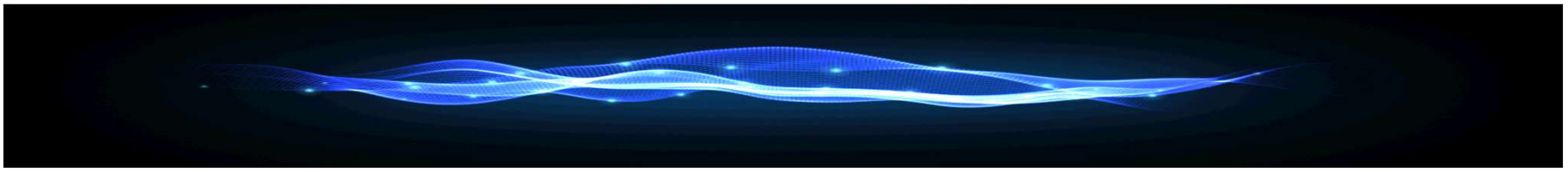
# Capstone Project

# Verbal Communication Quality Monitoring & Feedback System

## (speech Emotion Recognition)

-Hrithik Chourasia



# What is Speech Emotion Recognition (SER) ?

- Speech Emotion Recognition (SER), is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch.
- This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.
- SER is tough because emotions are subjective and annotating audio is challenging.

- Why we need it?

- 1. Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data.

- 2. SER(Speech Emotion Recognition) is used in call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on.. for helping companies improving their services.

- 3. It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen.

# Project Introduction & Problem Statement

Verbal communication includes sounds, words, language, and speech.
Speaking is an effective way of communicating and helps in expressing our emotions in words.
Speech is the most natural way of expressing ourselves as humans.
It is only natural then to extend this communication medium to computer applications.
Verbal Communication is valuable and sought after in workplace and classroom environments alike.
Clear and comprehensive speech is the vital backbone of strong communication and presentation skills.
Millions of people are affected by stuttering and other speech disfluencies, with the majority of the world having experienced mild stutters while communicating under stressful conditions.
Research shows that mild disfluencies can be cured without medical help, just practicing speech regularly and constructive feedbacks are effective ways to improve.

**The above-mentioned problem solved by applying deep learning algorithms to audio/speech data.**
**The solution will be to identify emotions in speech.**

