CS536: Project Proposal
Speech Emotion Recognition

Harshini Bonam (NetID: sdb202) **Sai Mounica Pothuru** (NetID: sp1912)
Chaitanya Sharma Domudala (NetID: cd817)

April 15 2020

Introduction

Speech Emotion Recognition, abbreviated as SER, is a task of processing and classifying speech signals to recognize the embedded emotions. While this is efficiently performed by humans as a natural part of speech communication, the ability to conduct it automatically using programmable devices is still an ongoing subject of research. This is capitalizing on the fact that tone and pitch in the voice reflects underlying emotion.

SERs objective is to create efficient, real-time methods of detecting the emotions of diverse human-machine communication users. Adding emotions to machines has been recognized as a critical factor in making machines appear and act in a human-like manner (André et al., 2004). Machines capable of understanding emotions could provide appropriate emotional responses and exhibit dynamic personalities. Success in this field is defined by the machine's capability to conduct very natural and convincing conversations by appealing to human emotions.

Dataset

The data we are working with for this project is "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" [1] dataset which consists of 7356 files rated by 247 individuals 10 times on emotional validity, intensity, and genuineness. The entire dataset is 24.8GB from 24 actors but we are considered only audio file recordings of the 24 actors. The audio files of 24 actors alone constitutes of 1440 files of size 240MB. Each audio file has an unique name that defines various characteristics. A sample filename namely is 02-01-06-01-02-01-12.wav. Below are the characteristics that define the name.

- Modality
  1. full-AV
  2. video-only
  3. audio-only
- Vocal channel
  1. speech
  2. song
- Emotion
  1. neutral
  2. calm
  3. happy
  4. sad
  5. angry
  6. fearful
  7. disgust
  8. surprised
- Emotional intensity
  1. normal
  2. strong
- Statement
  1. "Kids are talking by the door"
  2. "Dogs are sitting by the door"
- Repetition
  1. 1st repetition
  2. repetition
- Actor
  1. (01 to 24. Odd numbered actors are male, even numbered actors are female

Data Preparation

Exploration

Using the Librosa library, the audio wave forms for different emotions by various actors with normal emotional intensity on the same sentence have been plotted along with the Spectogram frequencies.
Audio waveform for Actor 1 representing neutral emotion

