Analysis of Emotion Recognition using Speech

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Electronics and Communication Engineering

by

Sangeet Sagar - 15uec053

Under Guidance of
Dr. Joyeeta Singha
Dr. Navneet Upadhyay

Department of Electronics and Communication Engineering
The LNM Institute of Information Technology, Jaipur

May 2018

The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled Analysis of Emotion Recognition using Speech, submitted by Sangeet Sagar (15uec053) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by them at the Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2018-2019 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this report is of standard required for the award of the degree of Bachelor of Technology..

Date	Adviser: Dr. Joyeeta Singha
	Adviser: Dr. Navneet Upadhyay

Dedicated to My Family and Friends

Acknowledgments

In the present world of competition there is a race of existence in which those are having will to come forward succeed. A project work like this is similar a bridge between theoretical and practical learning. With this willing I joined this project. First of all I would like to express my sincere thanks to the supreme power the Almighty God who has always given me inner spirit and constant determination. Without his grace, this project could not become reality. Next to him are my parents, whom I am greatly indebted for my brought up with love and encouragement to this stage. I feel obliged in taking the opportunity to sincerely thanks Dr. Navneet Upadhyay and Dr. Joyeeta Singha. I am also indebted to the entire electronics and communication discipline for their attitude and friend behavior.

At last but not the least I am thankful to all my teachers and friends who have been always helping and encouraging me throughout the year. I am speechless when it comes to express my thanks, but my heart is still full of the favors received from every person.

Abstract

This work shown in this project is mainly concerned with the analysis and recognition of human emotion with the help of speech.

Earlier work in emotion recognition using speech includes algorithms for feature extraction like MFCC (Mel-frequency cepstral coefficients), delta-MFCC, double-delta-MFCC, LPCC (Linear predictive cepstral coefficients), pitch, energy, timber and many more. In this project we have tried to incorporate few of these existing techniques on standard databases like SAVEE (Surrey Audio-Visual Expressed Emotion) and Emo-DB (Berlin database of emotional speech) using classifiers like GMM (Gaussian mixture model), deep learning algorithm like CNN (convolutional neural networks) and MLPNN (Multilayer perceptron neural network). Apart from these using these above existing techniques we have implemented some new feature extraction algorithms like S-Transform (Stockwell transform) and image spectrogram of the speech signal. Both of these have shown outstanding classification accuracy as compared to the pre-existing ones. S-transform has an advantage that it represents the speech signal in time and frequency domain at once and thus it has proved to be a more dynamic feature extraction method. The spectrogram is one of the simplest and most effective ways of visualizing the time-frequency evolution of the spectral, prosodic and energy features.

In this work apart from classifying the emotions we have also attempted to characterize emotions on the basis of R G B spectrogram. We have computed statistical parameters for each emotional utterance for each of the R, G, B spectrogram.

Contents

Chapter *1*

Introduction

Emotion remains an unpredictable state of mind that largely depends on the people and the environment a person encounters during his/her daily life. It affects his/her living conditions, working efficiency, family life as well as social well-being. During the last few decades, many acoustic parameters describing speech emotions related to the pitch, duration, intensity, and their statistical characterization have been strenuously attempted for better SER accuracy. The major area of focus has been developing an efficient model using prosodic and spectral features with a better discriminating ability.

1.1 Objective

The fundamental objective of this work is to analyze how emotions can be perceived from the speech of an individual. We all know that emotions are the results of the mental activity going inside and speech is one of the sources that depict our mood. The study of emotions from voice communication remains significant when it is not possible to have a direct face-to-face conversation. This makes the domain a complex and challenging field of research. Besides this, there are a few emotions such as boredom and disgust which are difficult to perceive using facial or gestures modalities of expressive emotions. However, these emotional states can be identifiable by voice tone due to different speech prosody. This creates interest in the speech community and the same is our primary objective behind this work.

1.2 Motivation

Extraction of suitable feature to represent the emotional contents in a speech signal has been a long outstanding research area speech community has been exploring during the last few decades. The signal can be represented in time, frequency or both time-frequency plane and may contain different amount of energy corresponding to different emotional states. A common, simple and effective approach to represent the emotional speech signals in the time-frequency plane is the spectrogram. It is the best visual representation of the amount of the energy contained in an emotion speech utterance. Spectral features like LPCC and MFCC are quite popular in SER but it's quite old.

1.3 Application

Speech emotion recognition has quite an extensive field of application. They are used by the police department to examine the mental state of a criminal. Where criminal can always be deceptive from their statement but the emotion in their speech can describe a lot. SER can also be used in the car driving system to monitor the mental state of the driver and thus prevent possible rad accidents. They are also used in the health care industry to keep an eye on their patients and how well is he/she is recovering

Chapter *2*

Literature Survey

An accuracy of 73.76% has been claimed with K-nearest neighbors (KNN) classifier for stress identification [10]. The authors have applied the sigma-pi cells corresponding to the ERB (Equivalent Rectangular Bandwidth) spectrograms [10]. Current activities in SER emphasize on Deep Neural Network based speech emotion models using frame-based features speech spectrograms. The claimed accuracy has been 60.53% with eNTERFACE database and 59.7% with the SAVEE database [8]. Further progresses show the use of deep Convolutional Neural Networks (CNNs) to train RGB images of speech spectrograms [6, 11]. The authors have to train a huge number of labeled images known as fresh training using the complex CNN structures to achieve the desired label of accuracy. Such system desires a computationally intense and complex training unless extensively large labeled speech emotion dataset has been used for training which puts a limit on the study. To alleviate the issue, few authors have used fine-tuning to pre-train large number of images using the VGG-Face network, AlexNet or the Fast Region-based CNN (Fast R-CNN). Authors have used the RGB spectrogram of speech emotions using two image processing approaches such as AlexNet-SVM and FTAlexNet to classify speech emotions with efficient accuracy using CNN [9].

A paper on Emotional Speech Recognition using Optimized Features (2017) by H. K. Palo proposed a novel approach using Multilayer Perceptron (MLP) classifier for Spectral roll-off, Spectral centroid and Spectral flux features on SAVEE database. The average percentage of accuracy recorded for Spectral roll-off, Spectral centroid and Spectral flux features for three emotions sad, neutral and happy are 59.60%, 62.88%, and 65.95% respectively. In his another paper[3] Efficient feature combination techniques for emotional speech classification (2016), H. K Palo proposed the uses of reduced features sets LP VQC (Linear prediction vector quantizer coefficients) and pH VQC (Hurst parameter vector quantizer coefficients) using Radial Basis Function Neural Network (RBFNN) for classification on EMO-DB database.

The proposed feature extraction technique S-transform has been adopted from [1]. S-transform is an outstanding technique for representation of a signal in the time-frequency domain.

There have been many efficient classifiers used by researchers to develop suitable identification system modeling in the field of SER. The self-learning ability of NNs with their efficient regulation in representing complex inputs makes them well suited to model speech emotions [2, 3-4, 5]. In comparison to the stochastic Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM), the NNs found to be more versatile for nonlinear mapping of data. For small training samples, their recognition performance has been better than the GMM/ HMM in classifying speech emotions [6]. These NNs are parallel structured and do not require any assumption of input distribution as speech frequencies occur in parallel [7]. For multiple input streams, these networks provide flexible representation using inputs with a single frame as compared to the GMMs or HMMs that require multiple frames [8].

Chapter *3*

Proposed System: Emotion Recognition using Speech

3.1 Database

In this work, we have experimented with only two standard available databases. They are SAVEE and Emo-DB database.

Surrey Audio-Visual Expressed Emotion (SAVEE) database have been used to compute the S-transform (Stockwell Transform). The database has been collected by the University of Surrey, England. The recorded database consists of 4 male English speaking speakers. It has seven discrete states of emotions. These are disgust, sadness, happiness, fear, neutral, anger, surprise emotional states. A total of 60 utterances is contained in each emotion with each speaker recording 15 different voice samples for each emotion. So, collectively SAVEE consists of 420 speech utterances. These utterances are sampled at 41000 kHz.

The Berlin (EMO-DB) German database has been used to extract the spectrogram images of the chosen emotional utterances [14]. The database has been collected by the Technical University of Berlin. It has seven discrete states of emotions. These are disgust, sadness, happiness, fear, anger, boredom, and neutral emotional states. A total of ten professional actors that include five female and five males have mimicked these chosen emotional states. The chosen dataset comprises of five long as well as five shorter sentences in German native language [3]. Semantically neutral and everyday utterances with a duration between 1.5 to 4 seconds have been selected in this database. The utterances are sampled at 16 kHz. A precision of 16-bit is maintained for recording the utterances in an an-echoic chamber. A sum total of eight hundred utterances (10 actors 10 sentences 7 emotions + some second versions) have been simulated by these actors. Ten evaluators have recommended approximately five hundred utterances for experimental purpose. The selection of the utterances is made based on the clarity, and emotional relevance of the simulated sentences. Utterances based on more than 80% recognition accuracy and authenticated by above 60% evaluators are considered. The final database comprises 127 angry, 46 disgusting, 69 fear, 71 happy, 62 sad and 79 neutral emotional utterances in total.