# Data description:

## 1. Ryerson Audio-Visual Database of Emotional Speech (Ravdess):

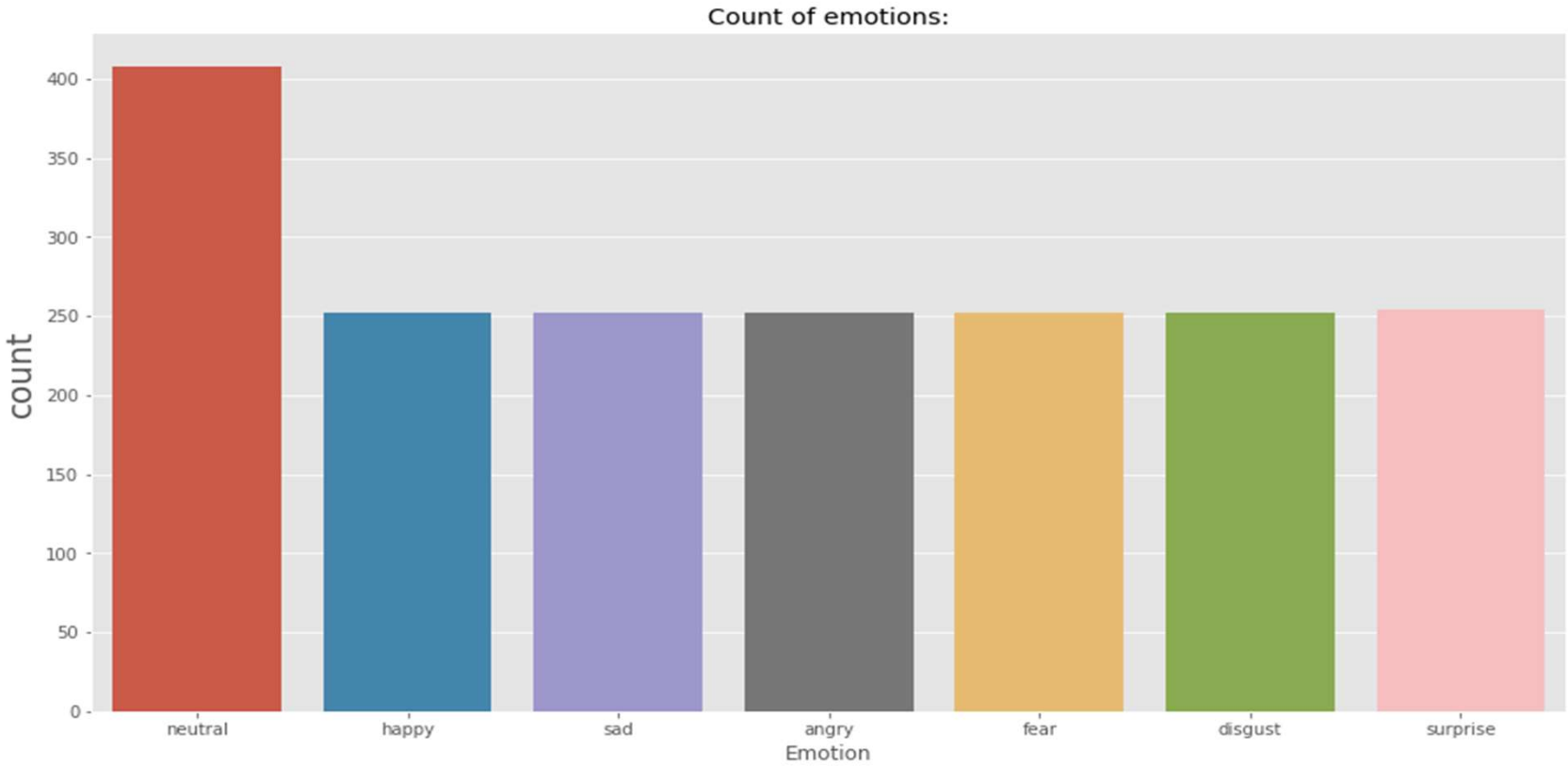
This dataset includes around 1500 audio file input from 24 different actors.

12 male and 12 female where these actors record short audios in 8 different emotions i.e 1 = neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7 = disgust, 8 = surprised.
Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

## 2. Surrey Audio-Visual Expressed Emotion (Savee):

This dataset contains around 500 audio files recorded by 4 different male actors.
The first two characters of the file name correspond to the different emotions that the portray.

