Sentimental Analysis of Audio samples

Team member:

1. Hemanth Reddy Yerramreddy
2. Chandrakanth Mandalapu
3. Sai Sri Harsha Chakravarthula
4. Narendar Reddy Nelakurthi

Goals and Objectives:

The primary objectives are processing all speech data and conducting in-depth exploratory data analysis. After EDA is finished, audio features are extracted using methods including ADA (analog-to-digital converter), temporal scope, and signal domain. These techniques give audio samples the essential depth and understanding, which improves model choice and hyperparameter optimization.

Motivation:

Many businesses can make a convincing argument for growing their firm. However, the organization needs a fundamental understanding of the target market to carry out these activities effectively. Market research should be conducted to understand consumer attitudes toward services and products and respond to client needs. The business used machine learning/deep learning models to accomplish this swiftly and effectively in the present. By accurately analyzing speech data and identifying each customer's attitude, we may gain priceless knowledge that advances our comprehension of the present and future markets. Additionally, by using voice data, this work enables us to evaluate various feature engineering strategies and comprehend their primary uses.

Significance:

Based on the mood analysis of the supplied data, various decisions are taken. This might have been brought on by a change in the restaurant's environment or improved service from the workers. The goal of businesses or individuals who employ insights from sentiment analysis is to raise the caliber of their goods or services. With the help of this technique, the business may grow steadily and provide superior customer service, which boosts client retention rates. Since many individuals in the target demographic have previously done requirements analyses, sentimental analysis initially reduces the effort involved in customer acquisition. Voice data is crucial in sentiment analysis since it allows real-time analysis of the source's voice. And in most cases, are substantially more qualified to do so.

Objectives:

The primary goals of this project are to:

- Conduct an exhaustive exploratory data analysis of the supplied data and show the results to assist you in comprehending the data.
- Following the EDA findings, the kind of preprocessing and procedures preserve the dataset's nuances while making the data suitable for sentiment analysis.
- Data preparation removes audio samples with excessive background noise or difficulty understanding, so they cannot be used.
- Using a range of feature engineering techniques, extract and choose pertinent characteristics. This data preparation phase guarantees that the model receives the critical training components.
- To train the model and solve problems like overfitting and underfitting, divide the data into training and test sets.
- During the model development stage, we experimented with various models, evaluated and contrasted their performance, and then chose the best model.
- The final criterion for evaluating model performance will be its capacity to recognize and categorize moods in audio recordings accurately.

Features:

The main feature of this project would be the feature extraction techniques we applied to the dataset. We first cleaned the data, followed by Exploratory Data Analysis. Then we explored various feature extraction techniques that can be done on audio signals. Upon doing that, we learned about techniques to derive the time domain features, such as RMS and Zero Crossing feature extraction techniques. Along with this, we have explored more techniques that can be used to get the frequency domain features. They include MFCC, Mel Spectrogram Features, Chroma Frequencies, Spectral Centroid Features, Spectral Roll-Off Features, Linear prediction coefficients, LPCC, LSF, HPCC, etc. We tried to implement most of the feature extraction techniques on the dataset. Later, we implemented a sequential model to train on the data. Later, we made the prediction on the test set, and the subsequent evaluation metrics were plotted.

Related work:

Audio Emotion Recognition is a System that identifies the emotion of different audio files. It is like the text sentimental analysis only difference is that the input data Here the input is Audio. Feature Extraction Techniques are used to Remove Noise and balance the Time-Frequency Ranges by converting the Digital and Analog Signals. Different Feature Extraction Techniques have been implemented to extract Features like Time Domain Feature Extraction and Frequency Domain Feature Extraction. In Time Domain, these are Extracted from the waveform of the raw audio Zero crossing Rate and RMS energy are examples. Frequency Domain mainly focuses on the Frequency Domain. Signals mainly Converted from the Time Domain to Frequency Domain using the Fourier Transform Mel spectrogram, Spectral Centroid, Mel Frequency Cepstral Coefficients, and Chroma Feature Extraction are Examples.

Dataset:

We have taken data from Kaggle. For developing a model with the best accuracy and Extracting the Features, we have taken 4 datasets Surrey Audiovisual expressed emotion (SAVEE), Ryerson audio-visual Database of emotional speech and song (RAVDESS), Toronto Emotional speech set (TESS), Crowd sourced emotional multimodal actors Dataset(CREMA-D).

SAVEE Dataset: It contains recordings of four male-actors in 7 different emotions and 480 British English utterances. The sentence was taken from the TMIT corpus and phonetically balanced for each emotion. The seven emotions are Anger, Happines, Sadness, surprise, fear, disgust, and Neutral.

RAVDESS Dataset: it is recorded by 24 professional actors, 12 females and 12 males, contains 7356 files, and 24.8 GB. Each expression contains two Emotional intensities (Strong, Neutral) Additionally the neutral Expression. The different emotions are Anger, Happiness, Sadness, surprise, fear, disgust, and Neutral.

TESS Dataset: it is recorded by two female actresses aged 26 and 64 Contains 2800 audio files. The best part of the dataset is that it contains only female voices and two of high quality. In this dataset they used 200 words were spoken by using the sentence say the word "–" by both actresses in different emotions like Anger, Happiness, Sadness, surprise, fear, disgust, and Neutral.

CREMA Dataset: is recorded by 91 actors and contains the 7442 Original clips. These clips are from 43 Female and 48 Males clips from different countries. The six different emotions are Anger, Happiness, Sadness, fear, disgust, Neutral, and surprise is missing.

Design of Features

We have designed several feature extraction techniques for the audio data in this project. These techniques would enable us to understand the data better. We have implemented various extraction techniques for features across two domains in the context of audio data. These are the frequency domain and the time domain. The techniques under these two domains have been implemented extensively.

We use a sequential model, a type of RNN model used for sequential data. This is done to create a baseline accuracy and improve it for future models. The baseline model would be evaluated on the following metrics precision, F-1 Score, Recall, and accuracy.
The final output would display the gender of the voice and their respective emotion for the above-mentioned metrics. We would also classify the male and female audio for each gender's overall precision and recall.

