

A PROJECT REPORT ON
“SPEECH EMOTION RECOGNITION”

Submitted

In the partial fulfilment of the requirements for

the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

By

V. VEERA BRAHMA CHARI (171FA04308)

Y.ANAND (171FA04315)

Under the guidance of

Mr. Modigari Narendra, Assistant Professor



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH

(Accredited by NAAC “A”grade)

Vadlamudi, Guntur.

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH

DEEMED TO BE UNIVERSITY

(Accredited by NAAC“A”grade)



CERTIFICATE

This is to certify that the Project Report entitled “**SPEECH EMOTION RECOGNITION**” that is being submitted by **V.Veera Brahma Chari (171FA04308)**, and **Y.Anand (171FA04315)** in partial fulfilment for the award of B.Tech degree in Computer Science and Engineering to the Vignan’s Foundation for Science, Technology and Research, Deemed to be University, is a record of bonafide work carried out by them under the supervision of **Mr. Modigari Narendra**.

Mr. Modigari Narendra
Asst. Professor

External Examiner

Dr. Venkatesulu Dondeti
HOD, CSE

DECLARATION

I hereby declare that the project entitled “**SPEECH EMOTION RECOGNITION**” submitted for the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**. This dissertation is our original work and the project has not formed the basis for the award of any degree, associate-ship and fellowship or any other similar titles and no part of it has been published or sent for publication at the time of submission.

By

V.Veera Brahma Chari

Y.Anand

Date:

ACKNOWLEDGEMENT

Semester long project is a golden opportunity for learning and self-development. Consider myself very lucky and honored to have so many wonderful people lead me through in the completion of this project.

We feel it is our responsibility to thank **Mr. Modigari Narendra** under whose valuable guidance that the project came out successfully after each stage, and also it is our responsibility to extend our thanks to **Mr. D. Yakobu, Mr. P. Vijaya Babu, Dr. Balaji Vijayan, Department Project Coordinators**, for extending their support towards the completion of project in **Department of CSE, in VFSTR**.

It is a great pleasure for us to express our sincere thanks to **Dr. Venkatesulu Dondeti, HOD, CSE** of VFSTR Deemed to be University, for providing me an opportunity to do our project at Department of CSE.

It is our pleasure to extend our sincere thanks to **Vice-Chancellor Dr. M.Y.S. Prasad** and we are very grateful to our beloved **Chairman Dr. Lavu. Rathaiah**, and **Vice Chairman Mr. Lavu. Krishna Devarayalu**, for their love and care.

We extend our whole hearted gratitude to all our **faculty members** of Department of Computer Science and Engineering who helped us in our academics throughout course.

Finally we wish to express thanks to our family members for the love and affection overseas and forbearance and cheerful depositions, which are vital for sustaining effort, required for completing this work.

With Sincere regards,

V.VEERA BRAHMA CHARI (171FA04308)

Y.ANAND (171FA04315)

TABLE OF CONTENTS

ACKNOWLEDGEMENT.....	4
LIST OF ABBREVIATIONS.....	6
ABSTRACT.....	7
1. INTRODUCTION.....	8
2. LITERATURE SURVEY	10
3. PROPOSED METHODOLOGY.....	15
3.1 SPEECH INPUT	15
3.2 FEATURE EXTRACTION.....	16
3.2.1 PROSODIC FEATURES.....	16
3.2.2 SPECRAL FEATURES.....	16
3.3 FEATURE SELECTION.....	16
3.4 CLASSIFICATION.....	17
3.5 ADVANTAGES.....	17
3.6 DISADVANTAGES.....	17
4. MODELS USED.....	18
4.1 DEEP LEARNING MODELS.....	18
4.2 ARTIFICIAL NEURAL NETWORKS.....	18
4.2.1 HOW ANN WORKS.....	18
4.2.2 ACTIVATION FUNCTION.....	19
4.3 MODES IN PERCEPTRON.....	19
4.4 APPLICATIONS OF ANN.....	20
4.5 MULTILAYER PERCEPTRON.....	20
4.5.1 STEPS TO UILD MLPCLASSIFIER.....	20
4.5.2 LAYERS IN MLP.....	21
5. DATASETS AND PACKAGES USED.....	22
5.1 DATASET USED.....	22
5.2 PACKAGES USED.....	24
6. SOFTWARE REQUIREMENTS.....	25
7. HARDWARE REQUIREMENTS.....	25
8. PERFORMANCE ANALYSIS.....	26
8.1 APPLICATION 1.....	26
8.2 APPLICATION 2.....	27
9. CONCLUSION.....	28
10. REFERENCES.....	29
11. APPENDIX.....	30
10.1 SOURCE CODE.....	30
10.1.1 APPLICATION 1.....	30
10.1.2 APPLICATION 2.....	33

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
ANN	: Artificial Neural Network
DL	: Deep Learning
SVM	: Support Vector Machine
KNN	: K-Nearest Neighbour
MLP	: Multilayer Perceptron
RAVDESS	: Ryerson Audio Visual Database of Emotional Speech and Song
PCA	: Principle Component Analysis
HMM	: Hidden Markov Model
GMM	: Guassian Mixture Model
MFCC	: Mel Frequency Cepstrum Coefficients
LPC	: Linear Predictive Coding
PLP	: Perceptual Acoustic Features

ABSTRACT

SPEECH EMOTION RECOGNITION is where emotions can be recognized from the speech. Speech is the most normal way to express yourself as human beings. Extending this means of communication to computer applications is only inevitable then. It describes speech emotion recognition (SER) systems as a set of methodologies that process and classify voice signals to detect the emotions embedded. It is used an MLPClassifier for this and made use of the sound file library to read the sound file, and the librosa library to extract features from it. Since emotions help us understand each other better, applying this understanding to computers is a natural outcome. Thanks to the smart mobile devices that are able to recognize and respond to voice commands with synthesized speech, speech recognition is already in our everyday lives. Recognition of speech emotions (SER) may also be used to enable them to detect our emotions.

CHAPTER 1

INTRODUCTION

Speech Emotion Recognition is a software used to recognize the emotions of humans. Attributes of human voice such as pitch, timbre, loudness and tone make human voice versatile for communication. It can be observed that humans can convey their emotions, even by changing the specified characteristics. This helps the human emotion to be defined by speech analysis. Speech Emotion Recognition recognizes the various emotions like happy, sad, anger, and many more.

