A

Honors Mini Project Report

On

Speech Emotion Recognition System (SER)

Submitted by

Ms. Vaibhavi Gaikwad (GL)

Ms. Laharika Gudur

Ms. Shravani Dudhyal

Under the guidance of

Dr. A. V. Thalange

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

WALCHAND INSTITUTE OF TECHNOLOGY, SOLAPUR
2024-25

CERTIFICATE

This is to certify that the Honors Mini Project entitled

Speech Emotion Recognition System (SER)

Submitted by

Name	Roll no.	Exam Seat No.
Vaibhavi Gaikwad	07	2104111154
Laharika Gudur	08	2104111160
Shravani Dudhyal	09	2204112010

has been approved as the partial fulfilment for the award of Final year of Electronics and Telecommunication Engineering in **Walchand Institute of Technology**, Solapur in the academic year 2024-25.

Dr. A. V. Thalange
Project Guide

Dr. A. V. Thalange
Head, E&TC Department

Dr. V.A. Athavale
Principal

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING WALCHAND INSTITUTE OF TECHNOLOGY, SOLAPUR, 2024-2025

ACKNOWLEDGEMENT

The project has certainly enlightened us with the modern era of Technologies and it has boosted our confidence. The project work has certainly rendered us tremendous learning as well as practical experience.

We are thankful to **Dr. V. A. Athavale**, Principal of W.I.T College, **Dr. A. V. Thalange** Head of Electronics & Telecommunication Engineering Department for granting permission to undertake this project and their valuable guidance about implementation and programming.

At last, but not least we are thankful to staff of Electronics & Telecommunication Engineering Department W.I.T. Solapur.

INDEX

SR. NO.	CONTENTS	PAGE NO.
	Abstract	1
Chapter 1	Introduction	2
Chapter 2	Literature Review	4
Chapter 3	System Description	6
Chapter 4	Implementation Details	11
Chapter 5	Results	12
Chapter 6	Advantages and Disadvantages	13
Chapter 7	Application	15
Chapter 8	Conclusion and Future Scope	16
	References	17

ABSTRACT

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. Emotion recognition is a rapidly growing research domain in recent years. Unlike humans, machines lack the abilities to perceive and show emotions. But human-computer interaction can be improved by implementing automated emotion recognition, thereby reducing the need of human intervention.

In this project, basic emotions like calm, happy, fearful, disgust etc. are analysed from emotional speech signals. We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non-linearly separated. Mel-frequency cepstrum coefficients (MFCC), chroma and mel features are extracted from the speech signals and used to train the MLP classifier. For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, numpy and sound file to analyse the speech modulations and recognize the emotion. [1]

Using RAVDESS dataset which contains around 1500 audio file inputs from 24 different actors (12 male and 12 female) who recorded short audios in 8 different emotions, we will train an NLP- based model which will be able to detect among the 8 basic emotions.

CHAPTER 1: INTRODUCTION

1.1 Introduction

The most elementary way of communication in humans is Speech. To enrich interaction, one needs to know and understand the emotion of another person and how to react to it. Unlike machines, we humans can naturally recognize the nature and emotion of the speech. Can a machine also detect the emotion from a speech? Well, this could be made possible using machine learning. Machines need a specific model for detecting the emotions of a speech and such a model can be implemented using machine learning.

Speech emotion recognition is a very useful and important topic in today's world. A machine detecting the emotion of a human speech can be proved useful in various industries. A very basic usage of speech recognition is in the health sector where it can be used in detecting depression, anxiety, stress etc. in a patient. It can also be used in industries like the crime sector where emotions can be recognized from the speech to distinguish between victims and criminals.

Emotions can be of various types like happy, sad, angry, disguised etc. depending on the feeling and frame of mind of the person. In our study, we have used various datasets with different emotions. We have also combined four datasets to one dataset and then applied the model so that the efficiency of the model can be improved and there can be a variety in the data points. This has also resulted in eliminating the overfitting condition in our model.[2]

1.2 Purpose of the Project

- The purpose of this Speech Emotion Recognition System (SER) project is to develop a model that can accurately identify and classify human emotions based on speech. Emotions play a vital role in communication, and by enabling machines to detect emotions from speech, this project aims to enhance the interaction between humans and computers.
- The ability to understand emotions allows systems like virtual assistants, customer support bots, and other AI-driven tools to respond in more empathetic

and personalized ways.

