**Emotion Recognition from Speech and Beyond**

by

Team Member 1: Ranjit Patil

Team Member 2: Paritosh Mangrulkar

Team Member 3: Vijayalakshmi Padmanaban

Team Member 4: Balaji Raghavan

Team Member 5: Tamilselvan Sivanatham

Report

Submitted in partial fulfillment of the requirements for the

DLFA Program

Centre for Continuing Education

Indian Institute of Science

Bangalore – 560 012 India

# Abstract

This paper describes a method for Speech Emotion Recognition (or, classification) using Deep Learning with CNN Network, CNN transformer, CNN Attention, TimeDistributed convolution, batch normalization pattern, LSTM and fully connected layers. In this work, we detailed the architecture, which extracts Mel-frequency cepstral coefficients, chroma-gram, Log-Mel scale spectrogram, Tonnetz representation, and spectral contrast features from sound files and uses them as inputs for the Neural Network for the identification of emotions using samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets. We utilize an incremental method for modifying our initial model in order to improve classification accuracy. Based on experimental results, our best-performing model outperforms existing frameworks for RAVDESS.

**Acknowledgments**

Project Mentor – Mr. Karthik Raja

Professor Assistant – Mr. Soumya Dutta

Professor – Dr. Sriram Ganapathy

**Table of Contents**

[Abstract 2](#_Toc127121806)

[Chapter I: Introduction 6](#_Toc127121807)

[Problem Statement 6](#_Toc127121808)

[Purpose of the Study 6](#_Toc127121809)

[Research Questions 6](#_Toc127121810)

[What is Speech Emotion Recognition? 6](#_Toc127121811)

[Properties of an audio signal: 7](#_Toc127121812)

[Definition of Terms 8](#_Toc127121813)

[Assumptions and Limitations of the Study 10](#_Toc127121814)

[Overview 11](#_Toc127121815)

[Chapter II: Related Work 12](#_Toc127121816)

[Introduction 12](#_Toc127121817)

[Emotion Classifier 12](#_Toc127121818)

[Feature Extraction with original audio 15](#_Toc127121819)

[Audio with Data Augmentation 23](#_Toc127121820)

[Feature Extraction with models 24](#_Toc127121821)

[Summary 54](#_Toc127121822)

[Chapter III: Method/Experiment 55](#_Toc127121823)

[Introduction 55](#_Toc127121824)

[Research Question(s) 55](#_Toc127121825)

[Data Preprocessing, Feature Engineering and Visualization 56](#_Toc127121826)

[Choice of Model 56](#_Toc127121827)

[Training the model, Performance of the Model and Metrics 57](#_Toc127121828)

[Summary 61](#_Toc127121829)

[Chapter IV: Results 63](#_Toc127121830)

[Introduction 63](#_Toc127121831)

[We present a new method combining DCNNs with DTPM for automatic affective feature learning. A DCNN is used to learn discriminative segment-level features from three channels of log Mel-spectrograms similar to the RGB image representation. 63](#_Toc127121832)

[DTPM is designed to aggregate the learned segment-level features into the global utterance-level feature representation for emotion recognition. Extensive experiments on four data sets show that our method can yield promising performance in comparison 63](#_Toc127121833)

[with the state-of-the-arts. In addition, we also find that with our generated DCNN input, DCNN models pre-trained on the large-scale ImageNet data could be leveraged in speech affective feature extraction. This makes DCNN’s training with a limited amount of annotated speech data easier. The success of thisworkwarranties further investigation on using deep learning in speech emotion recognition. 63](#_Toc127121834)

[Summary 63](#_Toc127121835)

[Chapter V: Summary, Conclusions, and Recommendations 65](#_Toc127121836)

[Introduction 65](#_Toc127121837)

[Summary of the Results 65](#_Toc127121838)

[Conclusions 65](#_Toc127121839)

[Recommendations 65](#_Toc127121840)

[References 66](#_Toc127121841)

# 

# Chapter I: Introduction

## Problem Statement

With the growing demand for conversational agents and personal assistants, automatic recognition of human emotion has emerged as a key task in enabling enhanced user experience. Human emotion recognition using multi-modal data of text, speech and video has substantial impact on various applications like smartphones, wearable devices, smart speakers, driver monitoring in automotives, mood analysis and mental health. This area of developing emotional intelligence would allow machines to be more human-like in the interactions.