3.2 Feature Extraction Technique: Stockwell Transform (S-Transform)

The Stockwell transform popularly known was S-transform works quite well for the non-stationary signal. It is a novel method of representation of a signal in the time-frequency domain. Apart from this it also contains information about the phase data. It can be viewed as an extension to wavelet transform.

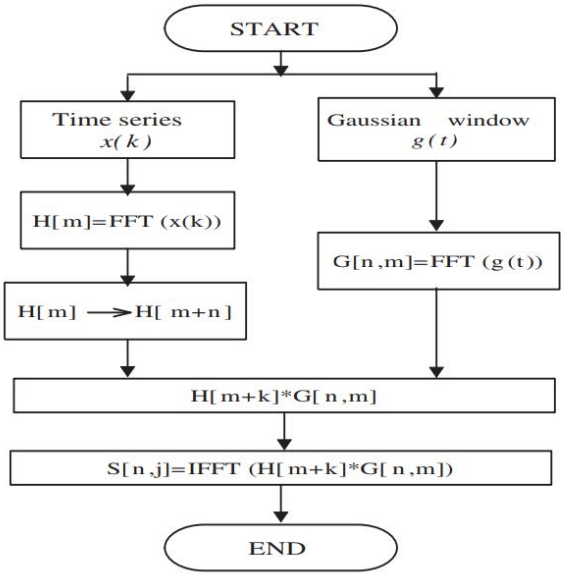


Figure 3.1: Block diagram representation of the S-transform

Mathematically, S-Transform is defined as

$$S_x(f,\tau) = \int_{-\infty}^{\infty} x(t)g(f,\tau,-t)e^{-j2\pi ft}dt \tag{1}$$

where $g(f,t))$ is a normalized Gaussian window. It is given by

$$g(f,t) = \frac{1}{\sigma(f)\sqrt{2\pi}}e^{\frac{t^2}{2\sigma^2(f)}} \tag{2}$$

The window width is decided with the help of $\sigma(f)$. It is inversely proportional to the frequency f. This can be represented as

$$\sigma(f) = \frac{1}{a+b|f|} \tag{3}$$

After multiplying the continuous wavelet transform with the phase correction error, we get the generalized S-transform

$$S_x(f,\tau) = e^{-j2\pi ft}W(f,\tau) \tag{4}$$

3.3 Feature Extraction Technique: Spectrogram

Acoustic speech characteristics vary in frequency, time and intensity. One of the most common ways of observing these variations is via the speech spectrogram. The spectrogram is one of the simplest and most effective ways of visualizing the time-frequency evolution of the spectral, prosodic and energy features. For each speech waveform, we first removed the silenced part and computed the short time Fourier transform. This gave three outputs namely time, the frequency and a complex matrix containing the computations of STFT. After taking the log of this complex matrix, we plot the spectrogram using the images MATLAB command.

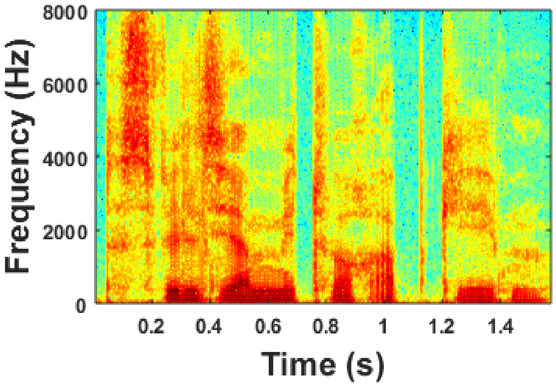


Figure 3.2: Sample Speech Spectrogram

The proposed characterization and classification of the Emotional Speech Recognition (ESR) has the following block diagram as shown in Fig.3.3.