- Beyond improving human-computer interaction, this system has practical applications in various fields, including healthcare, where it can be used to monitor patient emotions during telemedicine sessions or therapy.

1.3 Scope/Limitation of the Project

The scope of the Speech Emotion Recognition (SER) System project includes developing a machine learning model that accurately detects and classifies emotions like happy, sad, fearful, and calm from speech signals.

- By using the RAVDESS dataset and extracting key audio features such as MFCCs, chroma, and Mel spectrograms, the project aims to enhance human-computer interaction by enabling systems to understand emotional cues in speech.
- Emotion-labelled datasets are often limited, which can affect model training and lead to overfitting on smaller datasets like RAVDESS.
- SER systems are often language-dependent, requiring further customization for use with different languages and dialects.
- Recognizing complex or mixed emotions, such as sarcasm or ambivalence, remains challenging for SER systems.

CHAPTER 2: LITERATURE SURVEY

Md. Rayhan Ahmed et al. [3], used four deep neural network-based models built using LFABS. Model-A uses seven LFABs followed by FCN layers and a SoftMax layer for classification. Model-B uses LSTM and FCNs, Model-C uses GRU and FCNs and Model- D combines the three individual models by adjusting their weights. From each of these audio files, they hand-craft five categories of features- MFCC, LMS, ZCR, RMSE. did. data set. These features are used as inputs to a one-dimensional (1D) CNN architecture to further extract hidden local features in these speech files. To obtain additional contextual long-term representations of these learned local features via the 1D CNN block, we extended our experiment by incorporating LSTM and GRU after the CNN block, giving us more improved accuracy. After running DA, we observe that all four models perform very well on the SER task of detecting emotions from raw speech audio. Amongst all four models, the ensemble Model-D achieves the state-of-the-art weighted average accuracy of 99.46% for TESS dataset.

Dr. Nilesh Shelke et al. [4] used RAVDESS, TESS and SAVEE datasets for classification. Their purpose is to mandate the modernization of current plans and technology enabling EDS and to implement assistance in all areas of computers and technology. Analytics complement emotions extracted from databases, layers, and model libraries created for emotion recognition from speech. It mainly focuses on data collection, feature extraction, and automatic emotion detection results. The intermodal recognition computer system is considered a unimodal solution because it offers higher sorting accuracy. Accuracy depends on the number of emotions detected, the features extracted, the classification method, and the stability of the database.

A novel paradigm for emotion identification in the presence of noise and interference was put out by Shibani Hamsa et al. In order to examine the 21 speaker's emotions, our method takes into account the speaker's energy, time, and spectral factors.

To do. When tested on three different speech corpora in two different languages, our system-which combines this representation with a random forest classifier-performs better than other existing algorithms and is less prone to stressful noise. All metrics

(Accuracy, Precision, Recall, and F1 scores) in the RAVDESS and SUSAS datasets score above 80%.

A data imbalance processing approach based on the selective interpolation synthetic minority oversampling (SISMOTE) methodology is suggested by Zhen-Tao Liu et al. [5] to reduce the influence of sample imbalance on emotion identification outcomes. In order to minimise duplicate characteristics with inadequate emotional representation, a feature selection approach based on analysis of variance and gradient-enhanced decision trees (GBDT) is also provided. The results of speech emotion detection tests on the CASIA, Emo- DB, and SAVEE databases demonstrate that our technique produces an average of 90.28% (CASIA), 75.00% (SAVEE), and 85.82% (based on the findings) (Emo-DB). It demonstrates its precision in recognition. Utilizing voice emotion recognition is superior to some cutting- edge technologies.

To accomplish efficient speech emotion identification, Apeksha Aggarwal et al. [6] have presented two alternative feature extraction strategies. First, utilising super convergence to extract two sets of latent features from voice data, bidirectional feature extraction is presented. Principal Component Analysis (PCA) is used to produce the first set of features for the first set of features. A second method involves extracting the Mel spectrogram picture from the audio file and feeding the 2D image into his pre-trained VGG-16 model. In this study, several algorithms are used in comprehensive experimentation and rigorous comparative analysis of feature extraction approaches across two 22 datasets (RAVDESS AND TESS).

