Speech Emotion Recognition

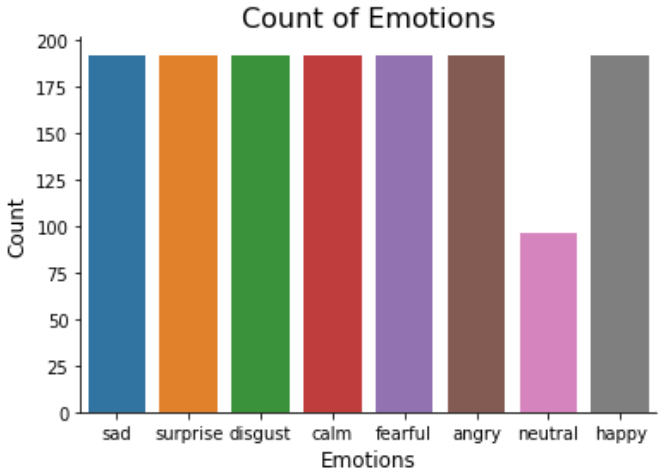
Abstract - This paper reports our experience with building an ML model to detect the emotion of a speech. Several classification models were applied to the dataset given. We compare the performance of different classification models on the data.

I. INTRODUCTION

As human beings speech is amongst the most natural way to express ourselves. As emotions play a vital role in communication, the detection and analysis of the same are of vital importance in today's digital world of remote communication. Emotion detection is a challenging task because emotions are subjective. There is no common consensus on how to measure or categorize them. In this paper, we will look at some simple yet interesting classification models to recognize the emotion of speech.
A meaningful exploration of the data was done.

A. About the Dataset

Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS.This portion of the RAVDESS [1] contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.



Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

B. Filename identifiers:

- *Modality (01 = full-AV, 02 = video-only, 03 = audio-only).*
- *Vocal channel (01 = speech, 02 = song).*
- *Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).*
- *Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.*
- *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).*

II. METHODOLOGY

A. Overview

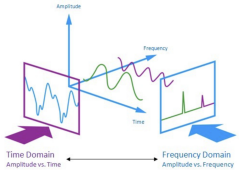
Among the various classification models present we will be applying these models and comparing the performances of the same:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. MLP Classifier
5. Kernel Support Vector Machine(SVM)
6. XGBoost
7. LightGBM

B. Feature Extraction and Pre-processing

The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency. Thus, we need to pre-process the data before we can actually apply classification ML models.
In this paper, we will be only extracting 3 features, and training our model, based on these features:

1. MFCC
2. MEL
3. CHROMA



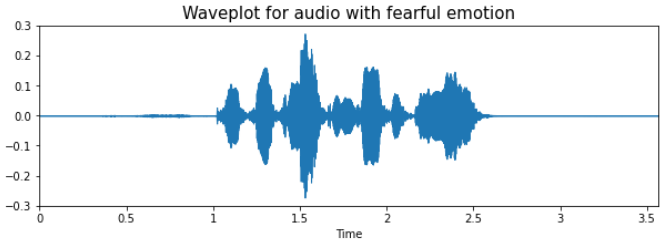
Source

The processed audio files are then stored in dataset format with their corresponding emotion in the other column which is derived from the file name identifiers.

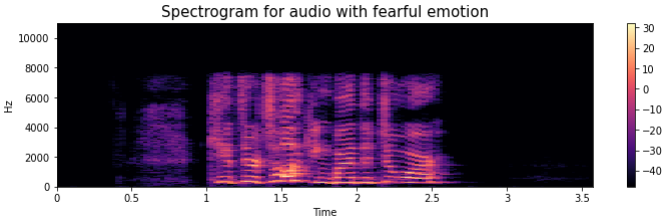
C. *Data Visualization and Analysis*

The data was visualized in three different formats: wave, audio, and spectrogram for each of the 8 emotions. One of them can be seen below, the others can be seen in the colab file submitted.

Fearful:



Waveplots let us know the loudness of the audio at a given time. The loudness of one of the "fearful" emotions can be seen above.



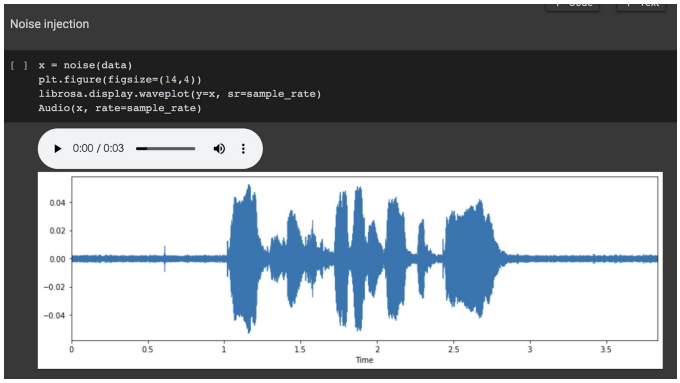
A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time. It's a representation of frequencies changing with respect to time for given audio/music signals. The spectrogram of one of the "fearful" emotions can be seen above. Also, an audio format of the audio was also created to listen to them.

The same visualization was done on all the eight emotions.

D. *Data Augmentation*

Data augmentation helps to generate synthetic data from existing datasets such that the generalization capability of the model can be improved. In audio this is done by applying noise injection, shifting time, changing pitch and speed. This is done so that our model is trained in such a way that it accounts for all the factors and so that the model can become more generalized and can be applied to various datasets.

