Speech Emotion Recognition

**Project proposal submitted by:**

* **Ramsha Gopchandani (021-19-0020)**
* **Depeeka Ghai (021-19-0027)**

**Project proposal submitted to:**

* **Prof Dr. Sher Muhammad Daudpota**

Table of Contents

[Introduction 2](#_Toc103549298)

[Problem statement 2](#_Toc103549299)

[Methodology 2](#_Toc103549300)

[Dataset Discussion 3](#_Toc103549301)

[Major outcomes 3](#_Toc103549302)

[Project Timeline 4](#_Toc103549303)

[Conclusion 4](#_Toc103549304)

[References 4](#_Toc103549305)

# Introduction

In daily interpersonal human relationships, emotion plays a crucial role. Emotion provide a wealth of information about a person's mental state. This has initiated a new branch of study known as automatic emotion recognition, whose primary purpose is to recognize and retrieve sentiments[1].

Emotion recognization mainly performed by three ways, first one is textual data next, facial expressions and last one is speech. Speech is the most basic protocol for exchanging information from one end to the other. It may express rich emotional information through emotions. The challenge of detecting the emotional qualities of speech, regardless of the semantic analysis, is known as Speech Emotion Recognition (SER)[2]. Moreover speech also gives the understanding in both cultures cognitive and emotional parallely. While people are capable of performing or getting the emotions efficiently as a natural aspect of spoken communication, but the ability to do automatically using programmable devices is currently under investigation[3]. According to some theories, emotions are not causal forces but simply syndromes of components such as motivation, feeling, behavior, and physiological changes[4]. Automatic emotion detection systems are being researched in order to develop efficient, real-time techniques of identifying the emotions of mobile phone users, call centre operators and consumers, car drivers, pilots, and a variety of other human-machine interface users.

It is known that some physiological changes occur in the body due to people's emotional state. Although changes in the person's condition cannot be identified without the use of a portable medical device, facial expressions and vocal signals can be received without the use of any device. As a result, the majority of research on this subject has been on the automatic recognition of emotions utilising visual and aural data. However, acoustic signals are the most used data after facial signs to identify a person's emotional state [5]. Through acoustic signals we can recognize emotions more clearly. As we know people are more comfortable in speech to convey their message than text more specifically when we talk about emotions. Its hard to recognize emotions through text and as well as facial expressions as we have to face some challenges like if we talk about text, words semantic can vary from context to context or one word can have more than one meanings and if we talk about facial expressions so people can made fake emotions which leads to the false results or we will fail to recognize actual emtions. Moreover in the world of marketing , customers are so central, Every product which is made will be on the basis of users use or we can say demand. One more issue with facial expressions is it has limited genralizability which means emotions are vary from culture to culture or we can say country to country like In japan they mostly focuses on eyes to get emotions where in USA they mostly focuses on mouth to get emotions. It is very hard challenge to manage all the issues with text and facial expressions.

To resolve above all mentioned issues SER came under the research. The aim of SER is to recognize the emotional state of a speaker from acoustic signals so that we can know the person’s mood and respond accordingly. SER is one of more interesting field of machine learning which is sub field of Artificial intelligence( AI) as we have to work with person’s mental state. Our aim is to build a model which will classify the emotions through the audio/voice and also resolve even some of the issues we have mentioned above.

# Problem statement

Most of the time, people understand speech rather than text or images. Person’s mental state can identified through acoustic behavior is most preferable way because through speech we will express real emotional state and it also shows true behaviour.

Emotions can be recognized through many aspects such as text, images, videos, facial expressions, speech etc. Among all aspects speech is considered better than others. To recognize emotions from text is difficult and less accurate because word’s semantics can vary from context to context and one word can have more than one meanings but this issue can be solved through speech emotion recognization (SER). Facial expressions varies from country to country so we can not make one model for detecting the meotions for all the cultures/countries specifically leads to limited generalizability.