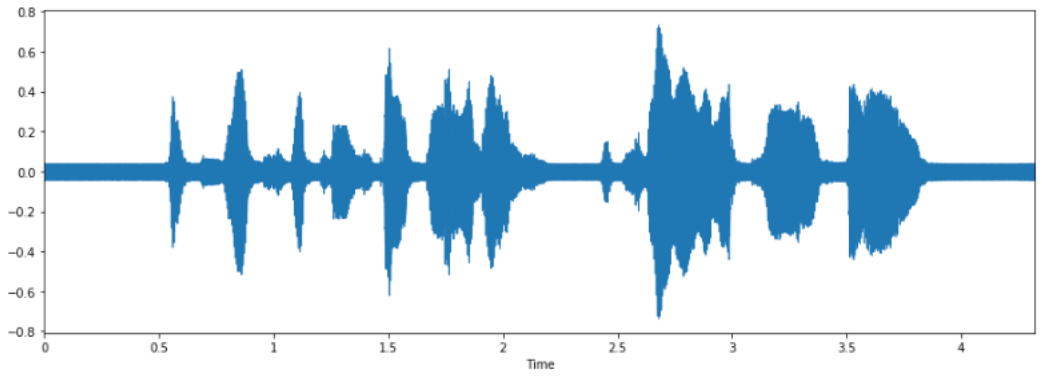
Analysis:

Exploratory Data Analysis:

Amongst all datasets, we have plotted the samples for a few data samples discussed below in detail.

Fear audio file:

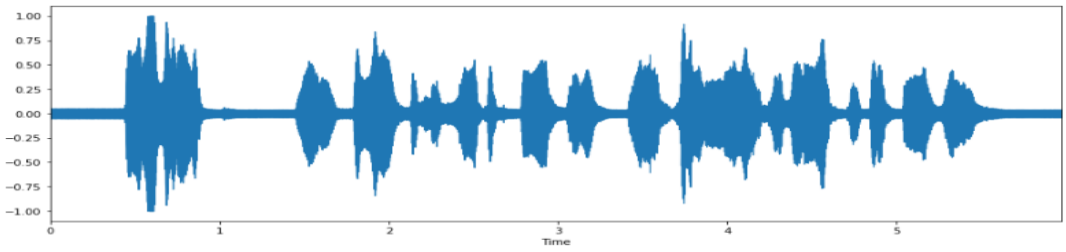
download.wav



So that's a Fear sampling. There isn't much background noise, and the dialogue is very clear. So, data is very useful in terms of quality. The plot shown above details how there is no noise in the speech. As you can see, the audio length is 3 mins long, so it’s easy to express the emotion in a statement and not too long to evaluate.

Now coming to the Happy audio file

download (1).wav

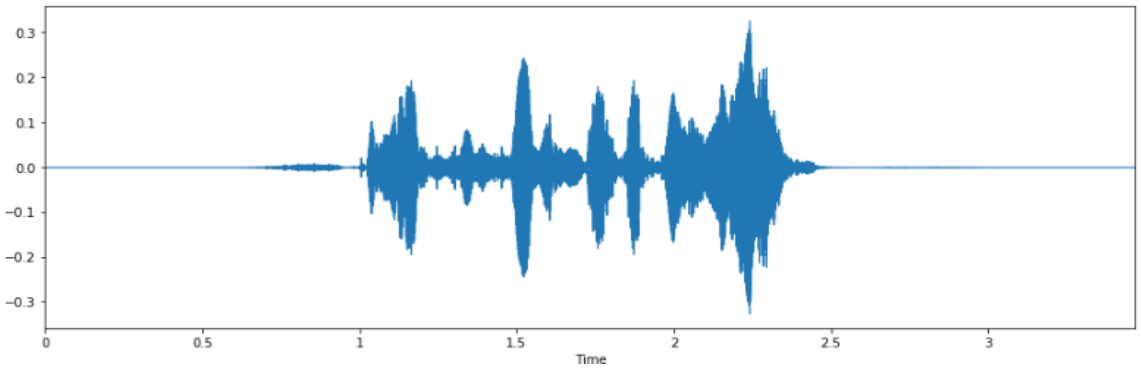


Here we can see from the wave that the data is different from the previous data, and the quality of the data is good. Here I could observe that the start and end silence period is comparatively shorter than the previous fear data.

Coming to the RAVDESS dataset, let us take the audio files like earlier and find out how the data is.

Taking the fear dataset

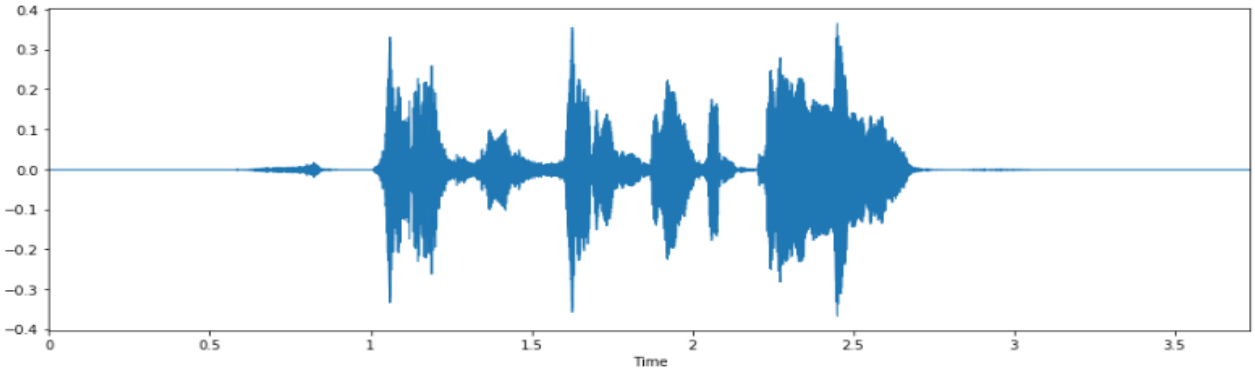
download (2).wav



The above audio shows that the start and end silence is long, and this data is also 3 seconds long. We may trim it later to improve the quality. Also, the stated language differs, so it's not a perfect comparison, but it gives us a good early indication of what we're dealing with. You can find from the audio that the person speaking is frightened.

