

**School of Computer Science and Engineering**

**J-component Report**

**on**

**Speech Emotion**

**Recognition**

**Course Title: Machine learning**

**“Department of**

**M.Tech integrated**

**CSE Specialization with Business Analytics”**

**By**

**SRIHARISH R -20MIA1027 – 8610171375**

**Email – sriharish.r2020@vitstudent.ac.in**

**Vasantha Kumaran C -20IA1098 – 8072924584**

**Email – vasanthakumaran.c2020@vitstudent.ac.in**

**Adhikesavan VS -20MIA1059 – 6379775835**

**Email – adhikesavan.vs2020@vitstudent.ac.in**

Under the supervision of

**Dr. G.Suganya**

# **Abstract**

*Communication is the key to express one’s thoughts and ideas clearly. Amongst all forms of communication, speech is the most preferred and powerful form of communications in human. The era of the Internet of Things is rapidly advancing in bringing more intelligent systems available for everyday use. These applications range from simple wearables and widgets to complex self-driving vehicles and automated systems employed in various fields. Intelligent applications are interactive and require minimum user effort to function, and mostly function on voice-based input. This creates the necessity for these computer applications to completely comprehend human speech. A speech percept can reveal information about the speaker including gender, age, language, and emotion. Several existing speech recognition systems used in Internet of things applications are integrated with an emotion detection system in order to analyze the emotional state of the speaker. The performance of the emotion detection system can greatly influence the overall performance of the Internet of things application in many ways and can provide many advantages over the functionalities of these applications. This research presents a speech emotion detection system with improvements over an existing system in terms of data, feature selection, and methodology that aims at classifying speech percepts based on emotions, more accurately*.

# **Keywords**

*Emotion recognition, speech analysis, machine learning, speech features, MFCC, Chroma, Mel, human-computer interaction*

## **Introduction**

*Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents. While humans can efficiently perform this task as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research for several years now, the growth in the field of Artificial Intelligence (AI) has been accelerated. AI, which was once a subject understood by computer scientists only, has now reached the house of a common man in the form of intelligent systems. The advancements of AI have engendered to several technologies involving Human-Computer Interaction. Aiming to develop and improve HCI methods is of paramount importance because HCI is the front-end of AI which millions of users experience. Some of the existing HCI methods involve communication through touch, movement, hand gestures, voice and facial gestures. Among the different methods, the voice-based intelligent devices are gaining popularity in a wide range of applications. In a voice-based system, a computer agent is required to completely comprehend the human’s speech percept in order to accurately pick up the commands given to it.*

#### **Background Study**

Emotion recognition in speech can be divided into two

main aspects of research. The first, where a machine is

designed to artificially produce emotional sounds, while

the second aspect deals with a machine that recognises

emotional states of a user. In order to succeed, a machine

has to learn about human emotions out of speech, using

different classifiers that discern from speech-based data

from which features are extracted.

An abstract conceptual architecture scheme of a

generic system that is capable of recognising emotions in

speech could be composed of a number of processes and

elements as presented in Fig. 2. It provides an overview of

the information system functions and processes that are

necessary to implement the recognition of emotions in

speech. The following paragraphs present some results

from published experiments with related techniques used

to recognise emotions in speech.