One such data augmentation technique for audio can be seen below:



The rest of the techniques can be seen in the colab file submitted. We can also notice that noise injection is a very good augmentation technique because it makes sure that our model doesn't overfit the given dataset. Similarly, stretching, shifting and pitching techniques were applied to the dataset.

E. *Evaluation of Models*

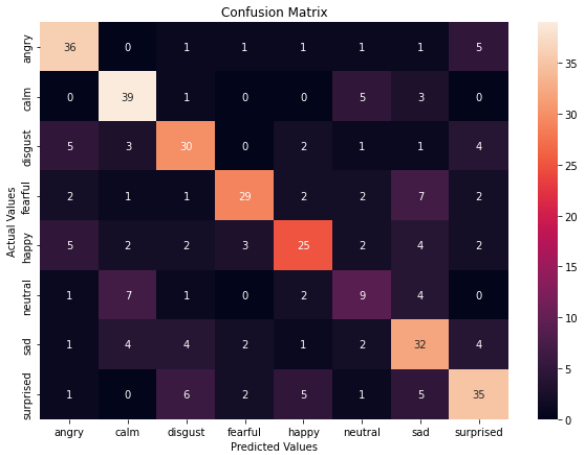
Model	*Accuracy*
MLP Classifier	***47.50%***
Logistic Regression	***48.88%***
Decision Tree Classifier	***33.05%***
Random Forest Classifier	***55.00%***
Kernel SVM	***56.11%***
XGBoost	***63.61%***
LightGBM	***65.27%***

An end-to-end pipeline of feature extraction, standard scaler, and classification ML model was made for each of the eight models. Further parameter tuning was performed on each of the eight models and the accuracies are reported above.

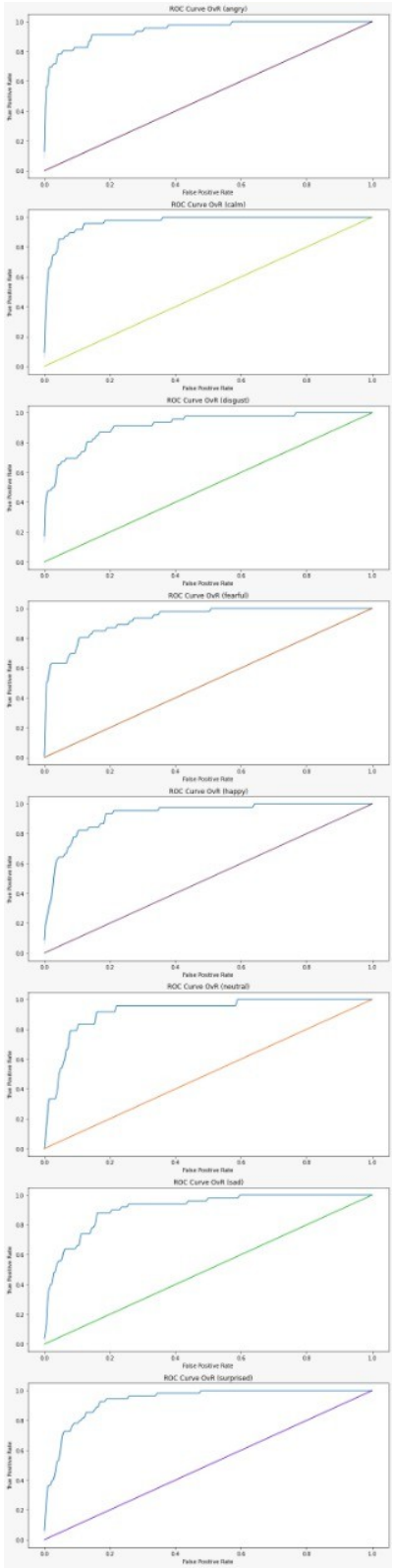
The final parameter of each mode after parameter tuning is:

- MLP Classifier:
  - alpha = 0.01
  - batch_size =256
  - epsilon = 10^-8
  - hidden_layer_sizes = (300,)
  - learning rate = ‘adaptive’
  - max_iter = 500
- Logistic regression
  - c =4
  - max_iter = 10,000
  - multiclass = ‘multinomial’
  - n_jobs = -1
- Decision Tree Classifier:
  - criterion = ‘Gini’
  - max_depth = 10
  - min_samples_split = 5
- Random Forest Classifier
  - bootstrap = False
  - max_depth = 10
  - max_features = sqrt
  - min_samples_leaf = 1
  - min_samples_split =2
  - n_estimators =800
- Kernel SVM:
  - kernel = ‘poly’
  - degree = 3
  - c = 100
- XGBoost:
  - colsample_bytree = 0.4
  - gamma = 0.0
  - learning rate = 0.25
  - max_depth = 5
  - min_child_weight = 5
  - objective = ‘multi:softprob’
- LightGBM
  - n_estimators = 500

We can see that LightGBM performs the best among all the classification models and hence we calculate the confusion matrix and ROC curves on the same.



Since we are dealing with a multi-label classification problem, the ROC Curves are plotted for each of the emotions using the technique of one versus the rest of the features, which can be seen below:



III. Results and Analysis

We performed feature extraction , pre-processed the data and applied seven different classification models and performed parameter tuning on each of the seven models, out of which lightGBM performed the best with an accuracy of 65.27%.

References

[1] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" by Livingstone & Russo is licensed under CC BY-NA-SC 4.0.

[2] https://medium.com/heuristics/audio-signal-feature-extraction-and-clustering-935319d2225