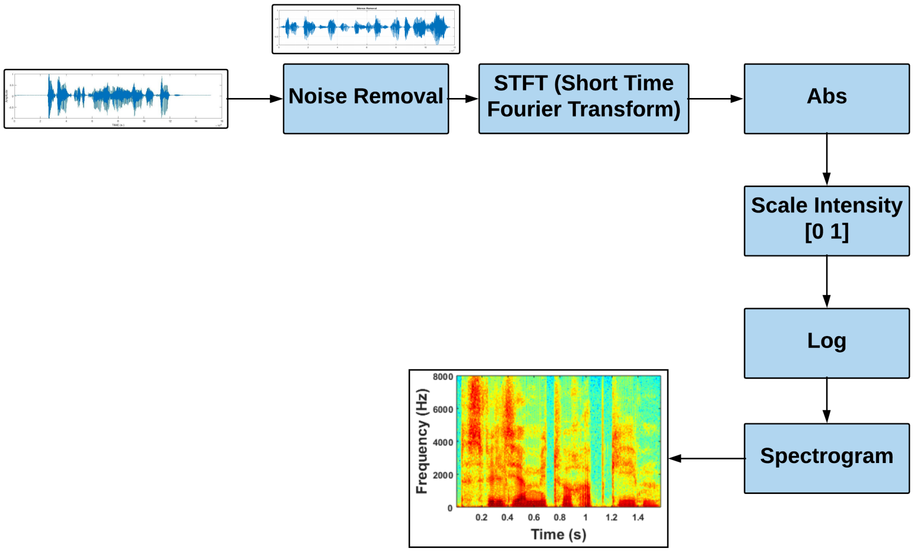


Figure 3.3: Block diagram on how to compute Spectrogram

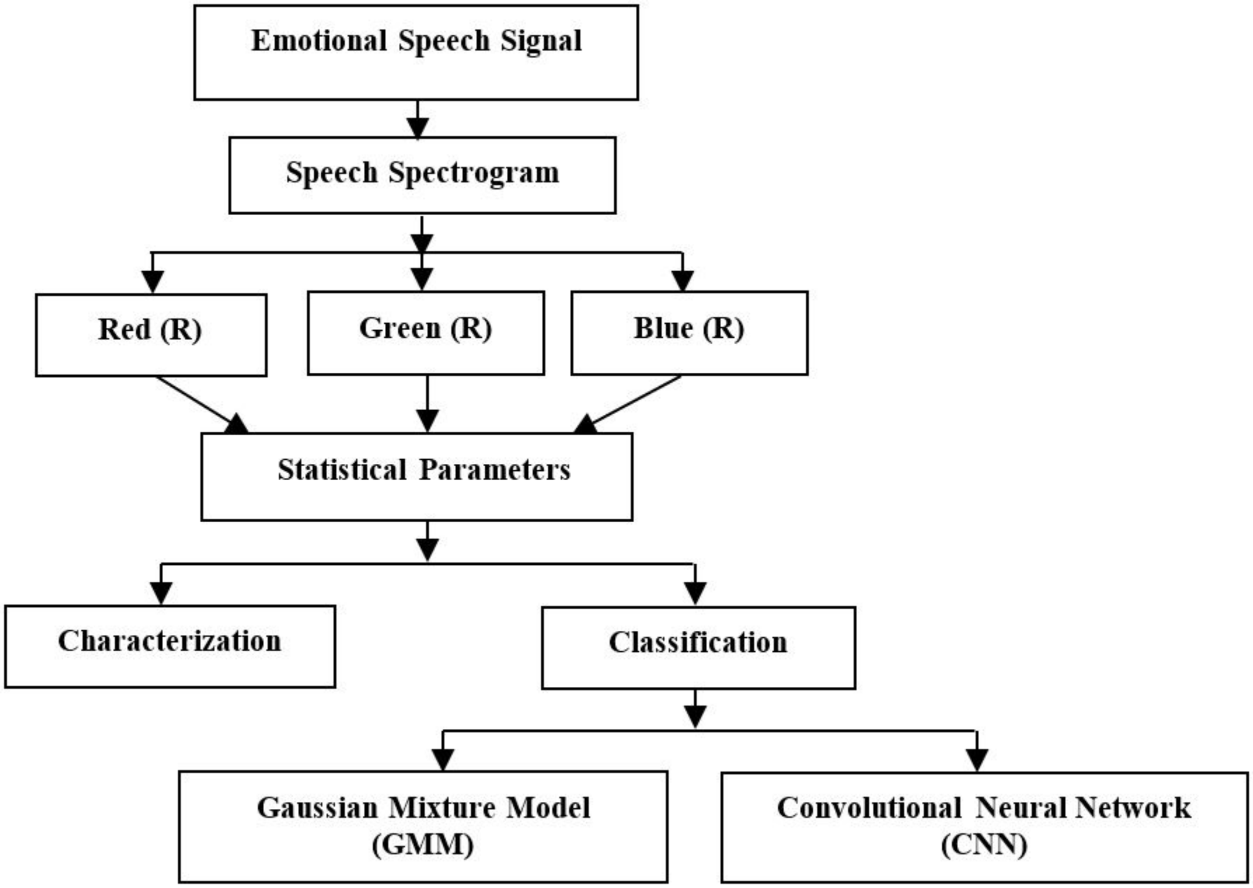


Figure 3.4: Block diagram of the proposed speech emotion recognition system

From each spectrogram, the corresponding R, G, B components are extracted. Statistical analysis related to the mean (average), variance, maximum (max), and minimum (min) values of individual R (Red), G (Green) and B (Blue) spectrogram array of an utterance are then estimated. Individual statistical parameters corresponding to the R, G, and B segments of the spectrogram corresponding to all the utterances of an intended emotion are stored. Individual statistical parameters related to the R, G, and B parameters collected from all the chosen emotions have been fed separately to the NN classifier for simulation and desired classification accuracy.

3.3.1 Estimation of Statistical Parameters from Speech Spectrogram

From each individual R, G, and B images of a spectrogram image matrix corresponding to each emotional utterance, different statistical parameters like mean, variance, maximum and minimum values are computed.

The computation of the mean or average of a signal helps in its spatial filtering. Correspondingly, it removes the noise component resides in a signal. In this context of image processing, the arithmetic

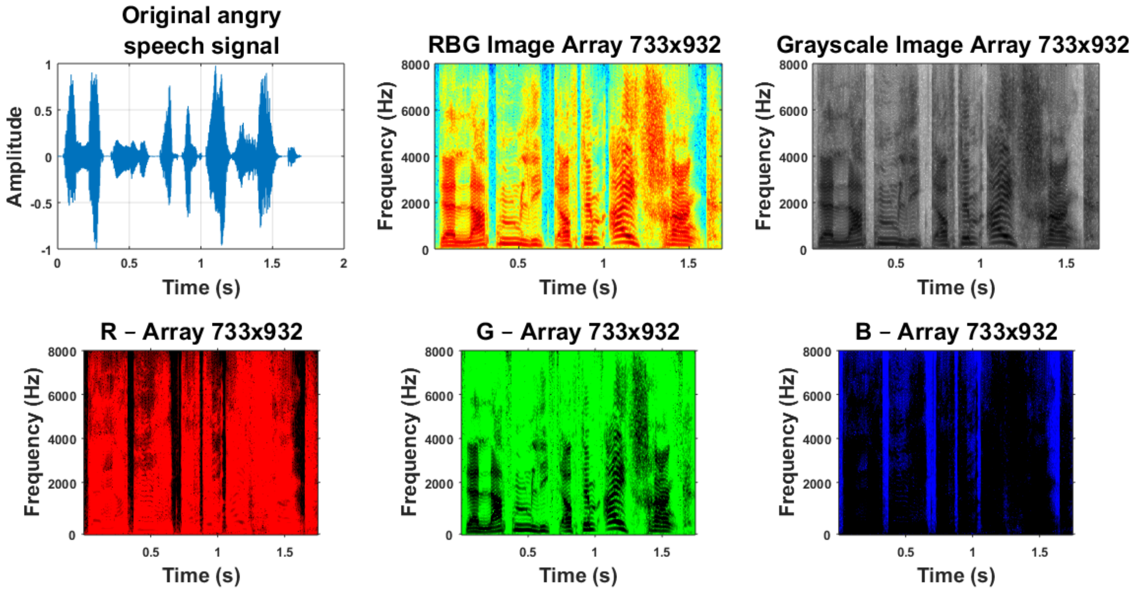


Figure 3.5: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally angry speech.

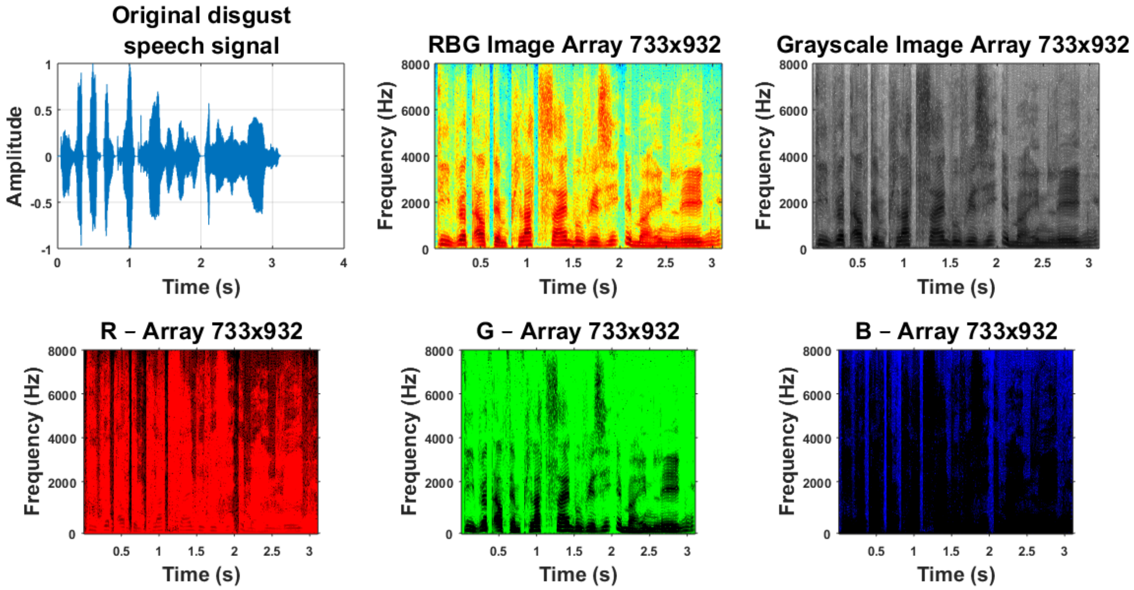


Figure 3.6: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally disgust speech.

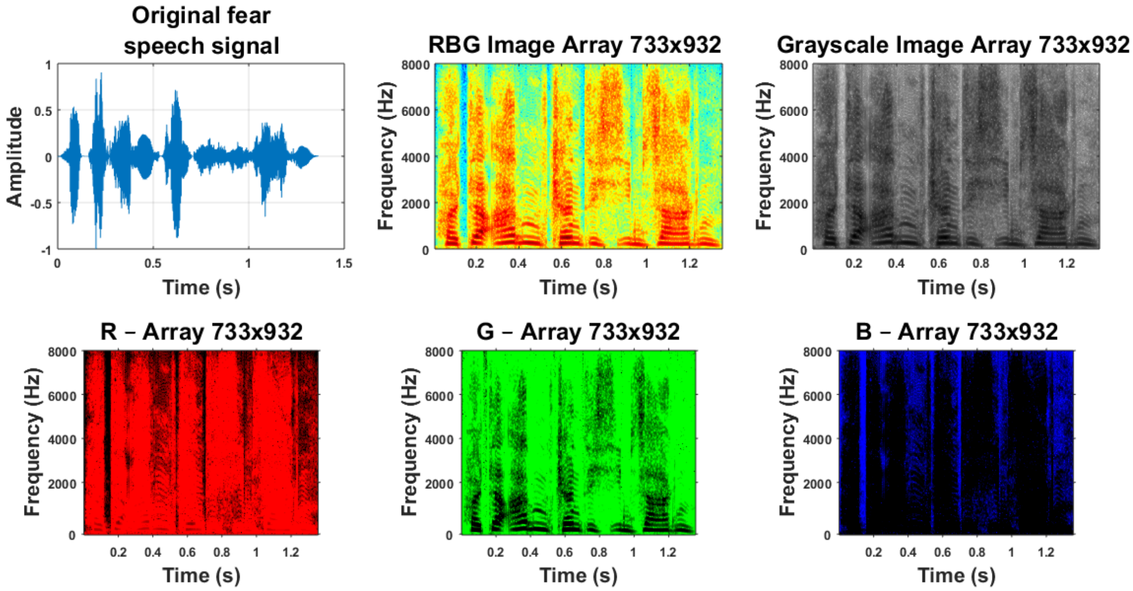


Figure 3.7: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally fear speech.

