Speech Emotion Recognition (SER) - Discrete classification of audio files into 8 different emotions using audio features

December 2020

Deepanshu Goyal

Naïve Bayes Model is able to classify speech emotion with an accuracy of 95.83%

Audio features used for machine learning model are:

Time Domain Features

- Root Mean Square (RMS) Energy
- Zero Crossing Rate (ZCR)

Time Frequency (Spectrogram) Features

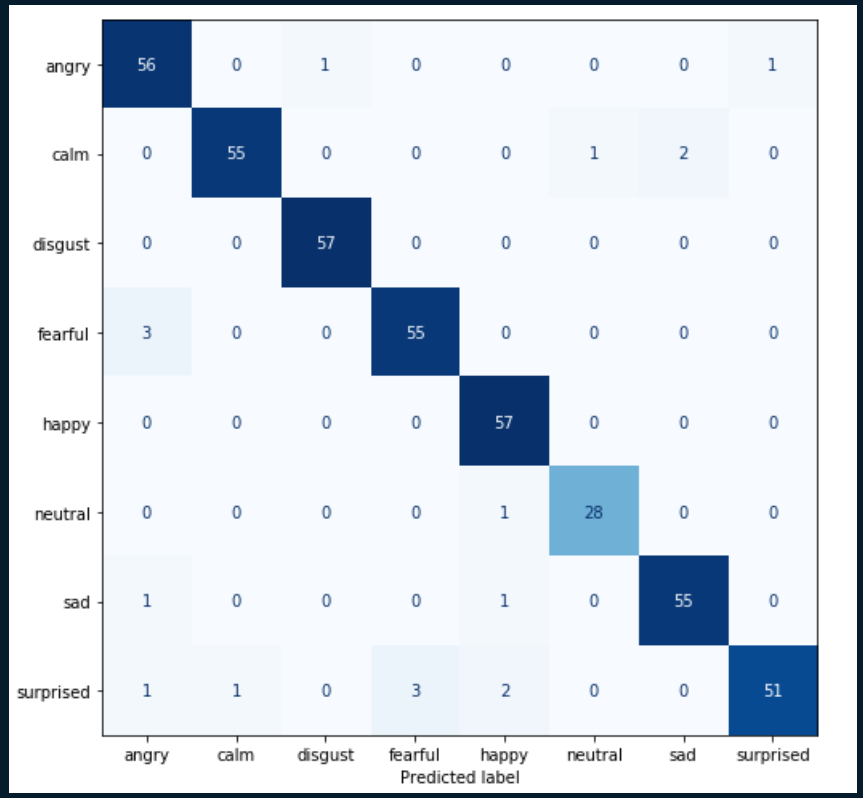
- Mel-spectrogram
- Chroma_stft
- Spectral Centroid
- Band Energy Ratio
- Bandwidth

Cepstrum (Spectrum of Spectrum) Features

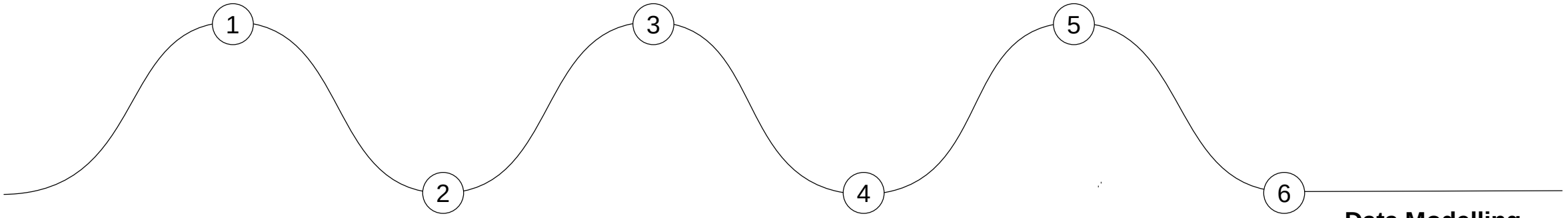
- MFCC

Mel-spectrogram & MFCC are important features of audio files as they have significantly improved the accuracy of the classification model

Model	Accuracy	F1 Score
Logistic Regression	95.37%	95.46%
Random Forest	91.67%	91.47%
Naïve Bayes	95.83%	95.87%
Gradient Boosting	91.90%	91.85%
XGBoost	93.52%	93.61%



Approach to Data Modelling



Exploratory Data Analysis

Understand & visualize the underlying data using plots, waveform and spectrogram

Data Cleaning

Remove noise from the audio data

Feature Extraction

Extract time domain, time frequency and cepstrum features using lebrosa library

Data Pre-processing

Transform features to follow normal distribution & scale data with mean = 0 and standard deviation = 1

Dimensionality Reduction

Perform LDA (Linear Discriminate Analysis) to reduce data dimensions from 569 to 7 features