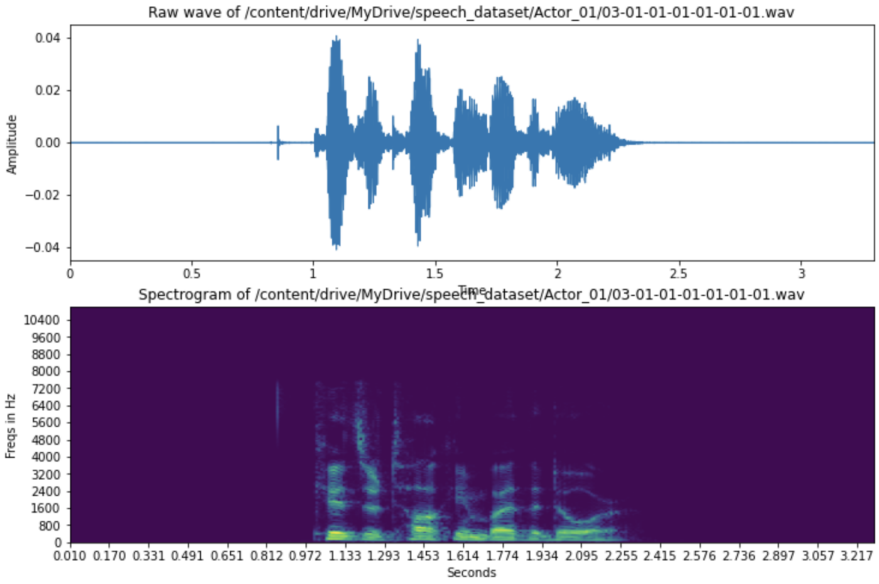

Figure 1: Emotion - Neutral

Audio waveform for Actor 12 representing calm emotion

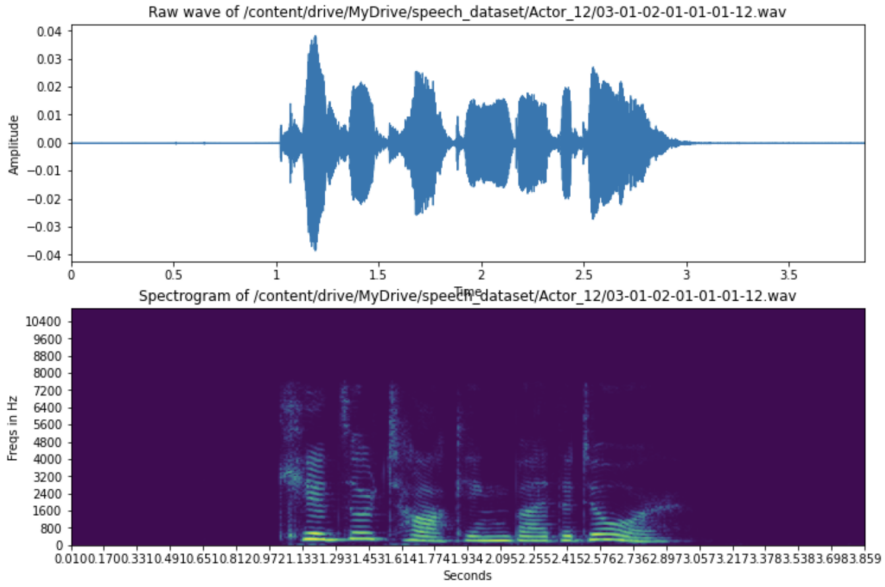

Figure 2: Emotion - Calm

Audio waveform for Actor 17 representing happy emotion

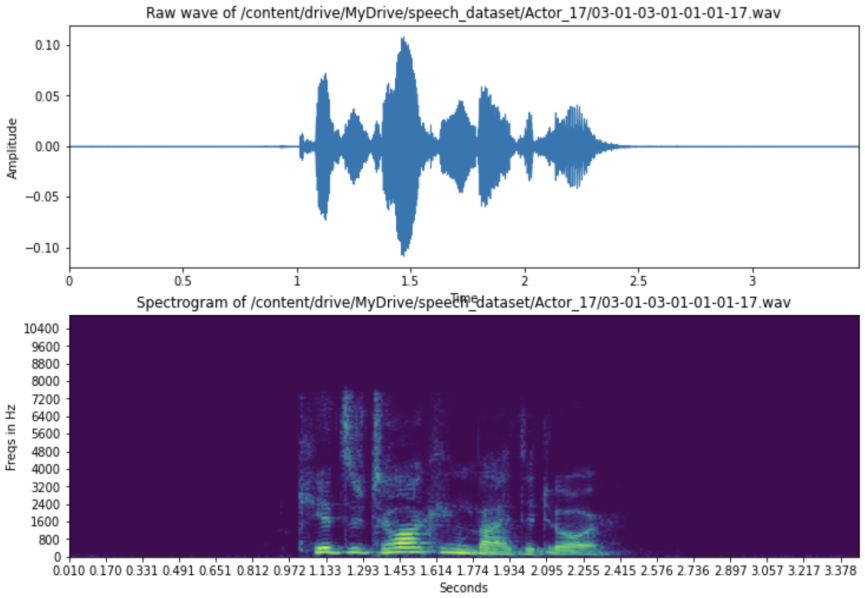

Figure 3: Emotion - Happy

Audio waveform for Actor 23 representing sad emotion

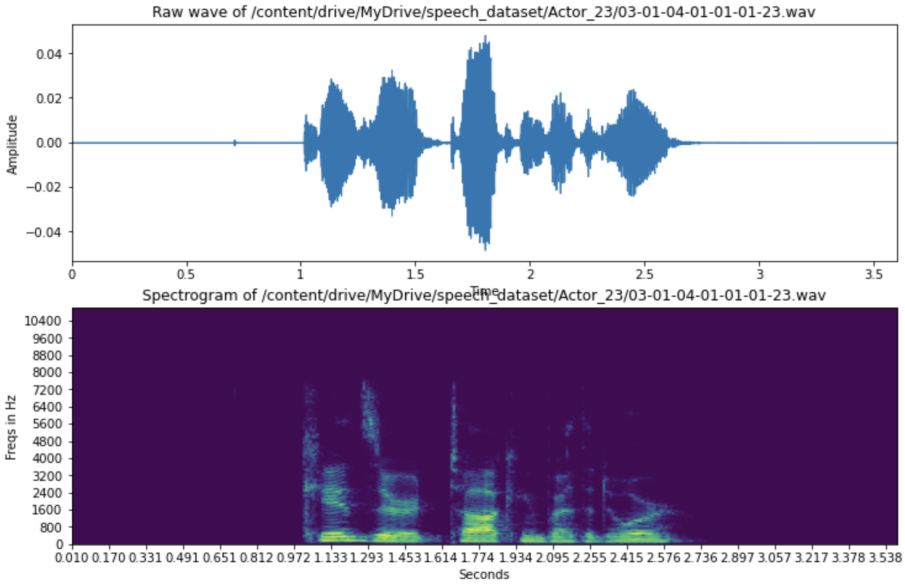

Figure 4: Emotion - Sad

Audio waveform for Actor 24 representing angry emotion

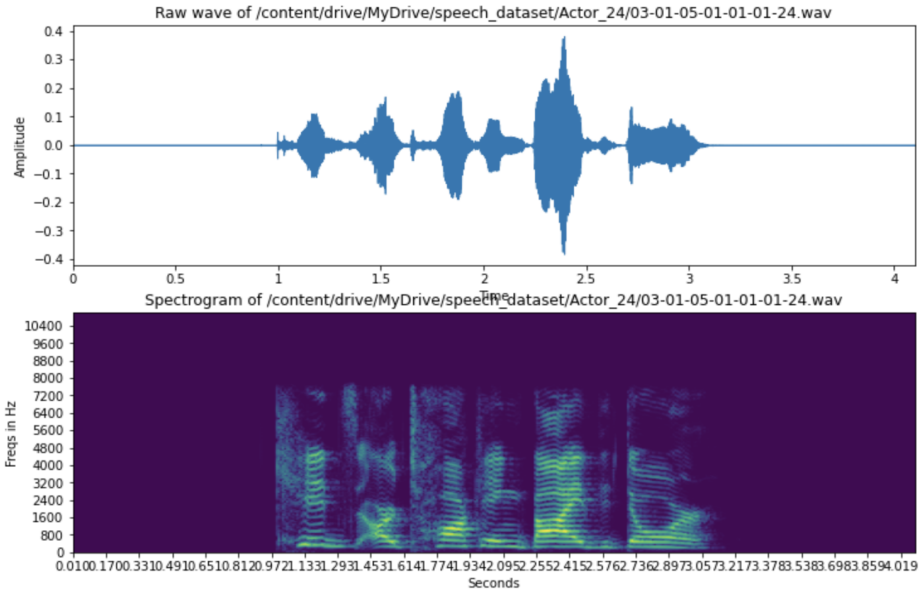

Figure 5: Emotion - Angry