- **Fear:** emotion comes with an unpleasant situation caused from pain, Anger or feeling afraid.
- **Anger:** involves a strong feeling of aggravation, uncomfortable situation stress, displeasure, or hospitality.
- **Sadness:** A feeling caused with disadvantage or loss due to anything.
- **Joy :** feeling happy. Other words are happiness, gladness.
- **Disgust:** A feeling with strong disapproval, nasty, dislike
- **Surprise:** occurred with an unexpected event or shock .

The tonal quality not only changes with different emotions and moods but the associated patterns of speech also shift. For example, when they are angry, people may tend to speak loudly and use shrill or high pitched voices while they are in an emotional state of fear or panic. Many people are likely to ramble when they are nervous or excited. Sound speech characteristics should be used in cases where face to face communication is not possible or where there is no readily accessible language constraint and proper model for lexicon based speech analysis.

Following are many conditions where speech characteristics can be used as a means of classifying human emotions:

- i. Play music and change ambient room lighting to the sound of the conversation.
- ii. Carrying out social science work
- iii. Customer service centers may gain insight into customer loyalty by actually

hearing the voice of their customers. Therefore, the scores collected as part of this analysis will be used to assess the overall opinion of a firm / product / services.

Speech Emotion recognition is particularly useful for applications which require man machine communication such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion.

Speech emotion recognition has also been used in call center applications and mobile communication. It is also useful for in car board system where information of the mental state of the driver may be provided to the system to initiate his / her safety.

The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice. So, it helps to know the state of speaker's emotion and results in best experience to the user.

CHAPTER 2

LITERATURE SURVEY

Over the last years, an excessive investigation has been completed to recognize emotions by using speech statistics refers to table 2.1.

Narayanan[1] proposed domain specific emotion recognition by utilizing speech signals from call center application. Detecting negative and nonnegative emotion (e.g. anger and happy) are the main focus of this research. Different types of information include acoustic, lexical, and discourse are used for emotion recognition. Both K-NN and linear discriminant classifier are used to work with different types of features. Experimental result confirms that the best results are achieved by combination of acoustic and language data. Outcomes demonstrates by combining three information source instead of one source, classification accuracy increases by 40.7% for males and 36.4% for females.

Yang & Lugger [2] presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory. First, beginning from predicted pitch of a speech signals, then computing spherical autocorrelation of pitch histogram. It calculate the incidence of dissimilar two pitch duration, which cause a harmonic or inharmonic impression. In Classification step, Bayesian classifier plays an important rule with a Gaussian class conditional likelihood. Experimental result in Berlin emotion database by using harmony features indicate an improvement in recognition performance. Recognition rate improved by 2% in average.

Cao et al.[3] 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 multiclass 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. In both acted data and the spontaneous data, which comprises neutral intense emotional utterances, ranking based SVM achieved higher accuracy in recognizing emotional utterance.

Chen et al.[4] aimed to improve speech emotion recognition in speaker independent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is an input parameters for multilevel SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Fisher is better than PCA and for classification, SVM is more expansible than ANN for emotion recognition in speaker independent is.

Nwe et al.[5] proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficients (LFPC) and discrete HMM to characterize the speech signals and classifier respectively. This method classified the emotion into six different categories then used the private dataset to train and test the new system. In order to evaluate the performance of the proposed method, LFPC is compared with the mel frequency Cepstral coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Result demonstrate the average and best classification accuracy achieved 78% and 96% respectively.

Rong et al.[6] presented an ensemble random forest to trees (ERFTrees) method with a high number of features for emotion recognition without referring any language remains an unclosed problem. This method is applied on small size of data with high number of features. ERFTrees performs better than popular dimension reduction methods such as PCA and multidimensional scaling (MDS) and recently developed ISOMap. The best accuracy with 16 features for female dataset achieved the maximum correct rate of 82.54%, while the worst is only 16% on 84 features with natural data set.

Albornoz et al.[7] 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. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75%.

Yeh et al.[8] proposed a segment based method for recognition of emotion in Mandarin speech. This approach is contain the following process. First, define the k parameter in weighted discrete K-NN classifier, the experimental testing of different k shows the best performance for K-NN is when k sets to 10. For selecting the foremost feature set, sequential forward selection (SFS) and sequential backward selection (SBS) are employed. SFS and SBS improves feature accuracy to 84% and 82% respectively. The highest accuracy in segment based method achieves 86%.

ai et al.[9] proposed a computational approach for recognition of emotion. 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 (LVSVR) model as well. The experimental results demonstrates the recognition rate for different emotion are different and the average rate of recognition achieves 82.43%

El Ayadi et al.[10] proposed a Guassian mixture vector autigressive(GMVAR) approach, which is mixture of GMM with vector autogressive for classification problem of speech emotion recognition. Berlin emotional dataset was used for evaluation of GMVAR. The experimental result shows classification accuracy achievees 76% when for HMM reached 71%, for K-NN 67% and 55% for feed forward neural networks.

REFERENCE	TYPE OF CLASSIFIER	TYPE OF FEATURES	TYPE OF DATASET	METHODS	RESULT
S.S.Narayanan[1]	Narayanan NN & linear discriminate	K- frequency(F0), energy duration	Private speech database from call centre	emotion recognition by KNN and linear discriminate classifier.	40.7% for males 36.4% for females
B.Yang and M.Lugger[2]	Bayesian learning framework	Energy, Pitch statistics, duration and zero crossing rate(ZCR)	Berlin emotion dataset	Harmony features with Bayesian classifier with Gaussian class	Best for sadness 87.6% and Overall 2% improvement
H.Cao, R.Verma and A.Nenkova[3]	SVM	SVM Prosodic and spectral features	Berlin & LDC	Ranking SVM	44.4%
L.Chen, X.Mao,Y,Xue and L.L.Cheng [4]	SVM & ANN	Energy, ZCR, pitch, spectrum cutoff frequency, correlation density	Beihang University Database of Emotional Speech (BHUDES)	Multilevel SVM classifier & ANN	86.5%,68.5% and 50.2% for Different level.
T.L.Nwe, S.W.Foo and L.C.De Silva [5]	HMM	Log Frequency Power Coefficients (LFPC) , MFCC	Two private speech dataset	Discrete HMM and LFPC	Avarage and best result is 78% and 96%.
J.Rong, G.Li and Y.P.P.Chen [6]	Decision tree & random forest	Linguistic, Spectral related, Tone based and /or vowel related features.	Chinese emotional dataset	Ensemble random forest to trees(ERFT trees)	Best 82.54% & worst 16%
E.M.Alborno, D.H.Milone And H.L.Rufiner[7]	HMM,GMM, MLP and hierarchial model	Mean of the Log spectrum (MLS), MFCCs and prosodic features	Berlin dataset	Spectral characteristics are used to group emotions based on acoustic rather than psychological	HMM 68.57%, Hierarchial model 71.75%

				considerations.	
J.H.Yeh, T.L.pao, C.Y.Lin, Y.W.Tsai and Y.T.Chen[8]	K-NN	Jitter, LPC, LPCC, MFCC, LFPC,PLP and Rasta PLP	Chinese emotional speech corpus we invited 18 males and 16 females.	Segment based method by employing KNN, Sequential forward selecting (SFC).	Best 86%