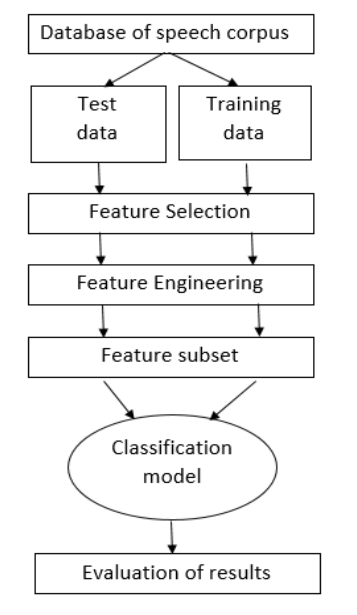
*As human beings speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emoji’s to express the emotions associated with the messages. As emotions 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 SER 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 study we attempt to detect underlying emotions in recorded speech by analyzing the acoustic features of the audio data of recordings. Emotion recognition in speech can be divided into two main aspects of research. The first, where a machine is designed to artificially produce emotional sounds, while the second aspect deals with a machine that recognizes emotional states of a user. In order to succeed, a machine has to learn about human emotions out of speech, using different classifiers that discern from speech-based data from which features are extracted. An abstract conceptual architecture scheme of a generic system that is capable of recognizing emotions in speech could be composed of a number of processes and elements as presented in Fig. 2. It provides an overview of the information system functions and processes that are necessary to implement the recognition of emotions in speech. The following paragraphs present some results from published experiments with related techniques used to recognize emotions in speech. The first step in getting machines to recognize emotions is to gather data from which a machine will learn. This could be done by using primary and secondary inputs. The primary input could be recordings of actors which express different emotions by reading the same text, while the secondary input could be using already existing databases, which were developed by other researchers. Some hybrid approaches are also possible.*

*The next step would be to extract features from the collected or acquired data.*

##### **Methodology**

*The speech emotion detection system is implemented as a Machine Learning model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better. The flowchart represents a pictorial overview of the process. 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.*

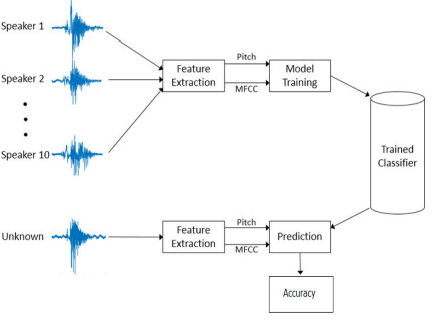
*Applications of simple speech recognition are widespread – YouTube auto-generated subtitles, live speech transcripts, transcripts for online courses, and intelligent voice-assisted chatbots like Alexa and Siri. Because of this, heavily dedicated research has yielded lucrative and fruitful results – YouTube auto-generated subtitles improve each year. However, applications of speech emotion recognition are more nuanced and add a newer dimension to the use of AI and how it can make our lives easier to improve them.*



Methodology Diagram

* **MLP Algorithm Approach**

*In the Speech Emotion Recognition System (SER), the audio files are given as the input. The data sets travels through a number of blocks of processes which makes it executable to help for the analysis of the speech parameters. The data is preprocessed to change it to the suitable format and the respective features from the audio files are extracted using various steps such as framing, hamming, windowing, etc. This process helps in breaking down the audio files into the numerical values which represents the frequency, time, amplitude or any other such parameters which can help in the analysis of the audio files. After the extraction of the required features from the audio files, the model is trained. We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. For the training, we store the numerical values of emotions and their respective features correspondingly in different arrays. These arrays are given as an input to the MLP. Classifier that has been initialized. The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing the performance of the model, if we enter the unknown test dataset as an input, it will retrieve the parameters and predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage which is the final result of our project.*



Feature Extraction Diagram

ADVANTAGES OF USING MLP FOR Speech Emotion Recognition

1. Provides the flexibility to work with nonlinear values

2. Less number of parameters required

3. Higher performance compared to previous systems

4. Better classification of parameters is shown.

5. Can handle missing values, model complex relationships and support multiple inputs.

DISADVANTAGES OF USING MLP FOR Speech Emotion Recognition

1. MLPs always need fixed number of inputs to be provided for fixed number of outputs, there is a fixed mapping function between the inputs and the outputs in these feed-forward neural networks that pose a problem when a sequence of inputs is provided to the model.

2. Network must be retrained when a new emotion is added to the system.

# **Implementation**

##### *MFCC (Mel Frequency Cepstral Coefficients)*

*In the conventional analysis of time signals, any periodic component (for example, echoes) shows up as sharp peaks in the corresponding frequency spectrum (i.e. Fourier spectrum. This is obtained by applying a Fourier transform on the time signal). Any cepstrum feature is obtained by applying Fourier Transform on a spectrogram. The special characteristic of 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. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.*

##### *Mel Spectrogram*

*A Fast Fourier Transform is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram. This is just a spectrogram that depicts amplitude which is mapped on a Mel scale.*