Audio waveform for Actor 24 representing fearful emotion

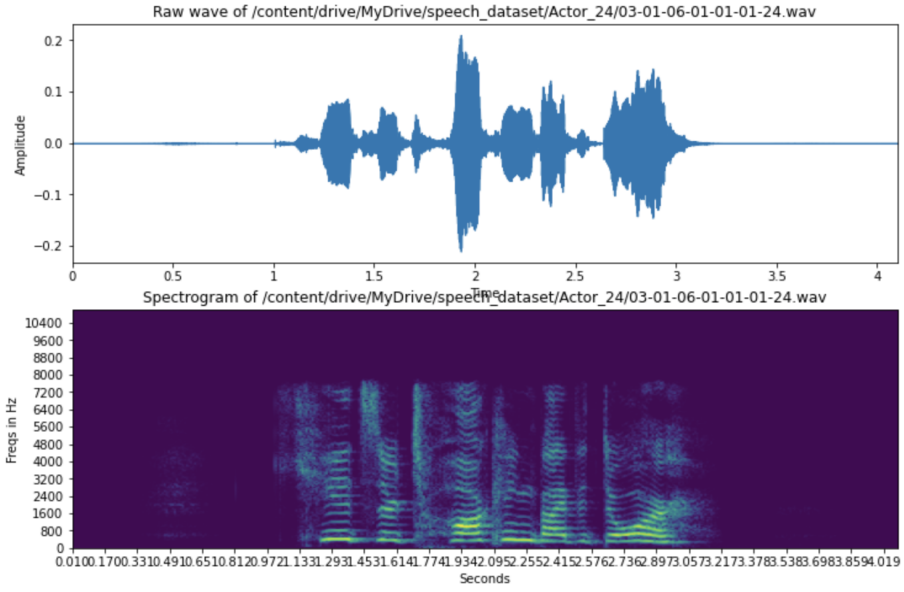

Figure 6: Emotion - Angry

Audio waveform for Actor 12 representing disgust emotion

Audio waveform for Actor 07 representing surprised emotion From a visual inspection it can be tricky to differentiate between some of the classes. Particularly, the waveforms for reptitive sounds for sad, neutral and calm are similar in shape. Likewise the peak in fearful emotion sample is similar in shape to the surprised emotion sample . Also, the car horn is similar too.

