Detection of Gender and Emotion from Audio Signal

1. Abstract

Human emotions help communicate the inner feelings of every being and most often it's easily done through speech, through semantics and voice attributes, such as pitch, loudness, etc. Human beings have innate ability to recognize these emotions expressed in speech, but this task is not easy for machines, yet! This project aims to build a simple neural network solutions to try and predicting some basic human emotions given audio speech data.

2. Introduction

The fast-paced tech assimilation seen today has necessitated numerous real-world projects with the goal of making interaction between humans and machines as fluent as possible. A user's reliance on machines would increase as well-designed machines can reliably understand user interactions and provide better service. Voice recognition, facial recognition, language translation, audio detection, gender detection etc. are some of such tasks based on which even simple applications can provide services to make our life easier and more convenient and many big tech companies on this.

Enhancing the machine-user interactions, through multi-dimensional ways, other than the traditional keyboard-based interactions, have many applications, including:

- Voice enabled commercial platforms and applications including OS's, web applications, apps on phones etc. Virtual assistant on phones and computers can be personalized to provide better services if it could understand the auditory emotions.
- In Law enforcement for identifying distress calls or in identifying criminals or human actions.
- Better marketing through targeted advertisements.
- Customer Relationship Management.

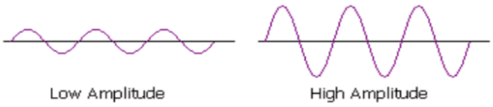
The aim of this project is to use Neural Network modelling for Speech Emotion Analysis, using the low-level audio features including usage of acoustic features and usage of the corresponding spectral information for the same. To satisfy the goal of this project we have resorted to use audio feature data of Mel-Frequency Cepstral Coefficients (MFCC) and by using Mel Spectrogram. We have used multiple architectures based on deep neural networks for the classification of emotions, including a 1D Convolutional Neural Network, LSTM Neural network and Transfer Learning based Neural Network for this project.

3. Brief Introduction to Audio Features

In physics, sound is a wave that is created by vibrating objects and it propagates through a medium from one location to another.

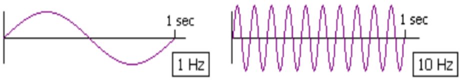
Amplitude

Amplitude is the size/strength of the vibration in a wave. The higher the amplitude of a wave, the louder the sound will be. Amplitude is important when balancing and controlling the loudness of sounds, such as with the volume control on a CD player.



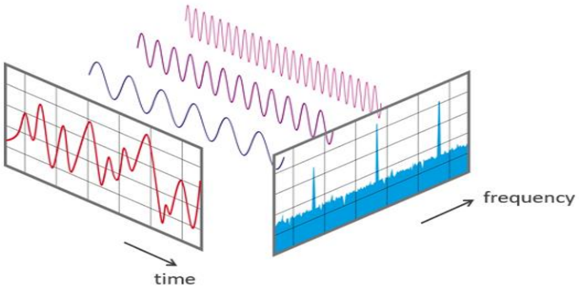
Frequency

Frequency is the speed of the vibration, and this determines the pitch of the sound. Frequency is measured as the number of wave cycles that occur in one second. The unit of frequency measurement is Hertz (Hz). A frequency of 1 Hz means one wave cycle per second. A frequency of 10 Hz means ten wave cycles per second, where the cycles are much shorter and closer together.



Fourier Transform

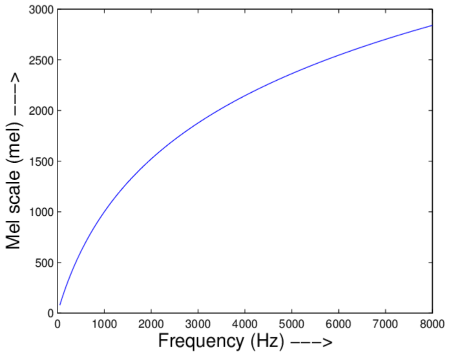
Fourier transform (FT) is a mathematical function which is used in frequency decomposition. FT treats signals with different frequency differently and is used to decompose a wave into its individual waveforms. Think of it like unmixing a bucket of paint into its individual colours. This concept is used heavily in audio analysis and decomposition specially in filtering out high pitch noises from audio.



Here we used Short-time Fourier Transform (STFT) to compute FT by dividing the audio signal into short-time windows and then computing FT on each window and then combining them back. In the below diagram we can see how the original waveform is divided into individual components.

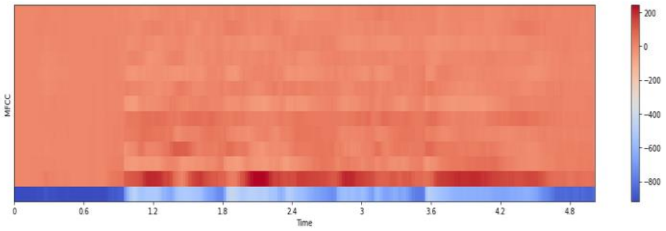
Mel-Scale

Humans don't perceive frequencies on a linear scale but rather on a logarithmic scale. What this means is that humans perceive the difference between lower and higher frequencies differently. Mel-scale perfectly represents the human perception of different frequencies. The below figure is a representation of the Mel-scale. On the lower frequencies the difference on the Mel-scale is higher while on higher frequencies the difference is not as much perceptible. This means we can detect small changes in lower frequencies but not so much in higher frequencies.