##### *Chroma*

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

### *Exploratory Data Analysis*

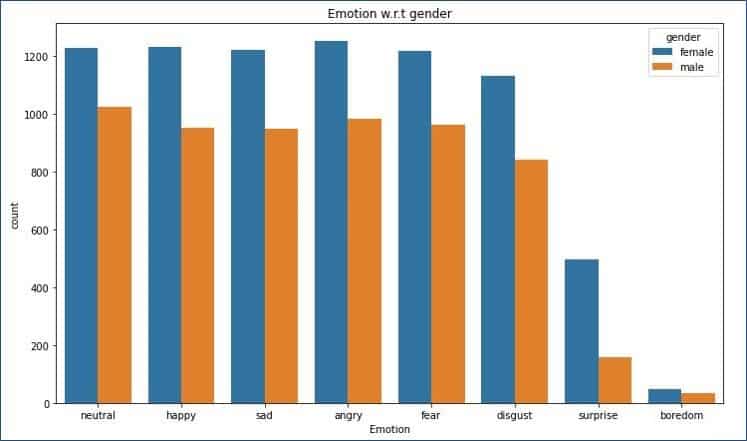
*The combined data set from the original 5 sources is thoroughly analyzed with respect to the following aspects*

* *Emotion distribution by gender*
* *Variation in energy across emotions*
* *Variation of relative pace and power across emotions*

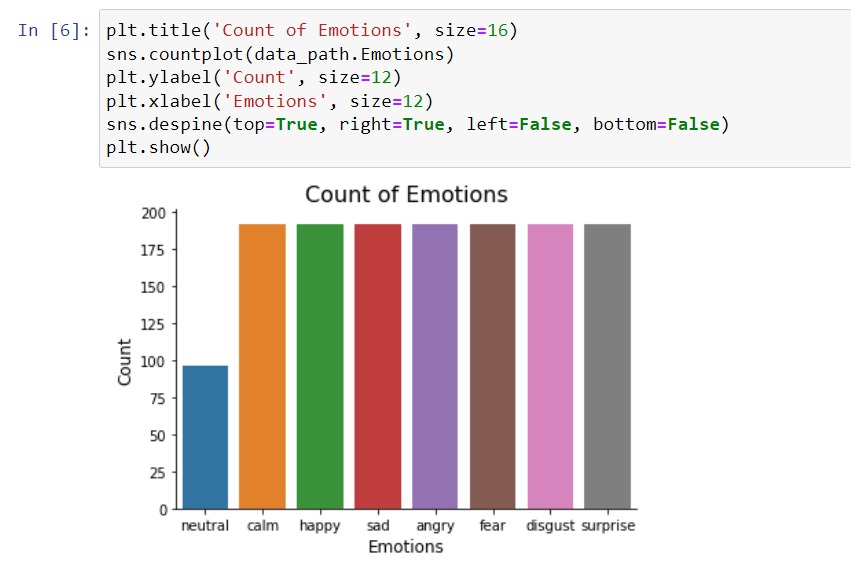
*We checked the distribution of labels with respect to emotions and gender and found that while the data is balanced for six emotions viz*. neutral, happy, sad, angry, fear*and*disgust*, the number of labels was slightly less for*surprise *and negligible for*boredom*. While the slightly fewer instances of surprise can be overlooked on account of it being a rarer emotion, the imbalance against boredom was rectified later by clubbing sadness and boredom together due to them being similar acoustically. It’s also worth noting that boredom could have been combined with neutral emotion but since both*sadness*and*boredom *are negative emotions, it made more sense to combine them.*

##### *Emotion Distribution of Gender*

*Regarding the distribution of gender, the number of female speakers was found to be slightly more than the male speakers, but the imbalance was not large enough to warrant any special attention.*



**Distributions of emotion with respect to gender**

****

*Librosa: We already explained this library in the above section. It helps with feature extraction.*

*Soundfile: Sound File can read and write sound files. This library helps us to open the sound file.*

*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 ndarray) for our sound file.*