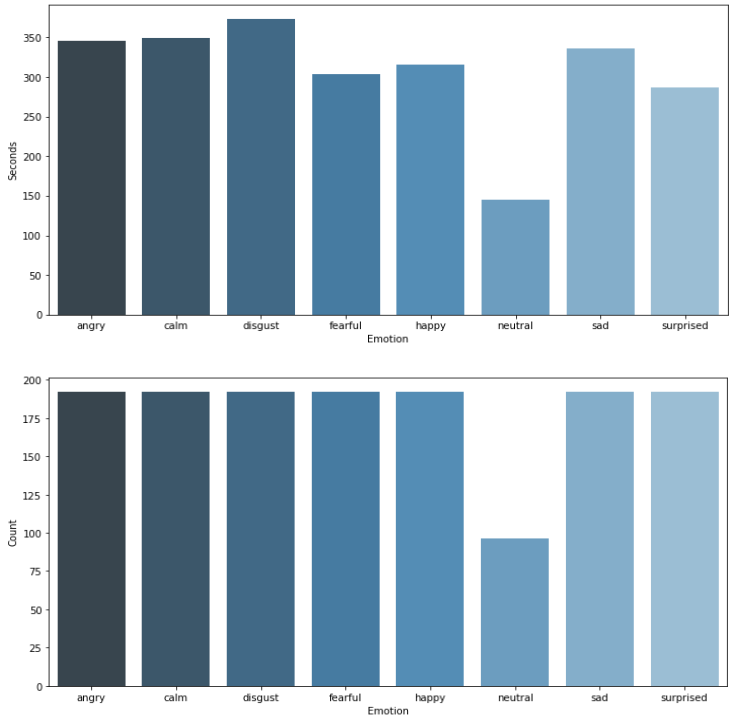
Data Modelling

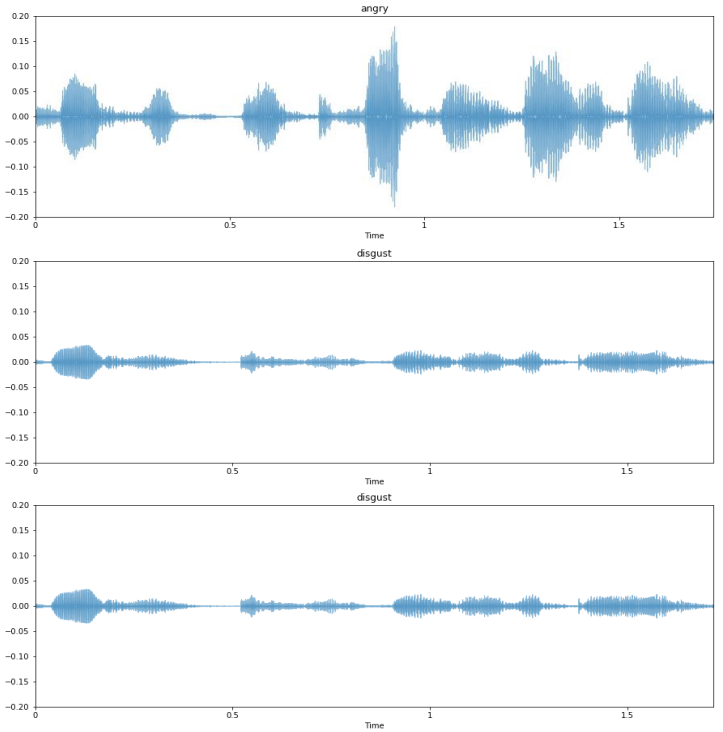
Apply & compare following models:

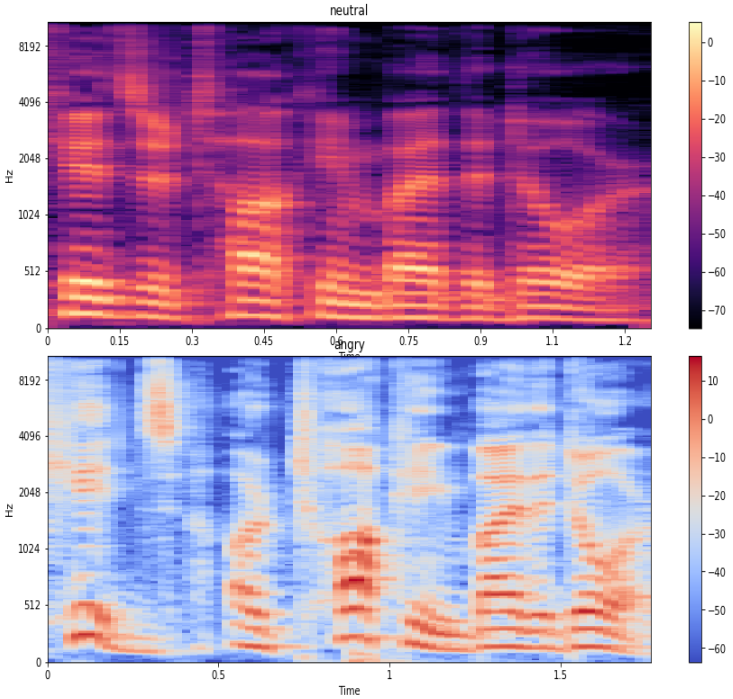
- Logistic Regression
- Random Forest
- Naive Bayes
- Gradient Boosting
- XGBoost

Exploratory Data Analysis

- There are total 1440 audio files categorized into 8 emotions
- Only “Neutral” has 96 audio files whereas all the other emotions have 192 audio files each
- There are ~2400 seconds of audio data (after removing the noise). On average each emotion has ~300 seconds of recording
- There are differences in the waveform and Mel-spectrogram. Hence both time-domain & time-frequency features are important