Table 2.1

Generally, analysis on Speech Emotion Recognition mainly aims to improve the recognition rate as well as accuracy. Table 2.1 presents the current available methods which are targeted the speech emotion recognition and these are evaluated with its classifier, features set and recognition rate on different dataset levels. Normally, SVM is employed as alone and also along with the combination of other classifier such as ANN and RBF to reduce dimensionality. It shows most accuracy according to Table 2.1, where it use the log frequency power coefficient (LEPC) and MFCC features set and achieve more than 85%. Besides SVM classifier, K-nearest neighbor (K-NN) is also one of the well-known classifier in speech emotion recognition. Yeh et al. [8] have proven that the best accuracy achieved is 86% of recognition rate using K-NN method. Nwe et al. [5] have proven that the best accuracy achieved is 96% of recognition rate by using HMM classifier and adopting LFPC as a feature parameters.

CHAPTER 3

PROPOSED METHODOLOGY

Speech Emotion Recognition, abbreviated as SER. Speech is a complex signal, It contains the information regarding the message, the speaker, the language and the emotions. Emotion makes speech more attractive, more effective and more expressive. So, Speech Emotion Recognition(SER) means to understand emotional state of a human by extracting features from his/her voice.

Speech Emotion Recognition contains five main modules as seen in figure 3.1 below.

1. Speech input
2. Feature Extraction
3. Feature Selection
4. Classification
5. Recognized Emotion

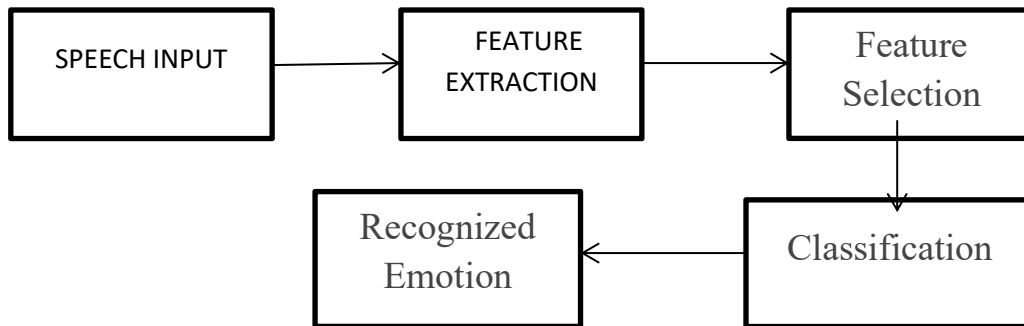


Figure 3.1: Modules in SER

3.1 SPEECH INPUT:

The evaluation of the speech emotion recognition system is based on the level of naturalness of the dataset which is used as an input to the speech emotion recognition system. If the inferior dataset is used as an input to the system then incorrect conclusion may be drawn. The dataset as an input to the speech emotion recognition system may contain the real world emotions or the acted ones.

3.2 FEATURE EXTRACTION AND SELECTION:

An extraction of these speech features which represents emotions is an important factor in speech emotion recognition system. Feature extraction means extracting the desirable features to extract emotion of human from the speech.

There are mainly two types of features.

1. Prosodic Features
2. Spectral Features

3.2.1. Prosodic Features:

Prosodic features are never deals with what we speak, but it deals with how we speak i.e., the loudness, the pitch, the energy, the stress or the rhythm given to speech we dealing.

3.2.2. Spectral Features:

Spectral features are frequency based features. Mel Frequency Cepstrum Coefficient(MFCC), Linear Predictive Coding(LPC), Perceptual Acoustic Features(PLP) are the some popular techniques to determine or to extract features from the speech.

- Typically anger has a higher mean value and variance of pitch and mean value of energy.
- In the happy state there is an improvement in mean value, variation range and variance of pitch and mean value of energy.
- On the other hand the mean value, variation range and variance of pitch is decreases in sadness, also the energy is weak, speak rate is slow and decrease in spectrum of high frequency components.
- The feature of fear has a high mean value and variation range of pitch, improvement of spectrum in high frequency components. Therefore statistics of pitch, energy and some spectrum feature can be extracted to recognize emotions from speech.

FEATURES USED:

MFCC:

MFCC is that it is taken on a Mel scale which is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear.

Syntax : mfcc(audioIN, sample rate, value)

CHROMA:

A Chroma vector is typically a 12element feature vector indicating how much energy of each pitch class is present in the signal in a standard chromatic scale.

Syntax: chroma_stft(stft , sample rate)

MEL:

Mel spectrogram plots amplitude on frequency vs time graph on a “Mel” scale. As the project is on emotion recognition, a purely subjective item, we found it better to plot the amplitude on Mel scale as Mel scale changes the recorded frequency to “perceived frequency”.

Syntax : melspectrogram(audioIN, sample rate)

3.4 CLASSIFICATION:

Once the Features are extracted using extracting techniques, next step is to map these features to corresponding emotions using classifiers. Gaussian Mixture Model(GMM), Hidden Markov Model(HMM), Artificial Neural Networks(ANN), Support Vector Machine(SVM), are some of the traditional classifiers to extract emotions from the speech.

3.5 ADVANTAGES:

- Helps HR and Employees of any company to manage stress levels. This will create healthy work environment in the company and increase productivity.
- It used in customer care service centers to know the customer emotion and helps in finding the best way of answering and clearing the ticket raised.
- It is used in security, medicine, education and so on.

3.6 DISADVANTAGES:

- 1.It is a challenge to make emotion available in different languages.
- 2.It cannot work efficiently in noise environment.
- 3.Privacy
- 4.Overlapping of utterances

CHAPTER 4

MODELS USED

4.1 DEEP LEARNING MODELS

This section is about what is deep learning and Artificial Neural Networks which performs both Feature Extraction and Classification and some popular training models that are used for Feature Extraction.

DEEP LEARNING:

It is an advanced field of Machine Learning that uses the concepts of Neural Networks to solve highly computational use cases that involves the analysis of multidimensional data. The main use of this deep learning is it directly automates the process of feature extraction, classification asking sure that very minimal human intervention is needed.

