Speech Emotion Recognition

Bhuminjay Soni(B20CS009), Ayush Gangwar(B20CS008)

soni.14@iitj.ac.in

gangwar.5@iitj.ac.in

***Abstract*— This paper reports an emotion recognition system of speech signals in two-stage approach, namely feature extraction and classification. Firstly, three sets of features are investigated which are: the first one- time domain features, we extracted a 4-dimensional vector of audio features including Zero Crossing Rate (ZCR) Mean, ZCR standard deviation, ZCR max, RMSE. Secondly- frequency domain features, we extracted a 152-dimensional vector of audio features including 12 Chroma features mean, 12 Chroma features std, 128 Mel spectrogram means. Lastly- Spectral shape-based features, we extracted a 46-dimensional vector of audio features including variance and mean of {Spectral centroid, Spectral rolloff, Spectral flux} and 40 coefficients of Mel Frequency Cepstral Coefficients, further a feature set formed by merging all 3 feature sets. Apart from the given dataset we also applied data augmentation technique to generate a new dataset on which a significant increase in the accuracy was observed. We applied various classification models on the above features and compared the results in this report.**

I. Introduction

Speech Emotion Recognition (SER) is a technology to extract emotional feature from speech signals. SER is a hot research topic in the filed of Human Computer Interaction (HCI). It has a wide range of applications, such as interaction with robots (Human Machine Interaction), Call centres, banking etc. In general, the SER is a task comprising of two major parts: feature extraction and emotion classification. The questions that arise here: What is the optimal feature set? What combination of acoustic features for a most robust automatic recognition of a speaker's emotion? Which method is the most appropriate for classification? In this report we analysed various feature sets for SER along with various pre-processing techniques like Augmentation etc. This report is organised as follows Section II describes Problem Definition and Algorithm. Section III provides the experimental Evaluation including methodology, results and discussion. Finally, a conclusion is given in Section IV.

II. Problem Definition & Algorithm

A. Dataset

We are given a speech corpus containing 1440 files; 60 trials per actor x 24 actors = 1440. The dataset contains 24 professional actors (12 female,12 male), vocalizing two lexically-matched statements. Speech emotions includes calm, happy, sad, angry, fearful, surprise and disgust expression.

File naming Convention:
The filename consists of a 7-part numerical identifier (e.g., 03-01-04-02-02-01-12)
Filename identifiers:

- Modality (03 = audio-only)
- Vocal channel (01 = speech)
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04=sad, 05 = angry, 06 = fearful, 07 = disgust, 08=surprised).
- Emotional intensity (01 = normal, 02 = strong)
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

We divided the dataset into train & test (test ratio = 0.25)

B.1 Pre-processing & Data Augmentation

Pre-processing: We categorically encoded the emotion labels and formed a data frame containing following columns = {'gender', 'emotion', 'wav-file'}

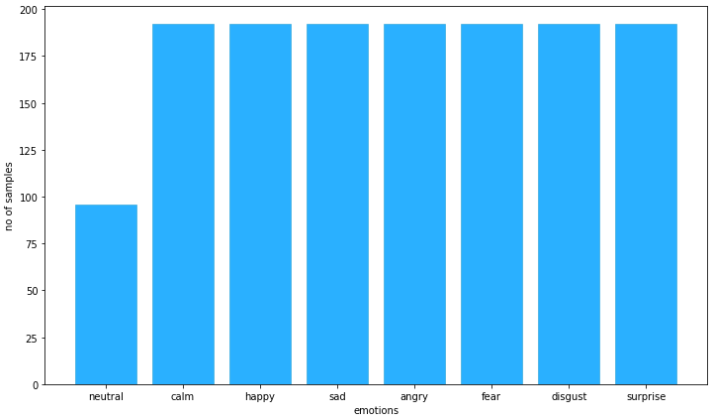

Fig. 1 Emotion wise distribution of the dataset

Data Augmentation: We implemented data augmentation methods including adding Gaussian white noise, normalizing signal and noise. Figure 2.a displays original sample signal while 2.b displays augmented signal.

