# AI-BASED EMOTION RECOGNITION FROM SPEECH/AUDIO FILE

COLLEGE: LINGARAJAPPA ENGINEERING COLLEGE, BIDAR

GUIDE (S) : Dr. SHARANBASAPPA SHETKAR
STUDENT (S) : Ms. ABHISHEK HONDALE
Ms. PRATIKSHA MASHETTY
Ms. SURIYA BEGUM

## KEYWORDS

SER, MFCC, LSTM, IoT, CTC, MLP, ML, AI, DFD, etc..

## BACKGROUND

Speech Emotion Recognition (SER) is the task of recognizing is the task of 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 a programmable device is still an ongoing of research. Studies of automatic emotion recognition systems aim to create efficient, real time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers , pilots ,and many other human and machine communication users .

Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human - like manner Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities. In some circumstances human could be replaced by computer generated character having the ability to conduct very natural and convincing conversations by appearing to humanemotions. Machines need to understand emotions conveyed by speech. Only with this capability, an entirely meaningful dialogue based on mutual human machine trust and understanding can be achieved. Traditionally, machine learning involves the calculation of feature parameters from the raw data (e.g., Video, speech, images, ECG, EEG).

The features are used to train a model that learns to produce desired output labels. A common issue faced by this approach is the choice of features. In general, it is not known which features can lead to the most efficient clustering of data in to different categories. Some insights can be gained by testing a large number of the different feature selection techniques. The quality of the resulting hand-crafted features can have a significant effect on classification performance.

# OBJECTIVES

- To build a model to recognize emotion from speech using the Librosa and sklearn libraries and the RAVDESS data set.
- To present a classification model of emotion elicited by speeches based on deep neural networks MLP classification based on acoustic features such as Mel Frequency Cepstral Coefficient (MFCC). The model has been trained to classify eight different emotions (calm, happy, fearful, angry, disgust, neutral, surprised, sad).
- Emotion Recognition from Bidar Kannada**.**

# METHODOLOGIES

## MFCC

For human speech, in particular, it sometimes helps to take one additional step and convert the Mel Spectrogram into MFCC (Mel Frequency Cepstral Coefficients).

## Data Augmentation of Spectrograms

We can now apply another data augmentation step on the Mel Spectrogram images, using a technique known as Spec Augment.

## CTC Algorithm

CTC is used to align the input and output sequences when the input is continuous and the output is discrete, and there are no clear element boundaries that can be used to map theinput to the elements of the output sequence

## CTC works in two modes:

- **CTC Loss** (during Training): It has a ground truth target transcript and tries to train the network to maximize the probability of outputting that correct transcript.
- **CTC Decoding** (during Inference): Here we don’t have a target transcript to refer to,and have to predict the most likely sequence of characters.

### TC Decoding

- Use the character probabilities to pick the most likely character for each frame, including blanks. e.g., “*-G-o-ood*”

