Speech Emotion Detection

ML-15
Ayush Awasthi
Atharva Nanoti
Shaurya Khetarpal
Jasmer Singh Sanjotra

Problem Statement

The objective of this project was to develop a robust Speech Emotion Detection system. The main task was to analyze the speech signals and classify them into various emotional states like ‘happy’, ‘angry’, ‘sad’ etc. We needed to train the model and make it as accurate as possible in order to apply it in real-time.

Methodology

- Preprocessing
- Feature Extraction
- Training
- Results

Preprocessing

1. We used kaggle notebook for preprocessing and training the model as it already had access to various datasets that we could try.

2. We combined the datasets and made a dataframe with the files and their respective emotions.

3. Then, we divided the dataset into training and testing parts.

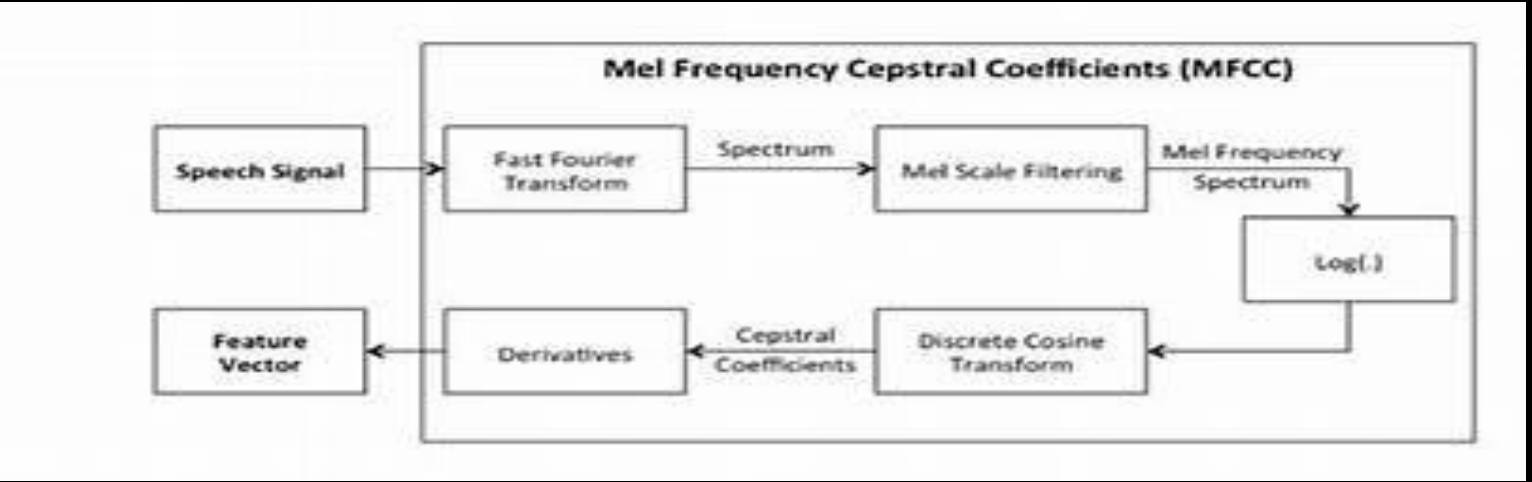
4. We applied the following augmentation techniques to the dataset: noise addition, pitch shift and time stretch. We used *numpy* and *librosa.effects* for this.

5. We combined the features into numpy arrays and appended the emotions by one-hot encoding them.

Feature Extraction

MFCCs are widely used in speech and audio processing tasks, including speech recognition, speaker identification and speech emotion recognition, due to their effectiveness in capturing relevant acoustic characteristics of the audio. MFCCs represent speech in a better way by taking the advantage of auditory perceptions in human speech.