## Purpose of the Study

To classify various emotions (calm, happy, sad, angry, fearful, surprise, and disgust) in Audio Files using deep learning. Essentially, it is a multiclass classification problem.

## What is Speech Emotion Recognition?

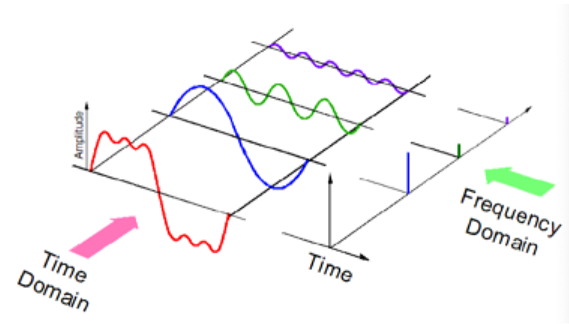
Speech Emotion Recognition (SER) is the task of recognizing the emotion from speech, irrespective of the semantics. Humans can efficiently perform this task as a natural part of speech communication, however, the ability to conduct it automatically using programmable devices is a field of active research.

Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers, pilots, and many other human-machine communication users. Adding emotions to machines forms an important aspect of making machines appear and act in a human-like manner.

With the growing demand for conversational agents and personal assistants, automatic recognition of human emotion has emerged as a key task in enabling enhanced user experience. Human emotion recognition using multi-modal data of text, speech and video has substantial impact on various applications like smartphones, wearable devices, smart speakers, driver monitoring in automotives, mood analysis and mental health. This area of developing emotional intelligence would allow machines to be more human-like in the interactions.

## Properties of an audio signal:

An audio signal is represented in the form of an audio signal having parameters such as frequency, bandwidth, decibel etc. A typical audio signal can be expressed as a function of Amplitude and Time.



## Definition of Terms

### Audio Signal

Representation of Sound.

1. High Level - Instruments, key, Chords, Melody, Rhythm, Tempo, Lyrics, Genre and Mood.
2. Mid Level - Pitch & Beat related descriptor such note, onsets, fluctuation pattern, MFCC
3. Low Level - Amplitude Envelope, Energy, Spectral Centroid, Spectral fluc, Zero crossing rate.

### Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of sound or other signals as they vary with time.

### MFCC

The Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice. MFCC coefficients represent the envelope of the time power spectrum of the speech signal. Frequency bands of this spectrum are spaced logarithmically according to the Mel scale.

### Mel Spectrogram

Standard Spectrogram in Mel scale (perceptual scale of pitches that listeners perceive to be equally spaced from one another. Low frequency contents differentiated more than the high frequency based on human audibility) gives poor representation of pitch, but captures timbre.

### Chroma

(potential for music audio) Full spectrum projected onto 12 bins (12 unique semitones/pitch). It captures harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation.

### Tonnetz

contain harmonic content of a given audio signal. An alternative representation of pitch and harmony can be obtained by the tonnetz function, which estimates tonal centroids - 6-dimensional basis representing the perfect fifth, minor third, and major third each as two-dimensional coordinates.

### power to db

specify the reference power (amplitude square) to dB ratio ex. max power to 0 dB

### Spectral Contrast

Spectral contrast is defined as the decibel difference between peaks and valleys in the spectrum. It considers the spectral peak, the spectral valley, and their difference in each frequency subband.

### Spectral Centroid

It indicates where the “centre of mass” for a sound is located and is calculated as the weighted mean of the frequencies present in the sound.

### Spectrogram vs Log scale spectrums

The difference between Spectrograms and log-scale spectrums, which are both being achieved by similar mathematical operations, is that while the first displays the frequencies and decibels over time, the latter shows the relation between the decibels and the frequencies.