CHAPTER 3: SYSTEM DESCRIPTION

Below is the flowchart for the Speech Emotion Recognition (SER) system, illustrating the process from audio input to emotion classification, including feature extraction and model prediction steps.

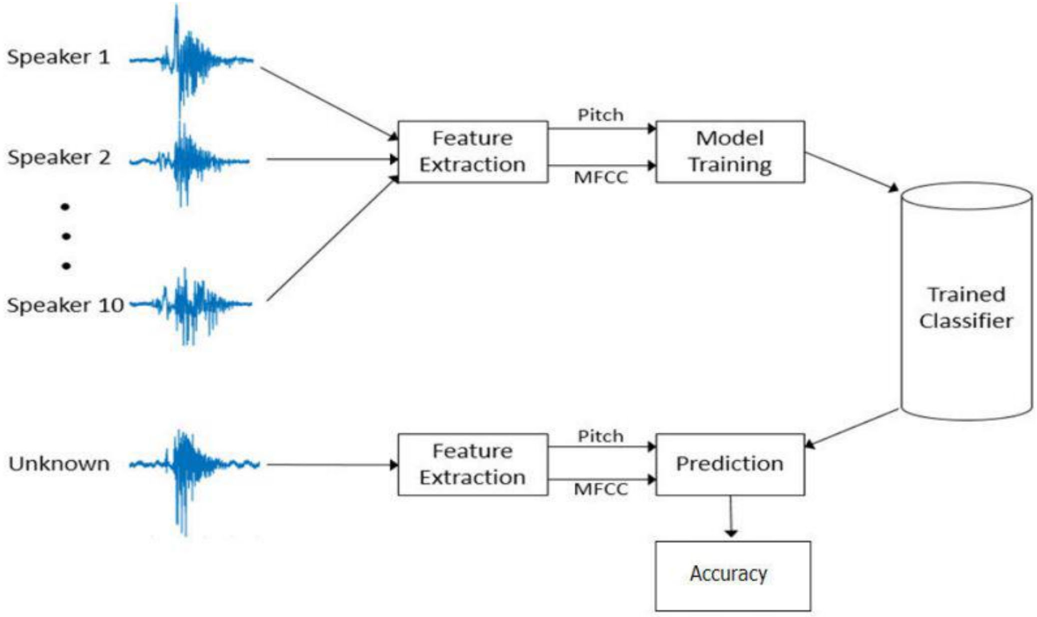


Figure 3.1: Speech Emotion Recognition System Flowchart

3.1. Brief Approach

Using RAVDESS dataset which contains around 1440 audio label inputs from 24 different actors, First, we analysed our dataset by plotting the spectrogram and waveforms of a sample audio file from the dataset. Then we extracted the features of the data using Mel Frequency Cepstral Coefficients (MFCCs), chroma frequencies, MeI spectrogram and Spectral centroid. Then we built a multi-layer perceptron (MLP) model to predict the emotions of the audio data as well as the gender of the speaker.

3.2. Dataset

This section gives a detailed description of RAVDESS dataset used in this project.[7] This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors. Speech emotions include calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). The dataset was split into **training and testing datasets** to facilitate model evaluation. Typically, 80% of the data (1,152 files) was allocated for training, allowing the model to learn emotion patterns, while the remaining 20% (288 files) was reserved for testing, providing a basis to assess the model's accuracy and generalizability. These identifiers define the stimulus characteristics:

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).
- 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: 03-01-06-01-02-01-12.wav

- Audio-only (03)
- 12th Actor (12)
- Speech (01)
- Normal intensity (01)
- 1st Repetition (01)
- Statement "dogs" (02)
- Fearful (06)

3.3. Feature extraction

It is very effective in increasing the accuracy and efficiency of machine learning algorithms in emotional speech recognition. The input audio file is reduced to a set of features which represent or summarise the original input. The extracted features are then fed into the neural network to identify the respective emotion. The features which have been used in the project are discussed below.

1.Mel Spectrogram.