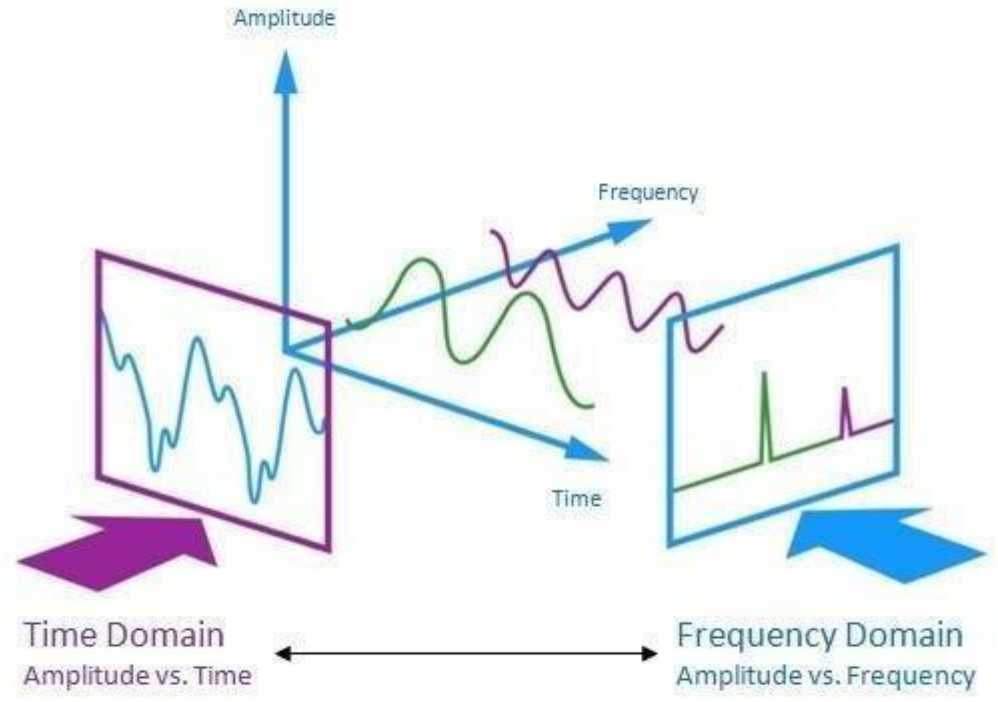
CTC Loss

The Loss is computed as the probability of the network predicting the correct sequence. To do this, the algorithm lists out all possible sequences the network can predict, and from that it selects the subset that matches the target transcript.

## 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.

The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.



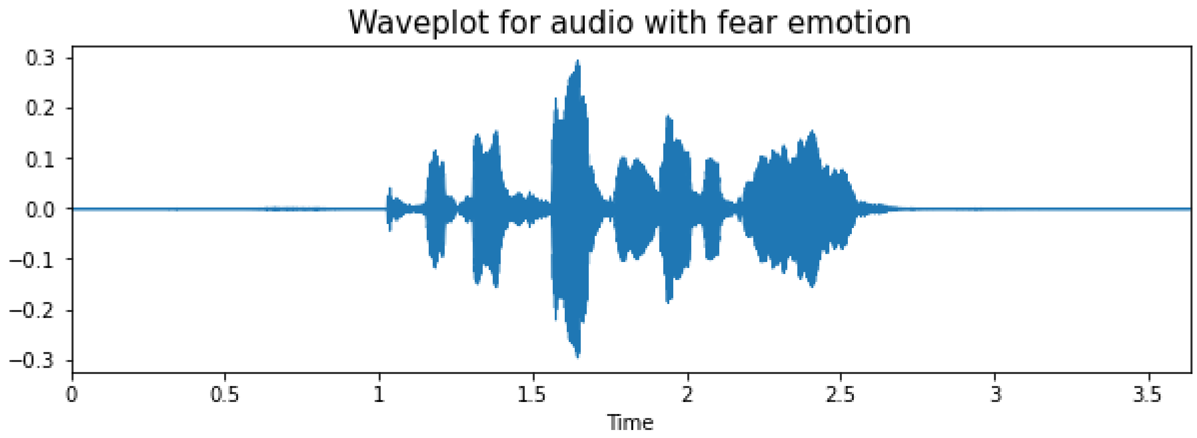
1. Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.
2. Energy : The sum of squares of the signal values, normalized by the respective framelength.
3. Entropy of Energy : The entropy of sub-frames’ normalized energies. It can beinterpreted as a measure of abrupt changes.
4. Spectral Centroid : The center of gravity of the spectrum.