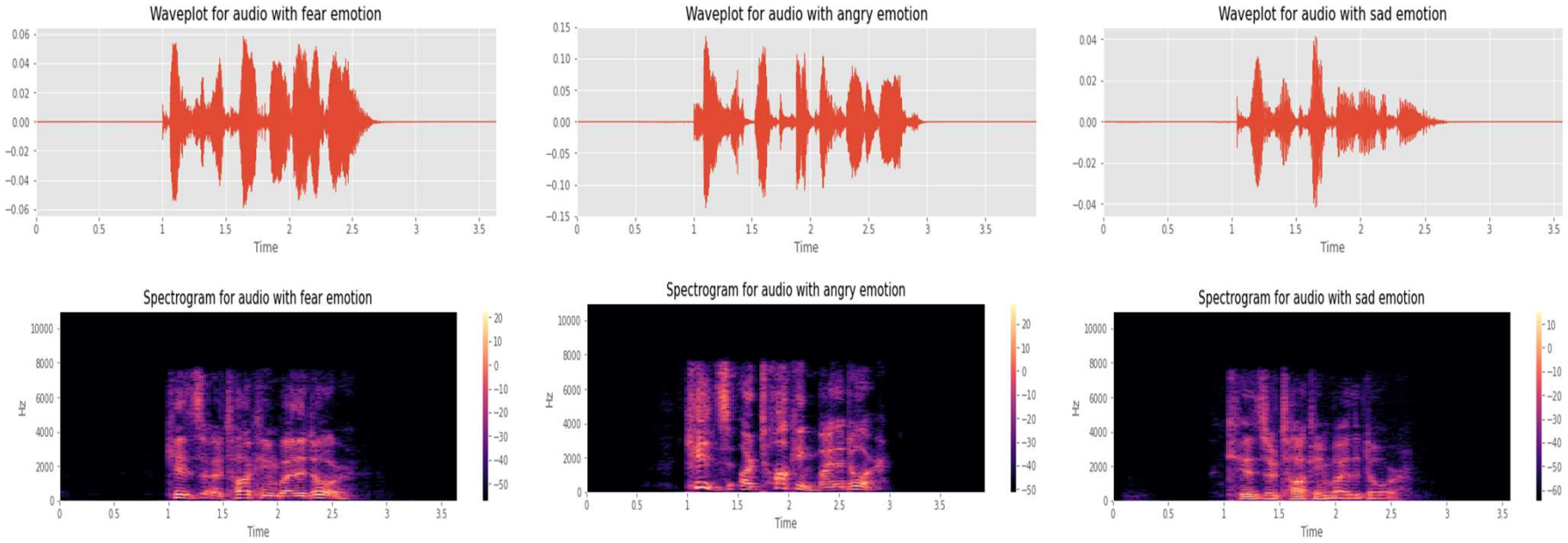
# Emotions in dataset



# Wave plots and Spectrograms for audio signals

**Waveplots** – Waveplots let us know the loudness of the audio at a given time.

**Spectrograms** – A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. It's a representation of frequencies changing with respect to time for given audio/music signals.



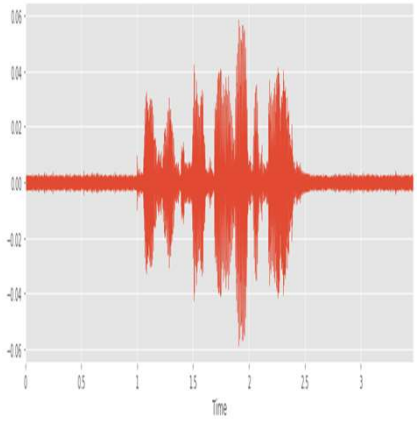
# Data Augmentation

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set.

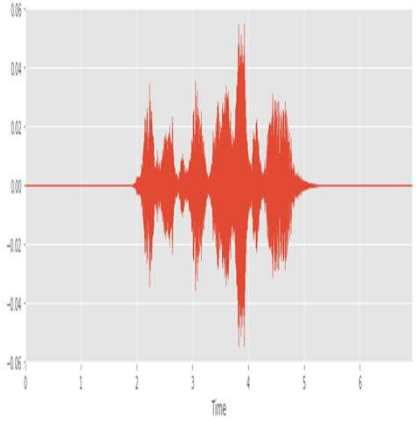
**Some ways for data augmentation in sound data:**

1. Noise injection - It simply adds some random value into data by using NumPy.
2. Stretching - Changing the speed/duration of sound without affecting the pitch of sound.
3. Shifting - Shift the wave with the help of sample rate & factor. This will move the wave to the right by given factor along time axis.
4. Pitching - It is a process of changing the pitch of sound without affect its speed.