An audio signal can be broken down into sine and cosine waves that form the original signal. The frequencies and amplitudes of these representative waves can be used to convert the input signal from the time to the frequency domain. The fast Fourier transform (FFT) is an algorithm that can be used to perform this conversion. The FFT is however performed on a single time window of the input signal. If the frequencies of the representative signals change over time (non-periodic), the change cannot be captured through a single time window conversion. Using the FFT over multiple overlapping windows can be used to construct a spectrogram representing the amplitude of the representative frequencies as they change over time.[8]

2. Mel-frequency cepstral coefficients (MFCCs).

The Fourier spectrum is obtained by using the FFT on an audio signal. Taking the log of the Fourier spectrum's magnitude and then taking the spectrum of this log through a cosine transformation allows us to calculate the cepstrum of the signal (ceps is the reverse of spec, called so due to being a non-linear `spectrum of a spectrum'). The amplitudes of this resulting cepstrum are the cepstral coefficients. Representing the frequencies in terms of the Mel scale (discussed above) instead of a linear scale converts the cepstrum to a Mel-frequency cepstrum and the cepstral coefficients to Mel-frequency cepstral coefficients (MFCCs). The cepstrum represents the rate of change in the different spectrum bands, and the coefficients are widely used in machine learning algorithms for sound processing.[9]

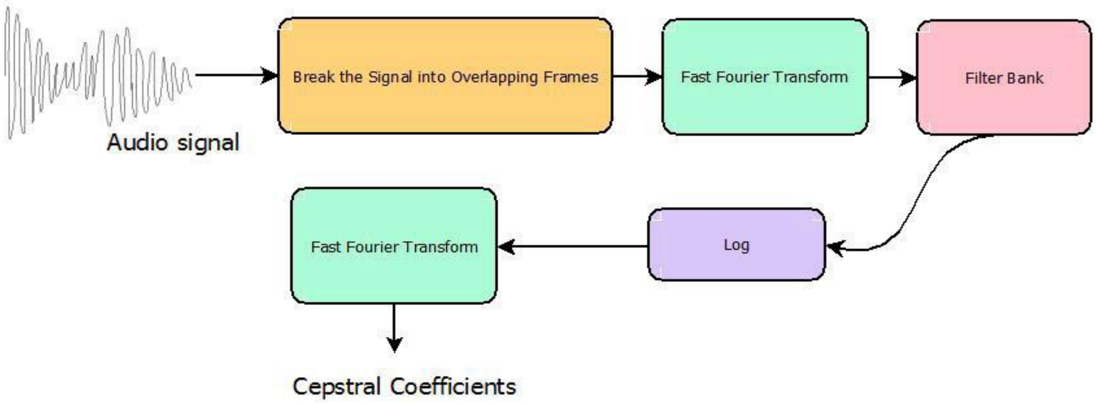


Fig. 3.2. Process of extracting MFCCs

3. Chroma

An octave is a musical interval or the distance between one note and another, which is twice its frequency: A3 (220 Hz) - A4 (440 Hz) {A5 (880 Hz). An octave is divided into 12 equal intervals, and each interval is a different note. The intervals are called chromas or semitones and are powerful representations of audio signals. The calculation of the chromogram is similar to the spectrogram using short-time Fourier transforms, but semitones/pitch classes are used instead of absolute frequencies to represent the signal. The unique representation can show musical properties of the signal which is not picked up by the spectrogram.[10]