# Proposed Methodology

We have used Google collab for the writing up the code simply model intialization and usage. After that we installed some libraries related to audio like librosa package for audio analysis , speech-features for getting the audio features which are (chroma , mel frequency cepstrals coefficient and mel spectrogram) and so on. This kit was designed to be the best collection for dataset for music and audio research. Firstly, we have loaded dataset named Ryerson-audio visual database of emotional speech and songs (RAVDESS). Which hac acutaully 7000+ instances it has 24 actors 12 male and 12 females means gender balanced and every actor has 60 different audio files in north American accent. Which is total 24\*60 = 1440 files(istances). Next, we have performed Feature extraction in which we have extracted exactly 180 features of 3 major features chroma, mel frequency and mel spectrogram performed in three domains like preceptual, frequency and time. Features like Mel-frequency cepstrum coefficient (MFCC) , chroma and mel-spectogram lies is preceptual and frequency domain. Among all, MFCC is most used representation of the spectral property of voice signals. After that, we have splited our dataset into train and test sets 80 for training and 20 for testing. After this, we have also performed the preprocessing techniques as we had class label in string so we wanted to convert it in neumeric. Moreover, we have tried 3 models KNN, deep neural network and Multi layer perceptron(MLP) classifiers but among all MLP performed well. So we have intialized it and train the model on trainig data . Finally, we have find the score , or evaluated the model it gives us 65.58% accuracy on RAVDESS dataset with MLP classifier. In last we have drawn the confusion matrix and classification report and save the model in model.pkl file so that we have can use it in with Flask(pthon web framework).

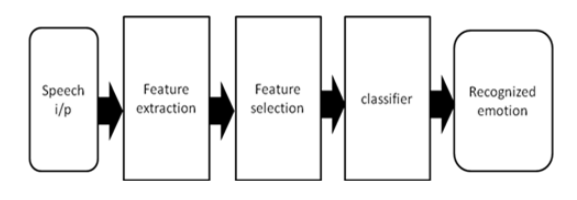


Figure 1: Proposed model

# 4. Dataset Discussion

RAVDESS actually stands for Reheryson audio visual database of emotional speech and songs. The RAVDESS dataset has selected which has actaull 7356 instanaces combination of audio and video where audio is also combination of speech and songs so as we are concern about speech only so we limitize the dataset upto 1440 files of speech of 24 actor 12 male and 12 females with different north american accent. This dataset categorized into eight different emotions by 247 untrained Americans: Happy, Sad, Fearful, anger, Disgust, Surprise and a neutral base for each performer.

* The data collection consists of 24 trained performers, 12 males and 12 females. Gender-balanced.
* In a standardized setting, the audio files were generated and each contains the same statements in an American focus. There are also two related file types:(o)The speech file contains 1440 files: 60 trials per actor x 24 actors = 1440 speech files.
* This dataset file name is consist of seven variable

Moreover this dataset filename has 7 identifiers as neumerics like

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

Example

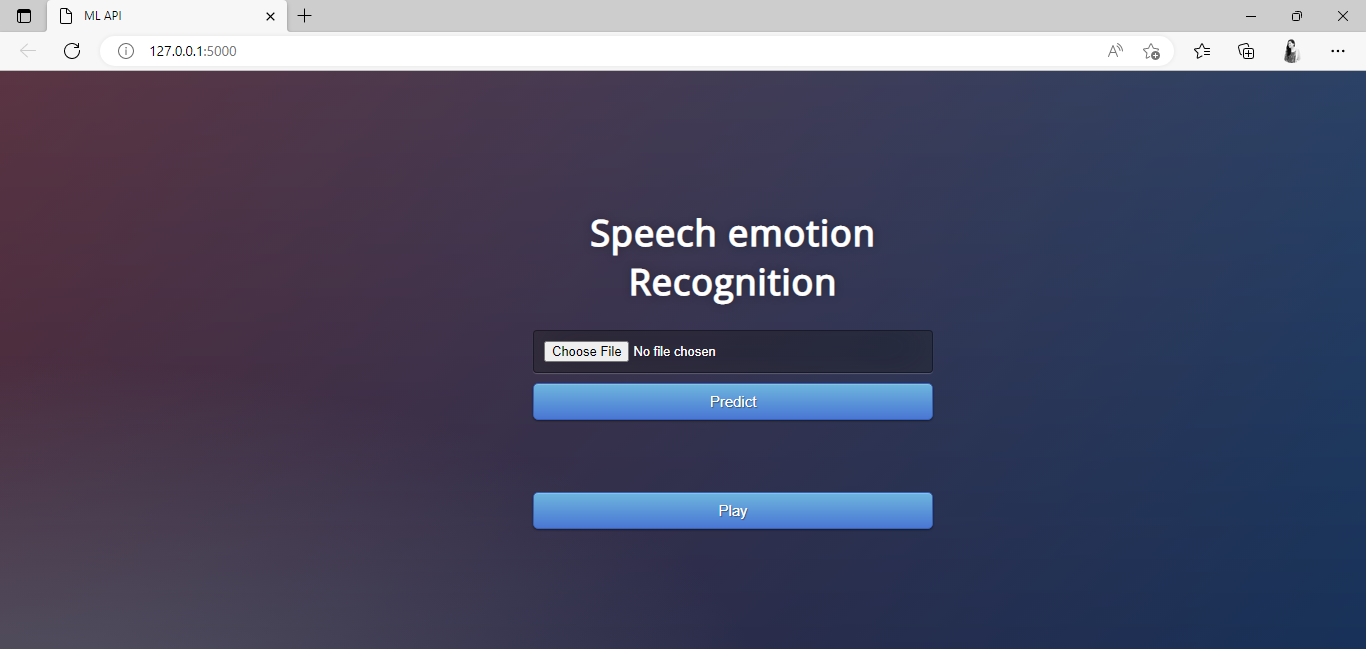
Filename example: 02-01-06-01-02-01-12.mp4

