**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 convolutional, pooling and fully connected layers. In this work, we detailed the architecture, which extracts mel-frequency cepstral coefficients, chroma-gram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features from sound files and uses them as inputs for the one-dimensional Convolutional Neural Network for the identification of emotions using samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB), and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. We utilize an incremental method for modifying our initial model in order to improve classification accuracy. All of the proposed models work directly with raw sound data without the need for conversion to visual representations, unlike some previous approaches. Based on experimental results, our best-performing model outperforms existing frameworks for RAVDESS and IEMOCAP.

**Acknowledgments**

Project Mentor – Mr. Karthik Raja

Professor Assistant – Mr. Soumya Dutta

Professor – Dr. Sriram Ganapathy

**Table of Contents**

[Abstract 2](#_Toc99296213)

[Chapter I: Introduction 5](#_Toc99296214)

[Problem Statement 5](#_Toc99296215)

[Purpose of the Study 5](#_Toc99296216)

[Research Questions 5](#_Toc99296217)

[Definition of Terms 5](#_Toc99296218)

[Assumptions and Limitations of the Study 5](#_Toc99296219)

[Overview 5](#_Toc99296220)

[Chapter II: Related Work 6](#_Toc99296221)

[Introduction 6](#_Toc99296222)

[Next Heading 6](#_Toc99296223)

[Summary 6](#_Toc99296224)

[Chapter III: Method/Experiment 7](#_Toc99296225)

[Introduction 7](#_Toc99296226)

[Research Question(s) 7](#_Toc99296227)

[Data Preprocessing, Feature Engineering and Visualization 7](#_Toc99296228)

[Choice of Model 7](#_Toc99296229)

[Training the model, Performance of the Model and Metrics 7](#_Toc99296230)

[Overall project and improvements and applications and results 7](#_Toc99296231)

[Summary 7](#_Toc99296232)

[Chapter IV: Results 8](#_Toc99296233)

[Introduction 8](#_Toc99296234)

[Summary 8](#_Toc99296235)

[Chapter V: Summary, Conclusions, and Recommendations 9](#_Toc99296236)

[Introduction 9](#_Toc99296237)

[Summary of the Results 9](#_Toc99296238)

[Conclusions 9](#_Toc99296239)

[Recommendations 9](#_Toc99296240)

[References 10](#_Toc99296241)

# 

# 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.

## Research Questions

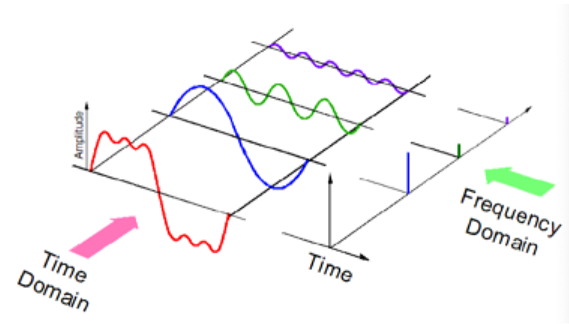
**What is Speech Emotion Recognition?**

**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.

a) High Level:Instruments, key, Chords, Melody, Rhythm, Tempo, Lyrics, Genre and Mood.

b) Mid Level: Pitch & Beat related descriptor such note, onsets, fluctuation pattern, MFCC

c) 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 (13 coeff) 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 freq contents differentiated more than the high freq 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. Consider two songs, one from a blues genre and the other belonging to metal.

**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).

Filename example: 03-01-06-01-02-01-12.wav

|  |
| --- |
| - Audio-only - 03  - Speech - 01  - Fearful - 06  - Normal intensity - 01  - Statement "dogs" - 02  - 1st Repetition - 01  - 12th Actor - 12 Female, as the actor ID number is even. |

## Overview

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

The audio files were shared as two parts 1. Train data 2. Test data

Both were classified and processed.

Chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence

## Next Heading

Add as many headings as needed.

## Summary

Write the chapter summary here.