## Assumptions and Limitations of the Study

### RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech 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. The conditions of the audio files are: 16bit, 48kHz .wav.

### File Naming Conventions

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

### File Name Identifiers

* Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
* Vocal channel (01 = speech, 02 = song).
* Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
* Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
* Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
* Repetition (01 = 1st repetition, 02 = 2nd repetition).
* Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

## Overview

(This has to be modified at the last) We extracted the IEEE paper which was less efficiency outcome. The paper suggested with Data process but could not extract properly train to get the result. Below, we showed the initial approach and the improved approach with greater efficient.

# Chapter II: Related Work

## Introduction

Generally, feature extraction and emotion classification are two key steps in speech emotion recognition. In this section, we first briefly review emotion classifiers and then focus on feature extraction since it is more relevant to our work.

## Emotion Classifier

For emotion classification various machine learning algorithms have been utilized to constitute a good classifier to distinguish the underlying emotion categories. Early emotion classifiers contain K-Nearest-Neighbor (KNN) and Artificial Neural Network (ANN). Then, a number of statistical pattern recognition approaches, such as Gaussian Mixture Model (GMM), Hidden Markov Models (HMM), and SVM, are widely adopted for speech emotion recognition. Recently, some advanced classifiers based on sparse representation

(Move this graph to Chap#3 – with another heading Data visualization RAVDESS)

have also been studied. Nevertheless, each classifier has its own advantages and disadvantages. To integrate the merits of different classifiers, ensembles of multiple classifiers have been investigated for speech emotion recognition.

* Map the labels for each .wav file as following

    neutral : 0

    calm : 1

   happy : 2

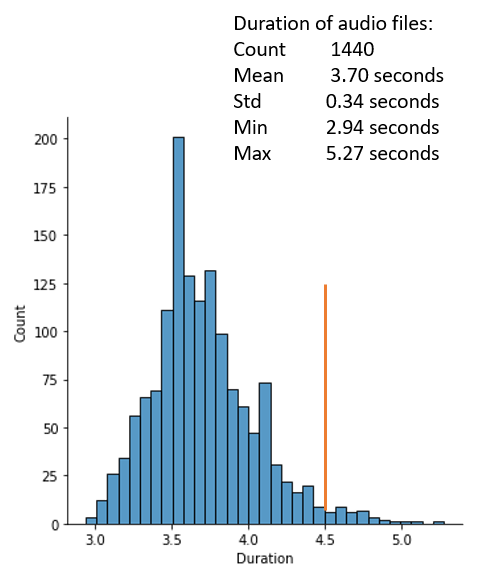
    sad : 3

angry : 4

    fearful : 5

    disgust : 6

    surprised : 7

* Maximum duration of 4.5 seconds is considered for calculating audio features e.g. MFCC, contrast etc.
* Based on duration of 4.5 seconds, dimension of features along time axis is 194
* Zero padding is done for features having less than 194 dimensions along time axis

## Feature Extraction with original audio

Affective speech features widely used for emotion recognition can be roughly divided into four categories: 1) acoustic features 2) language features, such as lexical information 3) context information, such as subject, gender, culture influences 4) hybrid features such

as the integration of two or three features above-mentioned. Acoustic features, as one of the most popular affective features, mainly contain prosody features, voice quality features, and spectral features. Pitch, loudness, and duration are commonly used as prosody features, since they express the stress and intonation patterns of spoken language. Voice quality features, as the characteristic auditory colouring of an individual voice, have been shown to be discriminative in expressing positive or negative emotions. The widely used voice quality features are the first three formants (F1, F2, F3), spectral energy distribution, harmonics-to-noise-ratio, pitch irregularity (jitter), amplitude irregularity (shimmer), and so on.