Now let us see Happy voice:

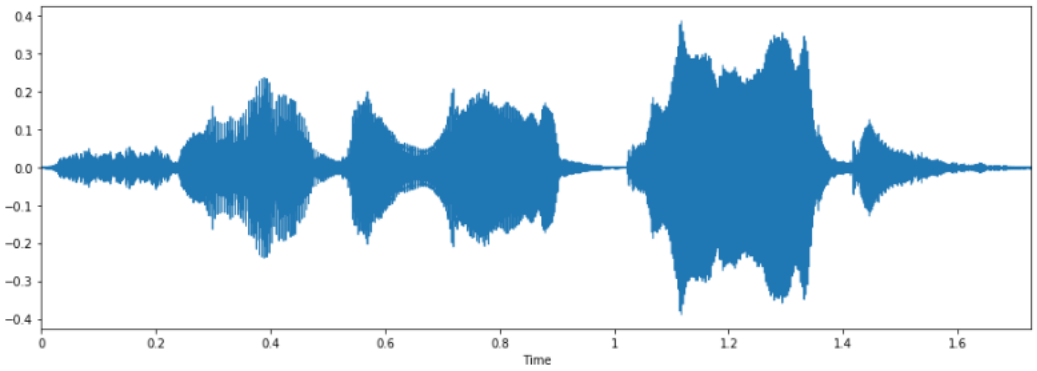
download (3).wav



The data here appears comparable to fear data. However, the amplitude is bigger in certain regions. The female voice is added to the data set, which is beneficial since having all the diverse variables in the dataset ensures that the model is accurate.

Coming to the Tess dataset

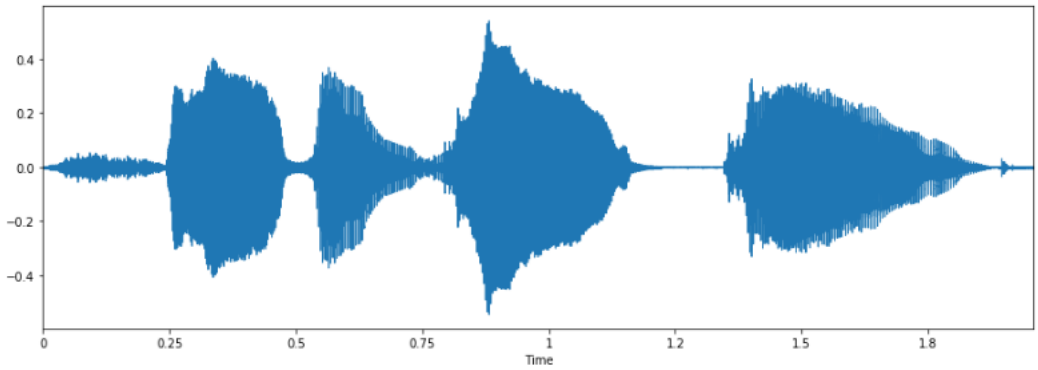
download (4).wav



The audio is like that of the RAVDESS dataset, but the starting and ending are very short. The amplitude looks much bigger, and the total time is approximately 2 secs. That is what we can find with this dataset.

Now coming to Happy sample.

download (5).wav

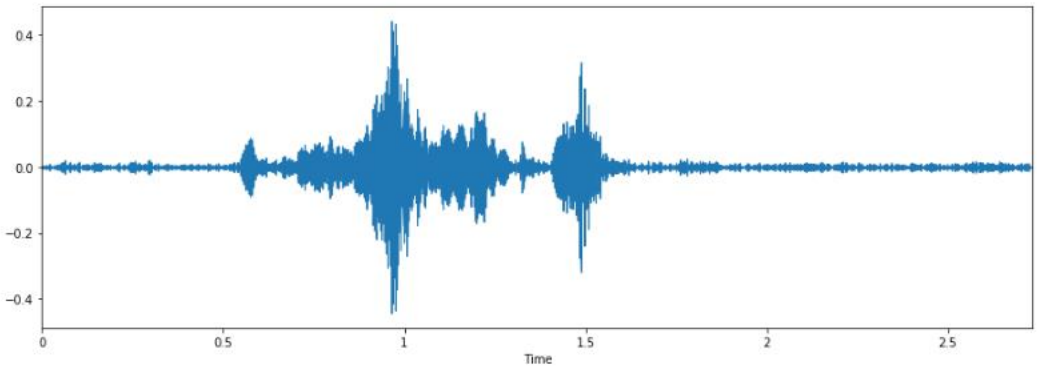


The amplitude in this sample skyrocketed, and you can see from the audio sample that it is happy data.

This data voice is like RAVDESS, but the amplitude has risen, and the sample is loud. Here you can see the data waveform seems different i.e., the short pitch is clear, and the high pitch has fewer samples.

Now Coming to CREMA Dataset.

download (6).wav



Here we can see that the data is shrilly and echoey, and the data has a lot of starts and ends of silence. It is not as clear. The audio is mostly neutral, and the data is 3 secs long.

Let us go through happy data.

download (7).wav

After listening to a few more random recordings, I found that the quality of this CREMA-D dataset varies greatly. Some sound clear, while others seem muffled or echoey. There is also a great deal of silence. The data was somewhat "dirtier" overall. But that's still high-quality data, and we'll put it to good use. On the plus side, a little noisy dataset would be a fantastic data supplement by adding noise to the pattern, something we now lack.

Feature Extraction:

Coming to the Feature extraction, we have both Time domain features and Frequency domain features.

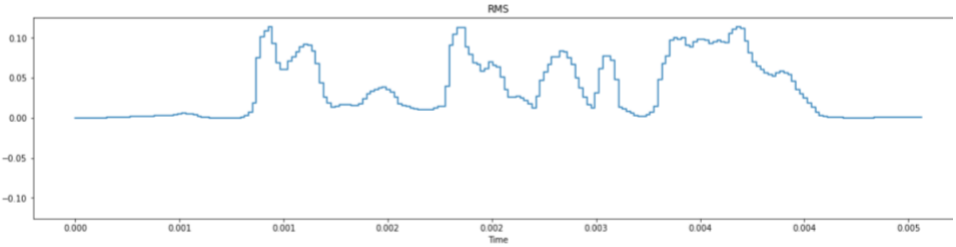
Time domain features:

RMS Features

The different features that we used are RMS feature extraction and Zero Crossing feature extraction.

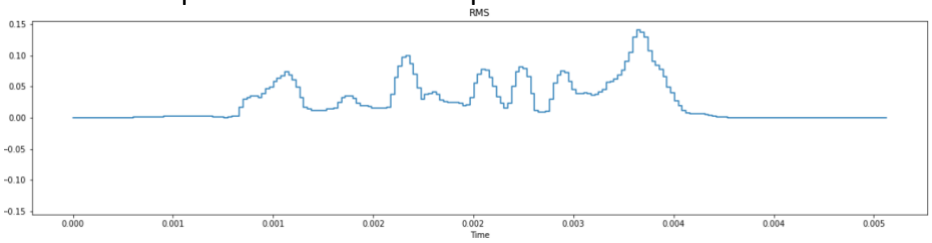
Let us investigate RMS features for various samples:

We took RMS Feature for the Happy sample from the audio corpus.