4.2 ARTIFICIAL NEURAL NETWORKS:

Artificial Neural Network is basically a computing system that is designed to simulate the way the human brain analysis and process the information. These are have self learning capabilities, that enable it to produce better results as more data become available. So, if you train the network on more data it will be more accurate. We can configure the neural network for specific applications also i.e., Pattern Recognition, data Classification etc.

4.2.1 How Artificial Neural Network Works:

This is basically a Artificial Neuron . This Artificial Neuron is also called as Perceptron. The combination of multiple Perceptrons is called as a Artificial Neural Network. For details refer below figure 4.2.1.1

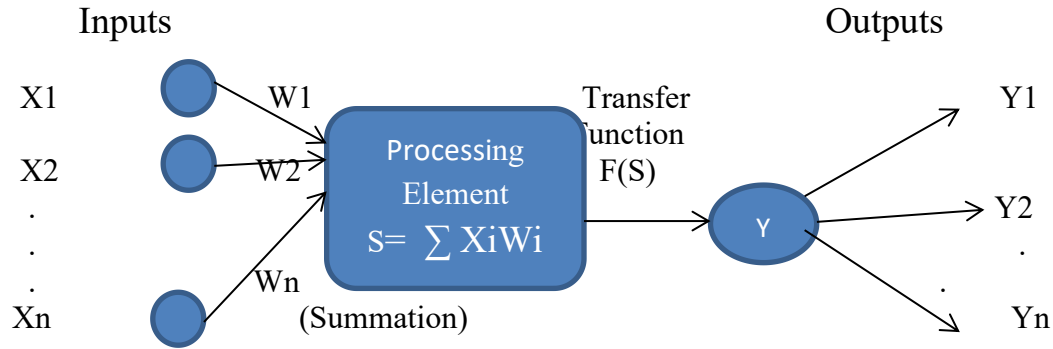


Figure 4.2.1.1: Perceptron in ANN

- Firstly we have multiple inputs X_1, X_2, \dots, X_n and we have weightage ($X_1=W_1, X_2=W_2, \dots, X_n=W_n$).
- Calculate the weightage sum of these inputs and pass it to Activation Function.

$$S = \sum X_i W_i$$

- Activation Function provides a Threshold Value.

$$S = W^T \cdot X = \sum_{i=1}^n W_i X_i$$

$$F(S) = (0 \text{ if } z < 0) \text{ or } (1 \text{ if } z \geq 0)$$

$$F(S) = F(W^T \cdot X)$$

- With that threshold value our output neuron will fire otherwise doesn't fire.

4.2.2 Activation Function :

The main of this function is to map the weighted sum to the output. Activation Functions such as Step Function, Sigmoid Function and Sign Function are the some of examples of transmission.

4.3 MODES IN PERCEPTRON:

Training Mode:

In the training Mode, the neuron can be trained to fire(or not), for a particular input patterns.

Using Mode:

In the Using mode, when a taught input pattern is detected as the input, its associated output becomes the current output.

4.4 APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS:

- Modeling and Diagnosing the Cardiovascular System
- Marketing

4.5 MULTILAYER PERCEPTRON CLASSIFIER:

Multilayer perceptron is applied for supervised learning problems. Multilayer Perceptron classifier relies on an underlying neural network to perform classification as shown in the figure 4.5.1.1.

4.5.1 Steps to build the MLP classifier:

- Initialize the MLP Classifier by defining and initiating the required Parameters.
- Data is given to the Neural Network to train it.
- The trained network is used to predict the output.
- Calculate the accuracy of the predictions.

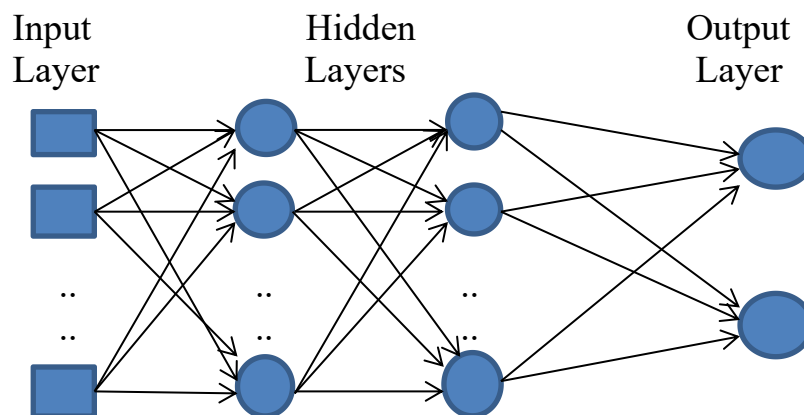


Figure 4.5.1.1: Multilayer Perceptron

4.5.2 A Multilayer Perceptron is a partition of three layers. They are:

1. *Input Layer:*

Input Layer is used to receive the input signal. This layer will take as input the three features Mel Frequency Cepstral Coefficients(MFCC), Mel Spectrogram Frequency and Chroma.

2. *Hidden Layer:*

The layers present between the input and output layer is called “Hidden Layer”. There can be many hidden layers, the number of hidden layers can be changed as per the requirement. This layer uses an activation function to act upon the input data and to process the data.

3. *Output Layer:*

Output Layer is used to make predictions or decisions for the given input. This layer classifies and gives the output of the predicted emotion, according to the computation performed by the hidden layer.

CHAPTER 5

DATASETS AND PACKAGES USED

This project aims to classify different types of emotions such as sad, happy, neutral, angry, disgust, surprised, fearful and calm. In this project, the emotions in speech are predicted using neural networks. This project uses MultiLayer Perceptron (MLPClassifier) for the classification of emotions. RAVDESS[11] (Ryerson Audio-Visual Database of Emotional Speech and Song Dataset) is the dataset used in this project to predict the emotions.

RAVDESS dataset has recordings of 24 actors (i.e., 12 male actors and 12 female actors), the actors are numbered from 01 to 24 as shown in the figure 5.1. The odd numbered are male actors and even numbered are female actors. The emotions contained in the dataset are sad, happy, neutral, angry, disgust, surprised, fearful and calm expressions. Thus, the part of the RAVDESS, that contains 60 trials for each of the 24 actors as shown in the figure 5.2, the we have 1440 files in total. The dataset is labelled in accordance with the decimal encoding and every file has a unique filename. The filename is made up of 7-part numerical identifier. the third numerical part of the filename denotes a label to the corresponding emotion. The emotions are labelled as follows: 01-'neutral', 02-'calm', 03-'happy', 04-'sad', 05-'angry', 06- 'fearful', 07-'disgust', 08 -'surprised'.