Chart, bar chart

Description automatically generated

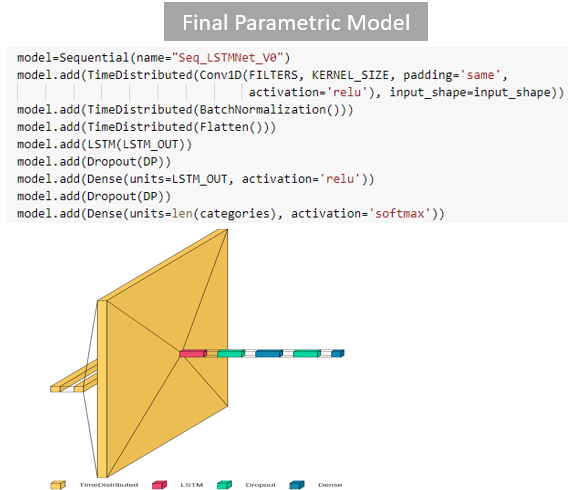
Chart, bar chart

Description automatically generated

* Implementation of model from reference paper got only 35% validation accuracy after 100 epochs against 71.61% claimed in paper after 700 epochs
* Various other networks e.g. CNNs including Conv1D and Conv2D, CNN Transformer, CNN Attention etc.
* Out of 14 different models, maximum validation accuracy of 66% achieved with two models using 1) Conv1D and, 2)Conv2D and Transformer Encoder

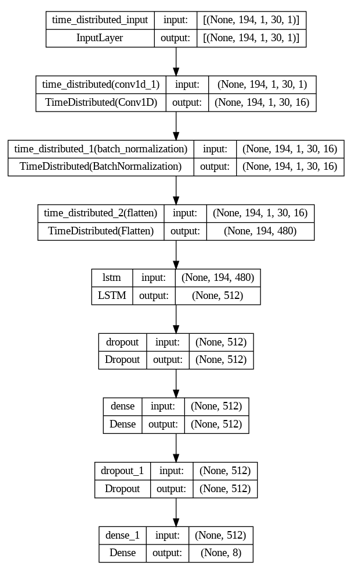
Combining prosody features and voice quality features shows better performance than using prosody features alone. In recent years, glottal features and voice source parameters have been used as more advanced voice quality features for speech emotion recognition. The third typical acoustic features are spectral features, computed from the short-term power spectrum of sound, such as Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC) and Melfrequency Cepstral Coefficients (MFCC). Among them, MFCC is the most popular spectral feature, since it is able to model the human auditory perception system. In recent years, modulation spectral features from an auditory-inspired long-term spectro-temporal representation, and weighted spectral features based on local Hu moments, have also been studied. In addition, the newly-developed Geneva minimalistic acoustic parameter set (GeMAPS), such as frequency, energy, spectral related features, has shown promising performance in speech emotion recognition.

Language features, which are computed based on the verbal contents of speech, are another important representation conveying emotion information. Note that, language features are usually combined with acoustic features for speech emotion recognition. In language features are extracted with the bag of n-gram and character n-gram approaches. Then the linguistic features are combined with acoustic features to predict dimensional emotions in a 3-D continuous space. By computing the weight of everyword, a four-dimensional emotion lexicon for four emotion classes, i.e., anger, joy, sadness and neutral, are obtained. Then, integrating these feature representations via early fusion and late fusion is employed for speech emotion recognition.



Table

Description automatically generated with medium confidence



* Parametric model of 8 layers was built using Keras Sequential function
* TimeDistributed function is used to apply same instance of Conv1D, BatchNormalization and Flatten functions to each of the 194 timesteps. **The same set of weights are used at each timestamp.**

Context information has also been investigated in recent literatures for emotion recognition. The authors present a context analysis of subject and text on speech emotion

recognition, and find that gender-based context information enhances recognition performance. The influences of cultural information on speech emotion recognition are explored. The authors claim that intra-cultural and multi-cultural emotion recognition paradigms give better performance than cross-cultural recognition.