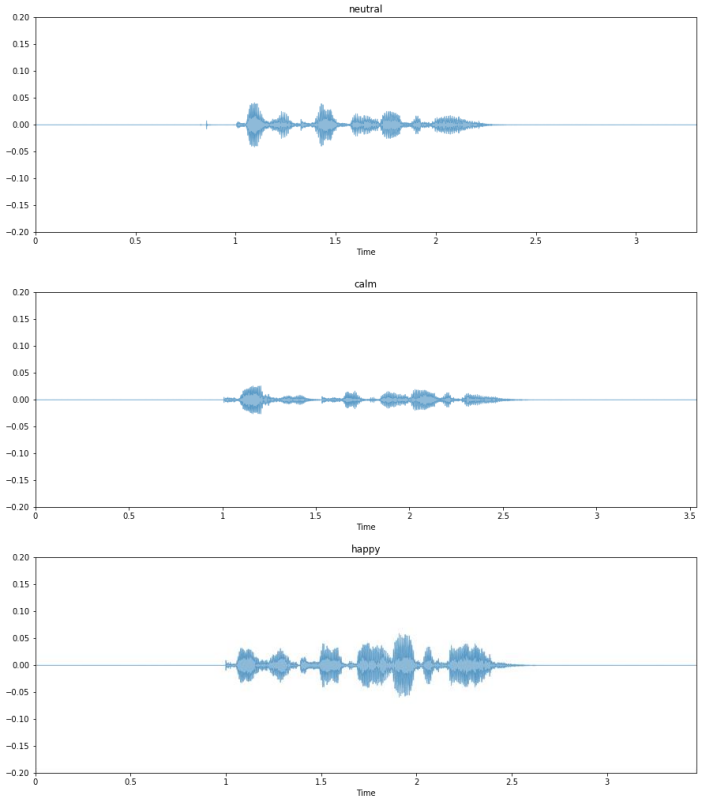




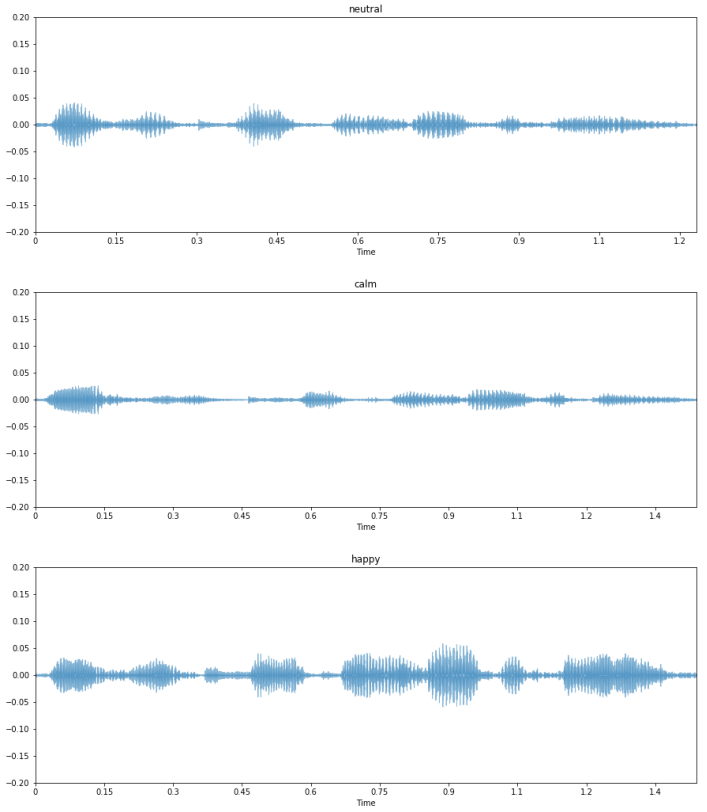
Data Cleaning

- Amplitude below 20db is removed with an intent to reduce the size of the data and for effective data processing
- Size of the data reduced from 1.4 hrs. of recordings to 0.48 hrs. of recordings

Before Noise Reduction



After Noise Reduction



Feature Extraction – Time Domain Feature

- There is significant difference in the RMS Energy of different emotions
- There is minor difference in the ZCR of different emotions. However it is still significant for data modelling

Root Mean Square (RMS) Energy

It is root mean square energy of all the samples in a frame and a measure of loudness

Emotion	
angry	0.055075
fearful	0.035712
happy	0.027128
surprised	0.019458
disgust	0.012517
sad	0.010528
neutral	0.007223
calm	0.005482

Zero Crossing Rate (ZCR)

Number of times a signal crosses the horizontal axis. This feature helps in identifying noise vs actual sound signal

Emotion	
angry	0.138542
disgust	0.137187
fearful	0.125682
surprised	0.118444
sad	0.114570
happy	0.113915
calm	0.108088
neutral	0.098019

Feature Extraction – Time Frequency Feature

- There is significant difference in the Band Energy Ratio of different emotions
- There is minor difference in the Spectral Centroid of different emotions
- Mel-spectrogram (512 Mel Bands) is an important feature as it has significantly improved the accuracy of the algorithm

Mel-spectrogram

Spectrogram with frequency is Mel Scale i.e. perceived frequency

Chroma_stft

Calculates the pitch class profile of an audio signal along 12 pitch classes

Spectral Centroid

Weighted mean of frequencies at each frame of an audio signal. Capture brightness of the sound

Emotion	
angry	2542.327939
fearful	2464.274262
disgust	2444.607684
surprised	2329.448815
happy	2247.294312
sad	2160.584775
calm	2119.902542
neutral	1988.823158