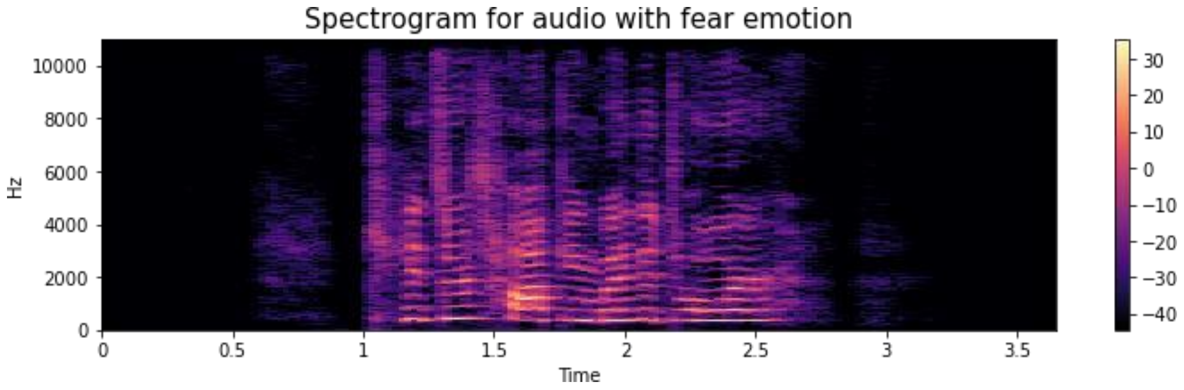
5. Spectral Spread : The second central moment of the spectrum.
6. Spectral Entropy : Entropy of the normalized spectral energies for a set of sub-frames.
7. Spectral Flux : The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8. Spectral Rolloff : The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9. MFCCs Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
10. Chroma Vector : A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
11. Chroma Deviation : The standard deviation of the 12 chroma coefficients.

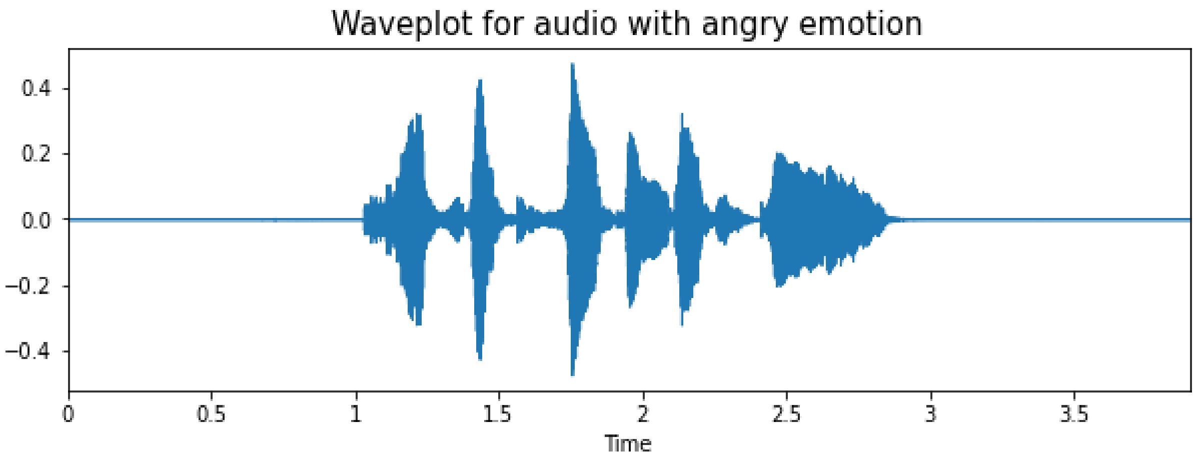
In this project i am not going deep in feature selection process to check which features are good for our dataset rather i am only extracting 5 features:

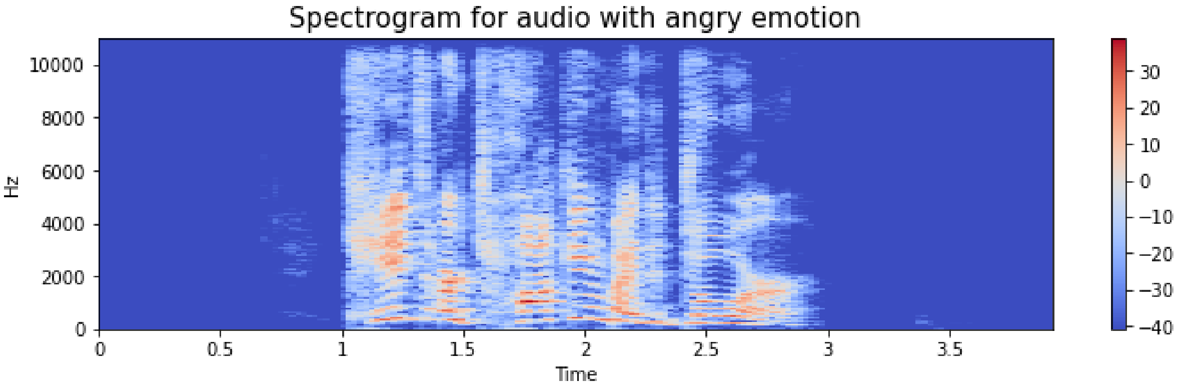
- Zero Crossing Rate
- sift
- Chroma_MFCC
- RMS(root mean square) value
- MelSpectogram to train our model