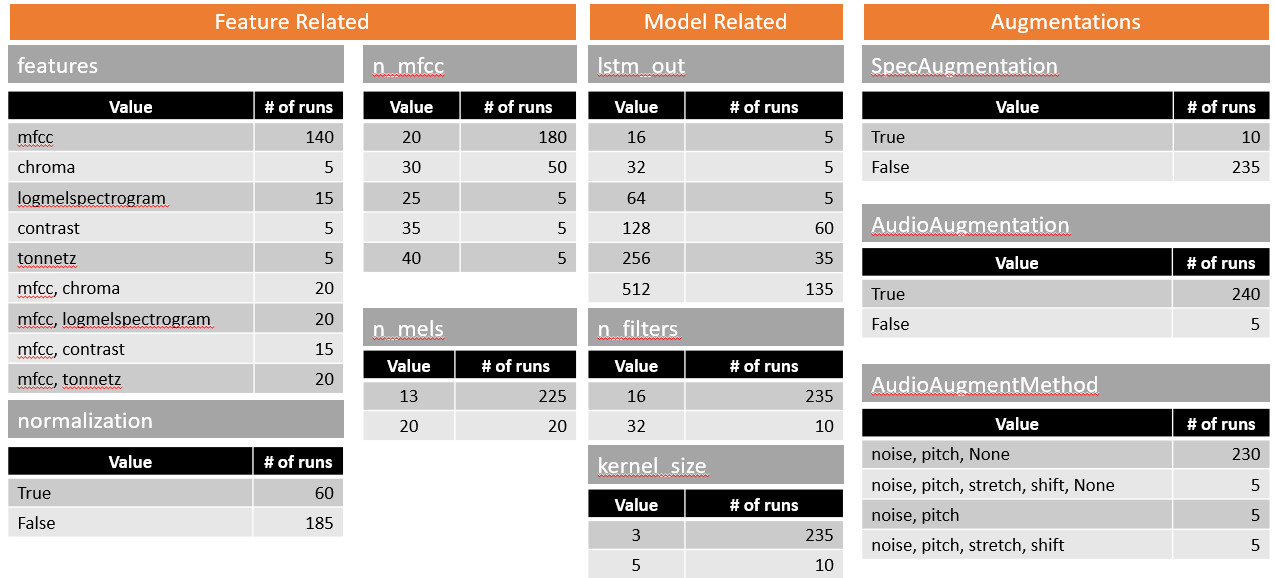
Note that, since these hand-designed features mentioned above are low-level, they may not be discriminative enough to identify the subjective emotions. To tackle this issue, it may be feasible to employ deep learning techniques to automatically learn high-level affective features for speech emotion recognition.

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Remark** |
| Dataset | Ravdess Speech |  |
| SampleRate | 22050 | Sampling rate |
| AudioDuration | 4.5 | Maximum duration of audio file |
| offset | 0.0 | Offset or start of audio file |
| HopLength | 512 | Hop Length |
| WinLength | 512 | Length of Window |
| Window | hann | Window function |
| n\_fft | 2048 | length of the FFT window |
| n\_chroma | 12 | Number of chroma bins |
| n\_bands | 6 | number of frequency bands |
| fmax | 1024 | highest frequency (in Hz) |

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Remark** |
| MaxInputLength | 194 |  |
| n\_folds | 5 | Number of folds |
| TestSize | 0.2 | Test dataset split size |
| BatchSize | 128 | Batch size |
| NumEpochs | 100 | Number of Epoch |
| InitialLearningRate | 0.001 | Initial learning rate |
| Dropout | 0.2 |  |
| LossFunction | categorical\_crossentropy |  |
| Optimizer | Adam |  |
| n\_classes | 8 | Number of categories |

Above mentioned parameters are kept constant during all the experiments

**Total of 245# DOE performed to select best hyperparameters**



|  |  |  |
| --- | --- | --- |
| Emotions | Spectral Centroids plot | Spectrogram |
| Neutral |  |  |
| Calm |  |  |
| Happy |  |  |
| Sad |  |  |
| Angry |  |  |
| Fearful |  |  |
| Disgust |  |  |
| Surprised |  |  |