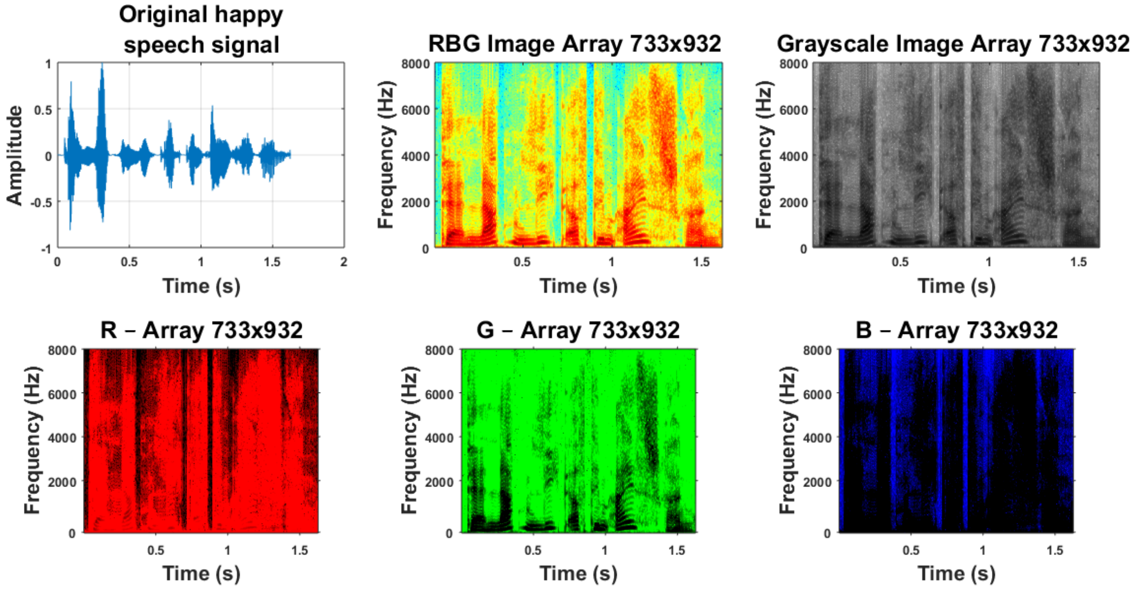


Figure 3.8: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally happy speech.

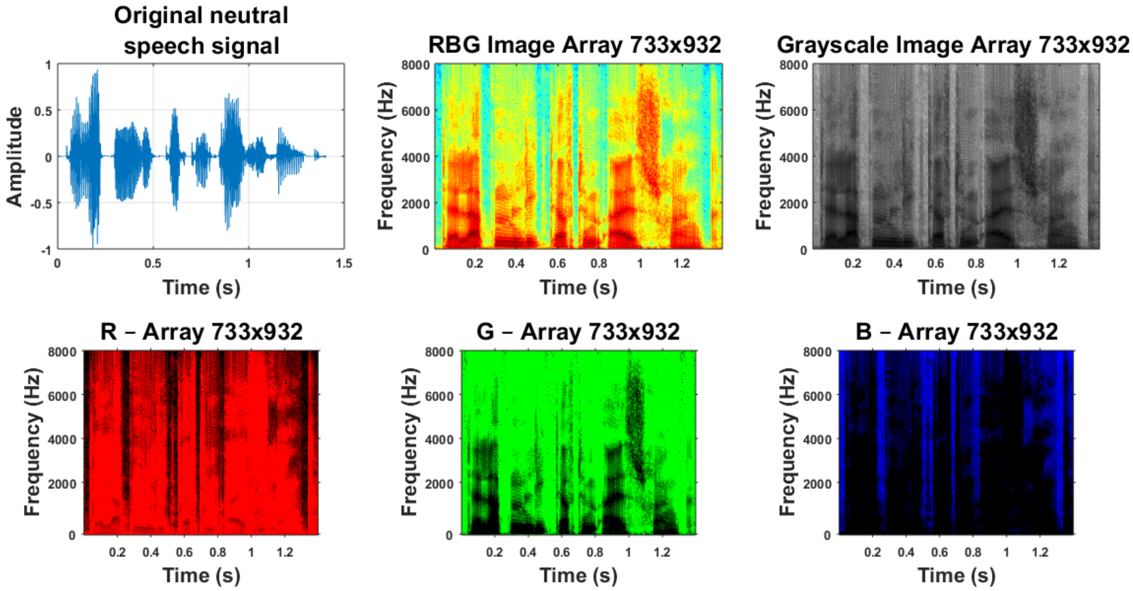


Figure 3.9: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally neutral speech.

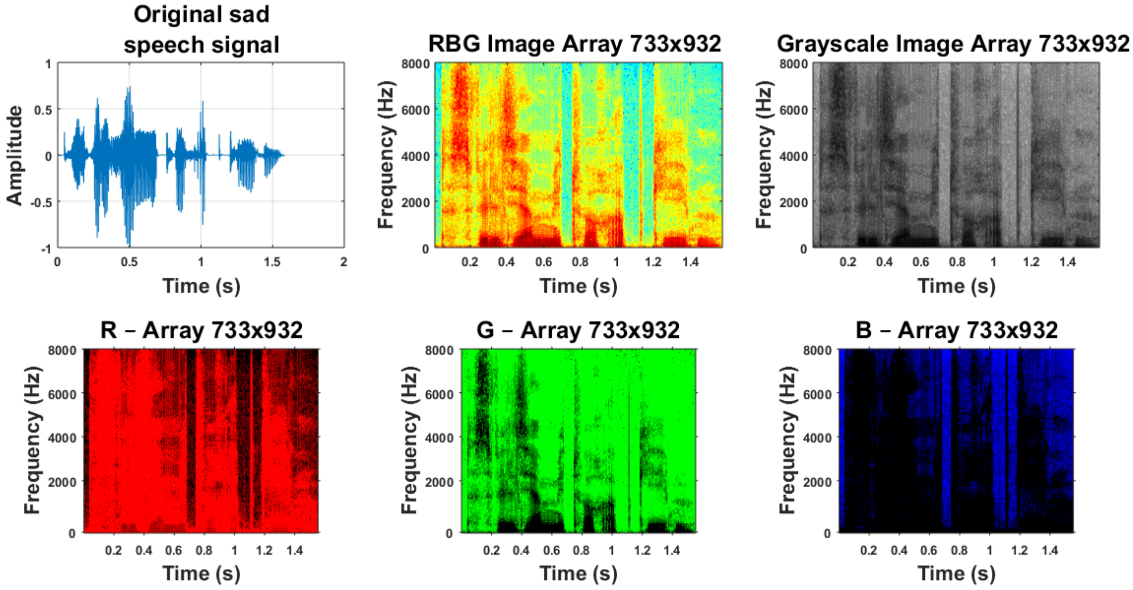


Figure 3.10: An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally sad speech.

means or averaging lter acts on a sliding window a x b. The filter computes the mean with respect to the pixel values available in a window. In the process, it replaces the pixel center in the end image with the designated outcome. Mathematically it is expressed for R the component of a spectrogram image as

$$f_R(m,n) = \frac{1}{ab}\sum_{x,y} i_R(x,y) \tag{5}$$

The term i_R in above the equation denotes the noisy red image whereas, $f(m,n)$ signifies the restored image. The window Z having the dimension ab, is used as the operating zone. The variables x and y denote the row and the column coordinates respectively.

The variance indicates the range in which a set of analyzed numbers spread out. It represents one of the descriptors corresponding to a certain probability distribution. As such, it indicates the layout of representing numbers and their distance from the expected or mean value. Thus, the computation of the variance signifies a systematic way that distinguishes different probability distributions. Among the many approaches, the calculation of variances based on moments is computationally simple and mathematically more relevant. Hence, it is preferable in the current scenario. The variance for the R component of a spectrogram image is given by

$$f_R(m,n) = \frac{1}{ab-1}(i_R(x,y) - \frac{1}{ab-1}\sum_{x,y} i_R(x,y))^2 \tag{6}$$

In case of processing an image, the variance filter computes the edge position.