# Chapter III: Method/Experiment

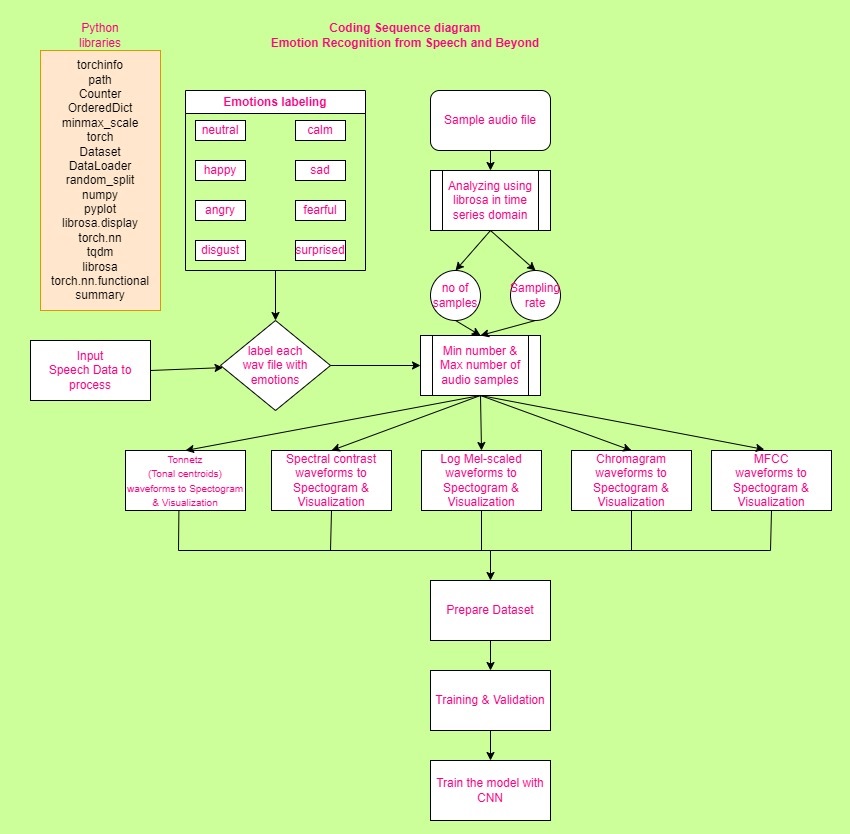
## Introduction

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).

## Data Preprocessing, Feature Engineering and Visualization

Describe the steps involved in EDA.

## 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

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 | CNNNetV2 | 1.00 | 0.58 | 1.27 | 1.69 |
| 5 | CNNNetV2\_1 | 0.90 | 0.32 | 1.48 | 1.94 |
| 6 | CNNAttentionNetV0 | 0.81 | 0.31 | 1.47 | 1.96 |
| 7 | CNNNetV2\_2 | 0.99 | 0.64 | 1.28 | 1.64 |
| 8 | CNNNetV2\_2 | 0.95 | 0.57 | 1.35 | 1.72 |
| 9 | CNNNetV2\_2 | 0.99 | 0.62 | 1.29 | 1.67 |
| 10 | CNNNetV2\_3 | 0.99 | 0.60 | 1.29 | 1.65 |
| 11 | CNNNetV2\_4 | 0.98 | 0.59 | 1.30 | 1.68 |
| 12 | Conv1D | 0.97 | 0.66 | 0.10 | 1.72 |
| 13 | Conv1D | 0.71 | 0.55 | 1.10 | 1.74 |
| 14 | Conv2D, TransformerEncoder | 0.98 | 0.66 | 0.053 | 1.52 |

|  |  |
| --- | --- |
| Parameters of the model | Value |
| Sample rate | 22050 |
| HopLength | 512 |
| WinLength | 512 |
| Window | Hann |
| n\_fft | 2048 |
| n\_mfcc | 13 |
| n\_chroma | 12 |
| n\_mels | 13 |
| n\_bands | 6 |
| n\_batchsize | 64 |
| NumEpochs | 25 to 100 |
| fmax | 1024 |

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

Talk about the instruments used.

<https://github.com/braghavan1/capstone7>

## Overall project and improvements and applications and results

Discuss methodological limitations or procedural weaknesses.

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Model | Limitations | Improvements |
| 1 | BaseCNNNetwork |  |  |
| 2 | CNNTransformerNet |  |  |
| 3 | CNNNetwork |  |  |
| 4 | CNNNetV2 |  |  |
| 5 | CNNNetV2\_1 |  |  |
| 6 | CNNAttentionNetV0 |  |  |
| 7 | CNNNetV2\_2 |  |  |
| 8 | CNNNetV2\_2 |  |  |
| 9 | CNNNetV2\_2 |  |  |
| 10 | CNNNetV2\_3 |  |  |
| 11 | CNNNetV2\_4 |  |  |
| 12 | Conv1D |  |  |
| 13 | Conv1D |  |  |
| 14 | Conv2D, TransformerEncoder |  |  |

## Summary

Summarize the main points of the methodology.

# Chapter IV: Results

## Introduction

Start with another introduction, you might briefly reiterate the purpose of the study and how it was conducted. The purpose is to provide the reader with *at a glance* information about the nature and scope of your paper/report.

## 

## Summary

Write a summary of the results.

# 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.