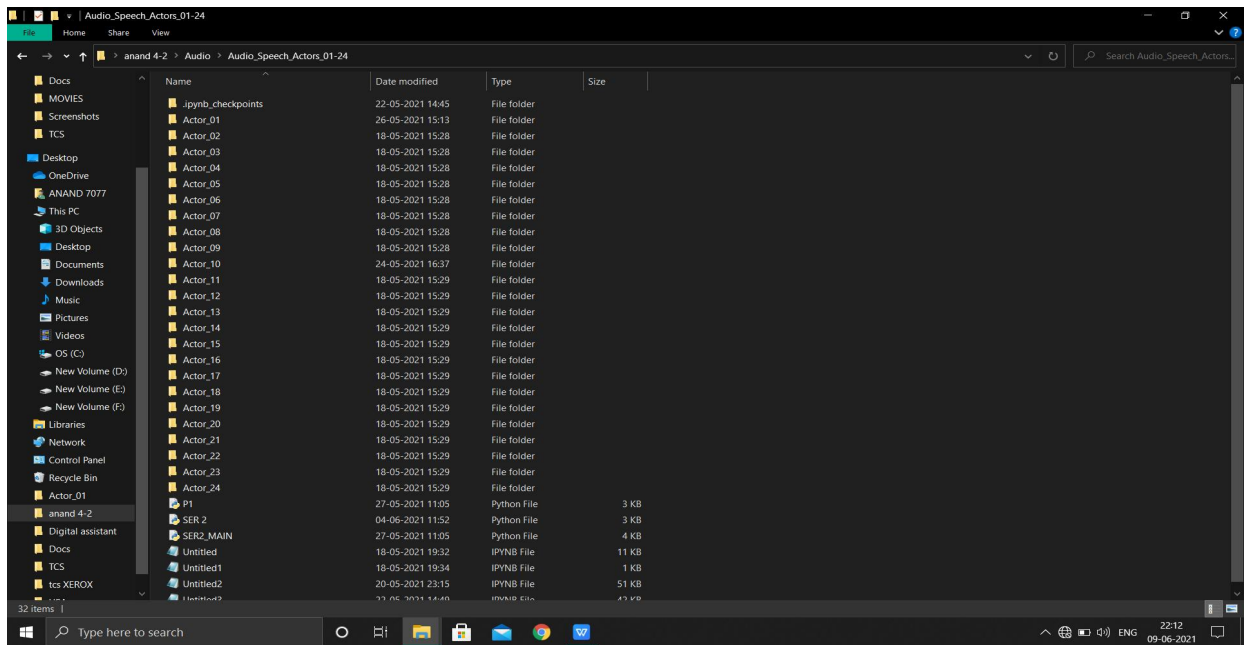


Figure 5.1

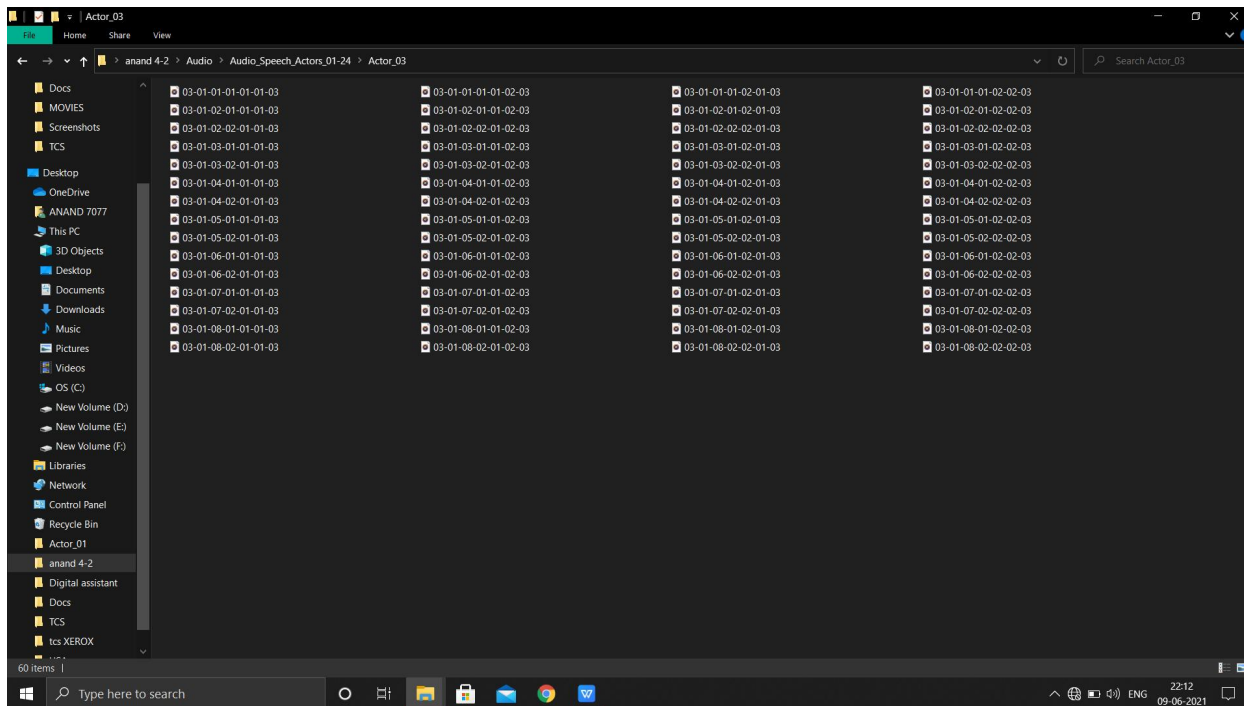


Figure 5.2

Our own Audio Dataset as follows:

