUIREMENTS:

◆ Python installed in the machine with following modules

- Librosa : Package for audio and music signal processing
- Numpy : It is a general-purpose array-processing package. It provides a high performance multidimensional array object and tools for working on them.
- Soundfile : Library to read the audio file
- Scikit-learn : Library for data mining and data analysis containing efficient tools for machine learning and statistical modelling
- TensorFlow : It is symbolic math library,and is also used for machine learning applications such as neural networks.
- PyAudio : Library to play and record audio on variety of platforms.
- Pandas : Library for data manipulation and analysis.
- Matplotlib :Visualization library for 2-D plotting of arrays
- Seaborn :Library for making statistical graphics.

CHAPTER V

FEASIBILITY:

A speech emotion detection systems accuracy fully depends on the datasets used for training the model, features used and the classifier algorithm.

If the number of features selected is very large ,there is a chance that the overfitting of features occurs. Also computational time and memory storage for processing the data increases.It should be noted that all features may not contain the useful speech information

The objective of feature selection in ML is to "reduce the number of features used to characterize a dataset so as to improve a learning algorithm's performance on a given task." The objective will be the maximization of the classification accuracy in a specific task for a certain learning algorithm; as a collateral effect, the number of features to induce the final classification model will be reduced. Feature selection (FS) aims to choose a subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to higher recognition accuracy . It can drastically reduce the running time of the learning algorithms.

Both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage.

CHAPTER VI

PLANNING:

PERT CHART: The project workflow for developing SER system is as shown in Figure 3

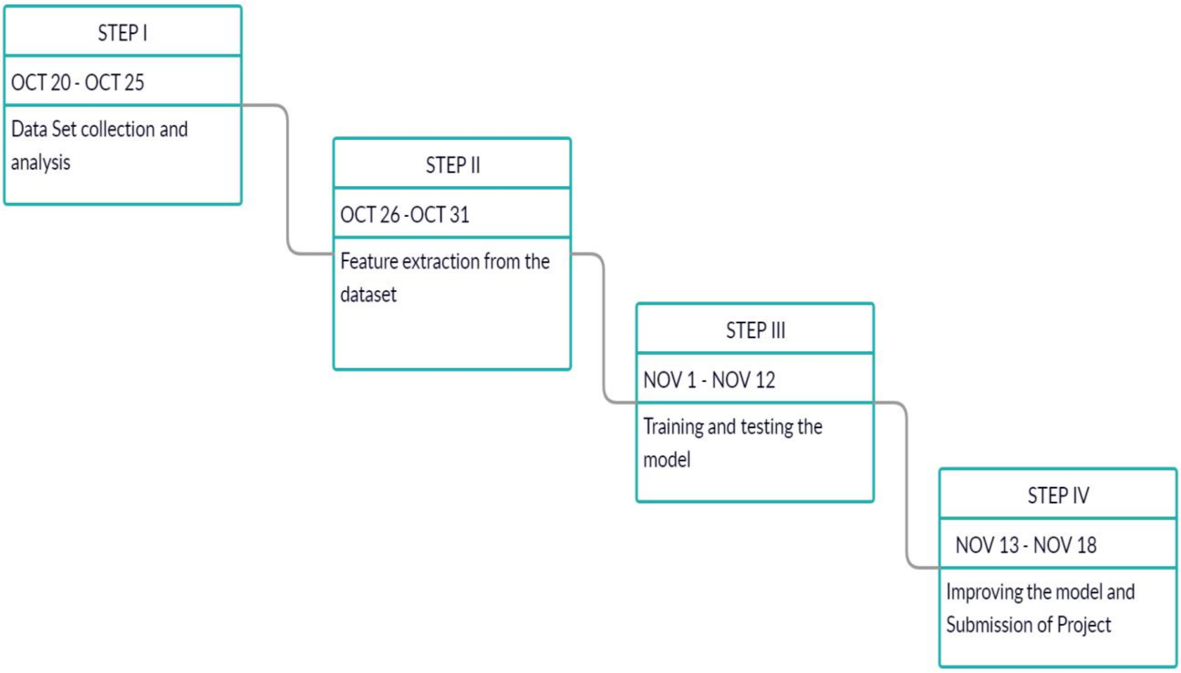

STEP I
OCT 20 - OCT 25
Data Set collection and analysis
STEP II
OCT 26 -OCT 31
Feature extraction from the dataset
STEP III
NOV 1 - NOV 12
Training and testing the model
STEP IV
NOV 13 - NOV 18
Improving the model and Submission of Project

Figure 3.Pert Chart

REFERENCES:

[**1**]. W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing on Vocal Social Media,”
Inf. Manag., Feb. 2015

[**2**]. H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and
spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[**3**]**.**C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binarYdecision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011

[**4**]. E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,”
Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011

[**5**].’Speech Emotion Recognition’S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh
Dept. of ECE, Amrita School of Engineering, Amrita Vishwa Vidyapeetam, Bangalore, India
IEEE- 2014 International Conference on Advances in Electronics Computers and Communications

[6]Speech emotion recognition based on deep belief network
IEEE- 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)
[**7**].Speech emotion recognition using Support Vector Machines,Authors:Thapanee Seehapoch and Sartra Wongthanavasu
Published in: 2013 5th International Conference on Knowledge and Smart Technology (KST)