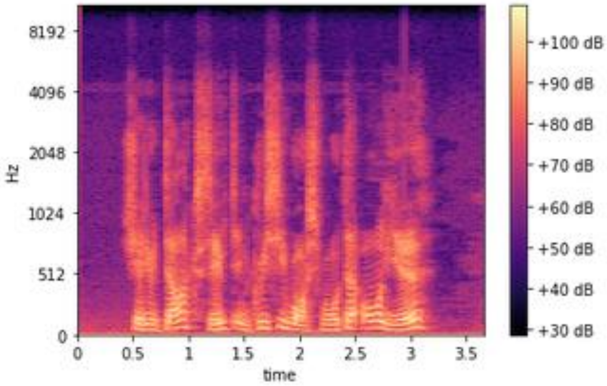
Mel-Frequency Cepstral Coefficient (MFCC)

The MFCC is considered as the finest feature to train models and is the primary feature to be used in automatic speech recognition. MFCC is a different interpretation of the Mel-Frequency Cepstrum (MFC). The MFC coefficients have mainly been used as the consequence of their capability to represent the amplitude spectrum of the sound wave in a compact vectorial form and are suitable candidates for input features to be used with neural networks.



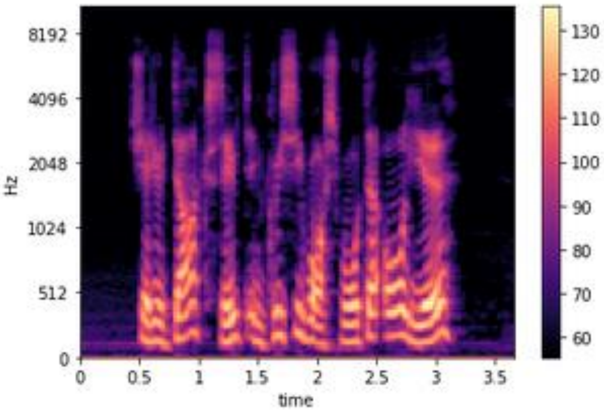
Spectrogram

In simple terms, spectrograms are a visual representation of sound. STFT is performed on raw audio data to convert them into a time v frequency scale to represent an image. A third dimension of amplitude which is generally indicated by a colour bar alongside the graph is used. The intensity of the colour tells us the amplitude at a particular time. Below is a spectrogram modified on the decibel scale.



Mel-Spectrogram

As the name suggests, when a spectrogram is converted to the mel-scale it is called Mel-spectrogram. It is an essential feature of Machine and Deep learning and is extensively used in audio analysis. Mel-spectrograms are preferred as stated above because of their scalability based on the human perception of sound. Below is a mel-spectrogram modified on the decibel scale.



4. Datasets

4.1. Introduction

We have used a combination of 4 datasets in this project.

RAVDESS:

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 2 Channel, 48 kHz).
In total there are about 1440 speech audio files, out of which 720 are from male and 720 female actors.

CREMA-D:

CREMA-D is a data set of 7,442 original clips (16 bits, PCM, 16 kHz .wav) from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).
Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad) and four different emotion levels (Low, Medium, High and Unspecified).
In total this dataset contains 7442 speech audio files, of which there are 3930 male audio and 3512 female audio file samples.

SAVEE:

Surrey Audio-Visual Expressed Emotion (SAVEE) database consists of recordings from 4 male actors in 7 different emotions (anger, happiness, sadness, surprise and neutral), 480 British English utterances in total (16 Bits, PCM 1 Channel, 44.1 kHz). The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion.

TESS:

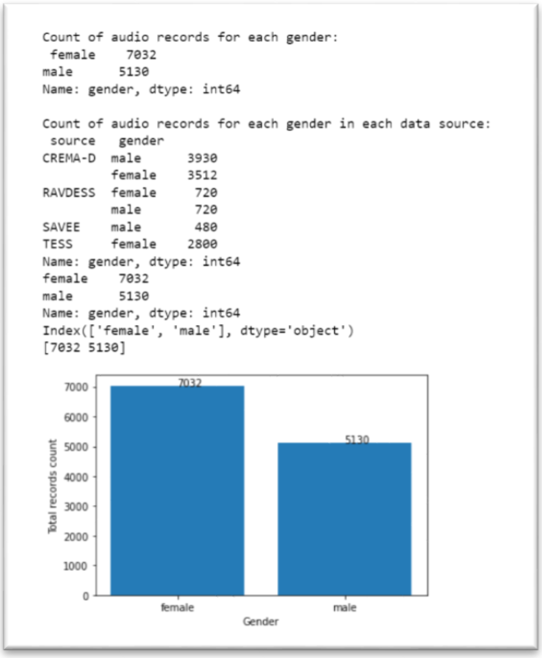
A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years). The recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total in .wav format (16 Bits, PCM 1 Channel, 24.414 kHz - 96 kHz).

NOTE:

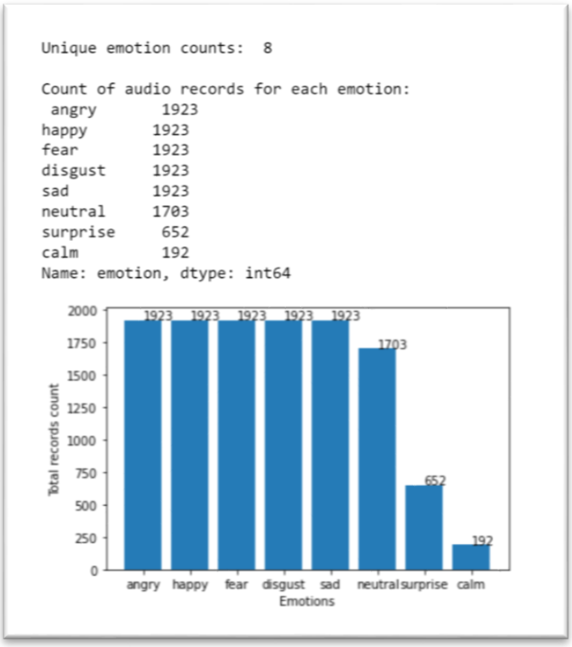
- RAVDESS and SAVEE Datasets are audio-visual database with both audio (speech and song) and the corresponding video data. For this project we have considered to use the audio speech data only.
- CREMA-D Dataset contains the same speech audio files in 2 formats ('.wav' and '.mp3' formats). For the project, only the '.wav' format files are considered.
- The combined datasets for the project include male and female open source speech sample audio files ('.wav' format). The speech audio length ranges from 1 seconds to maximum of 7 seconds.