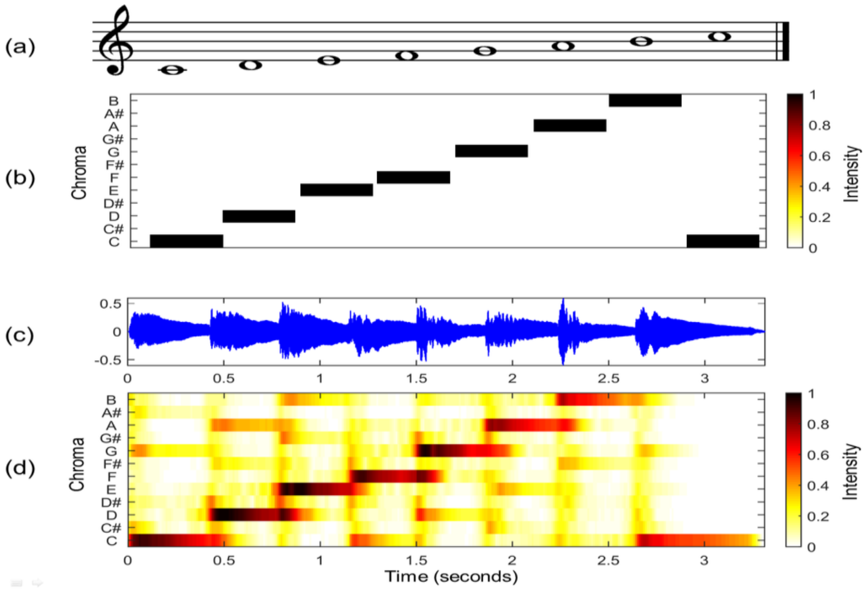


Fig. 3.2. Chromogram obtained from the audio recording.

3.4 Model Architecture

- **MLP-Model**

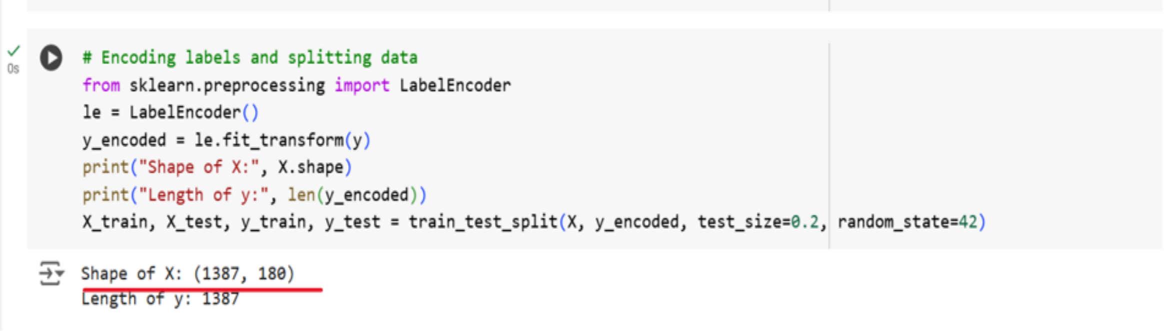
A fully connected neural network was deployed to predict the emotion from the features of the sound files. The model consists of 4 hidden layers with a dropout of 0.1 in the first three layers. The first and the second hidden layers consists of 512 neurons while the third layer contained 128 neurons and the fourth hidden layer contained 64 neurons. The Multi-Layer Perceptron (MLP) model is a type of neural network that is particularly effective for classifying complex, non-linear data, making it suitable for speech emotion recognition. In this project, the MLP model analyzes audio features like MFCC, chroma, and mel spectrograms, which capture various characteristics of speech sounds.

The model includes multiple layers—an input layer, hidden layers, and an output layer—allowing it to learn and identify patterns within the data. During training, the MLP model learns from labelled examples to distinguish between different emotions, such as happy, sad, and fearful, based on the extracted features. By using the MLP, the project achieves a structured approach to classify emotions in speech, paving the way for applications in interactive systems, mental health analysis, and customer service. [11]

CHAPTER 4: IMPLEMENTATION DETAILS

The results of the Speech Emotion Recognition (SER) system demonstrate its ability to classify emotions from speech data with a reasonable degree of accuracy. Using the RAVDESS dataset, which contains 1,440 audio files from 24 actors, the system was trained to detect basic emotions such as calm, happy, sad, angry, fearful, and surprised. Key findings include:

1.Feature Extraction Success: Features like MFCC (Mel Frequency Cepstral Coefficients), chroma frequencies, and Mel-spectrogram were successfully extracted and used as inputs to the model.



```
# Encoding labels and splitting data
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y_encoded = le.fit_transform(y)
print("Shape of X:", X.shape)
print("Length of y:", len(y_encoded))
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42)
```

```
Shape of X: (1387, 180)
Length of y: 1387
```