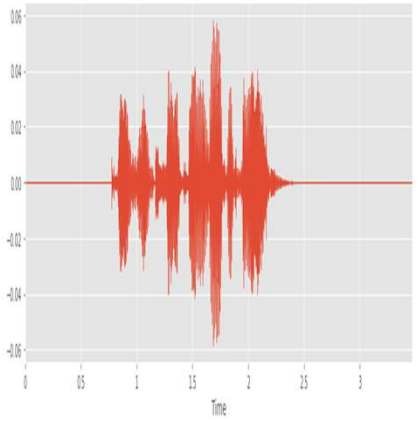
Noise injection



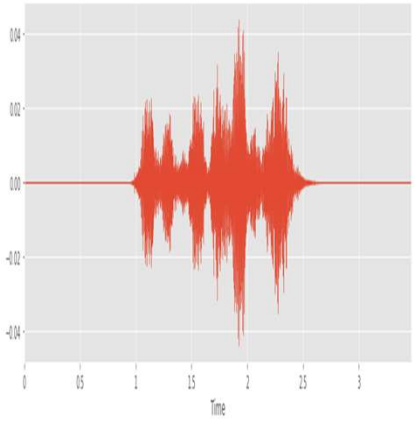

Stretching




Shifting




Pitching



# Feature Extraction

- Extraction of features is a very important part in analyzing and finding relations between different things.
  As we already know that the data provided of audio cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction is used.

- Used just 3 main features for this task after experimenting with many features: ZCR, RMS and MFCC

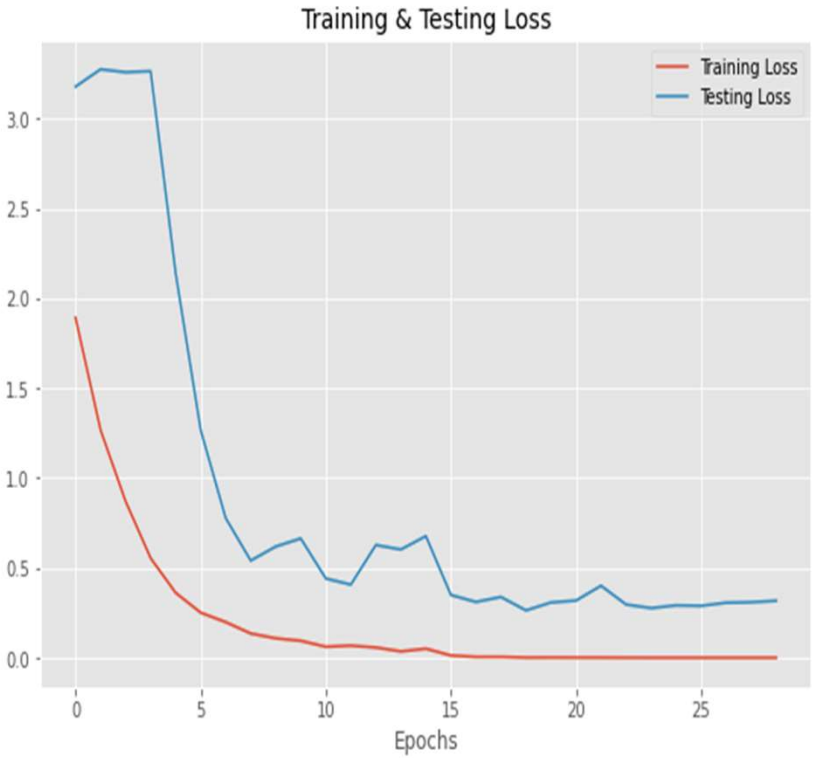
- **Zero Crossing Rate :** The rate of sign-changes of the signal during the duration of a particular frame. It indicates the number of times that a signal crosses the horizontal axis, i.e. the number of times that the amplitude reaches 0.

- **MFCCs :** Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
- The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency.