B.2 Feature Extraction

Feature extraction is the process of highlighting the most discriminating and impactful features of a signal. We extracted features in three domains namely: time domain, frequency domain and spectral shape of audio-based features.

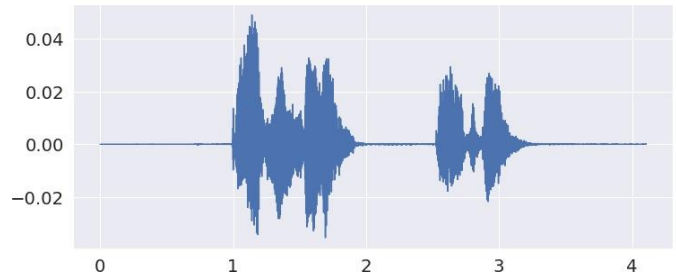

Fig. 2.a original audio sample

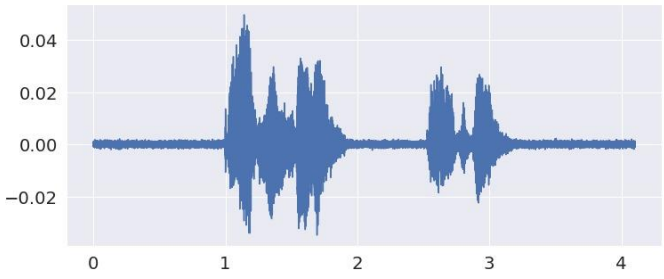

Fig. 2.b augmented audio sample

1. Time domain:

Zero Crossing Rate (ZCR)- The ZCR of an audio is defined as the rate at which the signal changes sign. ZCR is an efficient and simple way to detecting whether a speech frame is voice, unvoiced, or silent.

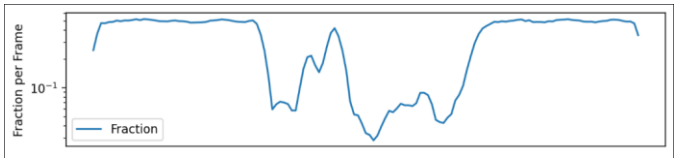

Fig. 3 y-axis - fraction per frame, x-axis - time

We extracted ZCR mean, ZCR max value and ZCR standard deviation values as features.

RMSE - It is instead the square root of the mean square (the average of the squares of magnitude of the audio frames).

2. Frequency domain:

Time domain plots signal variation with respect to time however to analyse a signal in terms of frequency, the time-domain signal is converted into frequency domain signal using Fourier transform.

Chroma energy distribution normalised statistics (CENS)- It is typically used to identify similarity between different interpretations of the music given. The main idea of CENS is that taking statistics over large windows of an audio file smooths local deviations in tempo, articulation, and musical ornaments.

We extracted CENS mean and CENS std as frequency domain features.

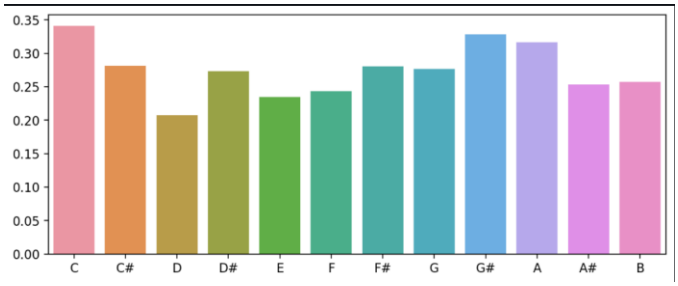

Fig. 4 CENS mean x-axis- octaves

Mel Spectrogram- It is a spectrogram where the frequencies are converted to the Mel scale.

We extracted the 128 Mel means from the Mel Spectrogram.