Band Energy Ratio

Ratio of lower frequency vs high frequency in each frame of audio signal

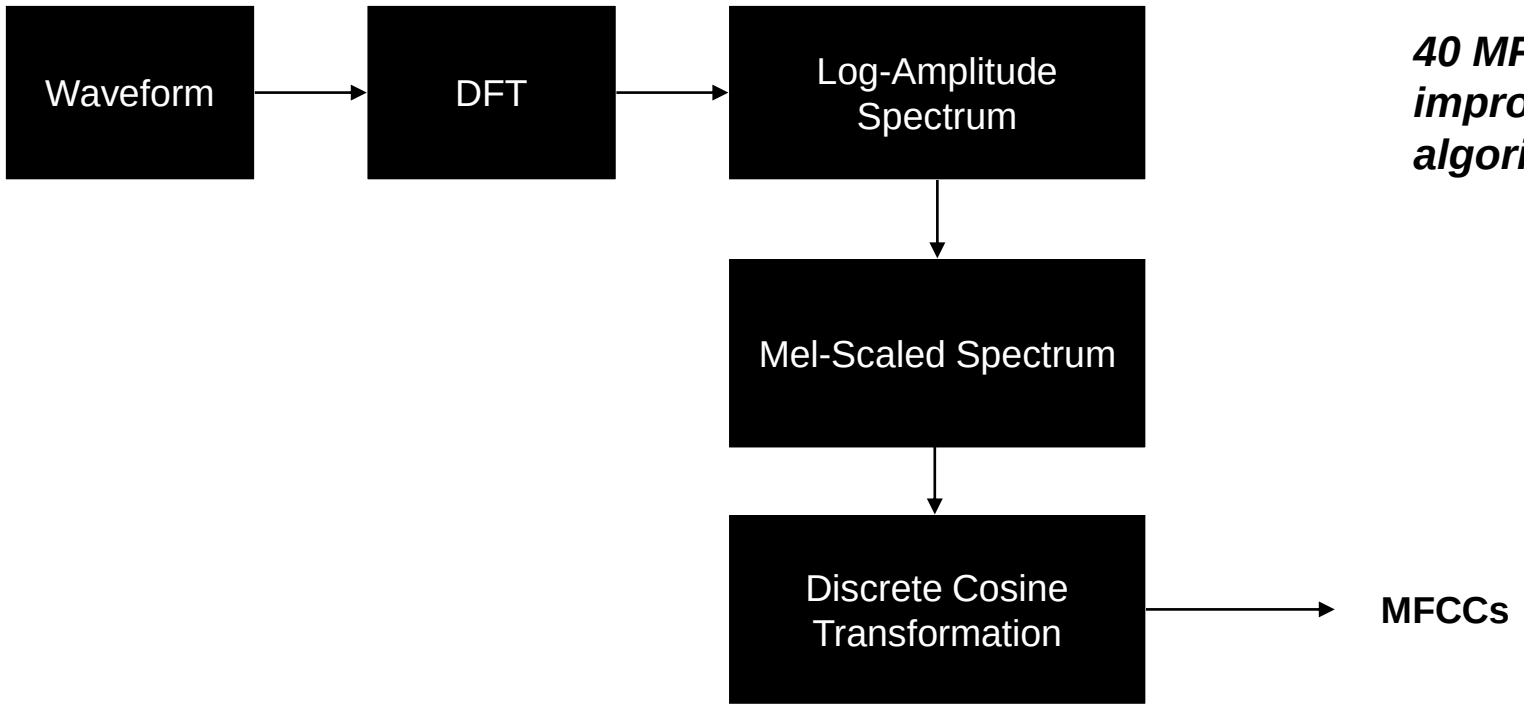
Emotion	
fearful	7834.325739
sad	4982.314455
disgust	3291.122136
calm	546.964972
surprised	514.072444
happy	403.385728
neutral	394.473646
angry	112.711930

Bandwidth

Weighted mean of the distances of frequency bands from Spectral Centroid. Generally correlated with Spectral Centroid

Feature Extraction – Cepstrum Feature

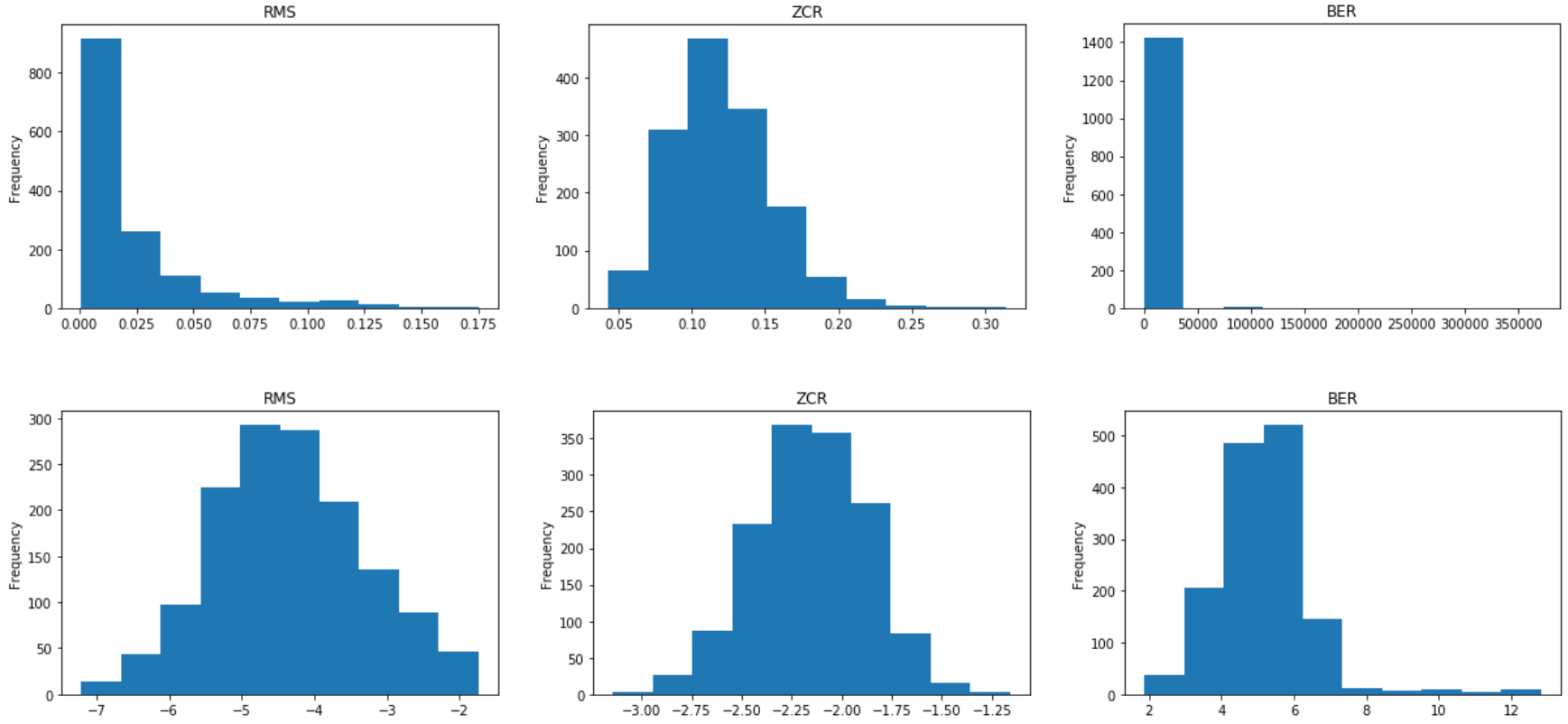
MFCC: These are coefficients of the discreate cosine transformation, which is decomposition of the log-amplitude and Mel-scaled version of the Discrete Fourier Transformation of the original signal