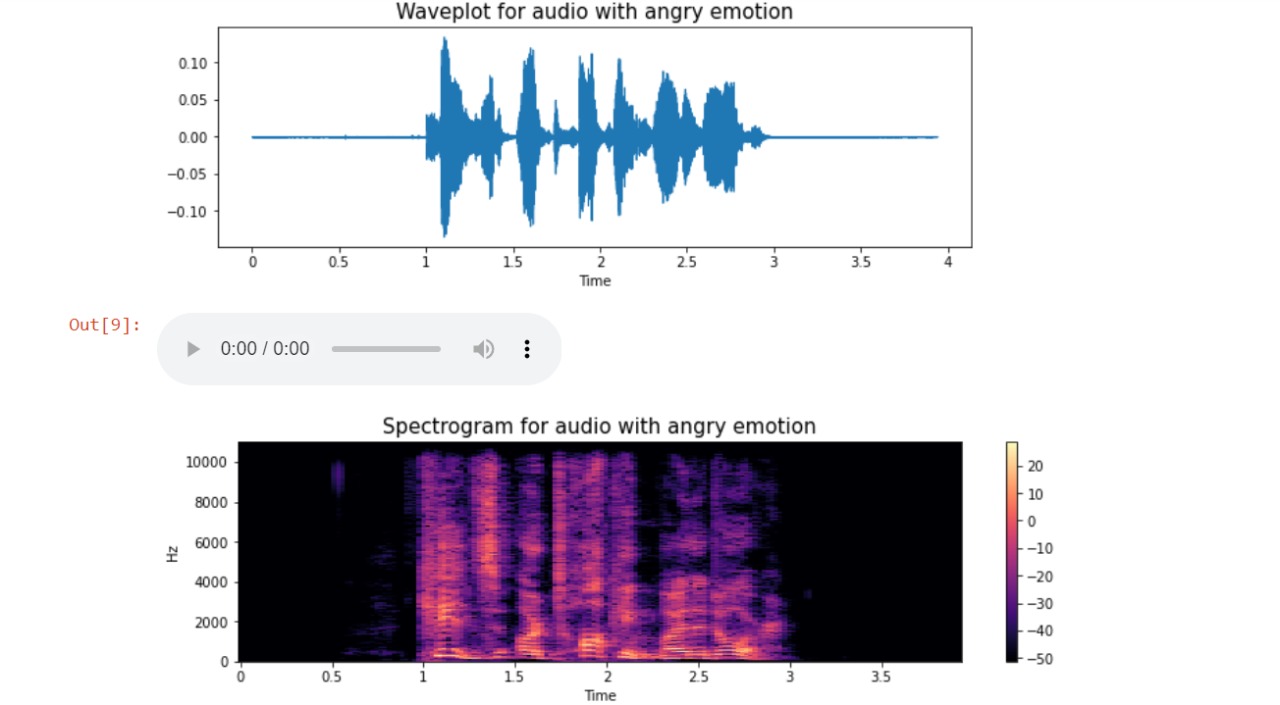
*Sklearn: Scikit-learn is an open-source Python library that has powerful tools for data analysis and data mining. It is used in our project mainly for training, testing and splitting our data then using it to make model data and finding the accuracy of our model.*

*Pyaudio: Pyaudio provides Python bindings for Port Audio, 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.*