Here we took the RMS values of the amplitude of the audio over the time period, and here we can see that for Happy sample has few peaks and a pattern that we could see, i.e., in the end, we could see the peak raises and falls gradually at the end of the sample.

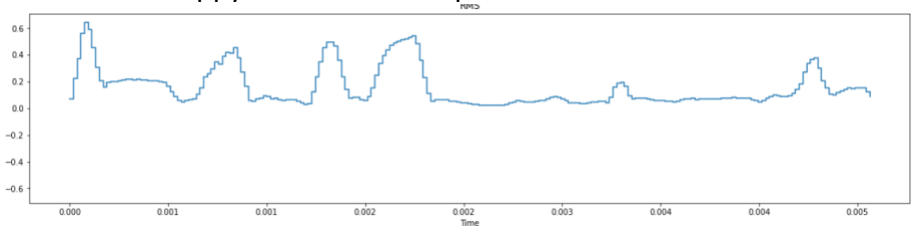
RMS Feature for Fear sample from the audio corpus



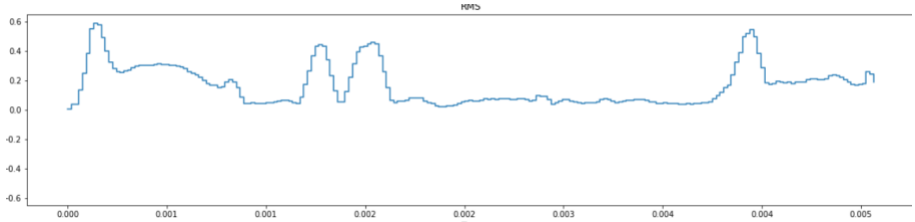
As we look at the fear sample, we can see that the sample, unlike the happy sample, has small continuous peaks and a higher peak at the end of the sample.
So, we can consider one of these patterns to differentiate between the Happy and fearful samples.

Zero Crossing feature:
Now coming to the Zero Crossing feature

Let us consider the same happy and fearful sample.



The above is the happy sample, and you can see that the first half of the sample has a large zero crossing, while the rest of the sample has a very low zero crossing.



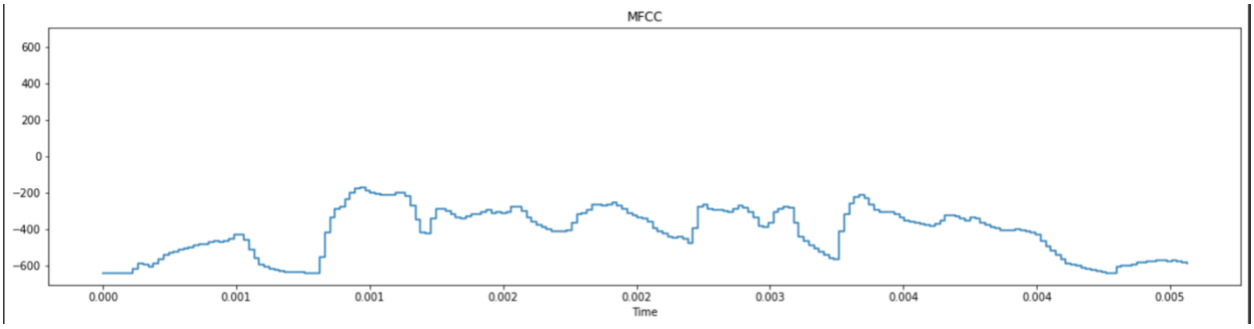
When we look at the fear sampling, we observe that it is comparable to the Happy sample. I couldn't see much difference in this characteristic between the happy and fearful samples.

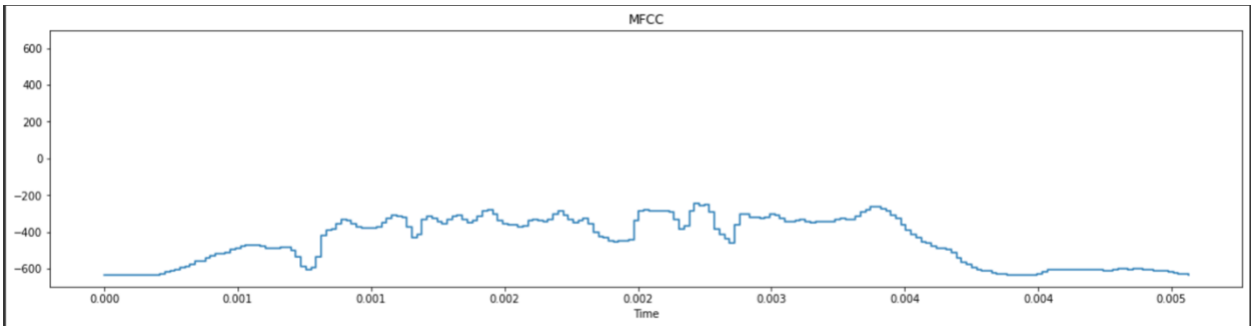
The only difference I can notice is that when comparing the two samples, the frightened sample has greater impulse than the joyful one. When compared to the happy sample, the peaks are a little sharper.

Frequency Domain Features:

Mel Frequency Cepstral Coefficients

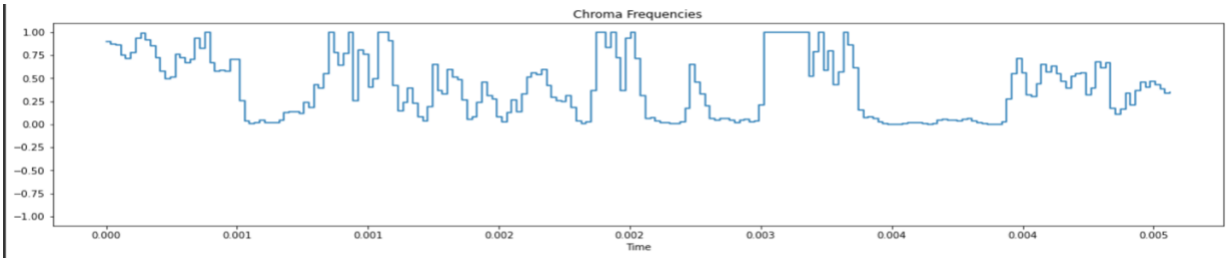
It is a feature-extracted technique in Speech recognition MFCC is used to represent the filters. The Feature extraction technique includes windowing the Signal and Applying the DFT, processing the amplitude log, wrapping frequencies on the Mel scale, and then applying the inverse DCT.