It is a way to visually represent a signal's loudness, or amplitude, as it varies over time at different frequencies.

The Mel scale is the result of some non-linear transformation of the frequency scale. Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also “sound” to humans as they are equal in distance from one another.

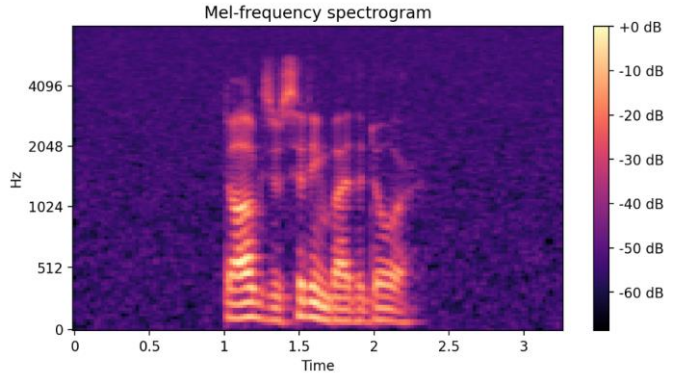

Fig. 4 Mel frequency spectrogram

3. Spectral Shape Based:

Spectral Centroid- It's a metric for determining the "centre of mass" of a spectrum. It has a strong perceptual link with the sensation of sound "brightness."

We extracted Spectral Centroid means and standard deviation as features.

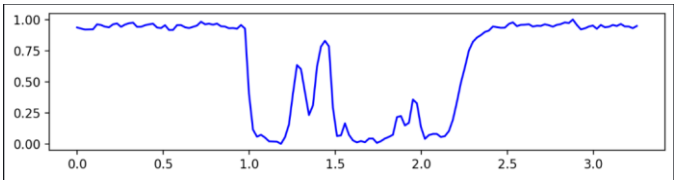

Fig. 5 Spectral Centroids at frame t (x-axis -t)

Spectral Roll off-The frequency below which the percent of the magnitude distribution is concentrated is known as the roll off point. The spectral roll off point is the frequency below which the majority of the signal energy is held.

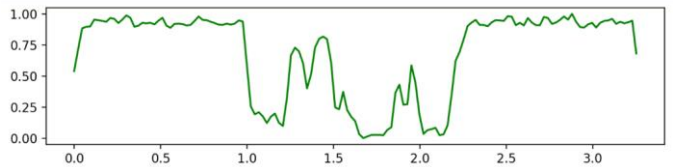

Fig. 6 Spectral roll off at frame t (x-axis -t)

Spectral roll off means and std were extracted as features.

Spectral Flux- Spectral flux is a measurement of how rapidly a signal's power spectrum changes, measured by comparing one frame's power spectrum to the preceding frame's power spectrum.

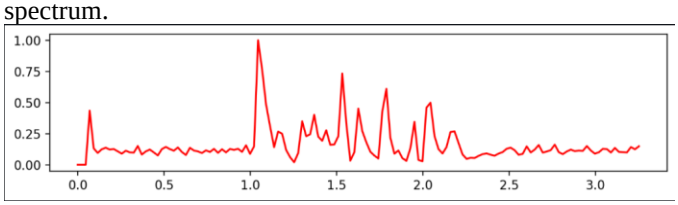

Fig. 7 Spectral flux at frame t (x-axis -t)

Spectral flux means and std were extracted as features.

Mel frequency Cepstral coefficients (MFCCs)- Mel-frequency cepstral coefficients are based on the discrete cosine transform of the log power spectrum on a non-linear Mel scale and indicate the short-time power of an audio clip.
MFCC feature extraction is done by windowing the signal, applying the DFT, obtaining the log of the magnitude, and then warping the frequencies on a Mel scale, followed by using the inverse DCT.