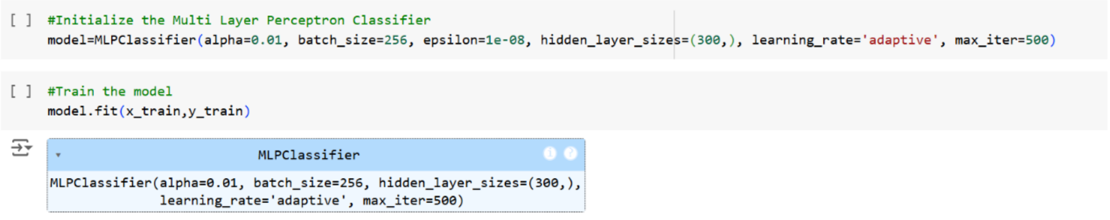
Here 1387 Likely Represents Total Number of Audio Files in the RAVDESS dataset for Training purpose and 180 Total Extracted Features including MFCCs (40), Chroma (12), and Mel spectrogram (128).

2. Model Performance:

- **MLP Model**

  - The Multi-Layer Perceptron (MLP) model achieved a good accuracy rate for emotion classification. However, the performance varied across different emotions.

  - Accuracy obtained is 62%

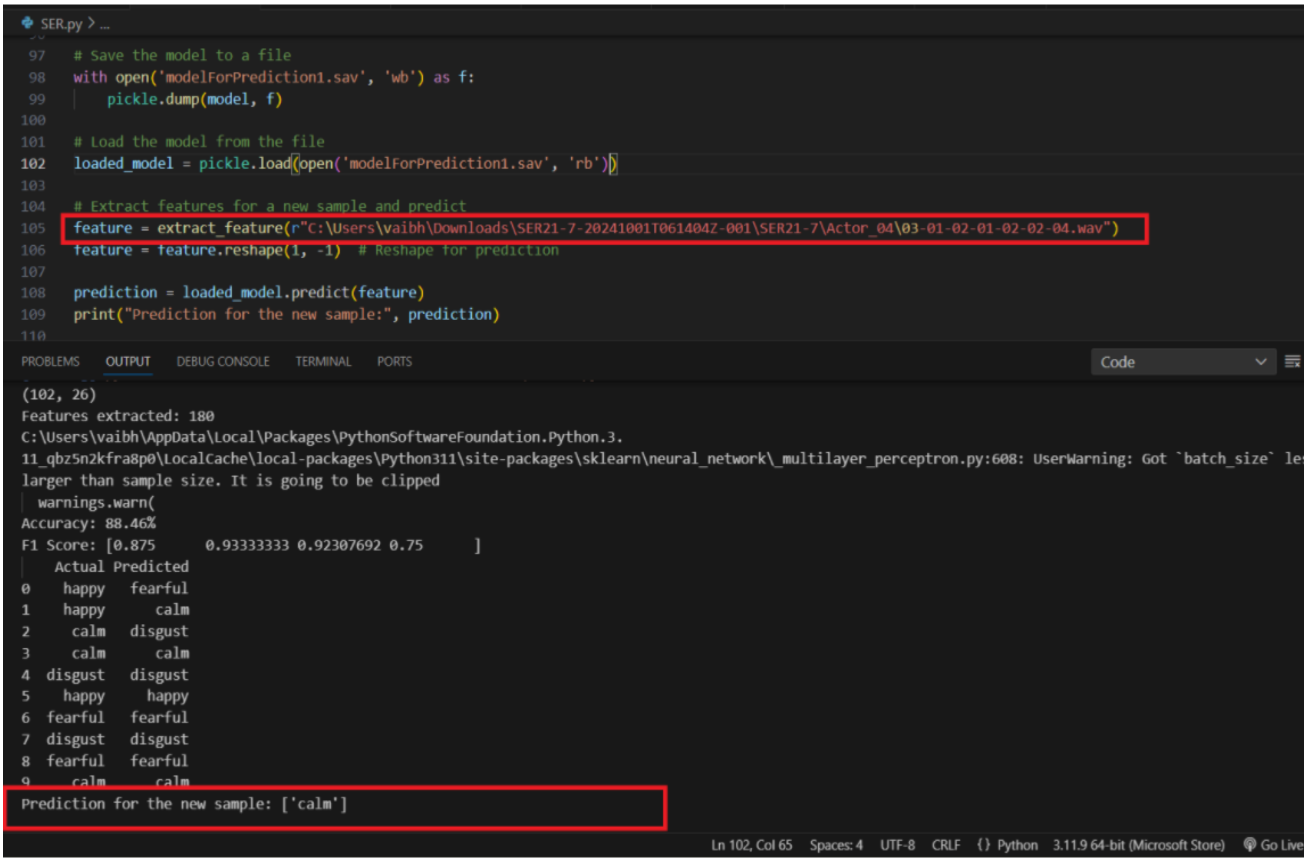


```
#Initialize the Multi Layer Perceptron Classifier
model=MLPClassifier(alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)
```

```
#Train the model
model.fit(x_train,y_train)
```

```
MLPClassifier(alpha=0.01, batch_size=256, hidden_layer_sizes=(300,),
              learning_rate='adaptive', max_iter=500)
```