4.2. Exploration

In total the combined dataset corpus contains 12162 speech audio sample. The gender and emotion class distribution of the dataset is as follows:



```
Count of audio records for each gender:
 female    7032
male      5130
Name: gender, dtype: int64

Count of audio records for each gender in each data source:
 source   gender
CREMA-D  male      3930
         female    3512
RAVDESS  female     720
         male       720
SAVEE    male       480
TESS     female    2800
Name: gender, dtype: int64
female    7032
male      5130
Name: gender, dtype: int64
Index(['female', 'male'], dtype='object')
[7032 5130]
```

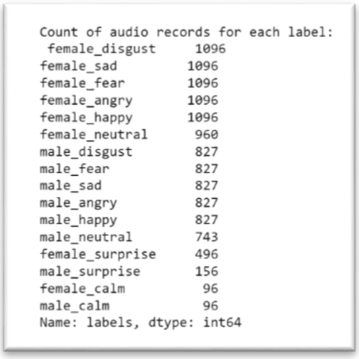


```
Unique emotion counts:  8

Count of audio records for each emotion:
 angry       1923
happy       1923
fear        1923
disgust     1923
sad         1923
neutral     1703
surprise     652
calm         192
Name: emotion, dtype: int64
```

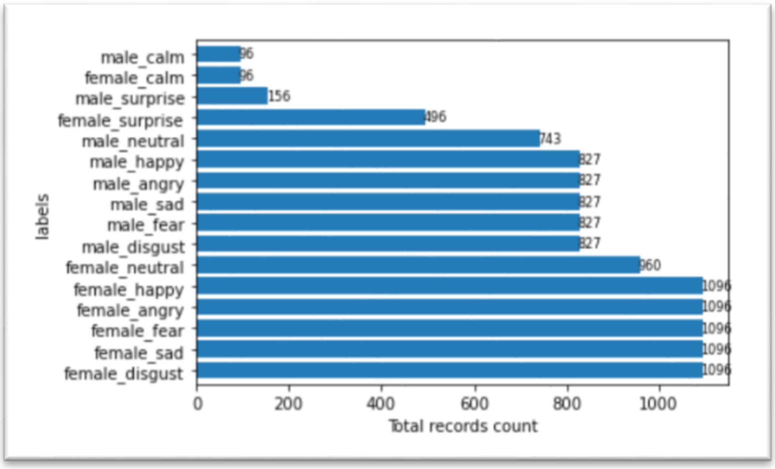
From the above image (left) it's clear that the gender distribution is skewed, with higher count (31.27%) of female emotion audio data than male emotion audio data. The speech audio datasets portray 8 basic emotions (shown in image on right). As is also evident, the total count of audio records for the emotions 'calm' and 'surprise' is substantially lower compared to the other emotion audio data.

4.3. Label Creation

Due to the difference in the pitch of female and male speaking voices, the generalizability of the model on different emotions on our approach went down. Hence it was necessary to keep the female and male emotions separate. 16 unique labels were created from the above data features to distinguish between male and female emotions. The labels are created by combining the gender (2) and the emotion (8) fields (in the format 'gender_emotion') to create the 16 unique labels.



```
Count of audio records for each label:
 female_disgust     1096
female_sad         1096
female_fear        1096
female_angry       1096
female_happy       1096
female_neutral      960
male_disgust        827
male_fear           827
male_sad            827
male_angry          827
male_happy          827
male_neutral        743
female_surprise     496
male_surprise       156
female_calm          96
male_calm            96
Name: labels, dtype: int64
```



4.4. Feature extraction

To extract audio features, the input data to the neural network framework, firstly we used the LibROSA python package which is detailed and easy to use for handling the audio data and converting the audio into machine readable code. Secondly, two separate approaches - MFCCs and Mel-spectrograms, were used as the input features.

MFCC Approach: The audio files were converted to MFCC values using the librosa API and the corresponding coefficients were update in the data frame for the corresponding audio label. These coefficient values were normalized and labels were converted to categorical form to be passed to the model.

Mel-spectrogram Approach: Here we used the image recognition method. Using the librosa API mel-spectrogram images were created from the audio data which was then loaded as a list of arrays which contain the values of each pixel. A corresponding list of labels is created simultaneously. The pixel values were normalized and labels were converted to categorical form to be passed to the model.

4.5. Dataset Augmentation:

Though the combined dataset does provide some amount of features diversity (since the audio files from different source datasets have different sampling rate), the deep learning models are still impacted with overfitting due to the comparatively small dataset (combined dataset of 12162 audio records). To increase the number of samples available for training and testing we have employed the data augmentation techniques using the SciPy library, LibROSA and Keras' ImageDataGenerator APIs. After extracting the audio data from the files, the SciPy library has been used to augment the audio data by adding noise, stretching and altering the speed and pitch of the audio. By this we were able to increase the total audio samples by 3 times. These techniques of augmentation are used in selected models only.

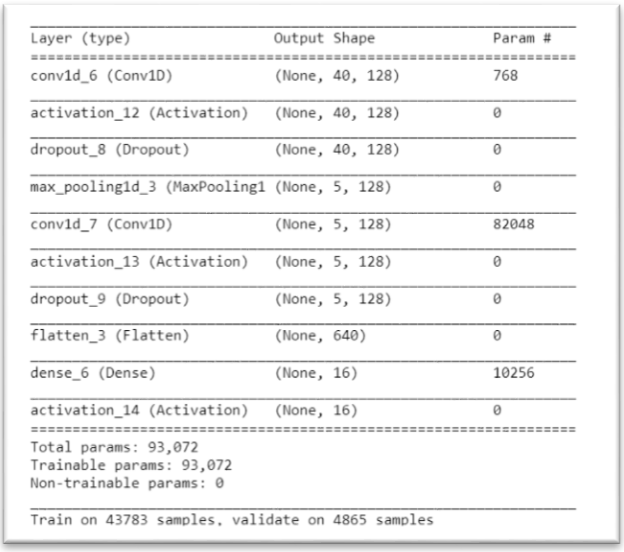
4.6. Training-Test Data Split

The audio file data have been randomly split into training and test datasets in ration 90%-10%. Consequently, for the first approach of using MFCC, the test set is composed by a 1217 MFCC vector of 40 features and rest is the training set containing 10945 MFCC vectors of 40 features. For the second approach of using Mel-Spectrogram images, the training set is composed of 10945 files, validation set with 973 files and test set with 244 files.