The maximum (max) and the minimum (min) values are also known as a rank-statistics filter as they operate on a ranked set of pixel values. The min filter takes place the position of the reference pixels with a lower value. It is thus designated as the zeroth percentile filter. This work utilizes the min filter to eliminate the salt noise present in the extracted image as it is fast and considers the minimum window value when these values are systematically arranged. Mathematically, it can be expressed as

$$f_R(m,n) = min\{i_R(x,y)|(x,y) \in Z\} \tag{7}$$

The maximum or max lter or the 100_{th} percentile lter replaces the reference pixel value with its highest value in a window. It is expressed as

$$f_R(m,n) = max\{i_R(x,y)|(x,y) \in Z\} \tag{8}$$

Similar to the min filter, the max filter eliminates the pepper noise. The formulation of the problem is as follows: Consider the mean values of the R, G, B spectrogram image of an original spectrogram corresponding to an emotional speech signal represents as R_{Au}, G_{Au}, B_{Au} where A denotes the mean. For an emotional state having $u = 1, 2, , U$ number of utterances, the mean values of the R, G, B matrix can be represented as vectors of R_A, G_A, B_A where $[R_A = R_{A1}, R_{A2}, ..., R_{AU}], [B_A = B_{A1}, B_{A2}, ..., B_{AU}]$ and $[G_A = G_{A1}, G_{A2}, ..., G_{AU}]$ respectively. Similarly, the variance, maximum and minimum values of R, G, and B components are formed for each emotional state. The individual classification model is formed each for mean, variance, maximum and minimum values of R, G, B components to check the recognition accuracy.

3.4 Classifier

3.4.1 Convolutional Neural Networks

As a classifier, we have implemented CNN (convolutional neural network) to classify emotions in a robust manner. R G B spectrogram being a higher dimension matrix size (227 X 227) the input data becomes quite large and thus is best suited for CNN. CNN works quite well when the input data is large, but that also increase our system requirements.

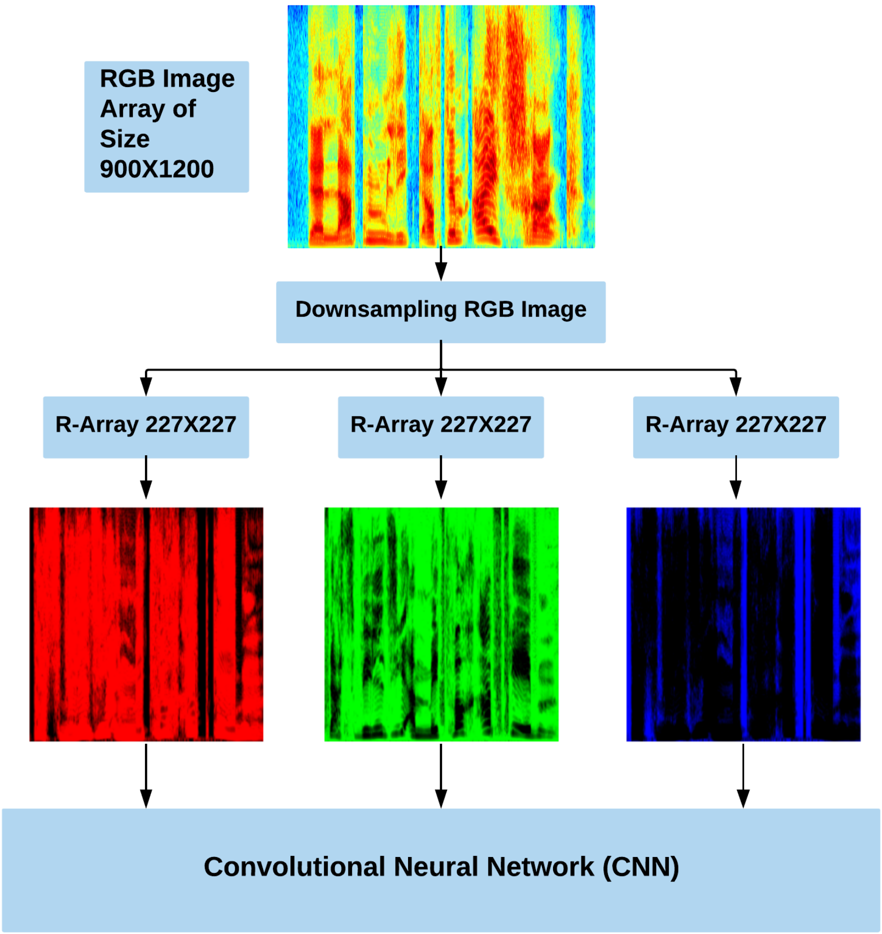


Figure 3.11: Conversion of the spectral magnitude arrays of speech into RGB image arrays provided as inputs into CNN

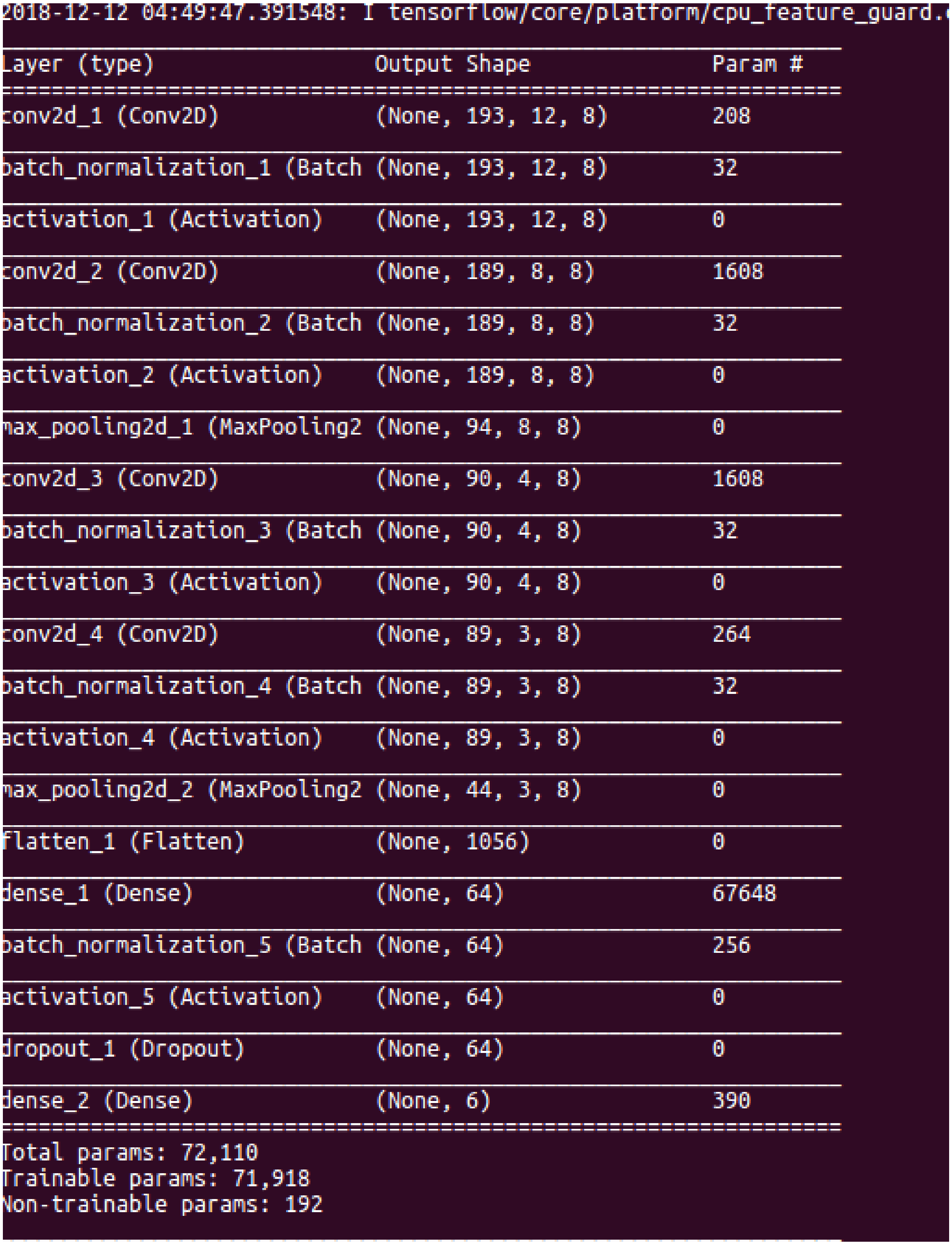
The proposed 2-d CNN model consists of five convolutional layers, four fully connected layers and an activation layer. The model is built in Keras and run on top of TensorFlow. Each convolution layer creates a convolution kernel (filter) that is convolved with the layer input to produce a tensor of outputs. The hyper parameters used are: **filters**(input an integer which defines the number of output filters in a convolution layer), **kernel_size**(a list of 2 integers specifying the height and width of 2-d convolution window), **activation**(activation to be used), **use_bias**(Boolean, whether the layer uses a bias vector or not) and many others. Other core layers in the model are: Batch Normalization Layer, Dense Layer, flatten Layer, Dropout layer. Batch Normalization is a method to improve performance and stability of Neural Network by initializing input to zero mean and unit variance. Dense Layer is responsible for densely connected CNN layer and implements the activation(function) of input. Flatten layer is used to flatten the input without affecting the batch_size of tensors. Dropout layer applies dropout to input by randomly setting a fraction of inputs to 0 at each update during the training time, helps in prevent overfitting of the model.