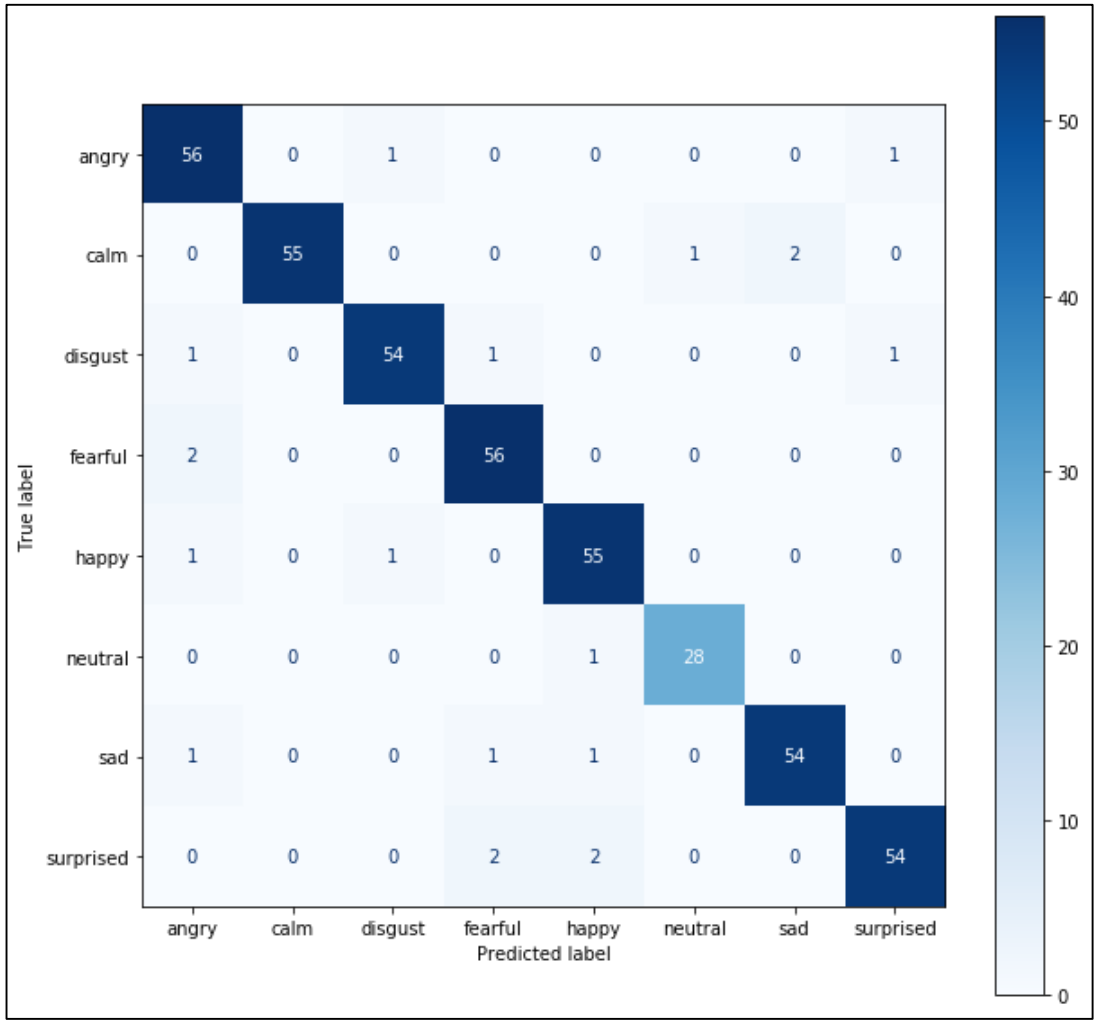
40 MFCCs has significantly improved the accuracy of the algorithm