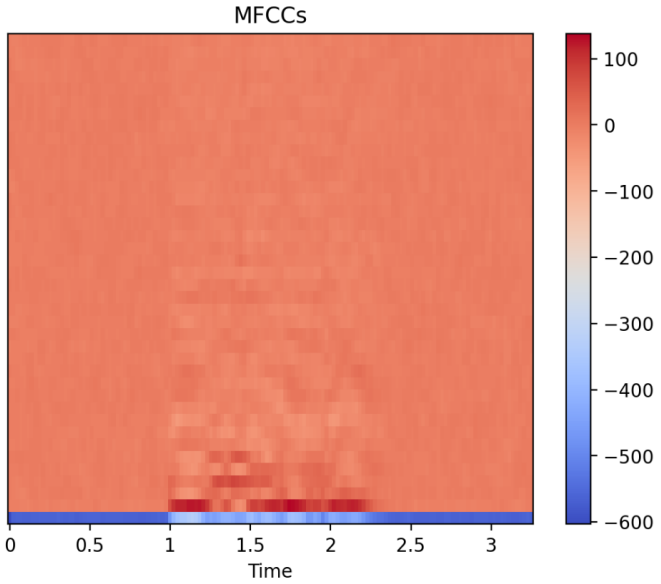

Fig. 8

III. Experimental Evaluation

For our Evaluation we used original dataset containing 1440 samples along with the augmented dataset.
Augmented dataset is created by first forming 4 augmented datasets of original dataset and finally concatenating them to finally form a dataset containing 5760 samples.
Here our augmented dataset addresses the problem of overfitting on original dataset and also have increased number of samples.

We applied various classification algorithms, Figure 9 represents the Accuracies% each feature sets for both original and augmented dataset.

Observations:
We observed that accuracies of every model increased on augmented dataset except AdaBoost and Gaussian Naïve bayes.

Figure 10 and Figure 11 Are confusion matrices of LightGBM model trained on spectral feature set with original dataset and augmented dataset respectively.

