Emotion Recognition From Speech and Beyond

Supervisor: Dr. Sriram Ganapathy (sriramg@iisc.ac.in)

# 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.

# Objective

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

# Details of the Datasets

## 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 convention

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:

## Filename identifiers

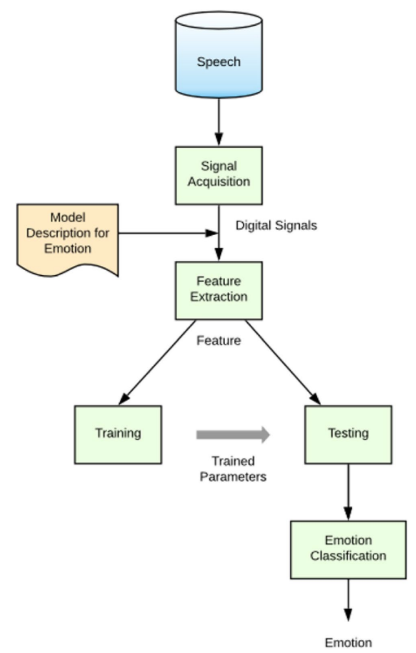
* Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
* Vocal channel (01 = speech, 02 = song).
* Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
* Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity 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. |

# Pre-processing steps

PreProcessing Steps



Any SER system consists of two components: a processing unit that extracts the appropriate features from the speech data and a classifer that ultimately decides the emotion from the underlying speech utterance. In this section, the methodology used for feature extraction, dimensionality reduction, and classifcation in the proposed model are presented. Also, the use of autoencoders for the purpose of dimensionality reduction, and its impact on classifcation is discussed.

# Method/Experiment highlighting the choice of your Model

## Classifiers

Each classifer has a unique set of advantages and limitations and therefore, the performance may vary with each classifer. The objective of this section is to provide an overview of the classifers used in this work.

# The Performance Metrics

# Conclusions