Data Pre-processing & Dimensionality Reduction (LDA)

- RMS, ZCR, Mel-spectrogram features and BER are not following normal distribution hence log transformed is done to convert it into normal distribution
- LDA is applied to reduce the dimension and at the same time maximize the separation between classes

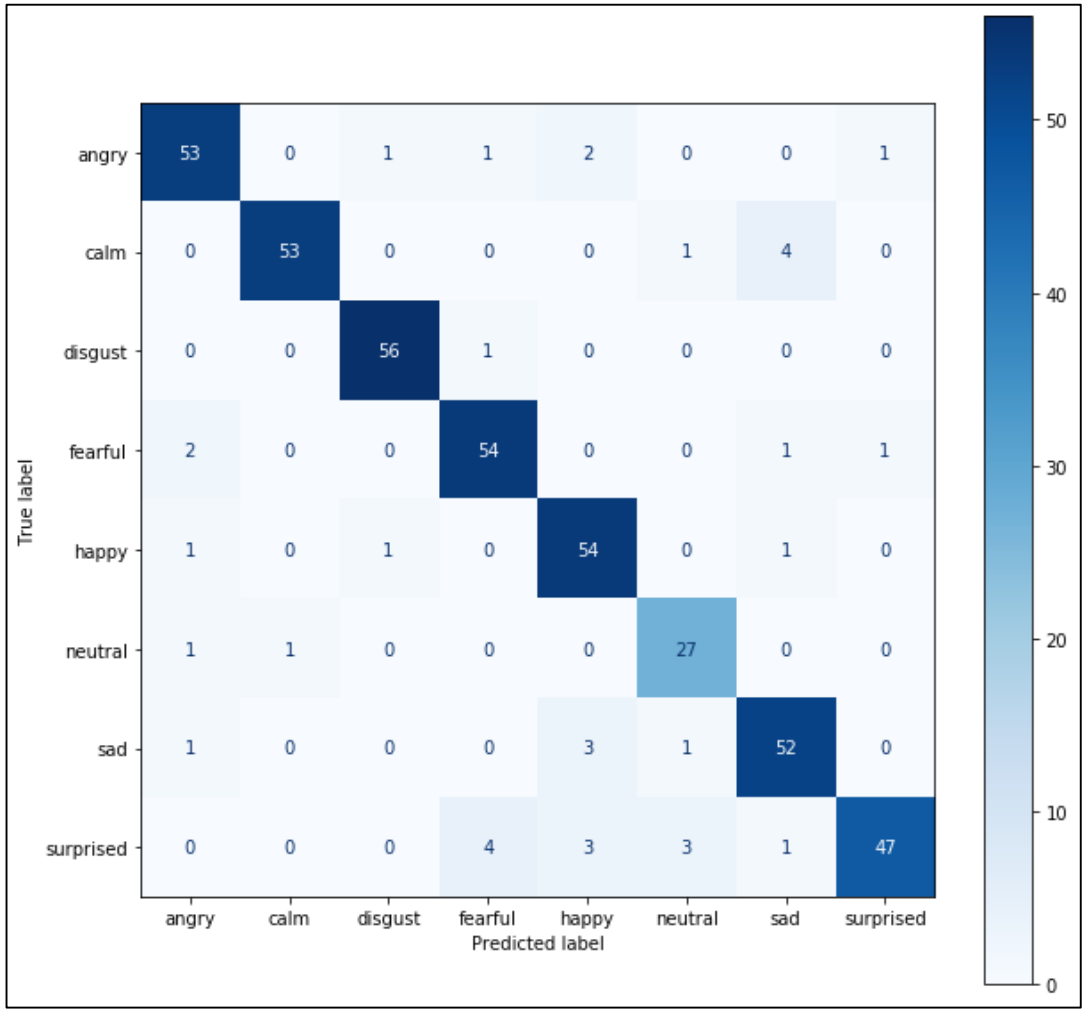


Data Pre-processing & Dimensionality Reduction (LDA)