## Summary

The results were improved with Data augmentation

# Chapter III: Method/Experiment

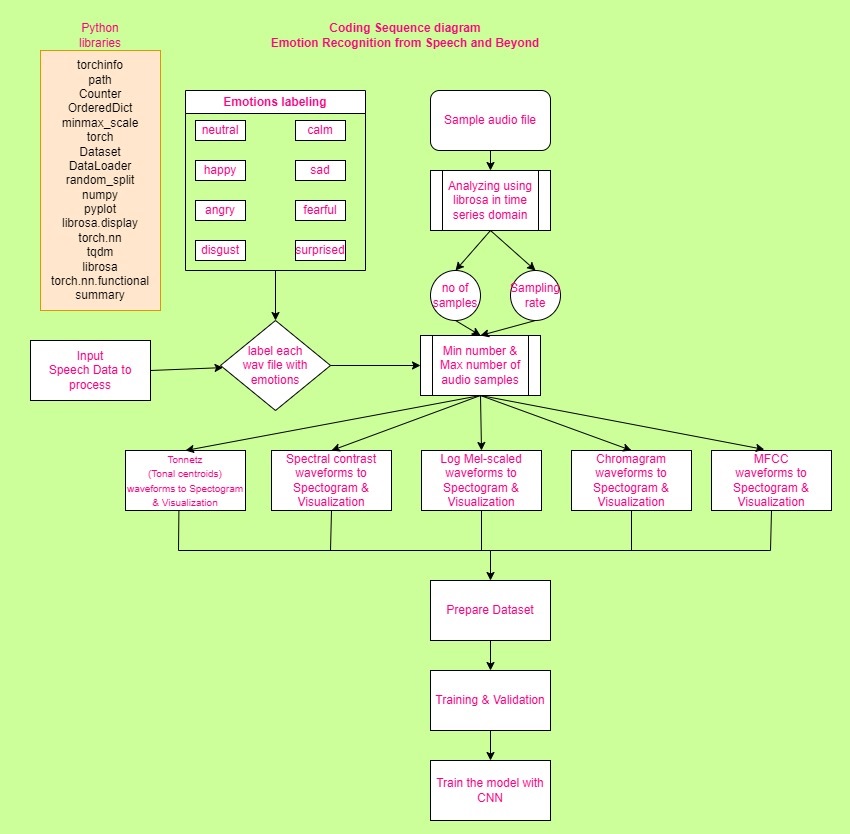
## Introduction

Begin with an introduction. Some suggestions include reiterating the statement of the problem and briefly discussing what this chapter will include. Sections to be addressed might include subject selection and description, Data Preprocessing, Feature Engineering and Visualization, Choice of Model, Training the Model, Performance of the Model and Metrics.

## Research Question(s)

State the research question or questions (if any). (Change the statement as earlier chapter)

## Data Preprocessing, Feature Engineering and Visualization



(Remove the last block – **Train the model**)

## EDA – Exploratory Data Analysis

(Move the graphs – Chapter 2 over here)

## Audio with Data Augmentation

The original audio is enhanced with Data Augmentation.

Some preparations for that

|  |  |  |
| --- | --- | --- |
| Category | Waveplot | Spectrogram |
| Original Audio |  |  |
| Noise addition |  |  |
| Time stretching |  |  |
| Time shifting |  |  |
| Pitch shifting |  |  |

## Feature Extraction with models

### MFCC (Need to be updated from latest pynb file – Stage 5 Feature Extraction MFCC)

|  |  |
| --- | --- |
| Emotions | MFCC Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

MFCC features represent better distinctions in between various emotions in terms of loudness, the spread of emotions, and frequency ranges. This is the reason why MFCC features are best suited for acoustic modeling and are used widely in the research work.

**Interprete MFCC charts** For emotions like Fear, Happy the sound level (dB) goes real low and observe much more black regions.

For emotions like Anger, Neutral the sound level (dB) doesn't go very low and some purple areas should be clearly visible.

Emotions like Fear or Happiness are impulsive and occurs for a shorter duration whereas Anger, Neutral and Sad emotions are much more spread out across the timeline.