5. Proposed Models

5.1. 1D Convolutional Neural Network based model

A simple Deep Neural Network was designed for the classification task is reported operationally in the below fig. It is a simple One-Dimensional Convolutional Neural Network, containing 2 hidden layers.

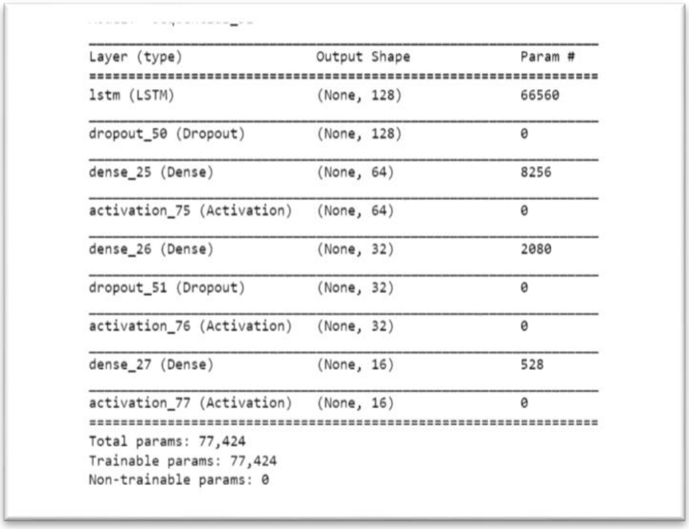


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_6 (Conv1D)            (None, 40, 128)           768
_________________________________________________________________
activation_12 (Activation)   (None, 40, 128)           0
_________________________________________________________________
dropout_8 (Dropout)          (None, 40, 128)           0
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 5, 128)            0
_________________________________________________________________
conv1d_7 (Conv1D)            (None, 5, 128)            82048
_________________________________________________________________
activation_13 (Activation)   (None, 5, 128)            0
_________________________________________________________________
dropout_9 (Dropout)          (None, 5, 128)            0
_________________________________________________________________
flatten_3 (Flatten)          (None, 640)               0
_________________________________________________________________
dense_6 (Dense)              (None, 16)                10256
_________________________________________________________________
activation_14 (Activation)   (None, 16)                0
=================================================================
Total params: 93,072
Trainable params: 93,072
Non-trainable params: 0
_________________________________________________________________
Train on 43783 samples, validate on 4865 samples
```

Uses Adam optimizer, and categorical cross entropy loss function. ReduceLROnPlateau and EarlyStopping call back functions are used.

5.2. LSTM Based Model

Among the variants of the RNN function, we LSTM based Neural Network. It is a simple network with 3 hidden layers as shown in the figure below:

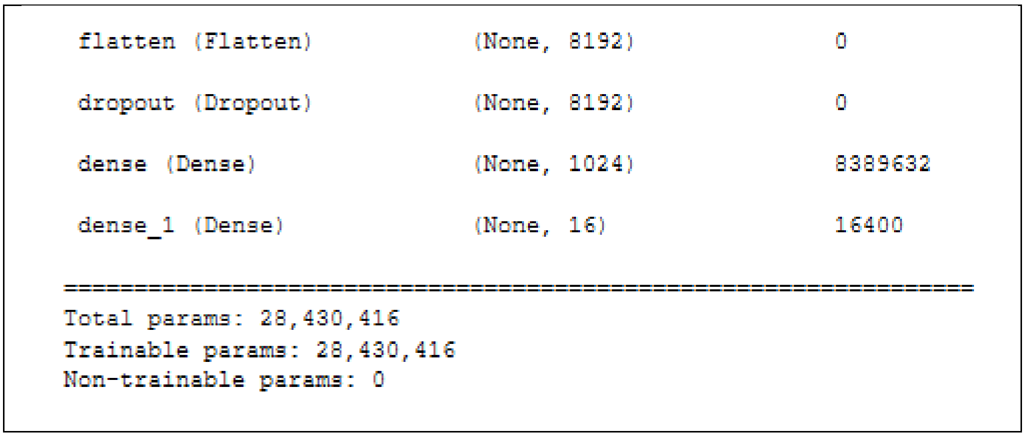


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 128)               66560
_________________________________________________________________
dropout_50 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_25 (Dense)             (None, 64)                8256
_________________________________________________________________
activation_75 (Activation)   (None, 64)                0
_________________________________________________________________
dense_26 (Dense)             (None, 32)                2080
_________________________________________________________________
dropout_51 (Dropout)         (None, 32)                0
_________________________________________________________________
activation_76 (Activation)   (None, 32)                0
_________________________________________________________________
dense_27 (Dense)             (None, 16)                528
_________________________________________________________________
activation_77 (Activation)   (None, 16)                0
=================================================================
Total params: 77,424
Trainable params: 77,424
Non-trainable params: 0
```

As an enhancement to the above LSTM model, the LSTM layer was updated to use BiLSTM or bidirectional LSTM, with the expectation of increasing the overall accuracy levels.
The model is trained with RMSProp optimizer, and categorical cross entropy loss function. ReduceLROnPlateau and EarlyStopping call back functions are used.