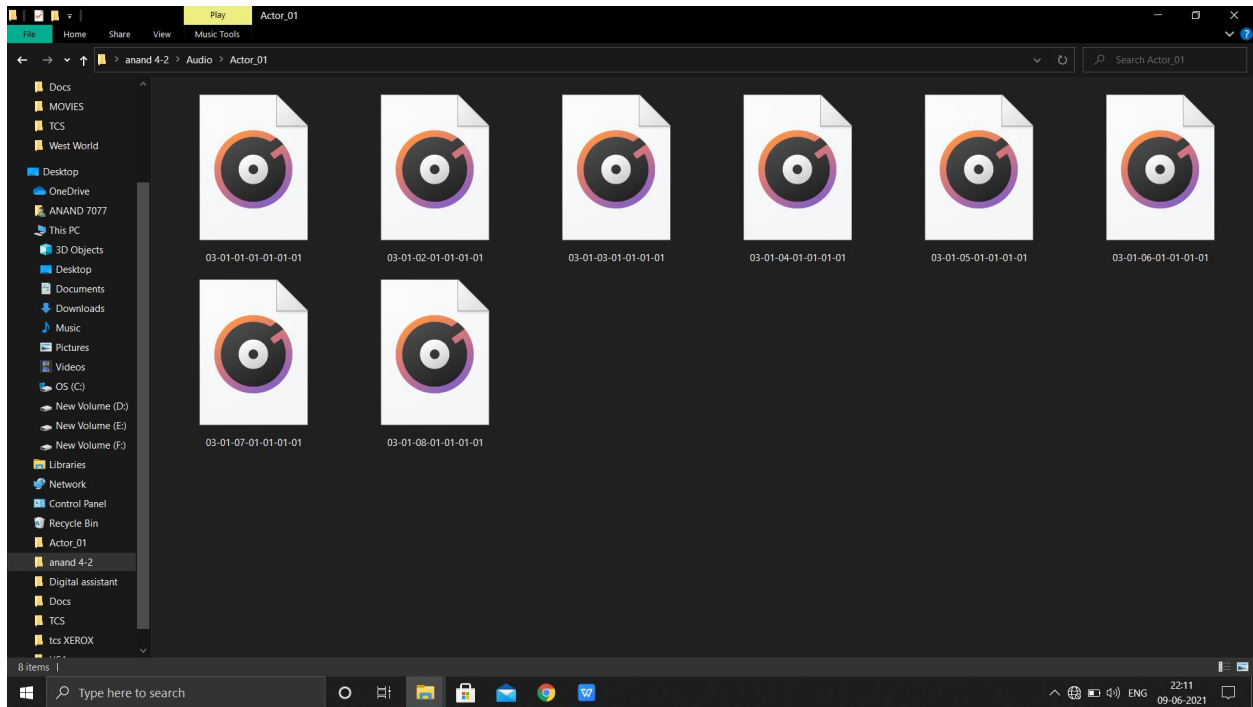


Figure 5.3

5.4 PACKAGES USED:

Librosa:

Librosa is a python library for analyzing audio and music. It provides the building blocks necessary to create music information retrieval systems. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

Numerical Python (Numpy):

Numpy is a fundamental package use or mathematical and numerical analysis. It is a fast, flexible container for large datasets in python.

Sound file:

Sound file is a python library used to read the sound file.

Glob:

Glob module is used to retrieve the all file paths that match a specific pattern.

Sci-kit Learn (sklearn):

Sci-kit Learn module is used to build machine learning models. This library contains a lot of efficient tools for machine learning and statistical modeling including classification , regression, clustering and dimensionality reduction.

Pickle:

Pickle is a python module used to serialize a python object into a binary format and deserialize it back to python object.

Pydub:

Pydub package is able to read and save wav file, but we need some type of audio package to actually play sounds.

6. SOFTWARE REQUIREMENTS:

1. Python
2. Visual Studio code or Anaconda

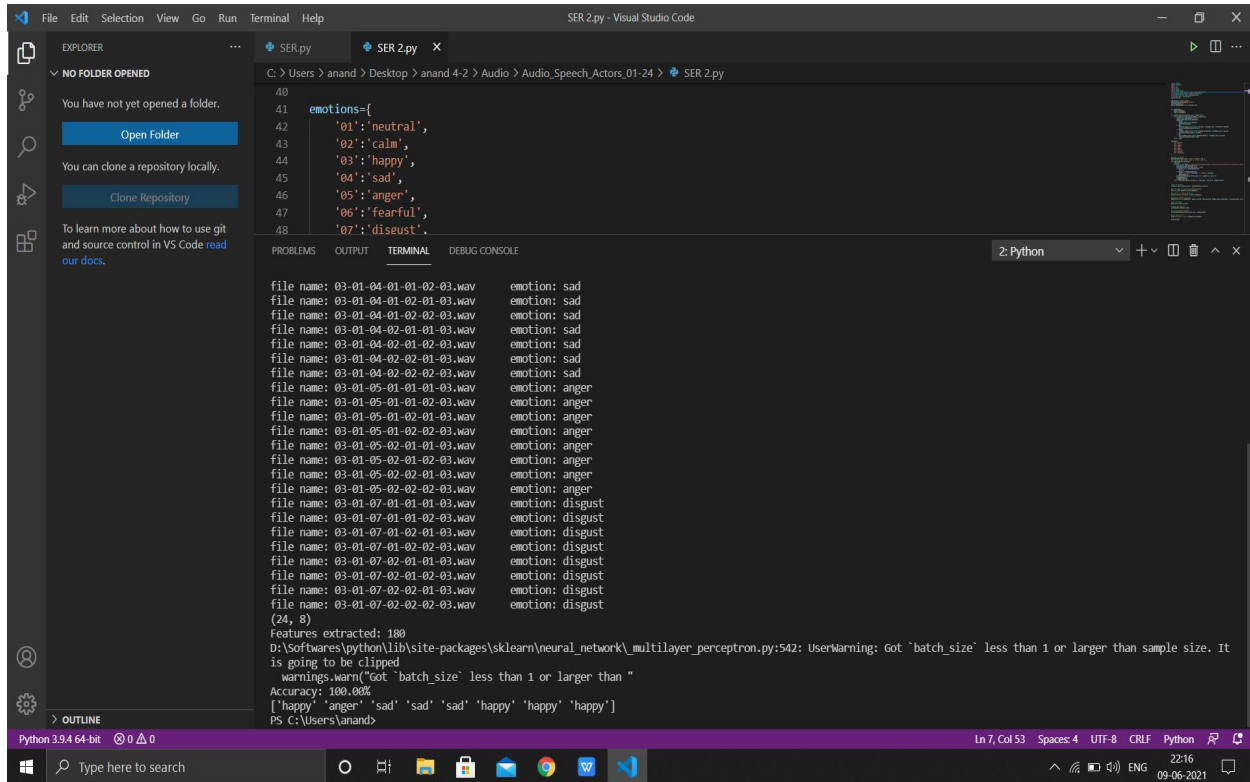
7. HARDWARE REQUIREMENTS:

1. Intel Pentium i3, i5 or i7
2. Microphone

CHAPTER 8

PERFORMANCE ANALYSIS

Application 1



The screenshot shows the Visual Studio Code interface. The Explorer panel on the left indicates 'NO FOLDER OPENED'. The main editor displays a file named 'SER 2.py' with the following Python code:

```
40
41 emotions={
42     '01':'neutral',
43     '02':'calm',
44     '03':'happy',
45     '04':'sad',
46     '05':'anger',
47     '06':'fearful',
48     '07':'disgust'.
```

The TERMINAL panel at the bottom shows the output of the script, listing 24 audio files and their corresponding emotions:

file name	emotion
03-01-04-01-01-02-03.wav	sad
03-01-04-01-02-01-03.wav	sad
03-01-04-01-02-02-03.wav	sad
03-01-04-02-01-01-03.wav	sad
03-01-04-02-01-02-03.wav	sad
03-01-04-02-02-01-03.wav	sad
03-01-04-02-02-02-03.wav	sad
03-01-05-01-01-01-03.wav	anger
03-01-05-01-01-02-03.wav	anger
03-01-05-01-02-01-03.wav	anger
03-01-05-02-01-01-03.wav	anger
03-01-05-02-01-02-03.wav	anger
03-01-05-02-02-01-03.wav	anger
03-01-05-02-02-02-03.wav	anger
03-01-07-01-01-01-03.wav	disgust
03-01-07-01-01-02-03.wav	disgust
03-01-07-01-02-01-03.wav	disgust
03-01-07-01-02-02-03.wav	disgust
03-01-07-02-01-01-03.wav	disgust
03-01-07-02-01-02-03.wav	disgust
03-01-07-02-01-03-03.wav	disgust
03-01-07-02-02-01-03.wav	disgust
03-01-07-02-02-02-03.wav	disgust

Below the table, the terminal shows the following output:

```
Features extracted: 180
D:\Software\python\lib\site-packages\sklearn\network\_multilayer_perceptron.py:542: UserWarning: Got 'batch_size' less than 1 or larger than sample size. It
is going to be clipped
warnings.warn("Got 'batch_size' less than 1 or larger than "
Accuracy: 100.00%
['happy' 'anger' 'sad' 'sad' 'happy' 'happy' 'happy']
PS C:\Users\anand>
```

Figure 8.1

Result:

So, finally figure 8.1 shows appropriate emotions of audio files which is present in the corresponding Actor directory with an average accuracy of 85%.

Application 2

OUTPUT : EMOTION “SAD”

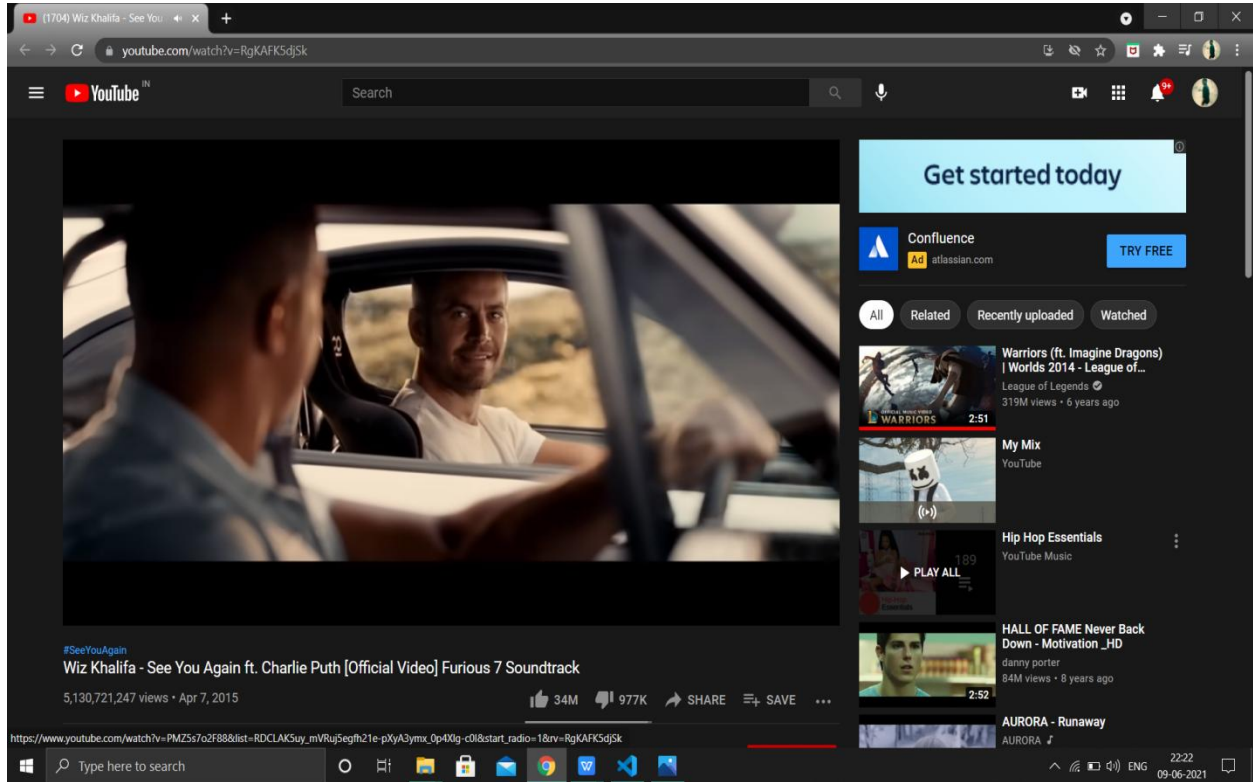


Figure 8.2

Result:

This application finds the emotion of a person and helps to play music according to the person's mood.

CHAPTER 9

CONCLUSION

Many databases available for Speech Sentiment Analysis have given rise to emotions. That is, it includes samples of speech formed in a given emotion by the equivalent utterances of a voice. Since these speeches are a deliberate effort, it may not always be like a more ordinary unprompted voice. The main downside in collecting unprompted speech samples, however, is that more human effort and time will be needed. This will also mean collecting speech samples all the time which may contribute to questions about privacy.

Most used methods of feature extraction and MLP classifier performances are reviewed. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. It can be seen that Integration of various features can give the better recognition rate. Classifier performance is need to be increased for recognition of speaker independent systems. The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can even provide emotion recognition of real time speech.