Fear and Anger emotions have much more yellow regions close to the 0 dB reference point which shows that these emotions are much louder as compared to other emotions

(Slide no# 5)

### Normalized MFCC

|  |  |
| --- | --- |
| Emotions | Normalized MFCC Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Chromagram

|  |  |
| --- | --- |
| Emotions | Chromagram Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Mel scaled

|  |  |
| --- | --- |
| Emotions | Mel scaled Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Normalized Mel

|  |  |
| --- | --- |
| Emotions | Normalized Mel Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Log Mel

|  |  |
| --- | --- |
| Emotions | Log Mel scale Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Normalized Log-Mel

|  |  |
| --- | --- |
| Emotions | Normalized Log-Mel Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Spectral Contrast

|  |  |
| --- | --- |
| Emotions | Spectral Contrast Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Normalized Spectral Contrast

|  |  |
| --- | --- |
| Emotions | Normalized Spectral Contrast Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Tonnetz

|  |  |
| --- | --- |
| Emotions | Tonnetz Spectrogram |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### Normalized Tonnetz

|  |  |
| --- | --- |
| Emotions | Spectral Centroids plot |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

### RMS Energy

|  |  |
| --- | --- |
| Emotions | RMS Energy plot |
| Neutral |  |
| Calm |  |
| Happy |  |
| Sad |  |
| Angry |  |
| Fearful |  |
| Disgust |  |
| Surprised |  |

## Choice of Model

Discuss the model you chose and why you chose to go ahead with that model. It’s limitations. Any other model you used

1. BaseCNNNetwork
2. CNNTransformerNet
3. CNNNetwork
4. CNNNetV2
5. CNNAttentionNetV0
6. Conv1D
7. Conv2D, TransformerEncoder
8. SequentialLSTMNetV0

## Training the model, Performance of the Model and Metrics

(Take it from PPT – Slide#8 – Replace the table with graph)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr. No. | Model | Training Accuracy | Validation Accuracy | Training Loss | Validation Loss |
| 1 | BaseCNNNetwork | 0.89 | 0.35 | 1.70 | 1.99 |
| 2 | CNNTransformerNet | 0.38 | 0.14 | 1.91 | 2.09 |
| 3 | CNNNetwork | 0.55 | 0.36 | 1.81 | 1.92 |
| 4 | CNNAttentionNetV0 | 0.81 | 0.31 | 1.47 | 1.96 |
| 5 | CNNNetV2\_4 | 0.98 | 0.59 | 1.30 | 1.68 |
| 6 | Conv1D | 0.97 | 0.66 | 0.10 | 1.72 |
| 7 | Conv2D, TransformerEncoder | 0.98 | 0.66 | 0.053 | 1.52 |

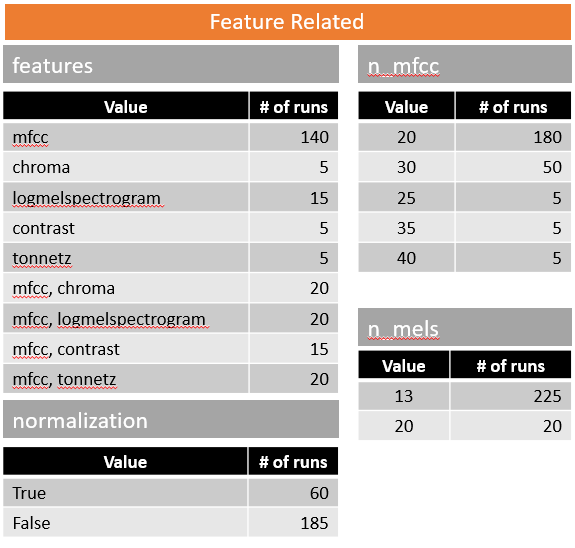
(Take it from PPT – Update slide#10 – add some text – fixed parameters)

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Remark** |
| Dataset | Ravdess Speech |  |
| SampleRate | 22050 | Sampling rate |
| AudioDuration | 4.5 | Maximum duration of audio file |
| offset | 0.0 | Offset or start of audio file |
| HopLength | 512 | Hop Length |
| WinLength | 512 | Length of Window |
| Window | hann | Window function |
| n\_fft | 2048 | length of the FFT window |
| n\_chroma | 12 | Number of chroma bins |
| n\_bands | 6 | number of frequency bands |
| Fmax | 1024 | highest frequency (in Hz) |