3.4.2 Multi-Layer Perceptron (MLP)

The conventional Multilayer Perceptron (MLP) classifier has been used for the proposed classification of speech emotions using the R, G, and B statistical spectrogram features. It is an Automatic Neural Network (ANN)-based classifier having basically three layers and uses the back-propagation algorithms for learning. These are the input, hidden and the output layer. Updating of weight takes place during the propagation of the extracted features through the hidden layer to output layer during the forward pass. Error calculation is done between the training and testing features during a backward pass based on the back-propagation algorithm. The correction corresponding to the synaptic weight is achieved by means of the negative gradient of the cost function (β) with respect to the designated synaptic weight. The relation is expressed as,

$$\Delta W = -\delta \frac{\partial \beta}{\partial w}$$

where δ indicates the learning rate of the back-propagation algorithm



```
2018-12-12 04:49:47.391548: I tensorflow/core/platform/cpu_feature_guard.c
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv2d_1 (Conv2D)            (None, 193, 12, 8)        208       
_________________________________________________________________
batch_normalization_1 (Batch (None, 193, 12, 8)        32        
_________________________________________________________________
activation_1 (Activation)    (None, 193, 12, 8)        0         
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 189, 8, 8)         1608      
_________________________________________________________________
batch_normalization_2 (Batch (None, 189, 8, 8)         32        
_________________________________________________________________
activation_2 (Activation)    (None, 189, 8, 8)         0         
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 94, 8, 8)          0         
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 90, 4, 8)          1608      
_________________________________________________________________
batch_normalization_3 (Batch (None, 90, 4, 8)          32        
_________________________________________________________________
activation_3 (Activation)    (None, 90, 4, 8)          0         
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 89, 3, 8)          264       
_________________________________________________________________
batch_normalization_4 (Batch (None, 89, 3, 8)          32        
_________________________________________________________________
activation_4 (Activation)    (None, 89, 3, 8)          0         
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 44, 3, 8)          0         
_________________________________________________________________
flatten_1 (Flatten)          (None, 1056)              0         
_________________________________________________________________
dense_1 (Dense)              (None, 64)                67648     
_________________________________________________________________
batch_normalization_5 (Batch (None, 64)                256       
_________________________________________________________________
activation_5 (Activation)    (None, 64)                0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0         
_________________________________________________________________
dense_2 (Dense)              (None, 6)                 390       
=================================================================
Total params: 72,110
Trainable params: 71,918
Non-trainable params: 192
_________________________________________________________________
```

Figure 3.12: Proposed architecture of SER using CNN

Chapter *4*

Results and Outcomes: Emotion Recognition using Speech

4.1 SAVEE Database

The following results were obtained using MFCC and S-transform as feature extraction technique and GMM as a classifier on SAVEE database.

A total of 6 emotions Angry, Disgust, Fear, Happy, Sad, Surprise were taken for classification. We evaluated a total of 60 utterances from each emotional state. The database was split into training and testing with 86.6% in training data and 13.7% testing data. All the testing data from each emotion was fed into GMM and an input model was prepared for each emotion. GMM further computed similarity measure between the prepared mode and the testing data and the following results were observed.

Feature dimension used for testing	**MFCC**	**S-Transform**
Total testing features	9600	96000
Features classified as Angry	986	8454
Features classified as Disgust	1490	14329
Features classified as Fear	1427	12623
Features classified as Happy	869	14588
Features classified as Sad	2693	28071
Features classified as Surprise	2135	17935
Average accuracy%	83.0417%	85.4104%

Table 4.1: Accuracy Table for SAVEE database

4.2 Emo-DB database

The characterization of the desired speech emotions using individual R, G and B spectrogram statistics have been graphically represented in Figure 14 through Figure 21. The emotions are compared based on the spectrogram magnitudes and their appearance corresponding to the range of intensities of these RGB components. The statistical values such as the mean or average, maximum, minimum and

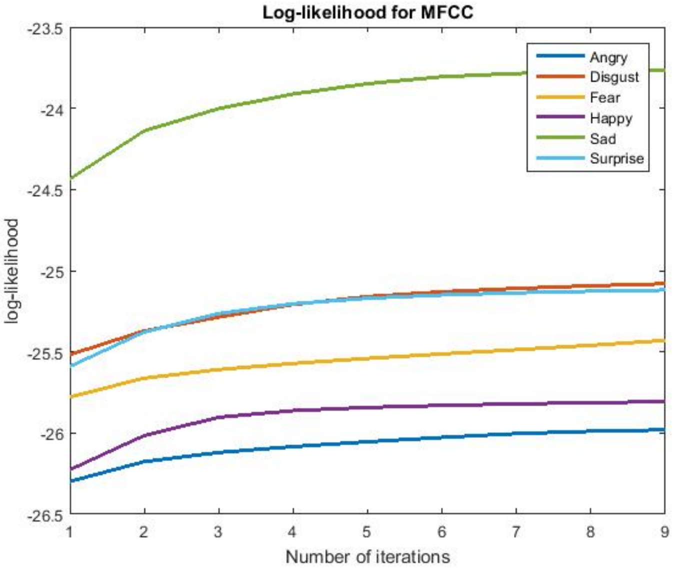


Figure 4.1: Comparison of goodness of fit model based on log likelihood ratio for Angry, Disgust, Fear, Happy, Sad, Surprise speech emotional states with MFCC features.

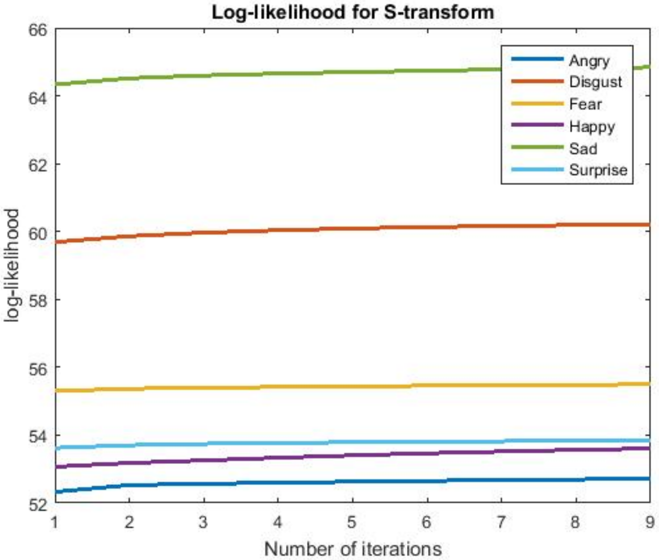


Figure 4.2: Comparison of goodness of fit model based on log likelihood ratio for Angry, Disgust, Fear, Happy, Sad, Surprise speech emotional states with S-Transform features.

variance of R-channel of forty-five chosen utterances of each emotional state have been computed. Similarly, these statistical parameters for the B-channel and G-channel for every utterance of an emotional state are estimated. These statistical values have been used to characterize the chosen emotional states as shown in the following Figures.

Fig. 4.3 shows the Maximum Green spectrogram magnitude across all the Emotions. The maximum values of sad lies below the maximum values of the neutral state. However, the maximum spectrogram magnitude values of happy and angry are higher than either the sad or neutral states. The emotional states such as anger or happiness have a higher arousal level with higher energy components. Thus, the G-spectrogram magnitudes lie opposite to the low arousal sad state with the neutral state act as a boundary between them.

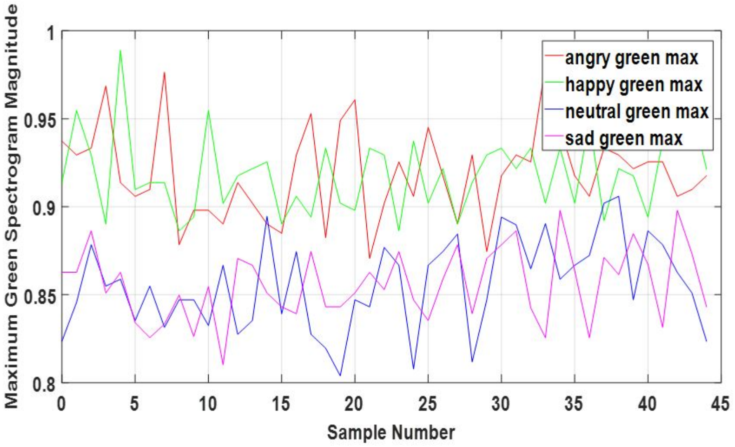


Figure 4.3: Maximum Green Spectrogram Magnitude across the Emotions

Fig. 4.4 provides the graphical representation indicating the variation of B-spectrogram components across the chosen emotional states. In this case, the angry state has shown a higher maximum spectrogram magnitude followed by the happy state as earlier. The low arousal sad state seems to possess the lowest B-spectrogram magnitude as observed in this Figure.