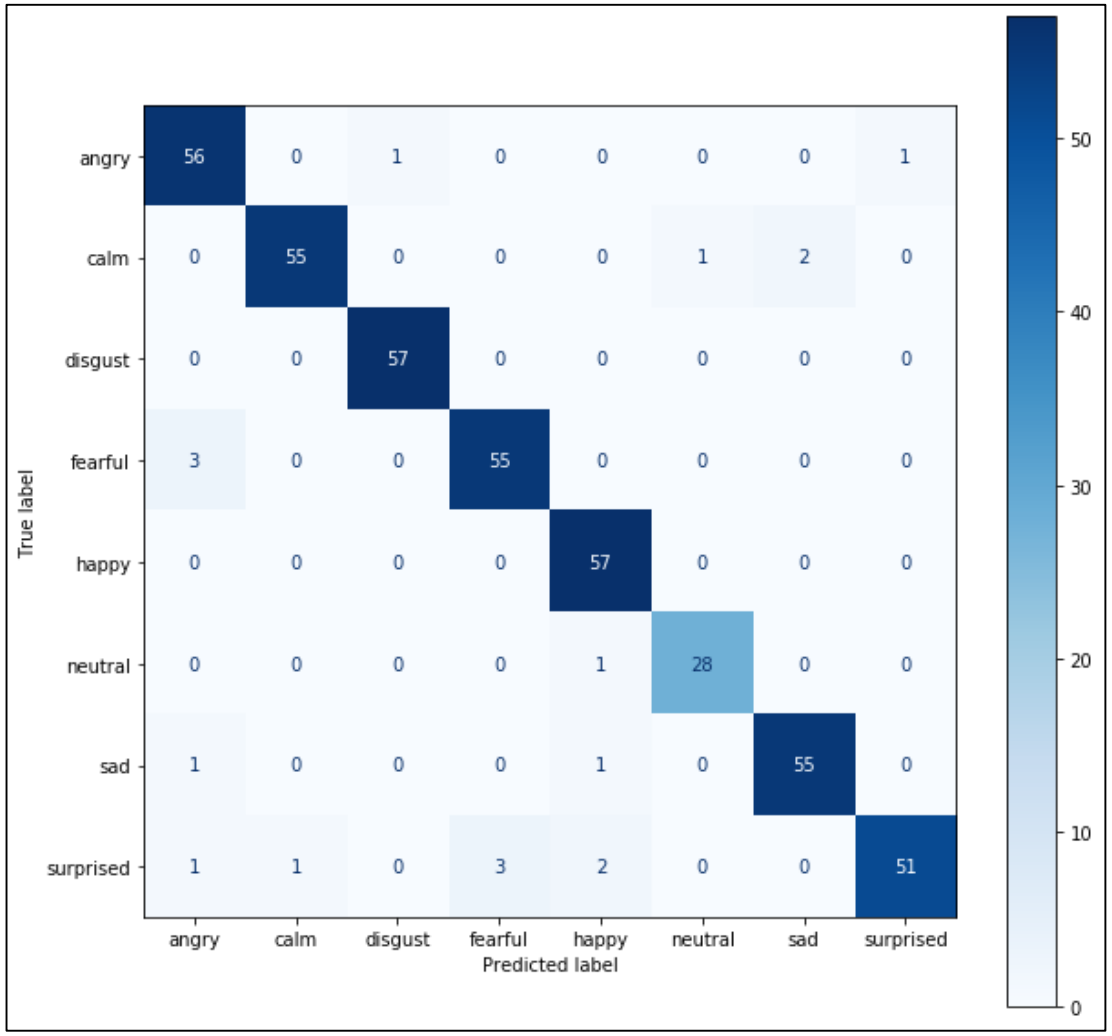
- **Logistic Regression has an accuracy of 95.37%**
- **F1 Score of 95.46%**

Data Modelling – Random Forest



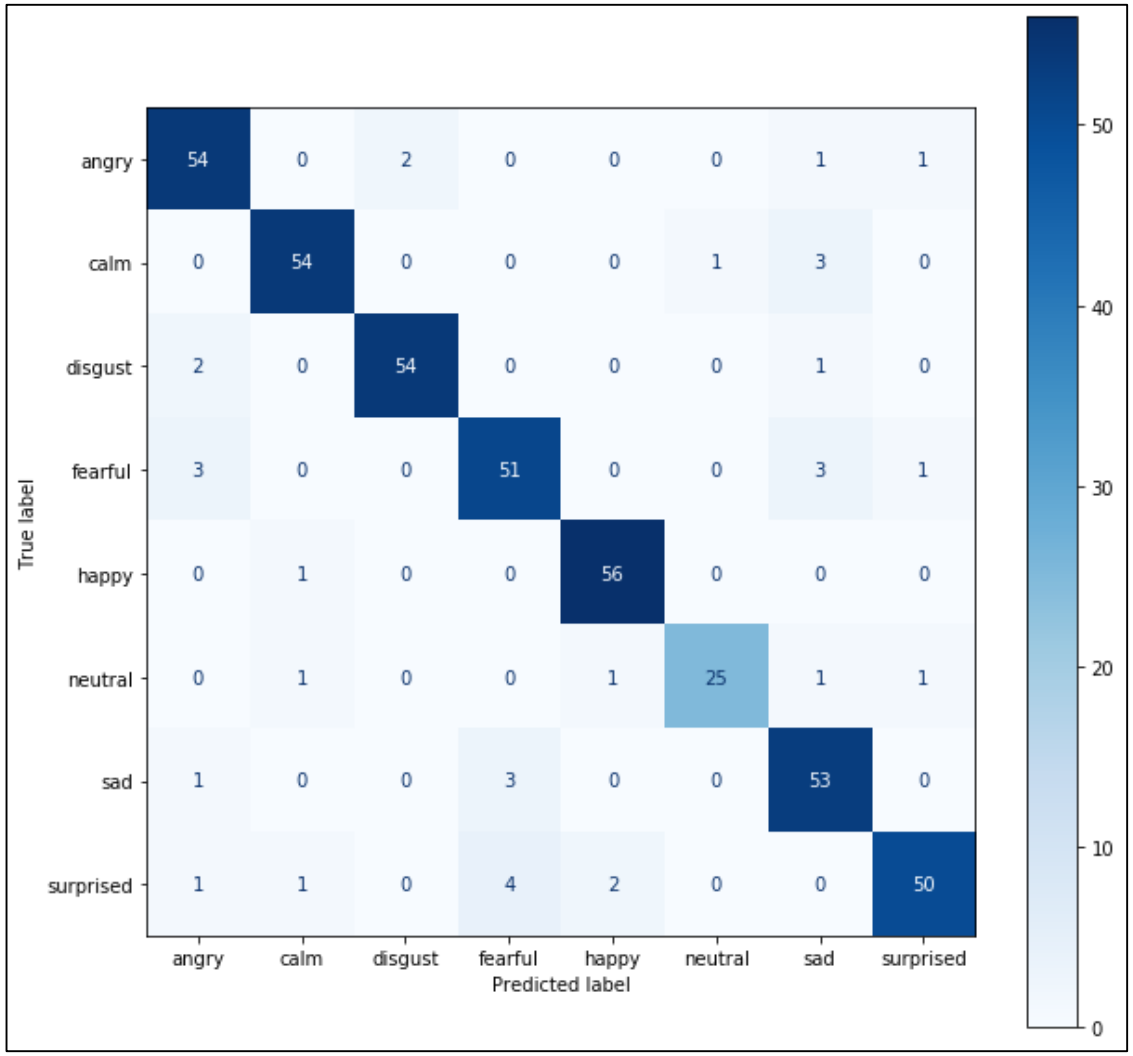
- **Random Forest has an accuracy of 91.67%**
- **F1 Score of 91.47%**