5.3. Mel-Spectrum based– Transfer Learning Model

Here we tested with pre-trained image recognition CNN models from the keras applications including VGG-16, VGG-19, Xception and Inception V3. Each of these models uses multiple 2D CNNs, pooling and Batch Normalization operations the layers. Instead of freezing the pre-trained layers, every layer was allowed to be trained (with a very low learning rate so as not to disturb the pre-trained weights too much) which provided the best results. Additional flatten, dropout and dense layers were added after the above model to regularize and to reconcile with the number of output classes in the output layer. Based on testing we have settled with VGG-19 model because of its higher accuracy scores with minimal overfitting.



```
 flatten (Flatten)            (None, 8192)          0

 dropout (Dropout)            (None, 8192)          0

 dense (Dense)                (None, 1024)          8389632

 dense_1 (Dense)              (None, 16)            16400

=================================================================
Total params: 28,430,416
Trainable params: 28,430,416
Non-trainable params: 0
```

Above are the additional layers added to the base VGG-19 model. The model is trained with RMSprop optimizer and categorical cross entropy loss function. ReduceLROnPlateau and EarlyStopping call back functions are used.

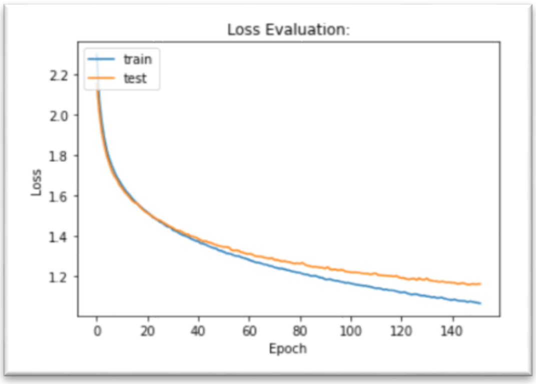
6. Evaluation

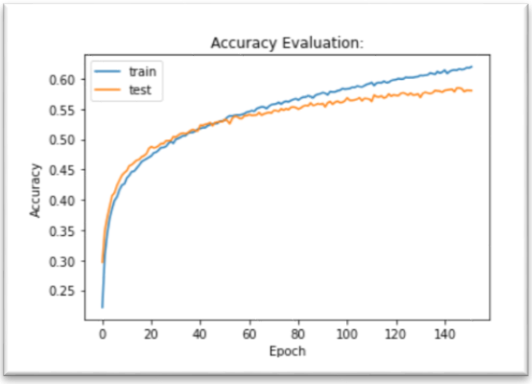
We used as evaluation metric the F1 score as a compact indicator of the quality of the classifier. Multiple testing iterations with different epochs were run and optimized values are chosen for each neural network. The number of batches has been set to 16 for optimization reasons. During the training we validate the model using the accuracy score as common in deep learning classification architectures. As apprised before, the total number of values of surprise and calm was substantially lower than other emotion classes, hence we didn't give these classes weight while evaluating the model.

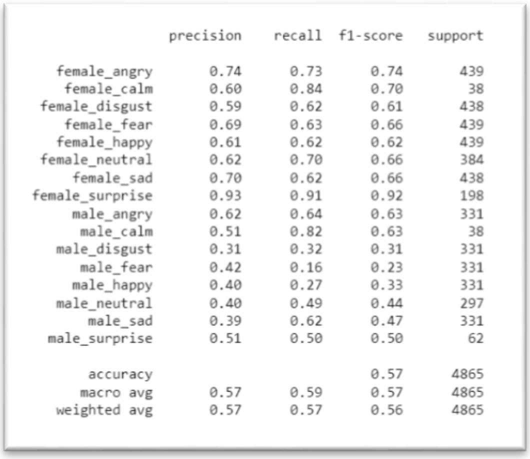
6.1. 1D Convolutional Neural Network based model

The above model was trained using the original datasets (without augmentation), with a batch size of 16 and epoch of 300. The initially model evaluation showed model overfitting with model's average validation accuracy of 57%+, and validation loss of 1.17.

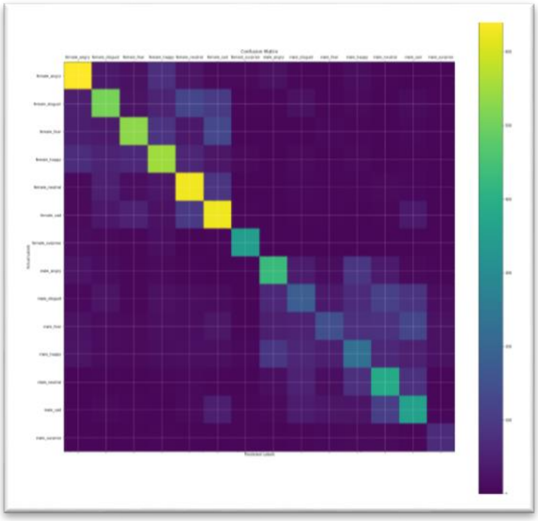
With application of data augmentation, dropout layers and early stopping, the overfitting reduced considerably. The model stopped training after 114th epoch due to early stopping. The overall evaluation validation accuracy is about 57%+ with a validation loss of 1.20. The overall overfitting reduced considerably as seen in the below plots:







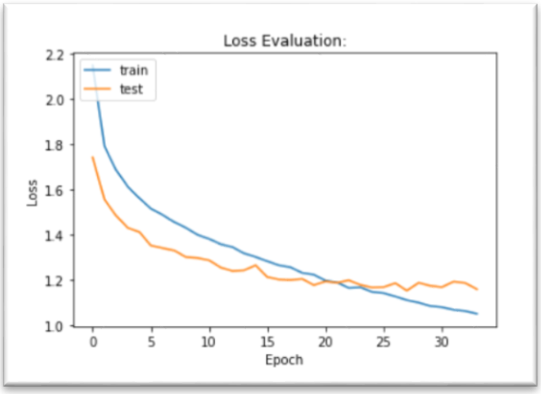
```
                 precision    recall  f1-score   support

   female_angry       0.74      0.73      0.74       439
    female_calm       0.60      0.84      0.70        38
 female_disgust       0.59      0.62      0.61       438
    female_fear       0.69      0.63      0.66       439
   female_happy       0.61      0.62      0.62       439
 female_neutral       0.62      0.70      0.66       384
     female_sad       0.70      0.62      0.66       438
female_surprise       0.93      0.91      0.92       198
     male_angry       0.62      0.64      0.63       331
      male_calm       0.51      0.82      0.63        38
   male_disgust       0.31      0.32      0.31       331
      male_fear       0.42      0.16      0.23       331
     male_happy       0.40      0.27      0.33       331
   male_neutral       0.40      0.49      0.44       297
       male_sad       0.39      0.62      0.47       331
  male_surprise       0.51      0.50      0.50        62

       accuracy                           0.57      4865
      macro avg       0.57      0.59      0.57      4865
   weighted avg       0.57      0.57      0.56      4865
```