CHAPTER 5: RESULT



```
# Save the model to a file
with open('modelForPrediction1.sav', 'wb') as f:
    pickle.dump(model, f)

# Load the model from the file
loaded_model = pickle.load(open('modelForPrediction1.sav', 'rb'))

# Extract features for a new sample and predict
feature = extract_feature(r"C:\Users\vaibh\Downloads\SER21-7-20241001T061404Z-001\SER21-7\Actor_04\03-01-02-01-02-02-04.wav")
feature = feature.reshape(1, -1)  # Reshape for prediction

prediction = loaded_model.predict(feature)
print("Prediction for the new sample:", prediction)
```

```
(102, 26)
Features extracted: 180
C:\Users\vaibh\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.
11_qbz5n2kfra8p0\LocalCache\local-packages\Python311\site-packages\sklearn\neural_network\_multilayer_perceptron.py:608: UserWarning: Got `batch_size` le
larger than sample size. It is going to be clipped
  warnings.warn(
Accuracy: 88.46%
F1 Score: [0.875      0.93333333 0.92307692 0.75      ]
    Actual Predicted
0    happy   fearful
1    happy      calm
2     calm   disgust
3     calm      calm
4  disgust   disgust
5    happy     happy
6  fearful   fearful
7  disgust   disgust
8  fearful   fearful
9     calm      calm
Prediction for the new sample: ['calm']
```

CHAPTER 6: ADVANTAGES AND DISADVANTAGES

6.1. Advantages

1. Improved User Experience: SER allows machines to recognize and respond to human emotions, making interactions with virtual assistants, chatbots, and other systems more personalized, empathetic, and user-friendly.

2. Real-Time Emotional Feedback: SER systems can analyse speech and detect emotions in real-time, enabling applications such as customer service or healthcare to adapt instantly to the user's emotional state, leading to more effective interactions.

3. Enhanced Communication in Human-Computer Interaction: By identifying emotions, computers and systems can adjust their responses or behaviour accordingly, improving communication and reducing frustration or misunderstandings.

4. Increased Efficiency in Customer Support: Emotion recognition in call centres helps agents identify distressed or frustrated customers, allowing them to tailor their responses and prioritize urgent issues, which enhances overall customer satisfaction and service quality.

5. Adaptability in E-learning and Education: In educational platforms, SER can monitor students' emotional responses to the content, helping educators tailor lessons based on engagement or confusion, ultimately improving learning outcomes.

6. Automation and Scalability: SER can automate the process of emotion detection, making it scalable for use in large systems such as customer support centres, security systems, or entertainment platforms, reducing the need for constant human intervention.

7. Objective Emotion Analysis: Unlike human observers, who may misinterpret emotions based on subjective biases, SER systems provide objective emotion analysis based on consistent data, improving accuracy in applications such as interviews, therapy, and customer feedback.

8. Enhanced Security and Safety: SER can detect fear, anxiety, or stress in public places or high-security environments, providing alerts for potential risks or suspicious behaviour, contributing to improved safety and security measures.

6.2. Disadvantages

1. Accuracy and Reliability: SER systems often struggle with accurately detecting emotions due to the complexity and variability in human emotional expression, leading to potential misinterpretation.

2. Dependence on Audio Quality: The performance of SER systems is highly dependent on the quality of the audio input. Background noise, poor recording conditions, or speech distortions can significantly affect the accuracy of emotion detection.