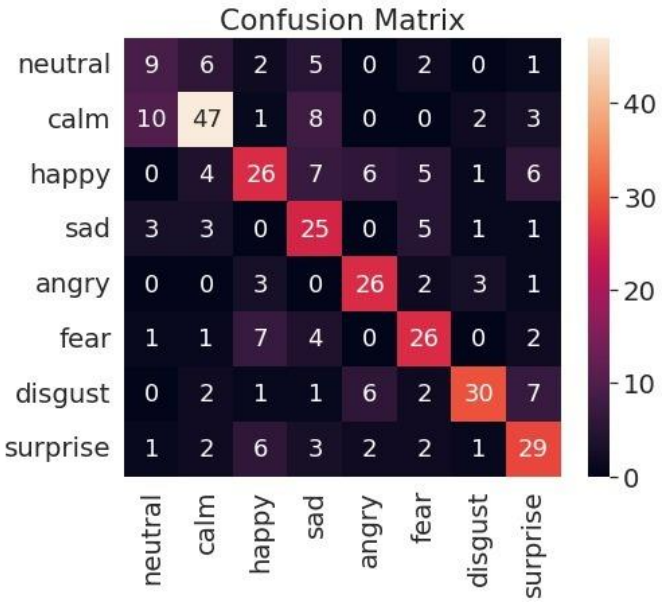

Fig. 10 Original Dataset

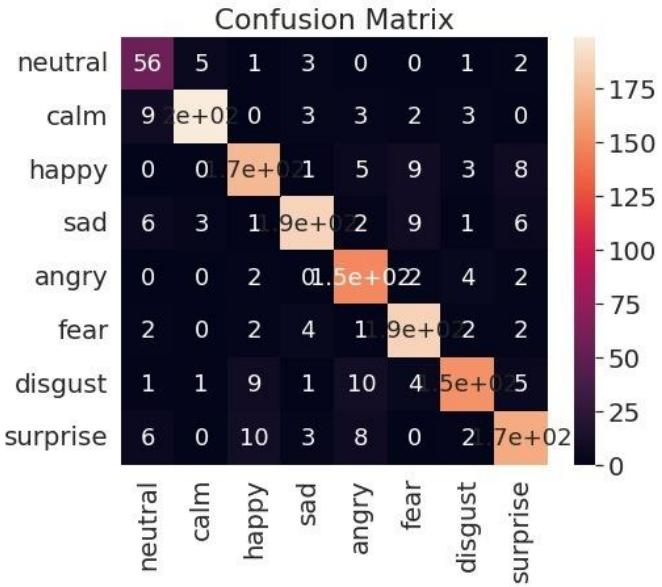

Fig. 11 Augmented Dataset

Performance on Precision, recall and F1 Score metrics

model	Original Dataset			Augmented Dataset		
metric	Precision	Recall	F1 Score	Precision	Recall	F1 Score
LightGBM (Spectral feature set)	0.61	0.61	0.60	0.88	0.87	0.88

Fig. 12 Performance Metrics

Model	Original Dataset					Augmented Dataset				
Features	Time Domain	Frequency Domain	Spectral Domain	Only MFCC	All features	Time Domain	Frequency Domain	Spectral Domain	Only MFCC	All features
Decision Tree	27.77	31.94	33.33	39.16	36.38	45.41	80.625	59.65	63.96	75.90
Random Forest	26.66	47.50	52.22	54.72	52.77	50.76	98.54	88.81	92.42	98.125
XGBoost	27.77	44.44	49.44	49.44	55.55	40.27	81.87	68.47	67.08	83.05
LightGBM	25.55	50.27	60.55	57.77	56.66	46.44	98.54	88.26	89.51	98.26
Adaboost	25.55	28.88	27.50	29.72	26.38	29.51	31.18	36.66	35.83	35.83
Gaussian Naïve bayes	29.72	29.72	25.00	32.22	27.50	33.33	30.83	26.31	36.52	27.78
QDA	30.55	14.72	56.94	56.94	19.44	31.94	78.54	77.63	70.48	87.98
LDA	23.88	33.88	46.94	41.38	47.44	29.72	46.94	51.11	47.91	64.86
MLP Classifier	28.33	38.88	15.00	45.55	15.00	33.68	88.12	19.09	60.41	07.56
KNN	29.66	40.27	23.61	36.66	23.61	33.19	77.56	23.75	55.69	23.75
SVC	28.05	36.11	25.00	51.66	25.27	36.31	71.87	28.33	65.97	28.33

Fig 9.

Further we observed that accuracy for MLP classifier increased only for Frequency Domain feature set and not much change was observed in other domains and there was decrease in the accuracy on All features.

Our top three models are LightGBM, Random Forest, and There is a Tie in between QDA and XGBoost.

IV. Conclusions

Data Augmentation leads to better classifications however in some cases it also leads to overfitting as in case of LightGBM.

Finally, from our experimentation we can conclude that Frequency Domain feature set and Only MFCC feature set are the optimal sets for Speech emotion Classification. Further we can conclude that only time domain feature set has worst performance. Also, we can see that on original dataset Only MFCC was the best feature set however on the augmented dataset Frequency domain feature set performs the best.

V. Future Scope

A few possible steps that can be implemented to make the models more robust and accurate are the following:

1. We can perform exhaustive feature selection to finally select best feature set.
2. Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition.
3. Further in this report we reported the accuracies on the base models which can further hyper tuned for better classification.

VI. Contribution

The learning, literature review, coding, deployment and report was done as a team both authors have equal contribution.

Further highlighted work of each author:

Bhuminjay Soni(B20CS009):

Decision Tree, Random Forest, XgBoost, LightGBM, QDA.

Ayush Gangwar(B20CS008):

SVC, KNN, MLP Classifier, LDA, Naïve Bayes.

VII. References

[1]https://rramnauth2220.github.io/blog/posts/code/200525-feature-extraction.html#centroids

[2]https://maelfabien.github.io/machinelearning/Speech9/#11-fundamental-frequency