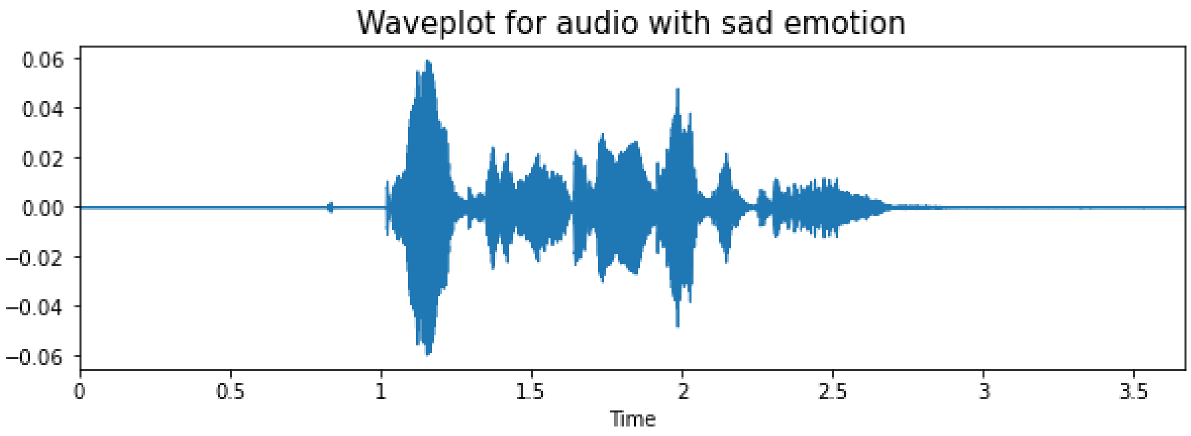
# RESULTS

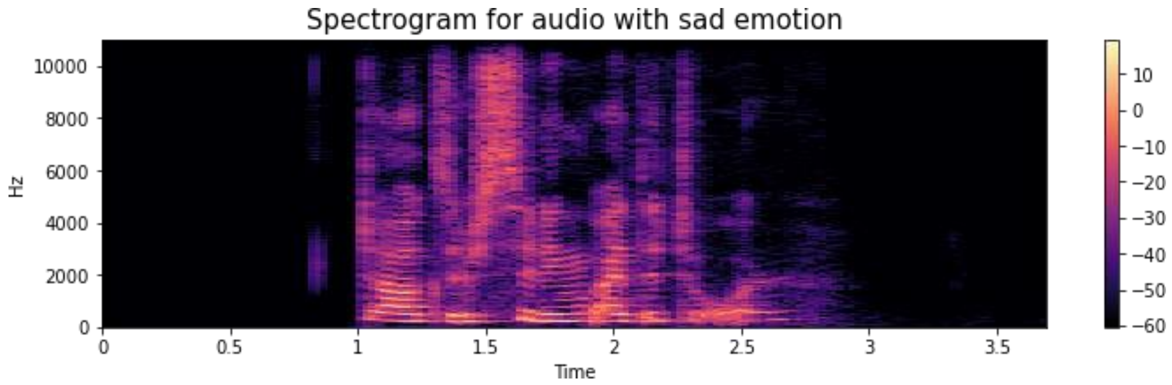

Waveplot for audio with fear emotion
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
0.5
1
1.5
2
2.5
3
3.5
Time