3. Limited Emotional Range: Most SER systems can only recognize basic emotions like happy, sad, and angry, but struggle to identify more nuanced or mixed emotions, reducing their effectiveness in capturing the full emotional spectrum.

4. Cultural and Linguistic Biases: Emotional expressions vary across cultures and languages, and SER systems may not account for these differences, leading to inaccurate results when applied to diverse populations.

5. Privacy and Ethical Concerns: The analysis of voice data raises privacy issues, as users might be uncomfortable with their emotions being monitored and recorded. Additionally, there are ethical concerns about the misuse of such sensitive data, especially in contexts like hiring or customer service.

CHAPTER 7: APPLICATION

1.Virtual Assistants: Enhances user interactions by enabling assistants like Siri or Alexa to respond empathetically based on the user's emotional state.

2. Customer Support: Helps identify the emotions of customers during calls or chats, allowing companies to tailor their responses and improve customer satisfaction.

3. Healthcare: Assists in monitoring patients’ emotional well-being, especially in telemedicine or mental health therapy, where emotion recognition can provide valuable insights for treatment.

4. Education: E-learning platforms can use emotion recognition to assess students' engagement and emotions, adapting teaching methods accordingly.

5. Security and Surveillance: Emotion recognition can be used in surveillance systems to detect stress, fear, or other emotional cues in public or sensitive areas, enhancing security measures.

6. Entertainment and Gaming: In gaming, emotion recognition can be used to modify in-game experiences based on the player's emotional state, creating more immersive experiences.

7.Human Resources: Emotion recognition during interviews or performance reviews can help assess candidates’ or employees' emotional responses and overall well-being.

8. Marketing and Advertising: Emotion recognition can be used to gauge consumer reactions to advertisements or products in real time, helping marketers understand the emotional impact of their campaigns and fine-tune their strategies to connect with their audience more effectively.

9. Assistive Technology: For individuals with communication disorders or emotional difficulties, SER can assist by interpreting their emotional state and relaying it to caregivers or devices, improving communication and support.

10. Social Robotics: In robotics, SER can enable social robots to better understand and interact with humans by responding to emotions, which can be used in elderly care, companionship robots, or interactive toys for children.

CHAPTER 8: CONCLUSION AND FUTURE SCOPE

5.1 Conclusions

In the project, we tried to use deep learning to analyse certain speech samples. In order to illustrate the various human emotions, we first loaded the datasets using the Librosa library and depicted them in the form of various wave plots and spectrograms. Then, we used the MFCC feature extraction method to analyse the acoustic characteristics of all of our samples and organised the sequential data obtained in the 3D array. Using the Matplotlib library, we put the data into a graphical form, then after some repeated Testing with various values reveals that the model's average accuracy is 71% at testing and 96% at the training phase.

5.2 Future Scope

So, the discourse feeling acknowledgment is an extremely fascinating subject and there is something else to find in the field, in our model the future work will incorporate the improvement of exactness of the model to come by improved results, we can likewise prepare the model to give aftereffects of the discourse that is longer in term, like in this model we can perceive the feeling just for brief length of time. In future we will ready to stack the more drawn-out example dataset and the model will arrange various feelings in various timeframe. Its future work can likewise incorporate the recording of on time information through a receiver with the goal that there is no need of stacking the dataset; we will just train the model and afterward information can be recorded to give the feelings of that individual's voice.

REFERENCES

We referred to the following links of some research papers and articles for our overall work:

1. https://ieeexplore.ieee.org/document/9453028
2. https://www.geeksforgeeks.org/what-is-speech-recognition/
3. https://scholar.google.com/citations?user=2Mh1CWwAAAAJ&hl=en
4. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Dr.+Nilesh+Shelke&btnG=
5. https://www.researchgate.net/publication/359381620_Two-Way_Feature_Extraction_for_Speech_Emotion_Recognition_Using_Deep_Learning
6. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
7. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
8. https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d
9. https://www.researchgate.net/publication/330796993_Chroma_Feature_Extraction
10. https://www.geeksforgeeks.org/classification-using-sklearn-multi-layer-perceptron/
11. https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/
12. https://colab.research.google.com/drive/1e5SU227zZkFfDQleJJCzjbmf6jx24XPX