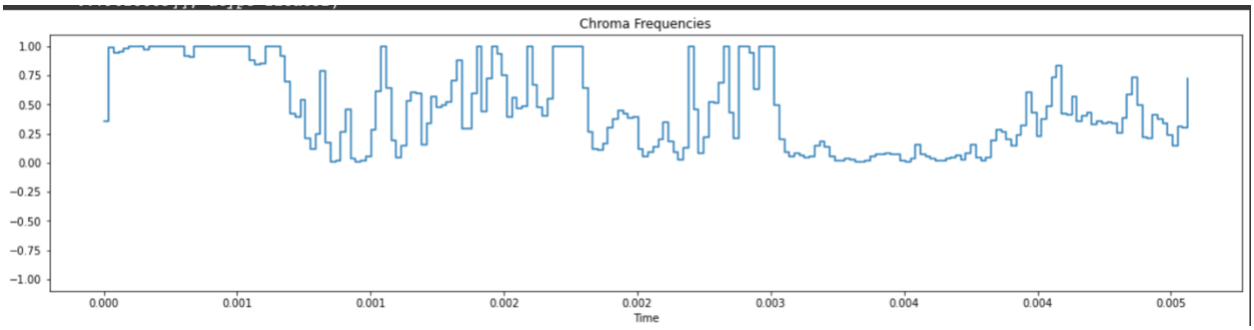




Chroma Feature Extraction

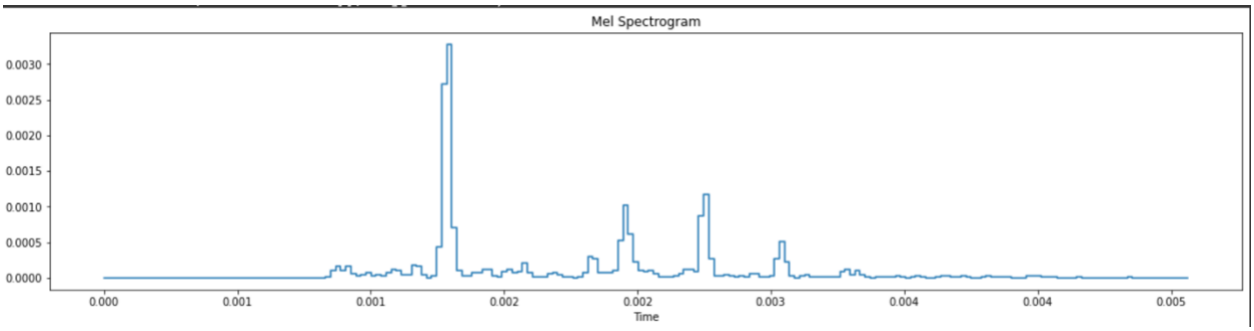
It is an important feature extraction technique in High-level sematic analysis. It is a descriptor that contains the tonal content of a musical audio signal in a condensed form.

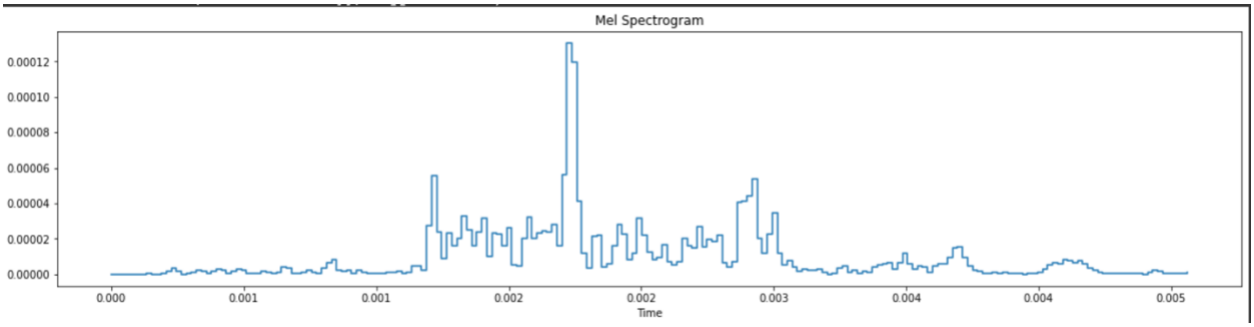




Mel Spectrogram

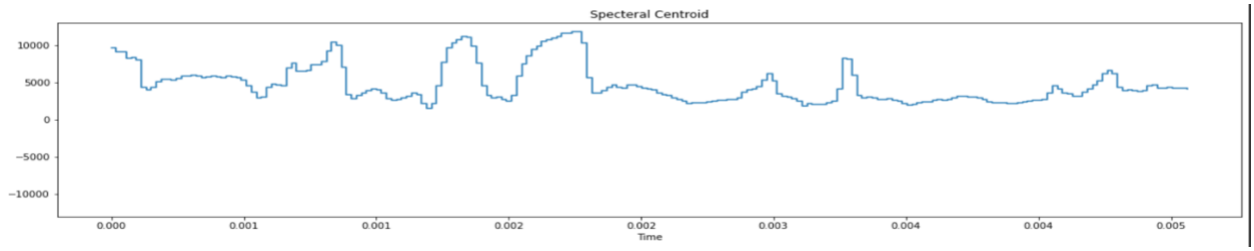
It is the Feature extraction technique in Frequency Domain. For Model building in Deep learning, we prefer to use the Mel spectrogram instead of the simple spectrogram. Mel spectrogram converts the Frequency scale in the y-axis to Mel Scale.