The human ear can naturally detect the difference between the harmonics, it will be interesting to see how well a deep learning model will be able to extract the necessary features to distinguish between these classes.

However, it is easy to differentiate from the waveform shape, the difference between certain classes such as angry and calm emotions.

Below are the distribution of emotions in the RAVDESS data.

According to the distribution, the emotions Calm, Happy, Disgust, Angry, and Fearful are only considered.

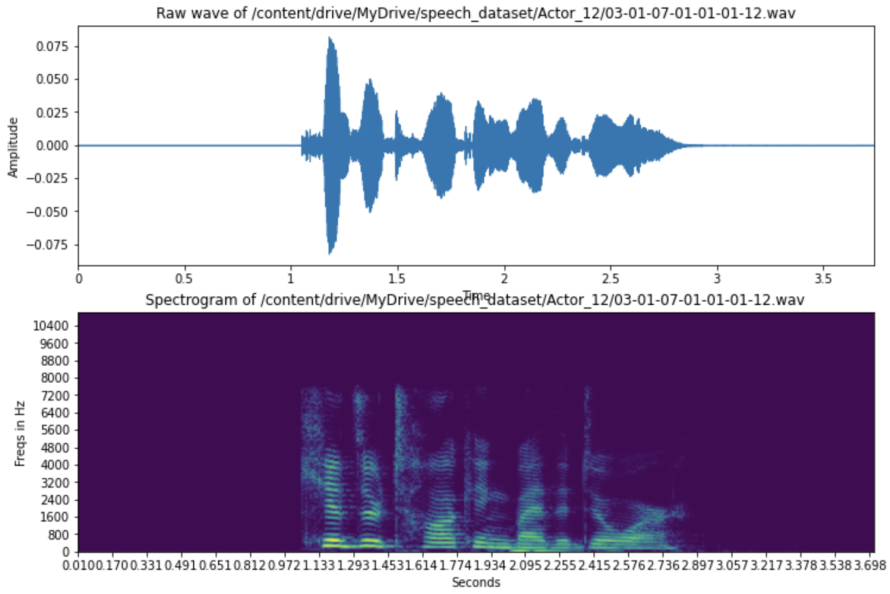

Figure 7: Emotion - Disgust

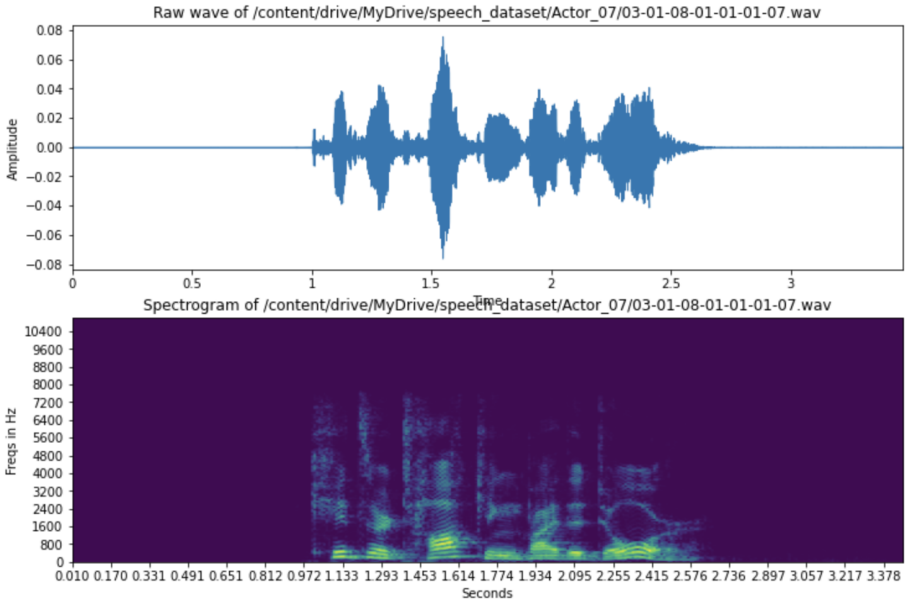

Figure 8: Emotion - Surprised

Feature Engineering

Audio files have various features that could be extracted using the Librosa library. The three most populat features are MFCC, Chroma and MEL.