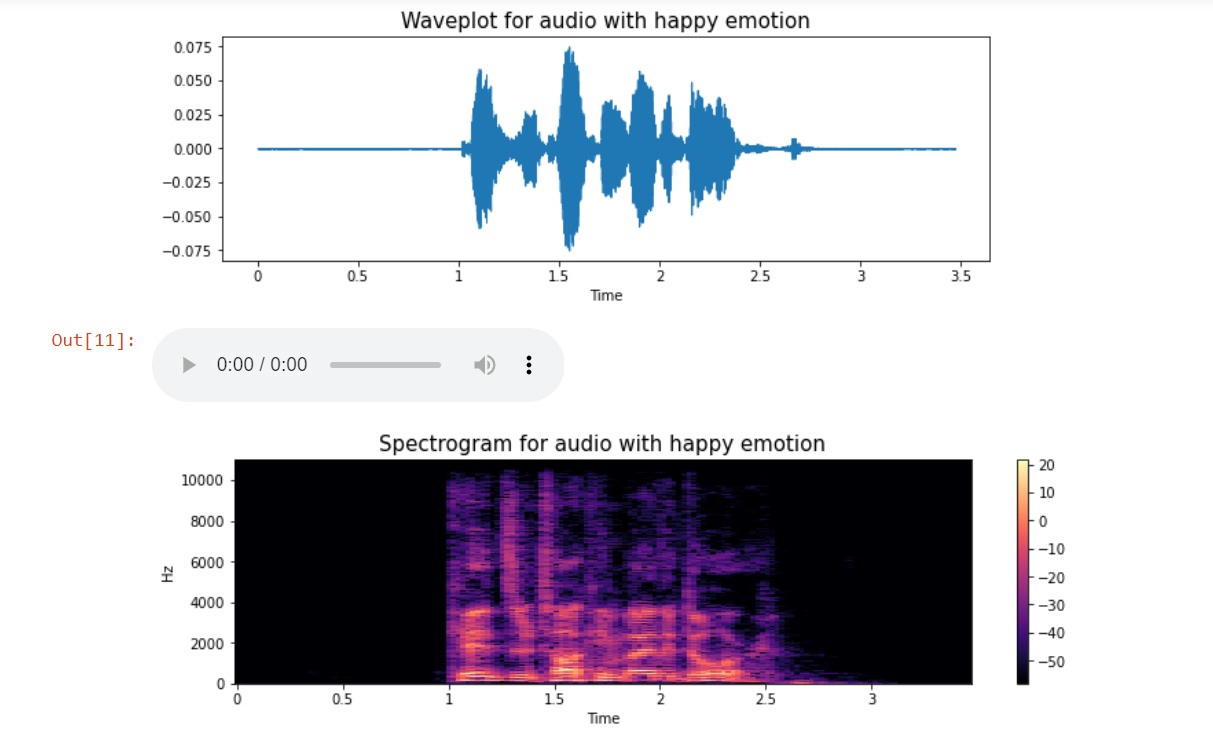
*Pickle: It is used for the saving model.*