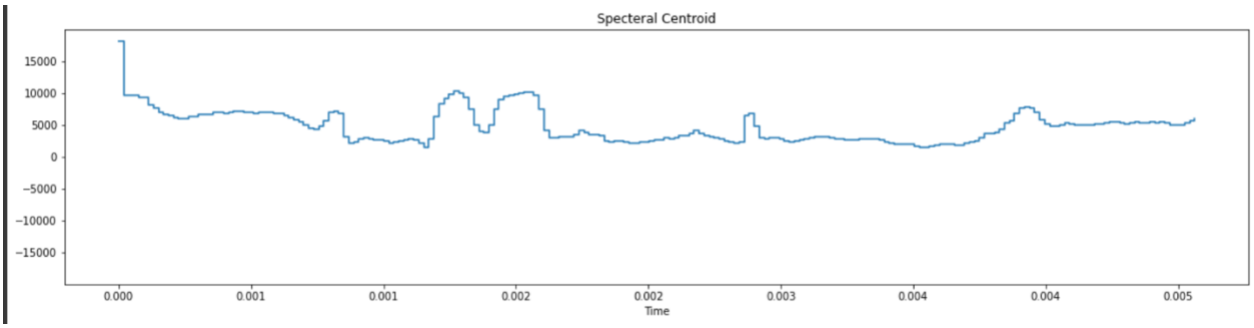




Spectral Centroid feature

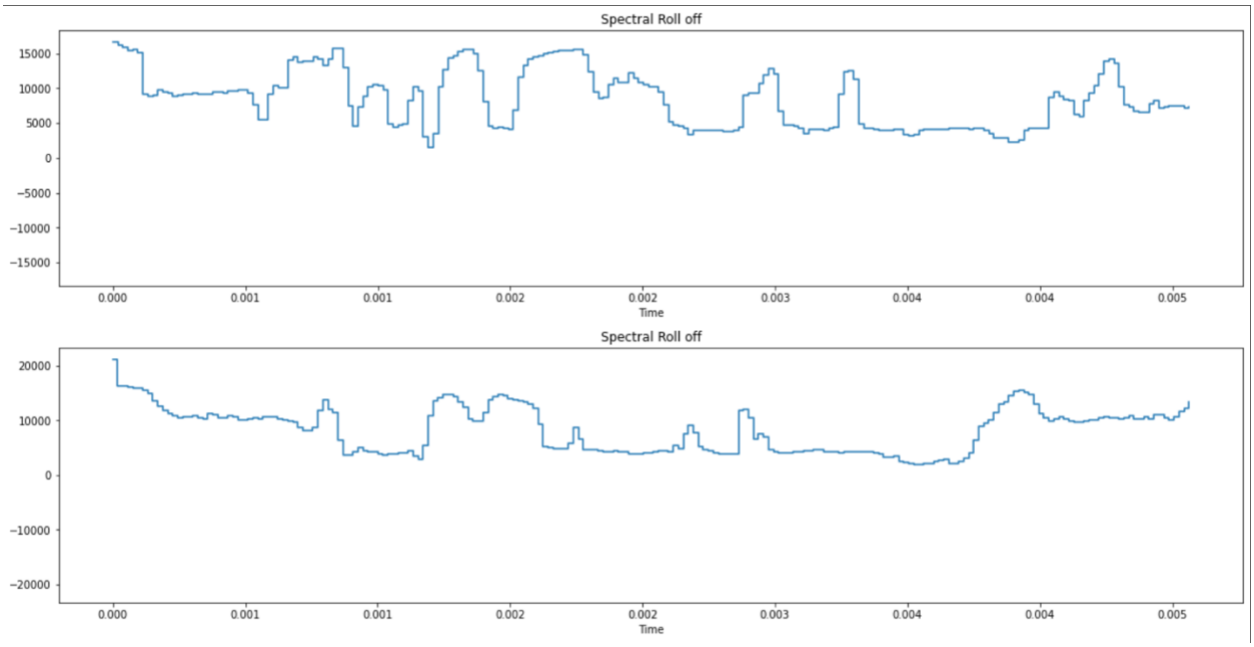
The spectral centroid is a measurement used in digital signal processing to describe a spectrum. It shows where the spectral center of mass is. It has a lot to do with how loud sounds are audibly perceived. Additionally, the spectral mass center is another name for it.





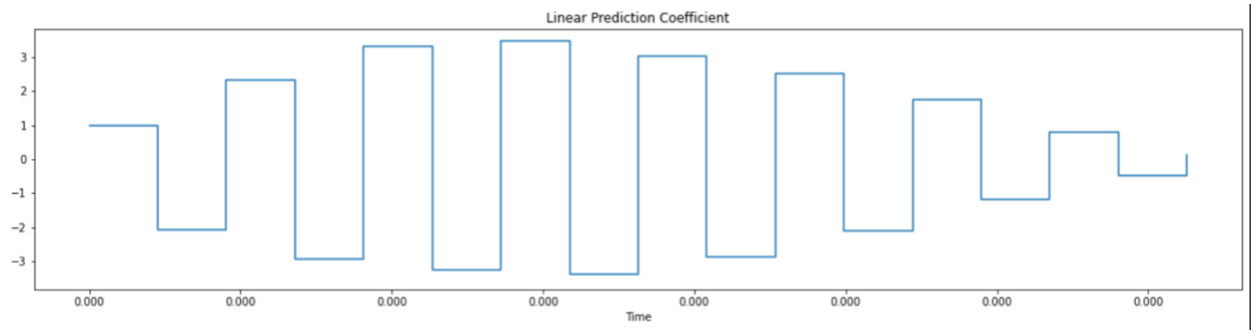
Spectral Roll Off feature

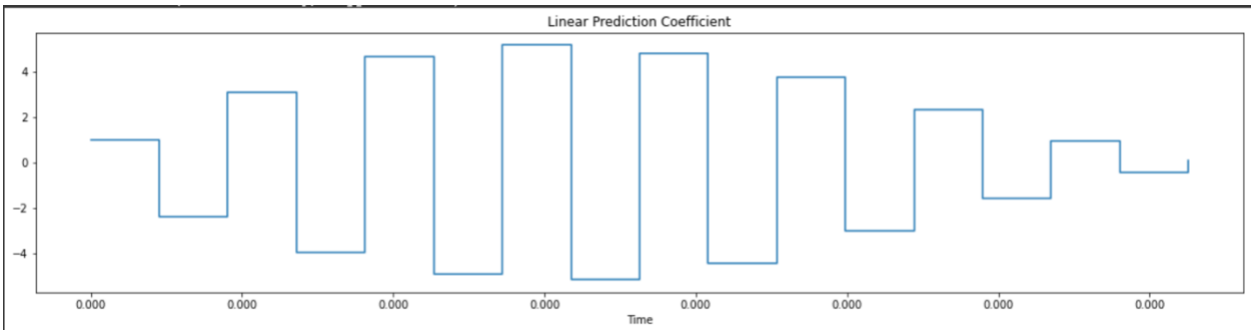
The frequency range that contains a fraction (cutoff) of the entire spectral energy is known as the roll-off frequency. It is possible to distinguish between harmonic and noisy sounds using the roll-off frequency.