An attempt is made to characterize the chosen emotional states using the mean spectrogram magnitude of G, and B-components. As shown in Fig. 4.5 through Fig. 4.7, the mean values of the sad state lies opposite to either happy or angry state. Thus, the representation of emotions using RGB images can demarcate high arousal states from low arousal states. The mean values of sad state are lower to the high arousal emotional states with the mean values of the neutral state remains in between. This result is found to be true for each individual G and B mean spectrogram components as observed in these Figures.

Similar results have been manifested when the variance of the spectrogram magnitude of the R, G and B segments of all the utterances of each emotion are computed and compared. This has been shown in Fig. 4.8 through Fig. 4.10.

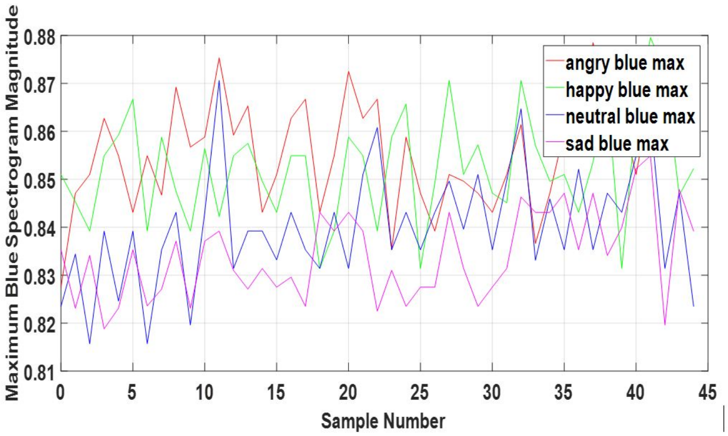


Figure 4.4: Maximum Blue Spectrogram Magnitude across the Emotions

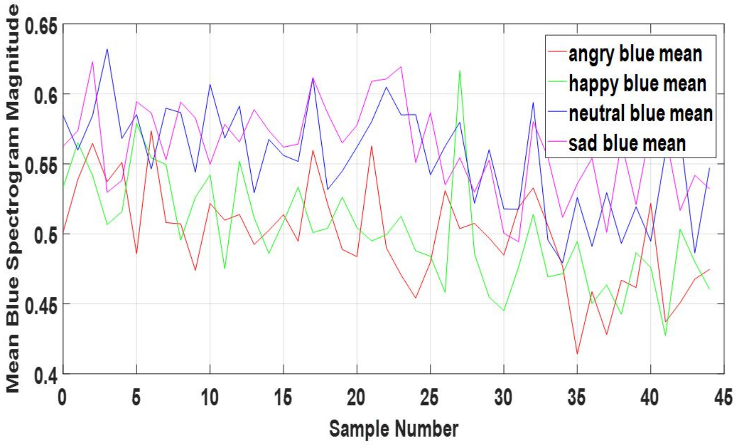


Figure 4.5: Mean Red Spectrogram Magnitude across the Emotions

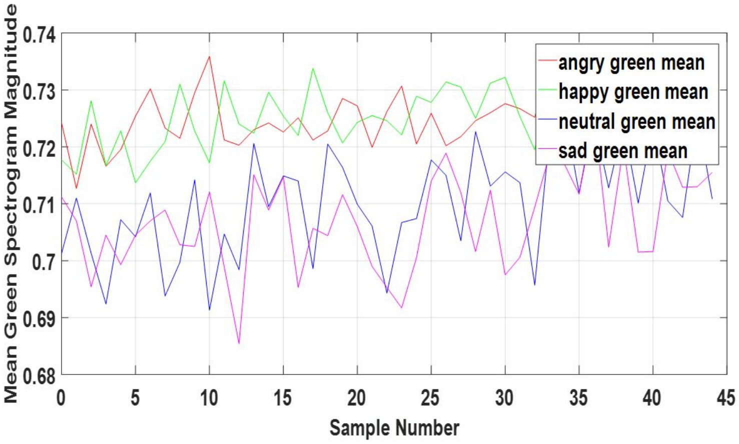


Figure 4.6: Mean Green Spectrogram Magnitude across the Emotions

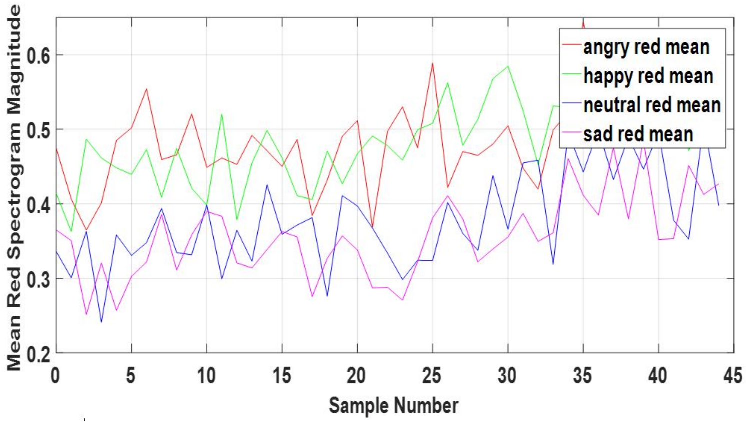


Figure 4.7: Mean Blue Spectrogram Magnitude across the Emotions

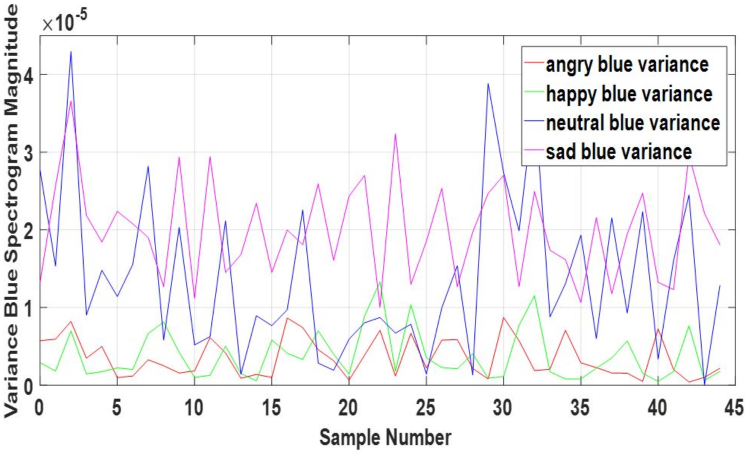


Figure 4.8: Variance Red Spectrogram Magnitude across the Emotions

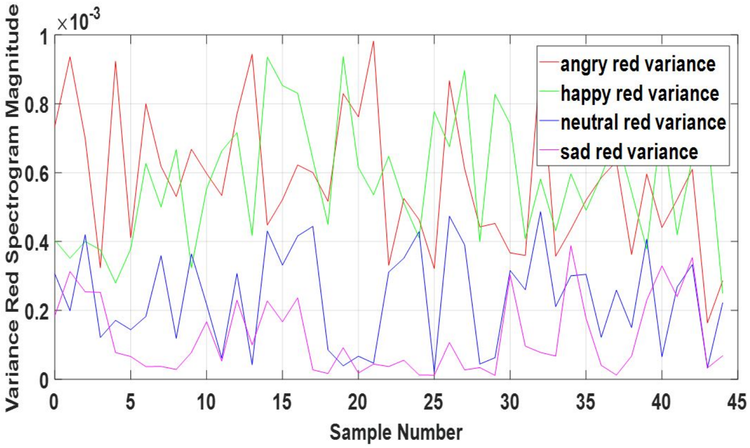


Figure 4.9: Variance Green Spectrogram Magnitude across the Emotions

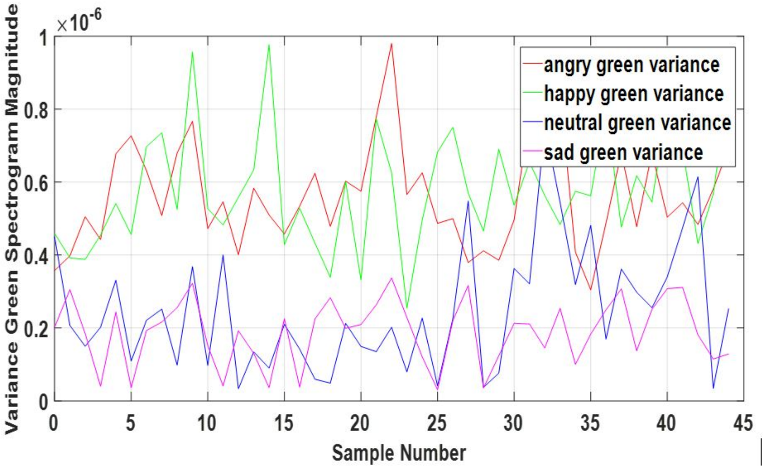


Figure 4.10: Variance Blue Spectrogram Magnitude across the Emotions

Features	Training Accuracy	Validation Accuracy	Testing Accuracy	Overall Accuracy
Spectrogram Mean	52.4%	44.4%	59.3%	57.2%
Spectrogram Variance	69.8%	48.1%	66.7%	68.1%
Spectrogram Maximum	56.3%	44.4%	40.7%	54.2%
Spectrogram Minimum	68.3%	63.0%	63.0%	67.7%