1. Video-only (02)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)
8. Female, as the actor ID number is even.

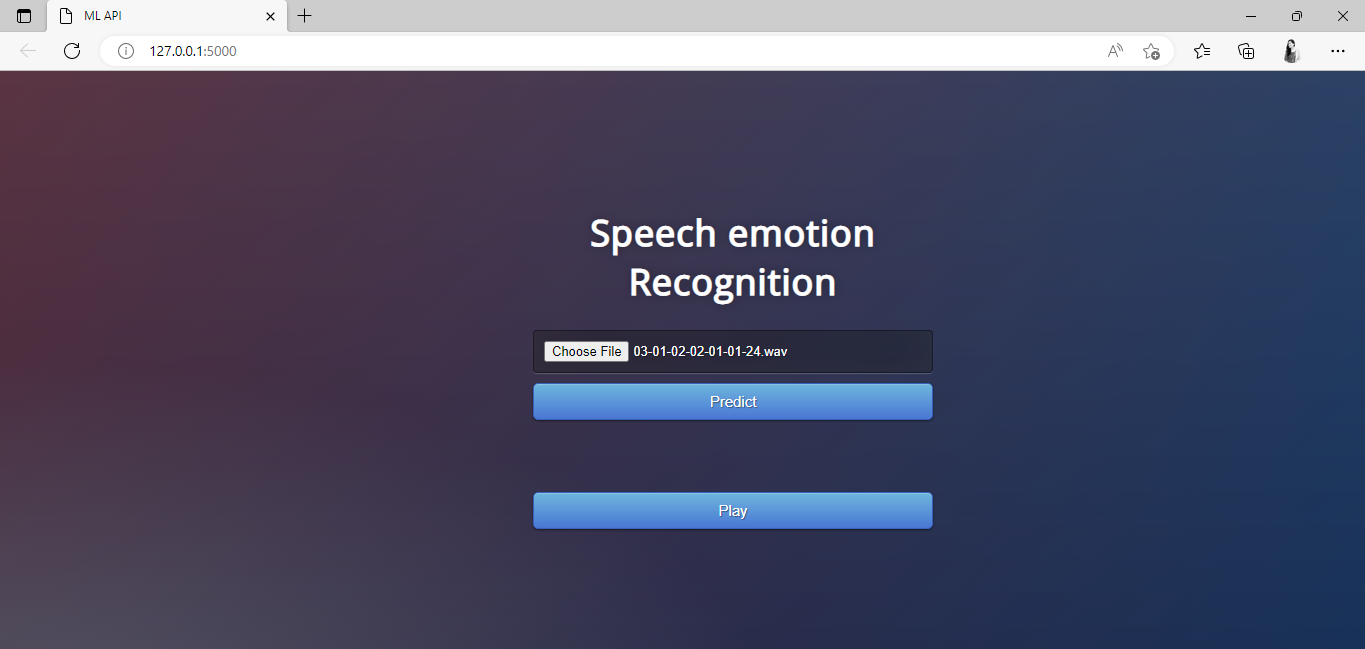
# 5. Major outcomes

After the completion of this project, We will have a web application developed with the help of  Flask and Python with the trained model on the RAVDESS dataset. That will recognize emotions through speech. If we talk about GUI then there will be one text field for speech input and four buttons for browsing, playing, Classifying, and exiting. If a user clicks on the browsing button pop-up of device storage will appear from which we can select the required audio file. If the user clicks on the playing button, the selected speech will be played. If the user clicks on the classifying button model will print the recognized emotion on the label. While clicking on the exiting button, the application will be closed. Moreover, this can be used in a variety of situations, including call centers for customer service or marketing, voice-based virtual assistants or chatbots, linguistic research, and so on [1].