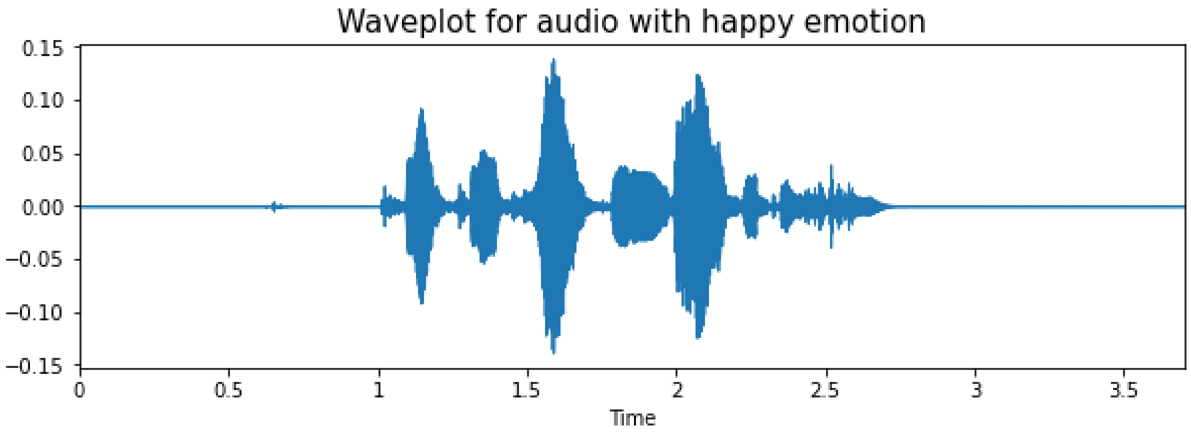

Spectrogram for audio with fear emotion
Hz
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
Time
30
20
10
0
−10
−20
−30
−40

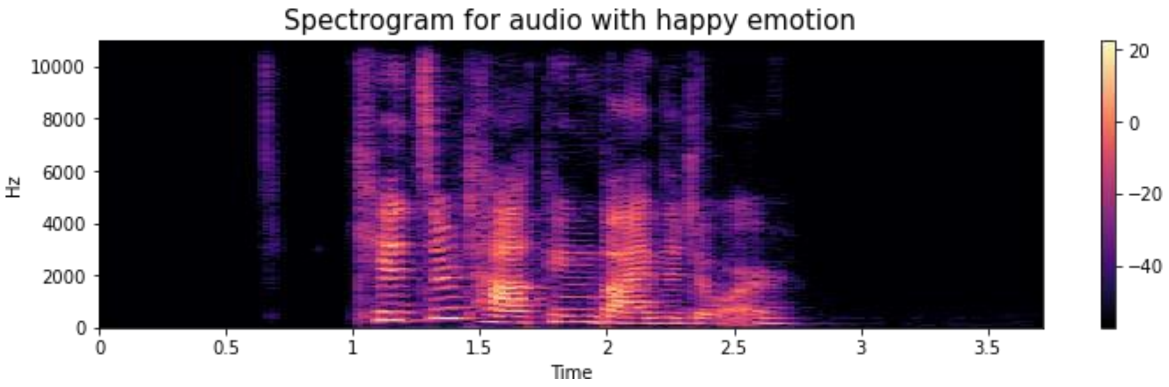

Waveplot for audio with angry emotion
0.4
0.2
0.0
−0.2
−0.4
0
0.5
1
1.5
2
2.5
3
3.5
Time


Spectrogram for audio with angry emotion
Hz
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
Time
30
20
10
0
−10
−20
−30
−40


Waveplot for audio with sad emotion
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
0.5
1
1.5
2
2.5
3
3.5
Time


Spectrogram for audio with sad emotion
Hz
10000
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
Time
10
0
−10
−20
−30
−40
−50
−60


Waveplot for audio with happy emotion
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
0.5
1
1.5
2
2.5
3
3.5
Time


Spectrogram for audio with happy emotion
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
2.5
3
3.5
Time
20
0
−20
−40

# CONCLUSION

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.

In this project, 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.

# SCOPE FOR FUTURE WORK

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