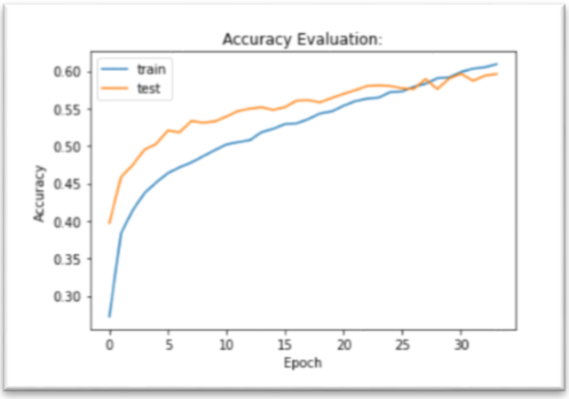


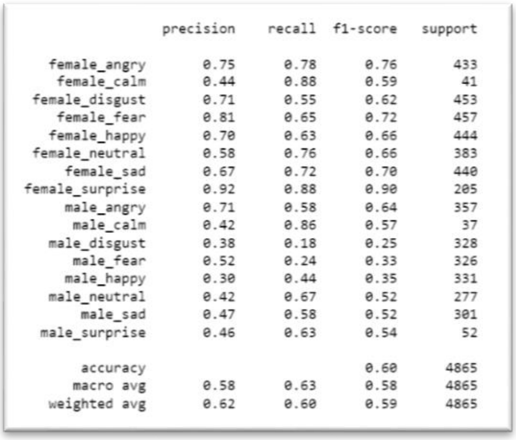
The highest f1 scores of 0.72 is seen for female_angry. The lowest f1 score 0.30 is of the male_fear label.

6.2. LSTM based model

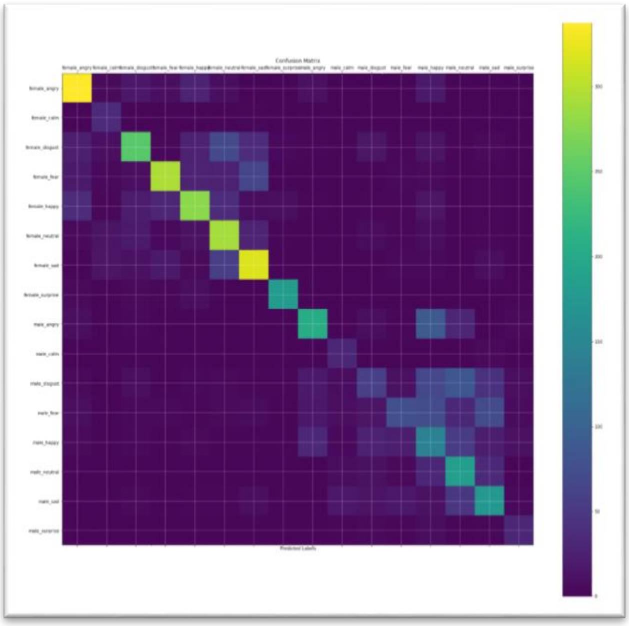
The LSTM model was trained using the complete dataset (original datasets + augmented data), with a batch size of 16. The training stopped after 36 epochs due to early stopping, and showed quicker learning (or generalization) compared to the previous model. The training had stopped just before the model started overfitting. The model evaluation showed average validation accuracy of 60%, and validation loss of 1.15.







```
                 precision    recall  f1-score   support

   female_angry       0.75      0.78      0.76       433
    female_calm       0.44      0.88      0.59        41
 female_disgust       0.71      0.55      0.62       453
    female_fear       0.81      0.65      0.72       457
   female_happy       0.70      0.63      0.66       444
 female_neutral       0.58      0.76      0.66       383
     female_sad       0.67      0.72      0.70       440
female_surprise       0.92      0.88      0.90       205
     male_angry       0.71      0.58      0.64       357
      male_calm       0.42      0.86      0.57        37
   male_disgust       0.38      0.18      0.25       328
      male_fear       0.52      0.24      0.33       326
     male_happy       0.30      0.44      0.35       331
   male_neutral       0.42      0.67      0.52       277
       male_sad       0.47      0.58      0.52       301
  male_surprise       0.46      0.63      0.54        52

       accuracy                           0.60      4865
      macro avg       0.58      0.63      0.58      4865
   weighted avg       0.62      0.60      0.59      4865
```



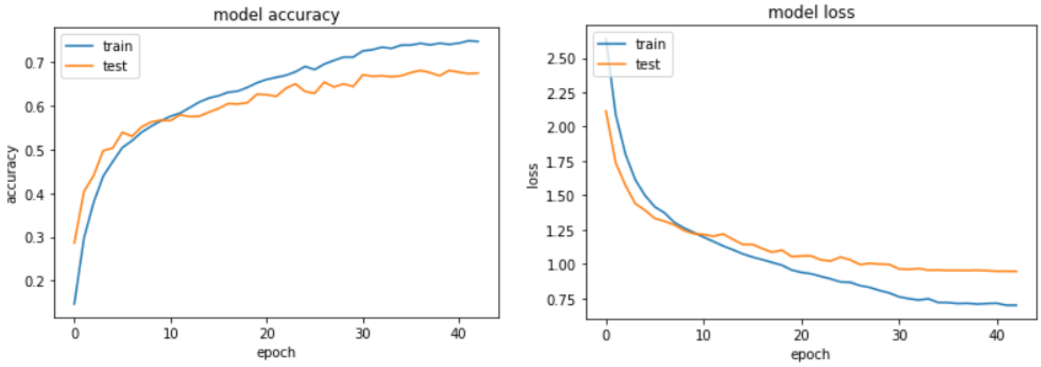
The highest f1 scores of 0.76 is seen for *female_angry* label. The lowest f1 score of 0.25 is of the *male_disgust* label.