Linear Prediction Coefficient feature

A linear predictor function, which is used to forecast the result of a dependent variable, is a linear function (linear combination) of a group of explanatory factors (independent variables) and coefficients. The units used to describe the coefficients in linear regression, the most prevalent setting for this type of function, are called regression coefficients. However, they do show up in several other linear classifiers, such as principal component and factor analysis, logistic regression, perceptrons, support vector machines, and linear discriminant analysis. The coefficients of several of these models are referred to as "weights."





Implementation:

We have performed several feature extraction techniques on our audio data in this project. This is done to identify the best technique and get an understanding of the important features that would be present in the audio file. We have implemented the Time and frequency domain features to do more from these for future increments.

The project is implemented in a way where first, the data set is loaded, as it consists of multiple audio files divided across multiple folders, each of which has its respective significance. After loading the dataset, we check for various attributes and show that the data is distributed with the dataset. We would display the amount of content present within each dataset folder and give a sample of the audio file. After this is performed, we would do the EDA of the dataset and get more information on the data distribution.

After the EDA is completed, the audio files are used for the various Time domain feature extraction techniques. This would allow us to complete an understanding of how each technique reacts or its effect on the given audio files. They are the RMS feature extraction technique and the Zero feature crossing technique for the time domain feature. These two are implemented as methods.

The method for the RMS feature is implemented with the path, plotting, and ret_len having false arguments. In the function, we use librosa, a python package for audio analysis; it loads the data and sets the sample rate. Using this, we would define the rms feature. Using the if-else structure, we would use it to plot the graph for the feature over the time domain. The method

for the Zero crossing feature is implemented similarly to the RMS feature. Still, the implementation differs because the librosa is used to invoke the zero-crossing feature. After these two features are implemented, we would use them to plot the graphs for each emotion, i.e., happy and fearful.

After implementing the time domain features, the next part of the implementation would be the frequency domain features. These, too, are implemented as methods. The style of implementing this method is like that of the Time Domain features, where we use librosa for invoking the required feature extraction techniques. The frequency domain features that we have implemented in this project are MFCC, Mel Spectrogram, and Chroma frequency features. We would display the waveform for each feature for the given audio file.

After completing these various feature extraction techniques and visualizing them successfully, we would move on to the model development phase. To train this data, we use the Sequential model as this model is used for various types of data that are sequential in nature, i.e., the input data and the output data are both sequential. The sequential model is a type of RNN widely used for sequential data. Since the audio data is sequential, we have selected this model to train and test the data.

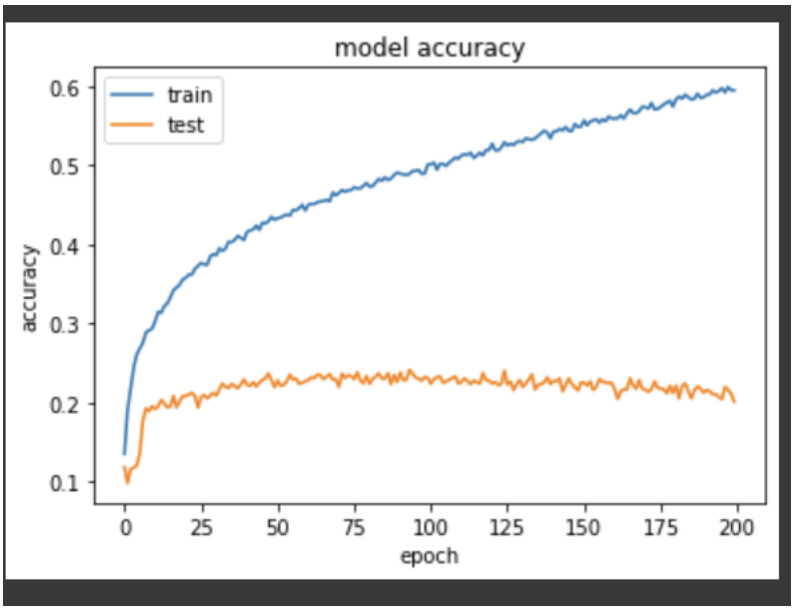
The sequential model is implemented by first loading the CSV file of the dataset, which would have all the attributes and their specific labels of the data. This CSV file is then converted into a data frame where it would be easier to create the train_test split on the given data and map them accordingly to the given audio files. After the train test split is created, we perform the data normalization. This is done to reduce the amount of redundant data and prevent modification errors from occurring. Doing these would enhance the workflow of the model. Also, implement the label encoder on the train set of the data to have the categorical values fit into the transform.

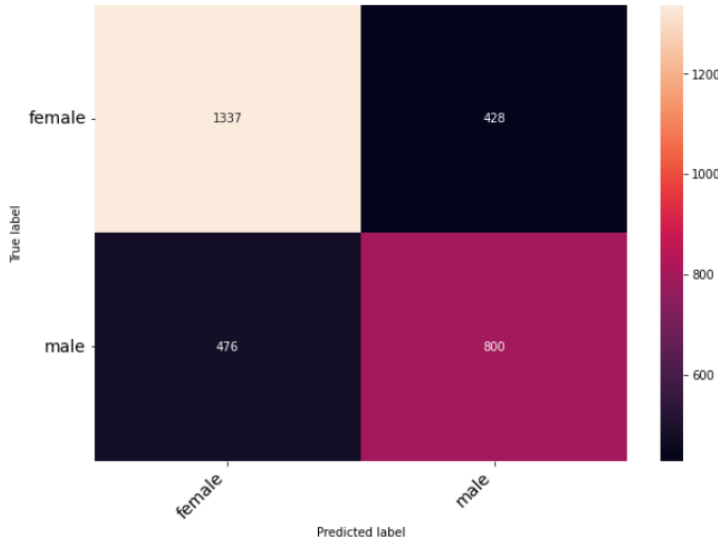
Finally, we would build the model. As the model is a sequential model, we would first load the dataset. The model is implemented with eight convolution layers starting with 2 convolution layers with 256 kernels, 4 with 128 kernels, and finally, in the end, we use 2 64 kernels. Each convolution layer is followed by an activation layer. The activation function is ‘relu’.The second convolution layer with 256 kernels is followed by an activation layer and has a batchnorimaliztion along with a dropout layer and ends with a max pooling layer. A similar implementation is done for the final Conv 128 kernels layer too. In the last convolution layer with 64 kernels, we use two activation functions, relu, and softmax; between these two, we would add a flattened and a dense layer. Once this is completed, we will generate the model summary, which will visualize the model's architecture.

