**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.

~~, thus setting the new state-of-the-art. For the EMO-DB dataset, it out performs all previous works except one but compares favorably with that one in terms of generality, simplicity, and applicability. Specifically, the proposed framework obtains 71.61%for RAVDESS with 8 classes, 86.1% for EMO-DB with 535 samples in 7 classes, 95.71% for EMO-DB with 520 samples in 7 classes, and 64.3% for IEMOCAP with 4 classes in speaker-independent audio classification tasks.~~

~~This paper describes a method for Speech Emotion Recognition (SER) using Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers. We used 3 class subset (angry, neutral, sad) of German Corpus (Berlin Database of Emotional Speech) containing 271 labeled recordings with total length of 783 seconds. Raw audio data were standardized so every audio file has zero mean and unit variance. Every file was split into 20 millisecond segments without overlap. We used Voice Activity Detection (VAD) algorithm to eliminate silent segments and divided all data into TRAIN (80%) VALIDATION (10%) and TESTING (10%) sets. DNN is optimized using Stochastic Gradient Descent. As input we used raw data without and feature selection. Our trained model achieved overall test accuracy of 96.97% on whole-file classification.~~

Signature??

Abstract to match with our paper submission, donot copy from the Reference paper. - completed**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

From the dataset we received, our approaches of research are listed below in reverse chronological as follows.

Phase 10:

We take

Final code available: <https://github.com/braghavan1/capstone7>

Phase 9:

We did

Phase 8:

We did

Phase 7:

We did

Phase 6:

We did

Phase 5:

We did

Phase 4:

We did

Phase 3:

We did

Phase 2:

We did

Phase 1:

We did

## Definition of Terms

### First term. The definition begins on the same line as the term. Include terms that may not be well understood by a layperson outside of your field. If you use the exact words of a definition - from a dictionary, for instance - indicate that it is quoted material by adding the citation.

### Next term. Type its definition, here. Format each new term just as you would any new paragraph.

### Next term. Definition ……..

## 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

Briefly outline the remainder of the paper and what is to come in the next chapter. Use the appropriate tense (past, present, or future) based on the context of the material being written. (Skipping for now)

# Chapter II: Related Work

## Introduction

To evaluate the data from the Emotions

## 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

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

Talk about the instruments used.

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

## Overall project and improvements and applications and results

Discuss methodological limitations or procedural weaknesses.

## 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.