Lets see how it works:



- Map the audio signal from the time domain to the frequency domain using the fast Fourier transform, perform this on overlapping windowed segments of the audio signal.
- Convert the y-axis (frequency) to a log scale and the color dimension (amplitude) to decibels to form the spectrogram.
- Map the y-axis (frequency) onto the mel scale to form the mel spectrogram.
- Mel scale: unit of pitch such that equal distances of pitch sound equally distant to listener.It can be obtained by above formula-

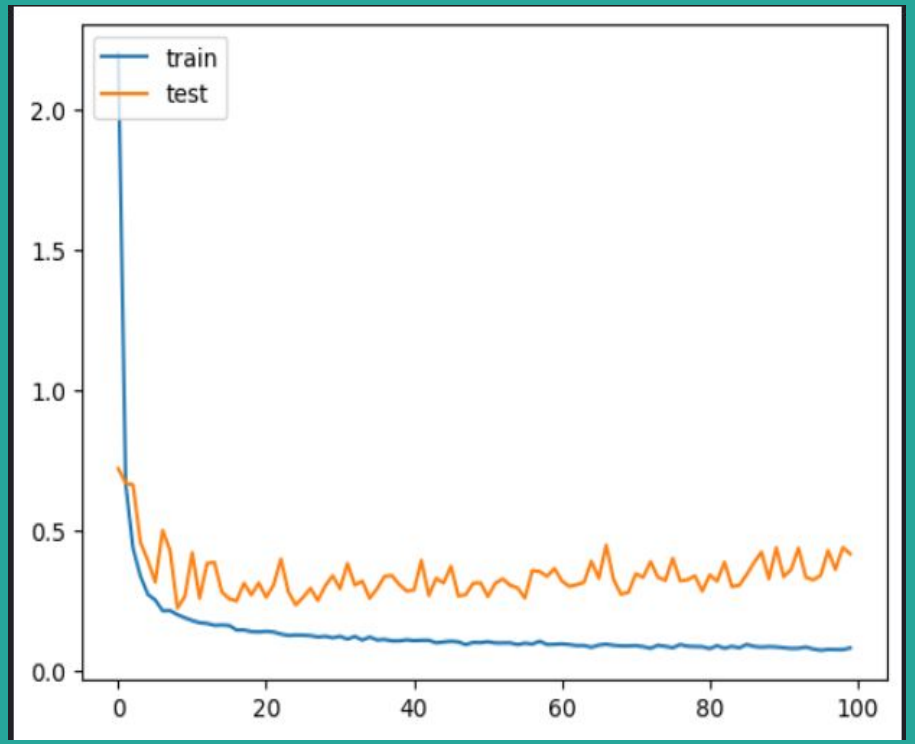
$$M(f) = 1125 \ln(1 + f/700)$$

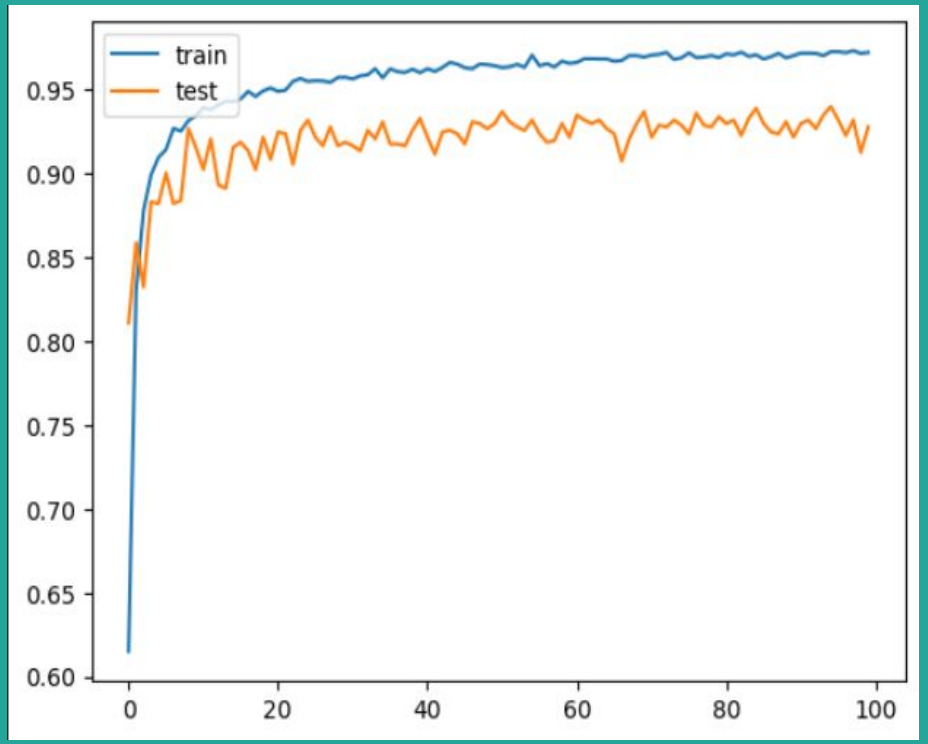
- Apply the Discrete Cosine Transform (DCT) to the log filter-bank energy.The DCT outputs the MFCC coefficients,representing the cepstral features.
- The mfcc removes fine spectral structure which is often less important.
- Other audio features such as chroma stft,mel spectogram,zero crossing rate,etc were also extracted and tested but mfcc was the one which gave the best results.

