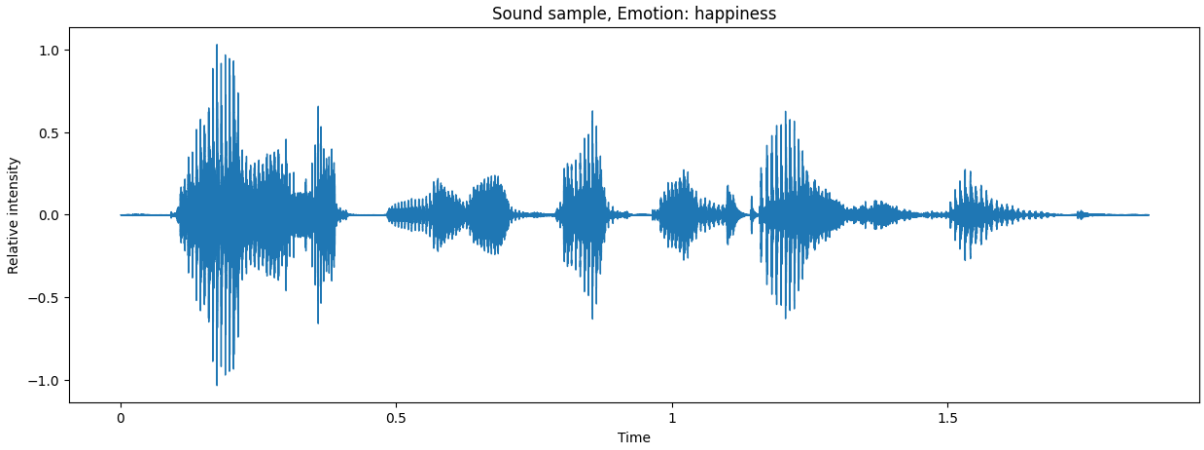
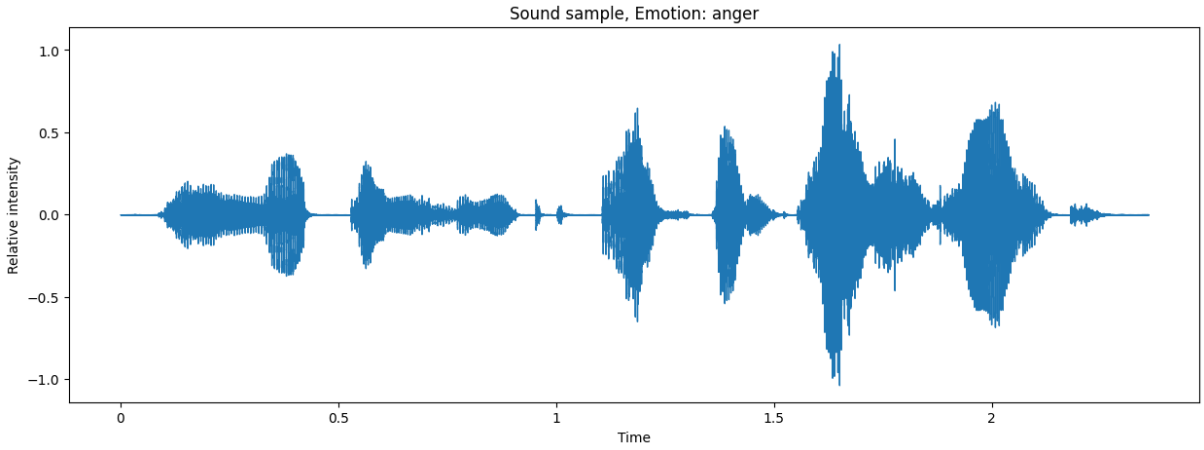
Task:

Given an audio signal, predict the emotion of the speaker.