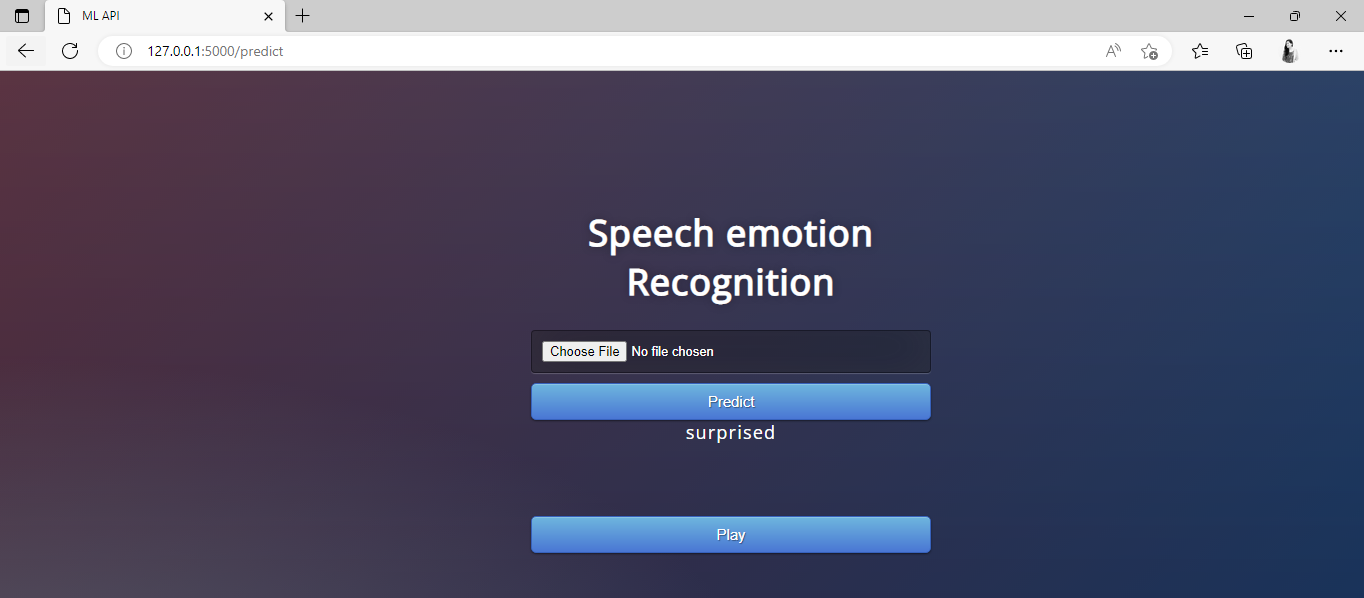
**Step:1 No file choosen , main page**



**step:2 Here by clicking on choose file button , it will open dailogue box of local pc where we can select the file.**



**step:3 Here by clicking on predict button , it will print the emotion on the screen and also clicking on play it will play the selected speech.**



# 6. Model accuracy

We have tried three models named as KNN whose accuracy is 61%, Deep neural network whose accuracy is 53% and MLPClassifier whose accuracy is 65.58% so we decided to choose this model and further use and deployed it.

# 6. Project Timeline

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **May** | | | **June** | | | |
| 1-10 | 11-20 | 21-31 | 1-10 | 11-17 | 18-25 | 25-31 |
| Literature Review |  |  |  |  |  |  |  |
| Problem Identification |  |  |  |  |  |  |  |
| System Requirements |  |  |  |  |  |  |  |
| System Design |  |  |  |  |  |  |  |
| System Development |  |  |  |  |  |  |  |
| System Testing |  |  |  |  |  |  |  |

# 7. Conclusion

Automatic speech emotion recognition is an emerging field that is increasing nowadays which results in better human and machine interaction [[6].](#_References) The system performs all steps like how we can use machine learning algorithms to extract underlying emotions from speech audio data, as well as some insights into human emotion through voice. Firstly we have loaded dataset and split it into train and test set. Next step feature extraction and feature selection is accomplished. Next, the MLP classifier is selected among KNN and SVM on the basis of accuracy to train a model and classify speech according to the respective emotion that speech contains. In last but not least we have integrated our model with Flask which is python web framework.

# 8. References

1. ("Introduction to Emotion Recognition 2021 | RecFaces", 2022)
2. (Kerkeni et al., 2022)
3. (Lech, Stolar, Best and Bolia, 2014)
4. (Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network, 2012)
5. (APA PsycNet, 2015)
6. (ijlaas 2016)