CHAPTER 10

REFERENCES

- [1]. S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13,no. 2, pp. 293–303, Mar. 2005.
- [2]. B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, May 2010.
- [3]. H. Cao, R. Verma, and A. Nenkova, “Speaker sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” *Comput. Speech Lang.*, vol. 28.1. pp. 186–202, Jan. 2015.
- [4]. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154– 1160, Dec. 2012.
- [5]. T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov.2003.
- [6]. J. Rong, G. Li, and Y.P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” *Inf. Process. Manag.*, vol. 45, no. 3, pp. 325-328 May 2009.
- [7]. 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.
- [8]. J.H. Yeh, T.L. Pao, C.Y. Lin, Y.W. Tsai, and Y.T. Chen, “Segment based emotion recognition from continuous Mandarin Chinese speech,” *Comput. Human Behav.*, vol. 27, no. 5, pp. 1545–1552, Sep. 2011.
- [9]. W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing on Vocal Social Media,” *Inf. Manag.*, Feb. 2015.
- [10]. M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '07-2007*, vol. 4, pp. IV-957-IV-960.
- [11]. <https://www.kaggle.com/uwrfkagglerravdess-emotional-speech-audio>

CHAPTER 11

APPENDIX

10. SOURCE CODE:

10.1 APPLICATION 1:

```
import librosa
import soundfile
import os
import glob
import pickle
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
from pydub import AudioSegment
import pytsx3

engine=pytsx3.init('sapi5')
voices=engine.getProperty('voices')
print(voices[0].id)
engine.setProperty('voice',voices[1].id)

def speak(audio):
    engine.say(audio)
    engine.runAndWait()

def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as sound_file:
        X=sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
```

```

    stft=np.abs(librosa.stft(X))
    result=np.array([])
    if mfcc:
        mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T,axis=0)
        result=np.hstack((result,mfccs))
    if chroma:
        chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
        result=np.hstack((result, chroma))
    if mel:
        mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
        result=np.hstack((result, mel))
    return result

emotions={
    '01':'neutral',
    '02':'calm',
    '03':'happy',
    '04':'sad',
    '05':'anger',
    '06':'fearful',
    '07':'disgust',
    '08':'surprised',
}

#emotions to observe
observed_emotions=['happy','anger','disgust','sad',]

#load the data and extract features for each sound file
def load_data(test_size=0.25):
    x,y=[],[]

```

```

for file in glob.glob(r"C:\\Users\\anand\\Desktop\\anand 4-2\\ Audio\\
Audio_Speech_Actors_01-24\\Actor_03\\*.wav");
    file_name=os.path.basename(file)
    emotion=emotions[file_name.split("-")[2]]
    if emotion not in observed_emotions:
        continue
    if emotion in observed_emotions:
        print("file name:",file_name,"","emotion:",emotion)
        speak(emotion)
    feature=extract_feature(file,mfcc=True, chroma=True, mel=True)
    x.append(feature)
    y.append(emotion)
return train_test_split(np.array(x),y, test_size = test_size, random_state=9)

#split the dataset
x_train,x_test,y_train,y_test = load_data(test_size=0.25)

#get the shape of training and testing datasets
print((x_train.shape[0],x_test.shape[0]))

#get the no of features extracted
print(f'Features extracted: {x_train.shape[1]}')

#initialize the MPL classifier
model=MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes=(300,),
learning_rate='adaptive',max_iter=500)

#train the model
model.fit(x_train,y_train)

```



```

#predict the test set
y_pred=model.predict(x_test)

#calc the accuracy of model
accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

#print the accuracy
print("Accuracy: {:.2f}%".format(accuracy*100))
print(y_pred)

```

10.2 APPLICATION 2:

```

import librosa
import soundfile
import os
import glob
import pickle
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
from pydub import AudioSegment
import webbrowser
import pytsx3

engine=pytsx3.init('sapi5')
voices=engine.getProperty('voices')
print(voices[0].id)
engine.setProperty('voice',voices[1].id)

def speak(audio):
    engine.say(audio)
    engine.runAndWait()

def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as sound_file:
        X=sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
            stft=np.abs(librosa.stft(X))
            result=np.array([])
            if mfcc:

```

```

        mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T,axis=0)
        result=np.hstack((result,mfccs))
    if chroma:
        chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
        result=np.hstack((result, chroma))
    if mel:
        mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
        result=np.hstack((result, mel))
    return result

emotions={
    '01':'neutral',
    '02':'calm',
    '03':'happy',
    '04':'sad',
    '05':'anger',
    '06':'fearful',
    '07':'disgust',
    '08':'surprised',
}

#emotions to observe
observed_emotions=['neutral','calm','happy','sad','anger','fearful','disgust','surprised']
#load the data and extract features for each sound file
def load_data(test_size=0.2):
    x,y=[],[]
    for file in glob.glob(r"C:\Users\anand\Desktop\anand 4-2\Audio\Actor_01\03-01-05-01-01-01-01.wav"):
        file_name=os.path.basename(file)
        emotion=emotions[file_name.split("-")[2]]
        if emotion not in observed_emotions:
            continue
        if emotion in observed_emotions:
            print("file name:",file_name,"","emotion:",emotion)
            if emotion=="happy":
                webbrowser.open("https://youtu.be/09R8_2nJtjg")
            elif emotion=="anger":
                webbrowser.open("https://youtu.be/r6zIGXun57U")
            elif emotion=="fearful":
                webbrowser.open("https://youtu.be/oNb8mqluP44")
            elif emotion=="disgust":
                webbrowser.open("https://youtu.be/YVkuVmdQ3HY")
            elif emotion=="neutral":
                webbrowser.open("https://youtu.be/sK7riqg2mr4")
            elif emotion=="sad":
                webbrowser.open("https://youtu.be/RgKAFK5djSk")

```

```

elif emotion=="calm":
    webbrowser.open("https://youtu.be/_sdh5h_zkkk")
elif emotion=="surprised":
    webbrowser.open("https://youtu.be/d_HIPboLRL8")
else:
    webbrowser.open("https://youtu.be/JGwWNGJdvx8")
speak(emotion)
continue
feature =extract_feature(file,mfcc=True, chroma=True, mel=True)
x.append(feature)
y.append(emotion)
return train_test_split(np.array(x),y, test_size = test_size, random_state=9)

```

```

#split the dataset

```

```

x_train,x_test,y_train,y_test = load_data(test_size=0.25)

```

```

#get the shape of training and testing datasets

```

```

print((x_train.shape[0],x_test.shape[0]))

```

```

#get the no of features extracted

```

```

print(f'Features extracted: {x_train.shape[1]}')

```

```

#initialize the MPL classifier

```

```

model=MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes=(300,),
learning_rate='adaptive',max_iter=500)

```

```

#train the model

```

```

model.fit(x_train,y_train)

```

```

#predict the test set

```

```

y_pred=model.predict(x_test)

```

```

print(y_pred)

```