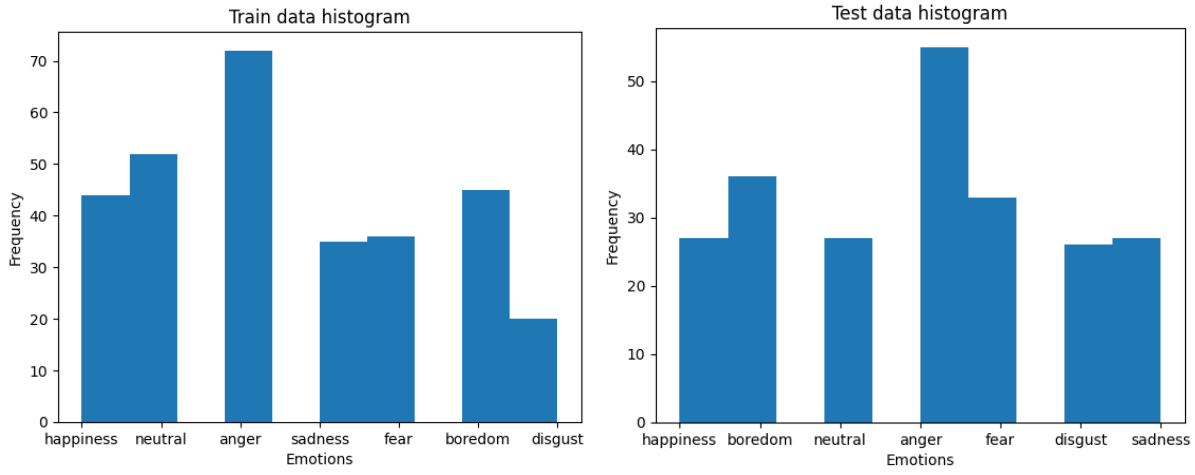
Audio samples: Two samples are given below.





Dataset:

Emodb: Audio samples spoken by actors. There are 535 audio files with 304 samples in the training set and 231 samples in the test set.



Model 1 using spectrogram features

Feature extraction: The spectral features such as spectral centroid, spectral contrast, spectral bandwidth, spectral roll off, crossing rate, mel-frequency cepstral coefficients and mel-scaled spectrogram features are extracted using librosa.

Preprocessing: Normalized the features and addressed class imbalancing

Model: Trained a convolutional neural network model

Model 2 using Raw Features

Feature extraction: The acoustic features are extracted using Opensmile package and extracted features using compare 2016 feature set.

Preprocessing: Normalized the features and addressed class imbalancing

Model: Trained a fully connected neural network model

Results

CNN model using Spectrogram

The model is evaluated using accuracy and average recall scores. The feature level study is conducted and observed that mel spectrogram features can predict as same as what all features together can predict.

Features	Accuracy	Average Recall
MFCC	61	59.3
Mel Spectrogram	63.2	61.9
MFCC, Mel Spectrogram, Crossing rate, Roll off, Bandwidth, Contrast	61	60.9
MFCC, Mel Spectrogram, Crossing rate, Roll off, Bandwidth, Centroid	61.9	60.6
MFCC, Mel Spectrogram, Crossing rate, Roll off, Contrast, Centroid	63.6	61.5
MFCC, Mel Spectrogram, Roll off, Bandwidth, Contrast, Centroid	63.6	62.3
MFCC, Mel Spectrogram, , Crossing rate, Bandwidth, Contrast, Centroid	61	59.3
Mel Spectrogram, Crossing rate, Roll off, Bandwidth, Contrast, Centroid	62.3	60.6

MFCC, Crossing rate, Roll off, Bandwidth, Contrast, Centroid	61.5	58.4
MFCC, Mel Spectrogram, Crossing rate, Roll off, Bandwidth, Contrast, Centroid (all features)	**64.5**	**62.3**

Model using raw features

The feed forward neural network model trained using raw features provided an accuracy of 75.76% and average recall of 74.46%.

The model based on raw features is performing much better than the spectrogram based model. This indicates that the time duration of audio is playing a key role in detecting the emotion.