1. MFCC (Mel-Frequency Cepstral Coefficients)
    - Summarises frequency distribution with the time characteristics of audio input.
    - Also called short term power spectrum of sound
    - State-of-the-art feature since it was invented in the 1980s.
    - Speech is the sounds that are generated by a human filtered by the shape of vocal tract including tongue, teeth. This shape if can be determined correctly, one can accurately represent what sound comes out.
    - The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC accurately represents this envelope.

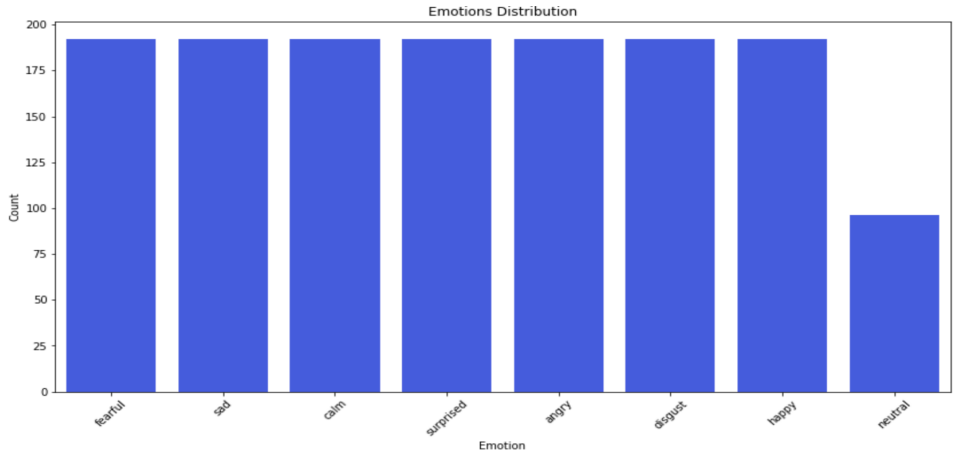

Figure 9: Emotions distribution

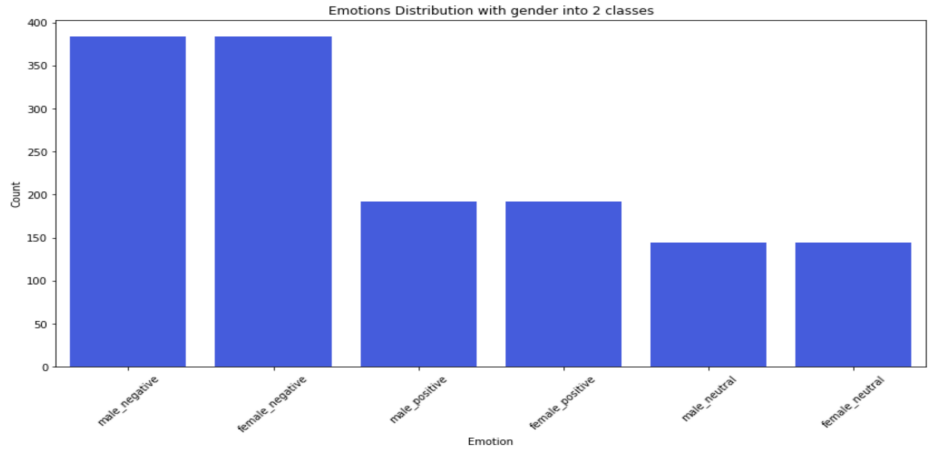

Figure 10: Emotions distribution with 2 classes with gender

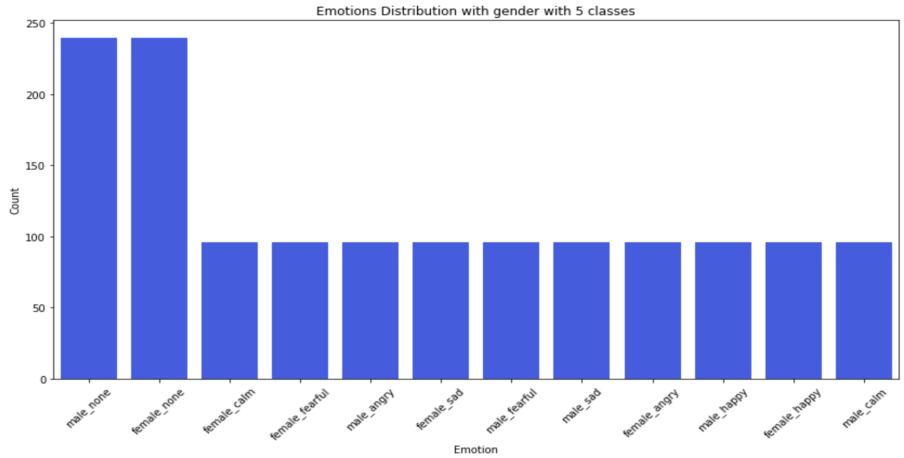

Figure 11: Emotions distribution with 5 classes with gender

2. Chroma
    - A descriptor which represents the tonal content of a musical audio signal in a condensed form.