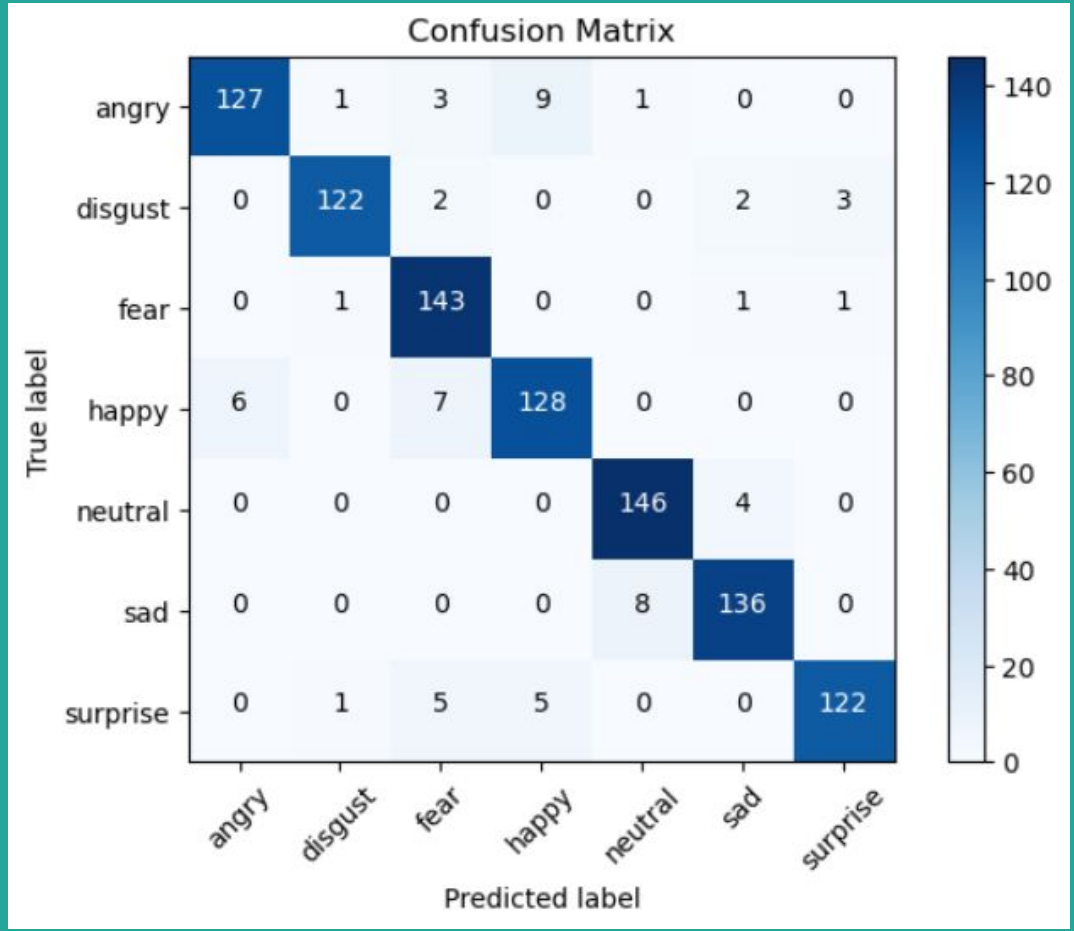
Model

1. We tried several model architectures like LSTM, Bi-directional LSTM, CNN etc.
2. CNN architecture gave the best accuracy among all the models and surprisingly, the model architecture which gave the best accuracy was fairly simple with one convolutional layer, one activation layer, and a flatten and dense(output layer).
3. We trained this model on the training data with 100 epochs.
4. We used two callbacks: ReduceLROnPlateau and EarlyStopping.
5. We used categorical_crossentropy as our loss function, as it is best suited for multiclass classification
6. Our model gave a validation accuracy of 92%.

Loss and Accuracy Plots

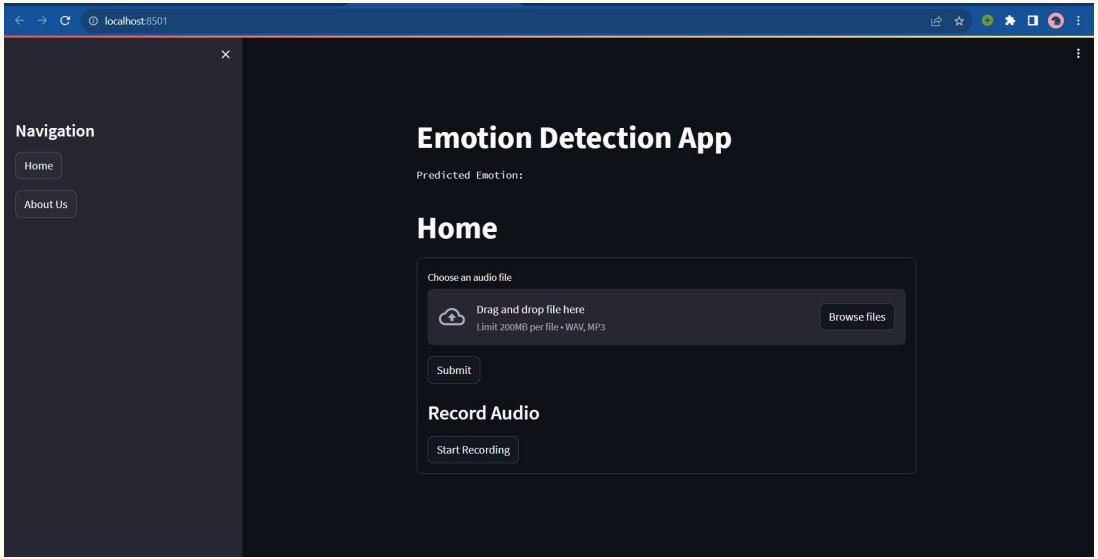





Confusion Matrix
angry
disgust
fear
happy
neutral
sad
surprise
True label
Predicted label
127 1 3 9 1 0 0
0 122 2 0 0 2 3
0 1 143 0 0 1 1
6 0 7 128 0 0 0
0 0 0 0 146 4 0
0 0 0 0 8 136 0
0 1 5 5 0 0 122
140
120
100
80
60
40
20
0

Website

- Used Streamlit to deploy our model locally.
- Audio files can be uploaded from the machine, or can be recorded directly as well.
- Audio is processed in real time and the result is displayed.



References:

1. SAVEE Dataset
2. TESS Dataset
3. MFCCs
4. Basics of CNN
5. CNN - Pooling Layers
6. CNN - Padding and Stride

Practical Applications of this project

The ability to detect the emotion of a speaker from their speech is a valuable tool in a variety of applications.

For example, in customer service,virtual assistants can be used to identify frustrated or angry customers, so that they can accordingly provide suggestions.

Speech emotion detection could be used in the security domain. It can detect whether someone is sounding frustrated or angry, and can accordingly determine whether the person can take hostile action.

Conclusion

The Speech Emotion Detection project successfully implemented a CNN-based model and deployed it on Streamlit for emotion recognition in speech signals. The model achieved a very good validation accuracy, demonstrating its potential in practical applications, such as sentiment analysis and human-computer interaction. With further improvements and enhancements, the system can be optimized for even more accurate and efficient emotion detection in speech.