*Matplotlib: It is used for plotting the graph and so on*

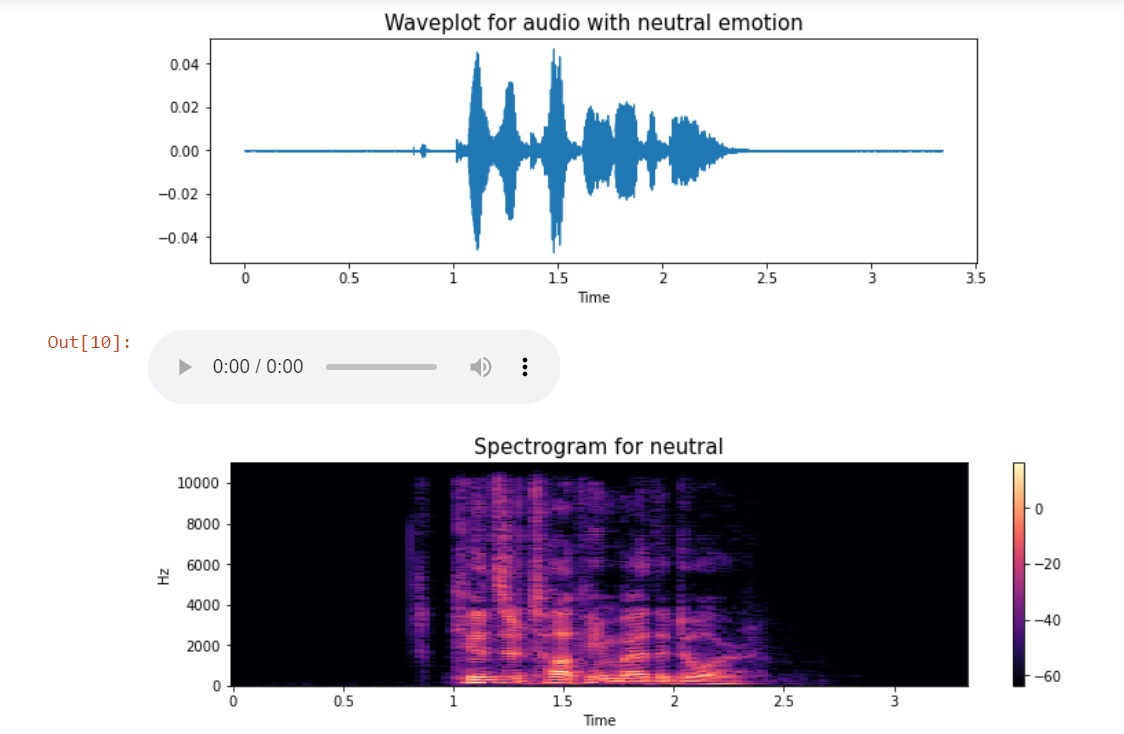
**Waveplot for Angry**

****

**Waveplot for Happy**

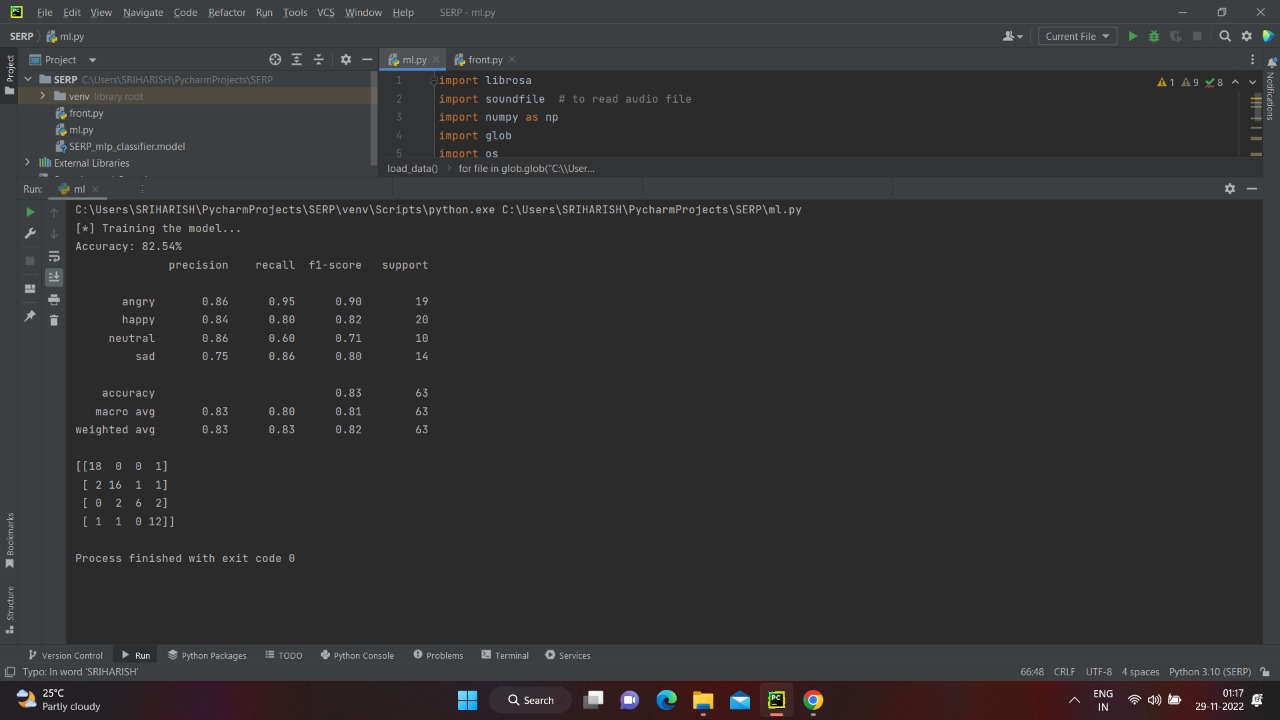
****

**Waveplot for Neutral**

****

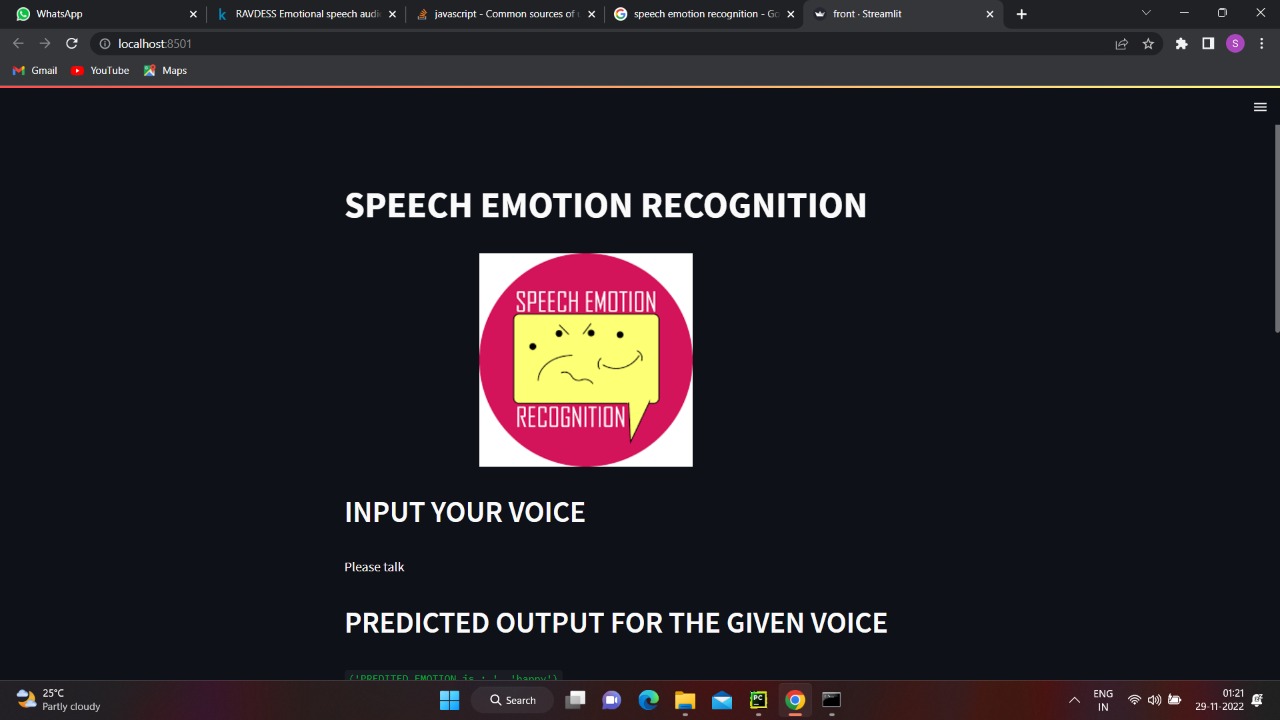
# **Results and Discussion**

**Confusion Matrix**

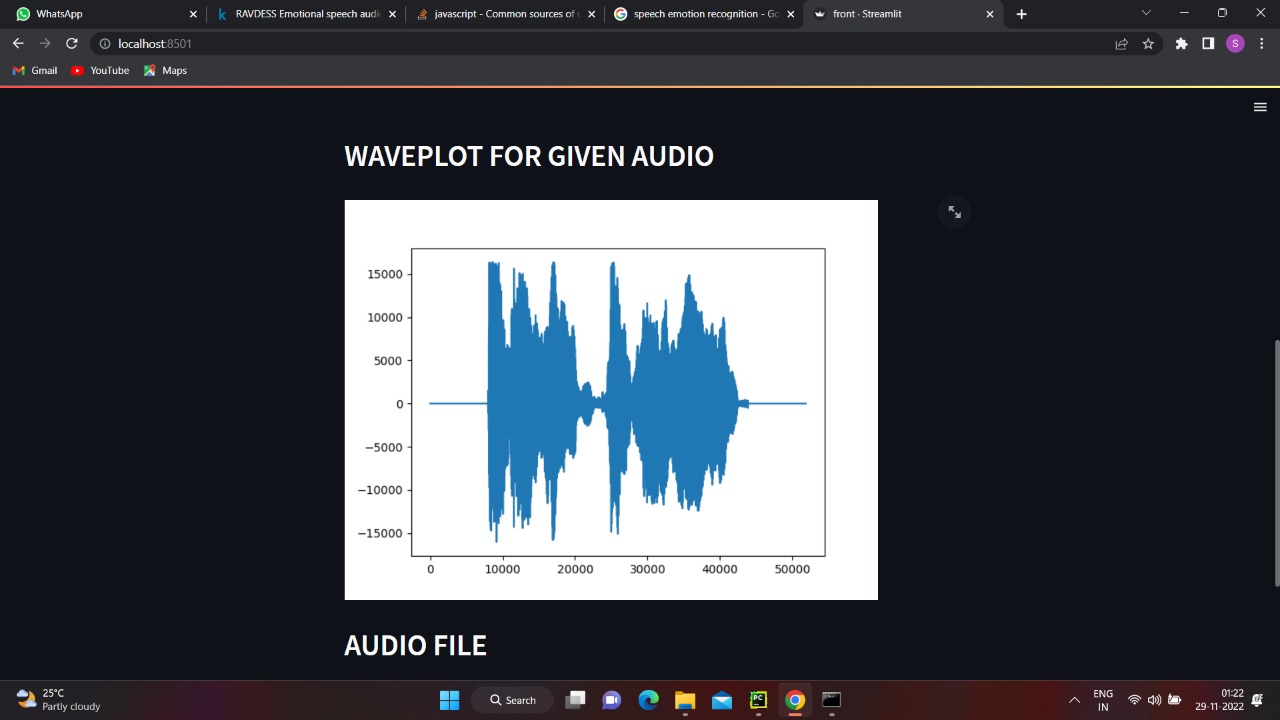


*Thus the given accuracy is 82.54%. Its look like an good accuracy so we can proceed….*

Frontend



# C:\Users\c . saravana guru\Downloads\WhatsApp Image 2022-11-29 at 1.22.25 AM.jpeg

****

*It recognize the voice and it will predict the emotion and it will print the waveform of the voice. If any error occurs we can check the waveform and rectify it. The waveform and the voice you stored is in the same data location.*

# **Conclusion & Future work**

*The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated devices work based on voice commands from the user. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system. In this thesis, the steps of building a speech emotion detection system were discussed in detail and some experiments were carried out to understand the impact of each step. Initially, the limited number of publically available speech database made it challenging to implement a well-trained model. Next, several novel approaches to feature extraction had been proposed in the earlier works, and selecting the best approach included performing many experiments. Finally, the classifier selection involved learning about the strength and weakness of each classifying algorithm with respect to emotion recognition. At the end of the experimentation, it can be concluded that an integrated feature space will produce a better recognition rate when compared to a single feature. For future advancements, the proposed project can be further modeled in terms of efficiency, accuracy, and usability. Additional to the emotions, the model can be extended to recognize feelings such as depression and mood changes. Such systems can be used by therapists to monitor the mood swings of the patients. A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem of emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. A sentiment detection using vocabulary, can be integrated with speech emotion detection to identify a possible sarcasm. Therefore, in the future, there would emerge many applications of a speech-based emotion recognition system.*

# **Reference**

1. https://stackoverflow.com
2. [https://www.kaggle.com](https://www.kaggle.com/)
3. https://extrudesign.com/speech-emotion-recognition/
4. https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/
5. <https://github.com/topics/speech-emotion-recognition>
6. https://www.stat.purdue.edu/~vishy/papers/VisMur02b. pdf
7. https://www.sciencedirect.com/science/article/pii/S111086 6512000345
8. [https://pdfs.semanticscholar.org/d945/29ae612b76287c7dc dccd5bf09f3c5e772af.pdf](https://pdfs.semanticscholar.org/d945/29ae612b76287c7dc%20dccd5bf09f3c5e772af.pdf)
9. <https://link.springer.com/article/10.1007/s10772-011-9125-1>