    - They capture harmonic and melodic characteristics of music.
    - Our dataset has 12 Chroma features
3. Mel Spectrogram
    - **Mel Scale:** The logarithmic transformation of a signal's frequency.
    - **Core idea:** Sounds of equal distance on the Mel Scale are perceived to be of equal distance to humans.

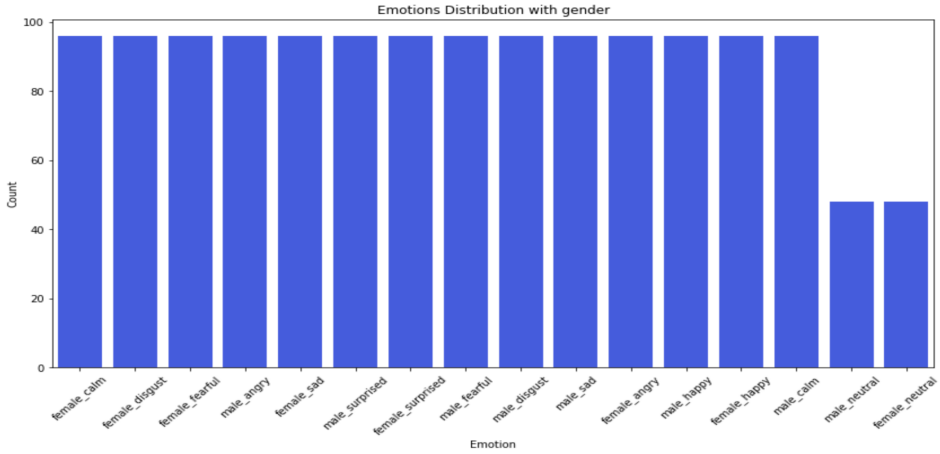

Figure 12: Emotions distribution with 8 classes with gender

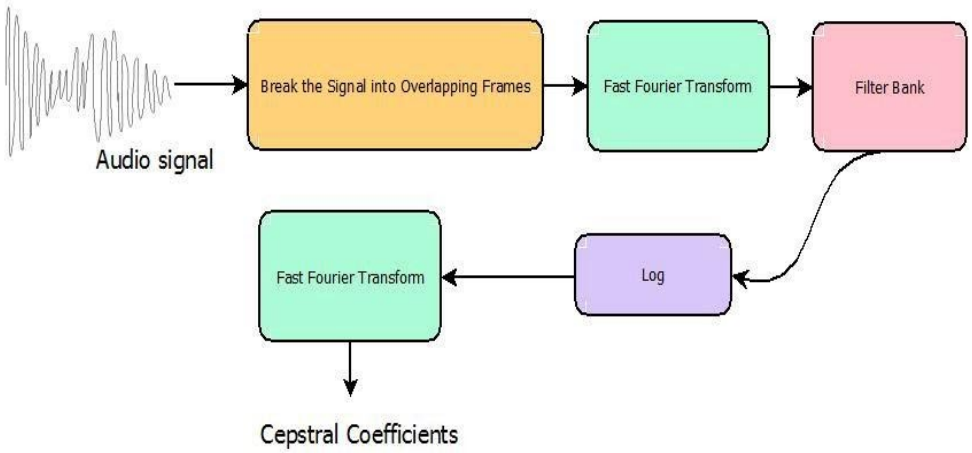

Figure 13: Flow Chart

- As opposed to the frequency domain, Mel Spectrograms are spectrograms that visualize sounds on the Mel scale.
- Our dataset has 128 Mel Spectogram frequencies

In the implementation, a function is written to extract the MFCC, Chroma and MEL features in the audio file by using sound file and librosa python libraries. Several combination of features have been used to model our algorithm, and conclude that MFCCs alone are better features for this dataset.