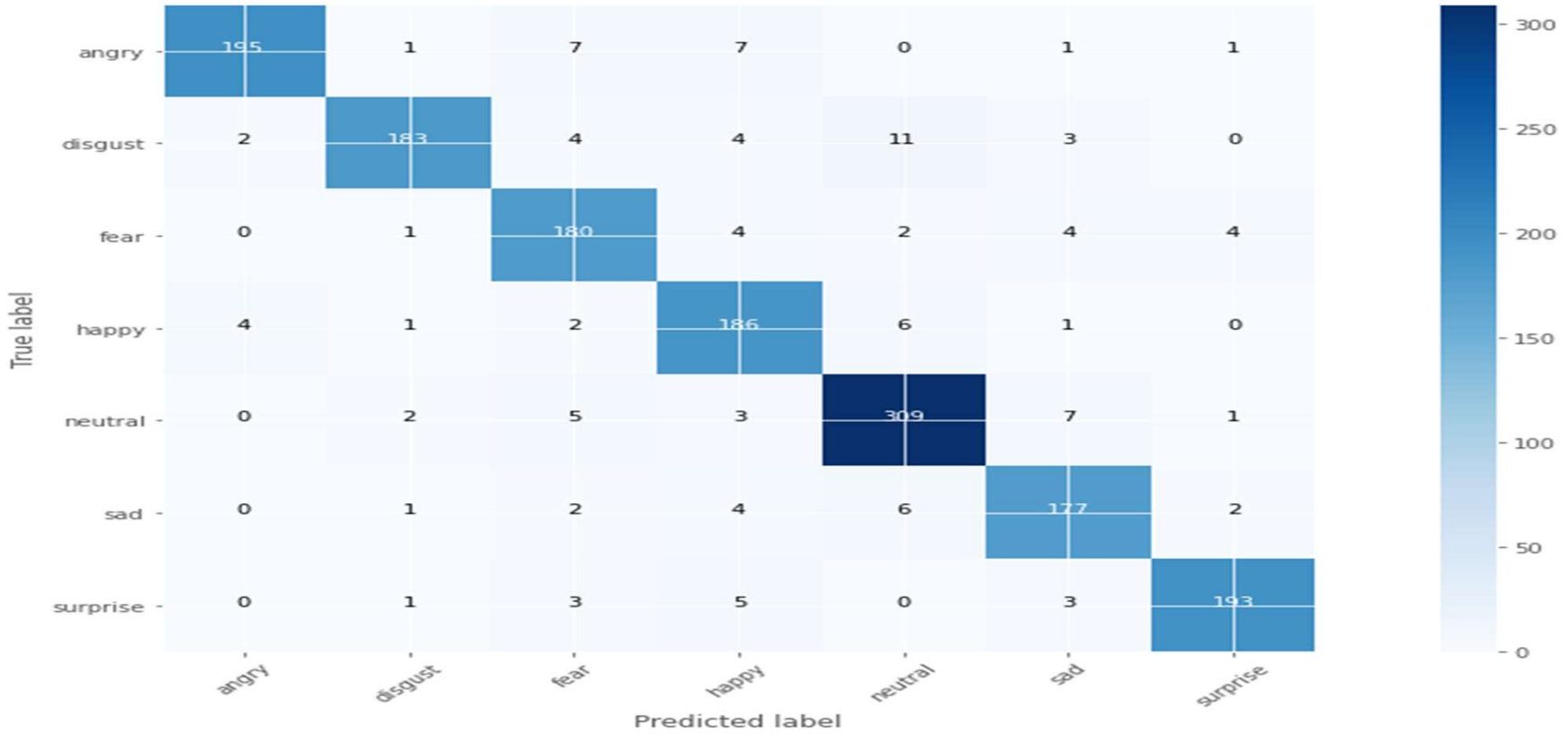
- **RMS** : The RMS Energy (RMSE) is simply the square root of the mean squared amplitude over a time window.

# Accuracy of our model on test data :

92.52275824546814 %




# Confusion Matrix



# Challenges:

- Emotions are subjective and it is hard to notate them.
- Lack of data is a crucial factor to achieve success in SER, however, it is complex and very expensive to build a good speech emotion dataset.
- Data provided by audio cannot be understood by the models directly so, Converted them into an understandable format with the help of feature extraction and decide which feature extraction is suitable for this audio data.
- Deciding the input for model: a sentence, a recording or an utterance.

# Conclusion:

- Deep learning can be used in Verbal Communication Quality Monitoring & Feedback System to process audio data in real time.
- With speech emotion recognition, It can identify or predict speech emotion after recording audio during speaking or drag & drop audio files.
- Speech emotion recognition system can identify the mental state of user, conversational analysis to improve customer satisfaction.
- Overall achieved 92% accuracy on test data but we can improve it more by adding more audio data, applying more augmentation techniques and using other feature extraction methods.