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Remark** |
| MaxInputLength | 194 |  |
| n\_folds | 5 | Number of folds |
| TestSize | 0.2 | Test dataset split size |
| BatchSize | 128 | Batch size |
| NumEpochs | 100 | Number of Epoch |
| InitialLearningRate | 0.001 | Initial learning rate |
| Dropout | 0.2 |  |
| LossFunction | categorical\_crossentropy |  |
| Optimizer | Adam |  |
| n\_classes | 8 | Number of categories |

**Tuned Hyper Parameters**

(add some text variable parameters – Take it from Slide#11)



Graphical user interface, table, timeline

Description automatically generated

The detailed statistics is available at: <https://github.com/braghavan1/capstone7/blob/documenting/03_ProjectDocumentation/ModelComparison.xlsx>

### Overall project and improvements and applications and results

* Increasing the number of LSTM units, increases the average validation accuracy. Maximum average validation accuracy is observed for 512 LSTM units
* Maximum average validation accuracy is observed for 16# of Filters and Kernel Size of 3

(Update the graphs & Text – slide#12 onwards )

Chart, bar chart

Description automatically generated

### Validation Accuracy for different Augmentations

* Spectrogram augmentation doesn’t improve the average validation accuracy
* Average validation accuracy is significantly improved by using augmented audio files in train dataset
* Maximum average validation accuracy is obtained by using all four types of audio data augmentations

Chart

Description automatically generated

## Summary

Tunned Model Performance Metrics

Graphical user interface, chart, application

Description automatically generated

# Chapter IV: Results

## Introduction

## We present a new method combining DCNNs with DTPM for automatic affective feature learning. A DCNN is used to learn discriminative segment-level features from three channels of log Mel-spectrograms similar to the RGB image representation.

## DTPM is designed to aggregate the learned segment-level features into the global utterance-level feature representation for emotion recognition. Extensive experiments on four data sets show that our method can yield promising performance in comparison

## with the state-of-the-arts. In addition, we also find that with our generated DCNN input, DCNN models pre-trained on the large-scale ImageNet data could be leveraged in speech affective feature extraction. This makes DCNN’s training with a limited amount of annotated speech data easier. The success of thisworkwarranties further investigation on using deep learning in speech emotion recognition.

## Summary

* Tuned final model achieved state of the art **93.61%** average validation accuracy in five fold cross validation
* Maximum average validation accuracy is observed using
  + MFCCs: 30
  + LSTM units: 512
  + Number of Filters: 16
  + Kernel Size: 3
  + Audio data augmentations: all four types (noise, pitch, stretch, shift) in train dataset
  + TimeDistributed function over 194 timesteps
* Time required to finish one iteration is about 15 mins on an average for final model, whereas other models used to take about 7 hours for one iteration. This is huge improvement in training times.

(Take the text from PPT Slide#17)

# Chapter V: Summary, Conclusions, and Recommendations

## Introduction

Again, start with an introduction. Summarize what has happened in your paper so far. This chapter will also vary considerably in headings and organization; what follows is a suggestion or possibility.

## Summary of the Results

State the results.

## Conclusions

Discuss the high points of your findings. This discussion should include a thorough discussion of the research question or questions, literature review, and the results. There should be a relationship to the literature review. Did your study correlate with previous research or did you find something different?

## Recommendations

Recommend some further research or a change in practices.

# References

Make sure that everything you cite in text is also in the reference list and vice versa. Below are examples of a journal and a book entry. Consult the current APA manual for additional examples.**Notice that entries use a hanging indent set at ½ inch, are single spaced, and have a blank line between each entry.**

Clough, M. (1992). Research is required reading. *The Science Teacher*, *59*(7), 36-39.

Cochran-Smith, M. (2001). Higher standards for prospective teachers. *Journal of Teacher Education, 52*(3), 179-181.