Data Modelling

In this project, two classifier models with the pre-processed data [6, 4, 5] have been implemented. 1. MLP, the base of Artificial Neural Networks(ANN) is used as the base model. 2. The same data is used to create a Conventional Neural Network(CNN) 1-d model. We concluded with CNN 2D model as the data is 2-Dimensional spectral information. 3. As a refinement, the 2-D CNN is model is trained with padded embeddings.

MLP Classifier

The MLP classifier uses the following configuration:

1. Data is loaded and split into train, and test with 0.25 ratio split.
2. Fed the training data to Sklearn library's MLPClassifier.
3. This Multi-layer Perceptron Classifier, a feed-forward ANN optimizes the log-loss function using L-BFGS or SGD.
4. We tuned a better model with the following hyper-parameters

- alpha = 0.01
- hidden_layer = 300
- learning_rate='adaptive'
- batch_size = 256
- epsilon=1e-07
- max_iter=1000

The best accuracy we could validate is 74.48% with this MLP Classifier. This beats the baseline accuracy as mentioned in [7, 2] with models trained with Speech audio and MFCC features.

CNN Classifier

Since the audio files are translated to 2-Dimensional spectograms, CNN's [3, 7] should refine the MLP model. hence a CNN has been implemented with the following configuration:

1. The features and emotions are the observations and labels and the data is split in the ratio of 0.2.
2. CNN model with Keras
3. Constructed with 4 Dense layers - 3 RELU activators and softmax activator, and Stochastic Gradient Descent optimizer, with a 0.1 dropout between each layer.
4. Compiled with 'categorical_crossentropy' loss method, accuracy as basis, and adam optimization algorithm for evaluation
5. Trained on a batch_size of 32 for 1000 epochs.

The validation accuracy has been significantly improved by using a CNN to 85.93%. However, our implementation could still be improved if padded embeddings were used.

CNN with Padding

In this implementation, every data token is embedded with a max padding of 174 and used the 2D-CNN with the following configuration:

1. CNN model with Keras.
2. Constructed with 7 layers - with 7 2D-ConvNNs each with RELU activators and softmax activator, with a 0.25 dropout between each layer.
3. Compiled with 'categorical_crossentropy' loss method, accuracy as basis , and adam optimization algorithm for evaluation
4. Trained on a batch_size of 100 for 500 epochs.

Results

1. The CNN Classification model attained a training accuracy of 100% and validation accuracy of 85.93%
2. The CNN classification model with padding obtained a training accuracy of 100.00% and testing accuracy of 88.54%
3. The MLP Classifier achieved an accuracy of 74.48%

Future Scope

Out of the three classification models generated, the CNN Classifier attains the best results when the data is padded. The implementation made sure that overfitting is avoided. The accuracy can further be attained by increasing the size of the dataset and by tweaking the hyper-parameters.

References

[1] he ryerson audio-visual database of emotional speech andsong (ravdess).

[2] Implementing baseline ser with mlp.

[3] Reza Chu. Speech emotion recognition with convolutional neural network, https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3, 2019.

[4] Shen L Pan Y, Shen P. Speech emotion recognition using support vector machine. int j smart home, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.390.8138rep=rep1type=pdf, 2012.

[5] Gaikwad VB Praksah C. Analysis of emotion recognition system through speech signal using knn, gmm svm classifier, https://www.ijecs.in/index.php/ijecs/article/view/3824, 2015.

[6] Neethu Sundarprasad. Speech emotion detection using machine learning techniques, https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1647context=etd_projects, 2018.

[7] Dezhong Peng Zhang Yi Yuni Zeng, Hua Mao. Spectrogram based multi-task audio classification, https://link.springer.com/article/10.1007/s11042-017-5539-3, 2017.