After the model is built successfully then, we start the training process. The model training parameters would be a batch size of 128 with 200 epochs. Also, a validation dataset is taken from the x and y tests. After successfully training the model, we save it at a checkpoint for future use or reference.

Preliminary Result:

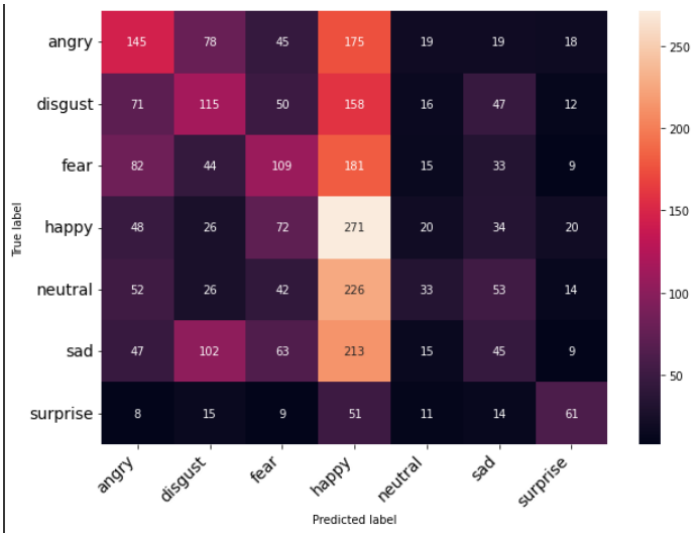
Here, as we can see, the model is a deep learning model with 8 input layers, several intermediate layers, and a softmax layer in the output. The model's accuracy is 70% for 200 epochs and a batch size of 128 for train data, while for the test data, it is 20%.





In this case, the model correctly predicted that the voice was female 1337 times out of 3041 samples, whereas the Male correctly identified male 800 times. We should further improve the model to get more accuracy.

And now for the emotional classification.



Here we can see that the number of samples in the happy dataset is greater than the others. Therefore, we need to decrease the samples to test them. However, as you can see from the model, most of the time, the model detects the correct samples in each sample. We are currently striving to improve the model's accuracy.

Project Management:

Implementation Status Report:

- **Work Completed:**
    - Description:
      So far, we have done the data exploration, cleaning, and plotting the preliminary Exploratory Data Analysis results. Then we have done a lot of research on the audio features and their techniques. We have derived the audio signals' time and frequency domain features. After that, we implemented a baseline sequential model to train and plot the results. We also tested this model on the test data to make predictions.

    - Responsibility:
        - Hemanth Reddy Yerramreddy: Frequency domain feature extraction, Implementing the Sequential Model for training.
        - Chandrakanth Mandalapu: Frequency domain feature extraction, Implementing the Sequential Model for training.
        - Sai Sri Harsha Chakravarthula: Exploration for a suitable dataset, Data Cleaning, Data Preprocessing, EDA, and Time Domain Feature Extraction.
        - Narendar Reddy Nelakurthi: Exploration for a suitable dataset, Data Cleaning, Data Preprocessing, EDA, and Time Domain Feature Extraction.

    - Contribution:

        - Hemanth Reddy Yerramreddy: 25%
        - Chandrakanth Mandalapu: 25%
        - Sai Sri Harsha Chakravarthula: 25%
        - Narendar Reddy Nelakurthi: 25%

- **Work to be Completed:**

    - Description:
      After increment 1, we plan to implement more audio feature extraction techniques and plot the subsequent results. After that, we plan to make a deep learning model from scratch and improve the performance compared to the baseline model. Also, this is now an Audio Sentiment analysis project. Still, we intend to make it more accurate by trying to train the model that will be able to classify between the genders and their emotions, respectively.

    - Responsibility:
        - Hemanth Reddy Yerramreddy: Fine-tuning the model and implementing a new model to compare it with the baseline model.
        - Chandrakanth Mandalapu: Fine-tuning and implementing a new model to compare it with the baseline model.
        - Sai Sri Harsha Chakravarthula: Exploring more feature extraction techniques and implementing them on the dataset.
        - Narendar Reddy Nelakurthi: Exploring more feature extraction techniques and implementing them on the dataset.

    - Issues/Concerns:
        - The dataset we got from Kaggle only had 250 samples of audio data. The dataset we got earlier is very small. So, the dataset's preprocessing and feature extraction techniques are not feasible.
        - Also, in the earlier dataset, the data samples showed similar spectrograms, which means that the learning algorithm will have difficulty classifying the audio sentiments correctly.
        - To investigate this problem, I listened to the audio clips and saw that the dataset is created by a single person same pitch and the same sound; the difference is in the word spoken. This shows that the dataset is not good for this project.
        - Due to these issues, we have decided to choose a different dataset with more data samples and classes to classify.

Reference

JIN LOK, E. (2019, September 20). Audio Emotion | Part 1 - Explore data. Retrieved from https://www.kaggle.com/code/ejlok1/audio-emotion-part-1-explore-data/data

savee | TensorFlow Datasets. (n.d.). Retrieved from https://www.tensorflow.org/datasets/catalog/savee

JIN LOK, E. (2019a, August 25). Toronto emotional speech set (TESS). Retrieved from https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

Livingstone, S. R. (2018, May 16). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391

Papers with Code - CREMA-D Dataset. (n.d.). Retrieved from https://paperswithcode.com/dataset/crema-d

Doshi, K. (2021, February). Audio Deep Learning Made Simple (Part 2): Why Mel Spectrograms perform better. Retrieved from https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505

Shah, A. K., Kattel, M., & Nepal, A. (n.d.). Chroma Feature Extraction. Retrieved from https://www.researchgate.net/publication/330796993_Chroma_Feature_Extraction

ProjectPro. (2022, June 14). Speech Emotion Recognition Project using Machine Learning. Retrieved from https://www.projectpro.io/article/speech-emotion-recognition-project-using-machine-learning/573

Github Link: https://github.com/HemanthReddy09/Sentiment-Analysis-of-Audio-Samples-Feature-Engineering

Dataset Link: https://www.kaggle.com/code/ejlok1/audio-emotion-part-1-explore-data/data

Video Link: https://www.youtube.com/watch?v=ggxlg7N45_w