Table 4.2: MLP Average Accuracy with different data division using R, G, B Spectrogram Statistics

A comparison of average recognition accuracy across all the chosen emotional states has been tabulated in Table I using the MLPNN classifier. The R, G, B Spectrogram statistical features have been used for this purpose.

As shown in Table 4.2, the spectrogram variance of RGB parameters has provided an overall average accuracy of 68.1%. This is the highest recognition accuracy as compared followed by the Spectrogram minimum values of RGB components. The spectrogram maximum values have provided the least recognition accuracy in our case.

The total input feature set has been divided into training, validation and testing ratio of 70%/ 15%/ 15%, 60%/ 20%/ 20% and 50%/ 25%/ 25% for simulation of the chosen classifier. However, a ratio of 70%/ 15%/ 15% have provided the maximum accuracy with the chosen feature set of EMO-DB database and retained for our result analysis. The training, testing and validation accuracy is also shown in Table II. As shown in this Table, the training accuracy has been highest for all the chosen feature sets as compared to either the validation or testing accuracy.

4.2.1 Classification results using GMM classifier

Table 4.3 gives the accuracy output for Emo-DB database on GMM classifier suing RGB features.

Feature dimension used for testing	**Spectrogram**
Total testing features	16186
Features classified as Angry	4570
Features classified as Happy	2996
Features classified as Neutral	4846
Features classified as Sad	3774
Average accuracy%	86.811%

Table 4.3: Accuracy Table RGB feature using GMM classifier

4.2.2 Classification results using the CNN classifier

A total of 454 spectrograms of size 227 X 227 were generated in MATLAB. Each spectrogram was further split into 3 channels - Red, Green, and Blue each of size 227*227. So, feature matrix for one speech sample after combining all the red, green and blue channel becomes 227 X 681. The dataset was split into training and testing of which 80% was used for training and 20% was used for testing. We ran the training process for 50 epochs with learning rate set as 0.1. In order to implement the architecture as shown in the figure, we have used our college GPU servers. The training was performed on single NVIDIA Quadro K4200 with 4GB as graphics memory. The complete training process took about 22 seconds in 10 epochs. While training a loss of 0.35 was observed. Table IV gives the extensive comparison between the classification accuracy as observed between two classifiers i.e. GMM and CNN on the same database.

Feature Extraction Technique	**GMM**	**CNN**
MFCC	76.08%	85.33%
RGB	86.811%	88.46%
MFCC+RGB	85.13%	91.54%

Table 4.4: Accuracy comparison on Emo-DB databse using different feature and classifiers

Chapter *5*

Conclusion and Future Work

After all the experimentations of different features extraction techniques like MFCC, S-Transform, and spectrogram and classifiers like GMM, MLPNN, CNN over two standard datasets we made few remarkable conclusions.

S-transform proved to be more dynamic feature extraction technique over MFCC on SAVEE dataset. The fundamental reason behind this is that S-transform is a time-frequency dependent feature. We achieved a total accuracy of 85.4% on the Surrey Audio-Visual Expressed Emotion (SAVEE) Database whereas MFCC gave a recognition accuracy of 83% for the same number of emotions. We also studied the log-likelihood plot for both MFCC and S-transform and conclude that sad emotion is the best fit model in both the cases.

Moving on to the feature extraction technique using speech spectrogram images on Emo-DB database we saw that it gave higher accuracy than MFCC irrespective of the classifiers used. Using spectrograms images in GMM gave classification accuracy of 86.81%. The reason being speech spectrograms using different frequency scales led to an interesting analysis showing how different emotions are coded into the amplitude-frequency characteristics of speech. RGB feature, when fed into CNN, gave 88.46% as average classification accuracy, thus making CNN more reliable over GMM.

As observed, it is possible to compare and demarcate different emotional arousal states with these RGB components. Further, the classification result shows the RGB spectrogram magnitudes are both reliable and discriminative in classifying speech emotions. Use of these features with other popular and effective classifiers may give better results. A further combination of these RGB components with other prosodic or spectral features may provide new insight into the characterization of emotional contents in the voice communication.

Bibliography

[1] R. G. Stockwell, L. Mansinha, and R. P. Lowe, Localization of the Complex Spectrum: The S-Transform, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 44, NO. 4, APRIL 1996.

[2] S. G. Koolagudi, Y. S.Murthy, and S. P.& Bhaskar, Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition, International Journal of Speech Technology, Vol. 21, No. 1, pp. 167183, March 2018.

[3] H. K. Palo, M. Chandra, and M. N. Mohanty, Recognition of Human Speech Emotion Using Variants of Mel-Frequency Cepstral Coefficients, In Advances in Systems, Control and Automation, Lecture Notes in Electrical Engineering, 442, pp. 491-498, Springer Singapore, 2018.

[4] S.Zhang, S. Zhang, T. Huang, and W. Gao, Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, IEEE Transactions on Multimedia, Vol. 20, No. 6, pp.1576-1590, June 2018.

[5] H. K. Palo, M. Chandra, and M. N. Mohanty, Emotion recognition using MLP and GMM for Oriya language, International Journal of Computational Vision and Robotics, vol.7, no. 4, pp. 426-442, 2017.

[6] W. Lim, D. Jang, and T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, IEEE Signal and information processing association annual summit and conference (APSIPA), Asia-Pacific, Jeju, December 13, pp. 1-4, .2016.

[7] S. Haykin, Neural Networks: A comprehensive foundation, 2nd Ed., Pearson Education, Delhi, India, 2006

[8] V. H. Do, X. Xiao, and E.S. Chng, Comparison and combination of multilayer perceptrons and deep belief networks in hybrid automatic speech recognition systems, Proceedings of Asia-Pacific Signal (APSIPA ASC), Xian, China, 2011

[9] M. N Stolar, M. Lech, R. S Bolia, and M. Skinner, Real-time speech emotion recognition using RGB image classification and transfer learning, 11th IEEE International Conference on Signal Processing and Communication Systems (ICSPCS), Dec 13, 2017.

[10] L. He, M. Lech, N. C. Maddage and N. Allen, Stress Detection Using Speech Spectrograms and Sigma-pi Neuron Units, 2009 Fifth International Conference on Natural Computation, Tianjin, 2009, pp. 260-264.

[11] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, Speech emotion recognition using CNN, Proceedings of the 22nd ACM international conference on Multimedia, Nov 3, pp. 801-804, 2014.

[12] H. K. Palo, and M. N. Mohanty, Modified-VQ Features for Speech Emotion Recognition, J. Applied Sci., vol. 16, no. 9, pp. 406-418, August 15, 2016.

[13] R. Ram, H. K. Palo and M. N. Mohanty, Recognition of fear from speech using adaptive algorithm with MLP classifier, 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, pp. 1-5, 2016.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F Sendlmeier, and B. Weiss, A database of German emotional speech, Ninth European Conference on Speech Communication and Technology, 2005.

[15] H. K. Palo, M. Chandra, and M. N. Mohanty, Emotion recognition using MLP and GMM for Oriya language, International Journal of Computational Vision and Robotics, vol.7, no. 4, pp. 426-442, 2017.

[16] Pao, T. L., Chen, Y. T., Yeh, J. H., Liao, W. Y. (2005). Combining acoustic features for improved emotion recognition in Mandarin speech. In J. Tao, T. Tan, & R. Picard (Eds.), ACII. LNCS (pp. 279285). Berlin: Springer.

[17] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in Sixth International Conference on Multimodal Interfaces ICMI 2004. State College, PA: ACM Press, October 2004, pp. 205-211.