6.3. Mel-Spectrum based– Transfer Learning Model

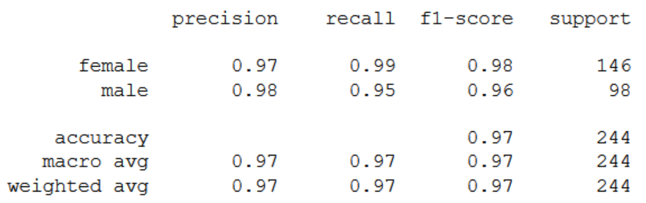
In identifying both the gender and emotion of the actor, the model resulted in a train accuracy of 73% with a loss of 0.71 and validation accuracy of 68% with a loss of 0.95.

Model was evaluated on test set with an overall average accuracy of 68% and a loss of 0.89. The model behaved particularly well in identifying the gender of the actor with an accuracy of 97%. Identifying only the emotion was a little less accurate with 70%.

Overall, the model is least accurate on the “disgust” class and most accurate on the “angry” class. The model generalizes better on female class against male.

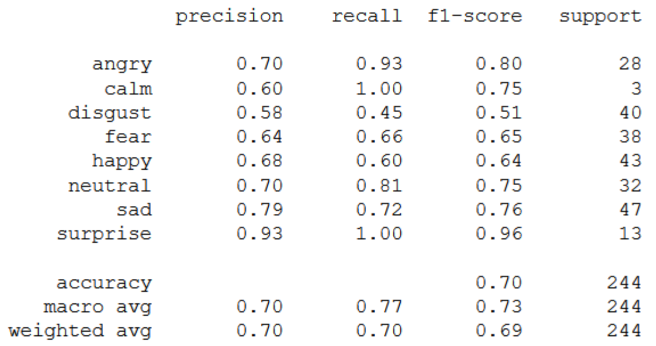


Gender evaluation



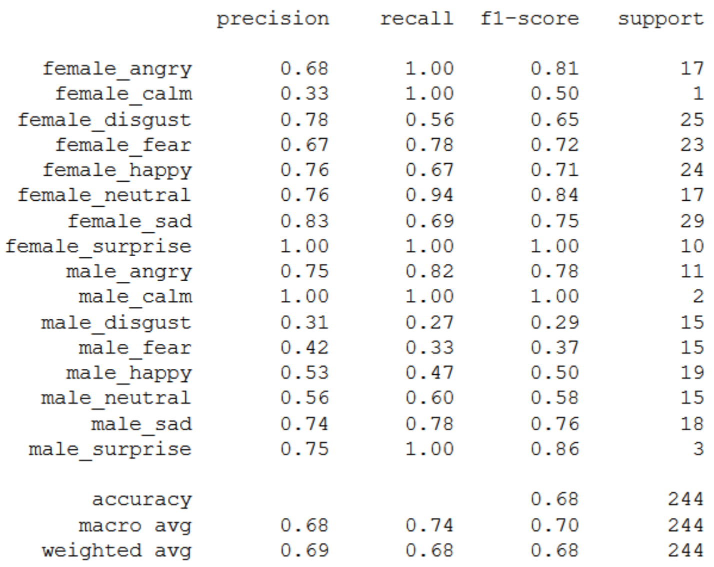
```
              precision    recall  f1-score   support

      female       0.97      0.99      0.98       146
        male       0.98      0.95      0.96        98

    accuracy                           0.97       244
   macro avg       0.97      0.97      0.97       244
weighted avg       0.97      0.97      0.97       244
```