Data Modelling – Naïve Bayes



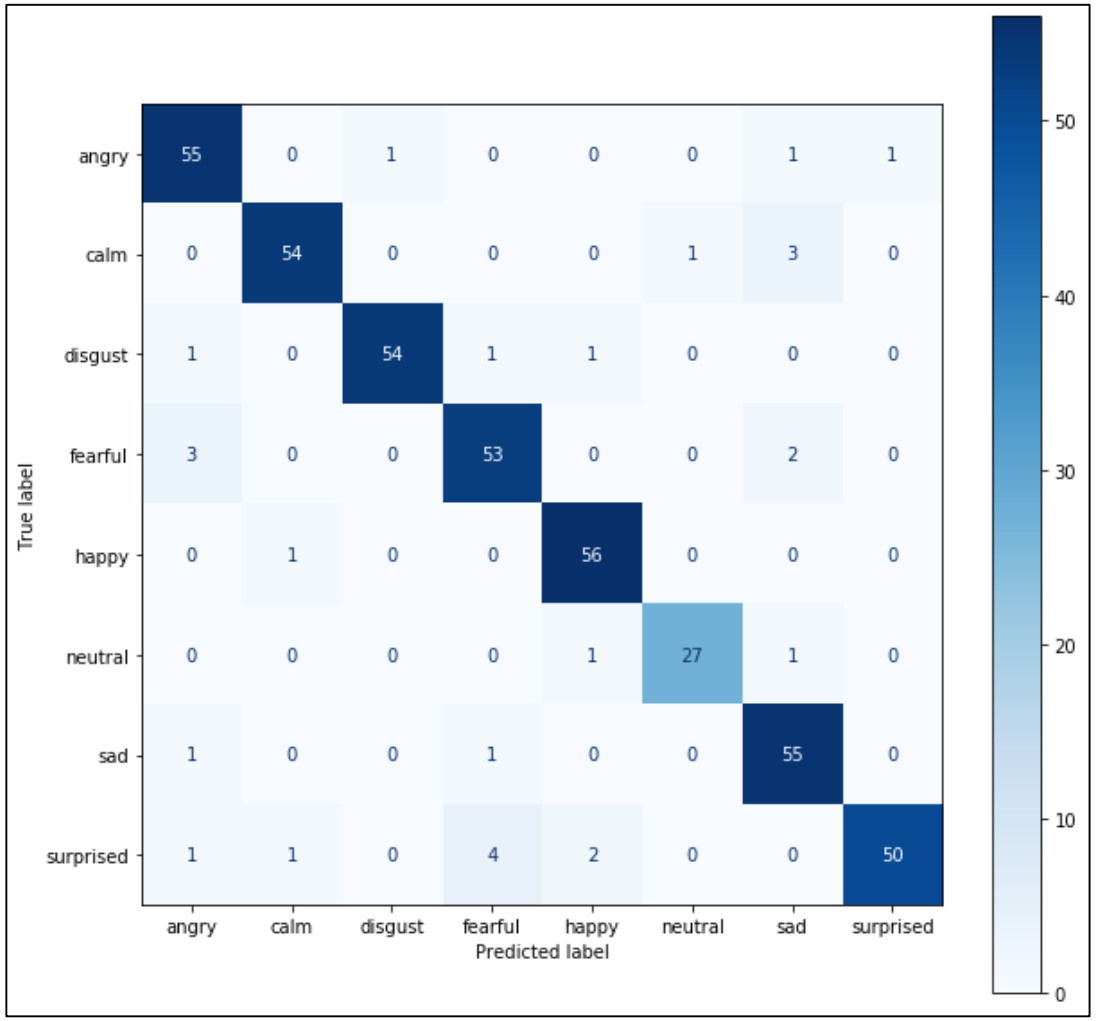
- **Naïve Bayes has an accuracy of 95.83%**
- **F1 Score of 95.87%**

Data Modelling – Gradient Boosting



- **Gradient Boosting has an accuracy of 91.90%**
- **F1 Score of 91.85%**

Data Modelling – XGBoost



- **XGBoost has an accuracy of 93.52%**
- **F1 Score of 93.61%**

Conclusion

- Naïve Bayes model has the highest accuracy and F1 score. Hence shall be used to predict speech emotion for any new data

- Mel-spectrogram (512 Mel Bands)and MFCCs (40 Coefficients) are two very important features as they have significantly improve the accuracy of the algorithm

External References

- https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/
- https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/
- https://librosa.org/doc/latest/index.html
- https://www.ijeat.org/wp-content/uploads/papers/v4i6/F4219084615.pdf
- https://www.youtube.com/watch?v=iCwMQJnKk2c&list=PL-wATfeyAMNqIee7cH3q1bh4QJFAaeNv0
- https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html
- https://www.analyticsvidhya.com/blog/2019/07/learn-build-first-speech-to-text-model-python/