Emotion evaluation

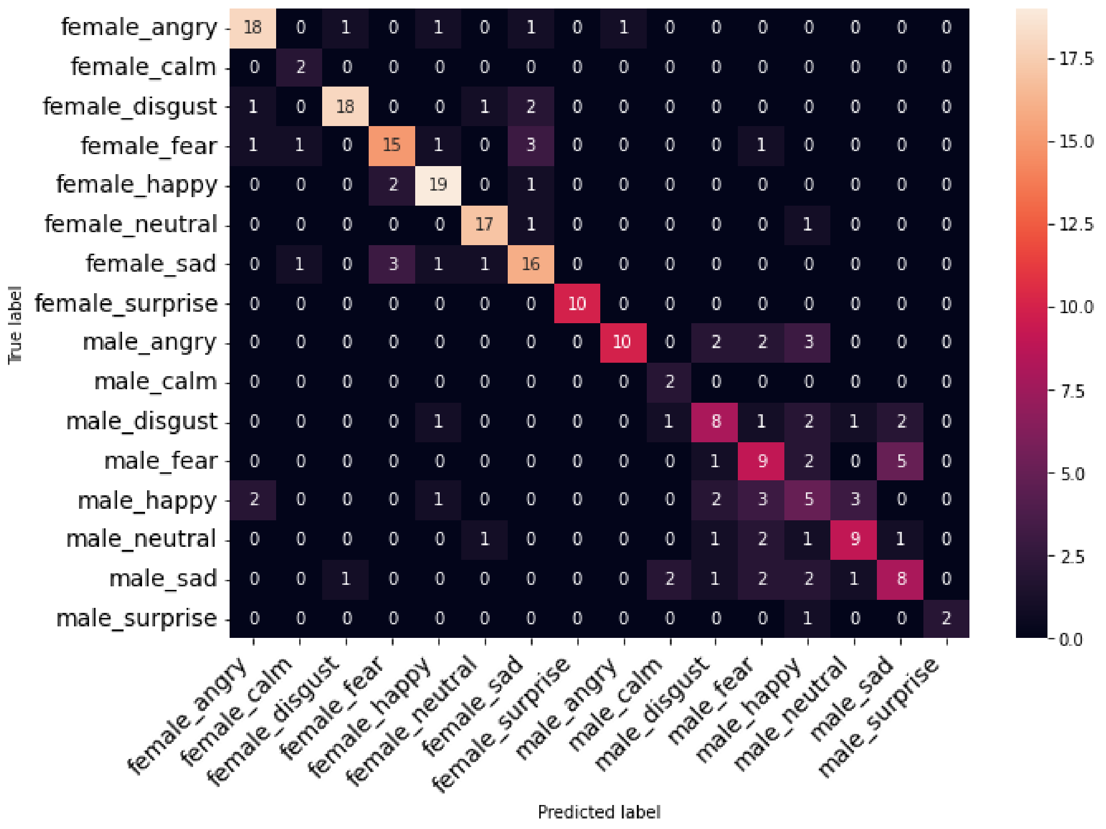


```
              precision    recall  f1-score   support

       angry       0.70      0.93      0.80        28
        calm       0.60      1.00      0.75         3
     disgust       0.58      0.45      0.51        40
        fear       0.64      0.66      0.65        38
       happy       0.68      0.60      0.64        43
     neutral       0.70      0.81      0.75        32
         sad       0.79      0.72      0.76        47
    surprise       0.93      1.00      0.96        13

    accuracy                           0.70       244
   macro avg       0.70      0.77      0.73       244
weighted avg       0.70      0.70      0.69       244
```

Overal evaluation



```
                 precision    recall  f1-score   support

   female_angry       0.68      1.00      0.81        17
    female_calm       0.33      1.00      0.50         1
 female_disgust       0.78      0.56      0.65        25
    female_fear       0.67      0.78      0.72        23
   female_happy       0.76      0.67      0.71        24
 female_neutral       0.76      0.94      0.84        17
     female_sad       0.83      0.69      0.75        29
female_surprise       1.00      1.00      1.00        10
     male_angry       0.75      0.82      0.78        11
      male_calm       1.00      1.00      1.00         2
   male_disgust       0.31      0.27      0.29        15
      male_fear       0.42      0.33      0.37        15
     male_happy       0.53      0.47      0.50        19
   male_neutral       0.56      0.60      0.58        15
       male_sad       0.74      0.78      0.76        18
  male_surprise       0.75      1.00      0.86         3

       accuracy                           0.68       244
      macro avg       0.68      0.74      0.70       244
   weighted avg       0.69      0.68      0.68       244
```



True label \ Predicted label	female_angry	female_calm	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_calm	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	18	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0
female_calm	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
female_disgust	1	0	18	0	0	1	2	0	0	0	0	0	0	0	0	0
female_fear	1	1	0	15	1	0	3	0	0	0	0	1	0	0	0	0
female_happy	0	0	0	2	19	0	1	0	0	0	0	0	0	0	0	0
female_neutral	0	0	0	0	0	17	1	0	0	0	0	0	1	0	0	0
female_sad	0	1	0	3	1	1	16	0	0	0	0	0	0	0	0	0
female_surprise	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
male_angry	0	0	0	0	0	0	0	0	10	0	2	2	3	0	0	0
male_calm	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
male_disgust	0	0	0	0	1	0	0	0	0	1	8	1	2	1	2	0
male_fear	0	0	0	0	0	0	0	0	0	0	1	9	2	0	5	0
male_happy	2	0	0	0	1	0	0	0	0	0	2	3	5	3	0	0
male_neutral	0	0	0	0	0	1	0	0	0	0	1	2	1	9	1	0
male_sad	0	0	1	0	0	0	0	0	0	2	1	2	2	1	8	0
male_surprise	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2

Given the size of the dataset with 12,612 audio files with an expected output of 16 classes leaves a very small subset of data for each class. To counter this further data augmentation was applied on the raw audio files to make the dataset bigger and with equal distribution of all classes. However, this approach did not have any functional effect on the model.

7. Results

7.1. Conclusion:

The performance of the neural network model does suggest that it is accurate enough in generalizing the emotions. This opens scope for further deeper analysis by subject matter experts or academic studies in this field. From the tests done, clearly the pre-trained CNN VGG-19 model applied on mel-spectrogram images performs better. The VGG-19 model, which was trained on imagenet database, was originally trained for object detection. However our project had utility in identifying and classifying the mel-spectrogram images and not objects. Hence the layers in the model were allowed to be retrained. Since all the layers were re-trained, the imagenet weights have been updated, however with minimal change as a very low learning rate was used.

7.2. Learnings and Future Enhancements:

We have trained many different types of Neural Networks and the corresponding results are shown. Based on this, we have listed out some limitations on achieving higher accuracy levels and better prediction results. To summarize them:

- Insufficient trainable data availability.
- The datasets used had audio recording samples for multiple enunciation and accents, but with insufficient samples for each labels.
- The overall f1 scores for the female emotion labels are higher than that of corresponding male emotion labels. Hence the model generalizes better for female emotion labels.
- The recorded speech by the actors in the audio files is dramatic to certain extent. This speech does not accurately represent normal speech pattern by the general public.

Further modifications and improvements with thorough *signal-processing* domain inclusive knowledge of pre-processing, feature selection and extraction methods are bound to work for a better accuracy.

Here we have used MFCC OR Mel-Spectrum as the audio features, which means the models generalizations to identify of emotions are learnt from the low-level speech signals. Generating audio feature models using low level speech audio feature data (MFCC, Mel Spectrogram) along with high level model based on speech text analysis (Speech-To-Text and NLP) should help achieve better accuracy levels.