Emotion Recognition using Speech Analysis

Kalyani Goyal, Muskan Jain, Rahul Kashyap, Rahul Baviskar

Indian Institute of Technology Goa

Objectives

Develop a Convolutional Neural Network to classify between different emotions using speech analysis. The classifier will classify it into seven emotional labels,i.e., Happy, Sad, Angry, Suprise, Disgust, Calm, Fearful

- Identify the underlying emotion in a male or a female voice in their speech
- Build a classifier that classifies the speech into the different emotions by feeding them to the classifier
- Model the classifier that evolves over time

Introduction

An *Emotion* is a strong feeling deriving from one's circumstances, mood, or relationships with others. Emotions are something that are innate to human beings. It is what separates us from the machines. This is where it gets fascinating, when a machine can classify an emotion better than a human, that is when we see the real power of Artificial Intelligence.

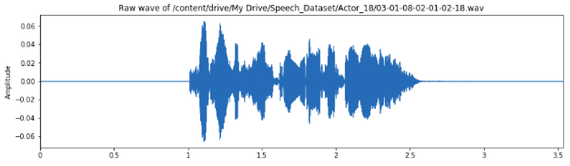


Figure 1:Waveform of female actor saying "Kids are talking by the door" in surprised tone.

- It might be impossible to understand human emotion but it is not impossible to detect them
- We do it all the time, e.g. while talking to a friend on the phone.
- Speech is nothing but audio signals with varying tonality and pitch with different emotions.
- Although the variable values might be different for every one, but the patterns emerge out to be same

The approach

Human voice is nothing, audio signals with a definite pattern. So,for classifying the audio signals we have developed a CNN(Convolutional Neural Network) classifier and trained it on the RAVDESS Dataset consisting of total 1440 voice/emotion samples.

- Each sample can be characterized into one of the eight different categories.

Contributions

- Visualized the difference between various classes based on the feature MFCC(Mel-frequency cepstral coefficients).
- Developed a Convolutional Neural Network classifier that can recognise emotions from various factors in the given voice input.
- Performed evaluation using RAVDESS Dataset.

Graphical work flow

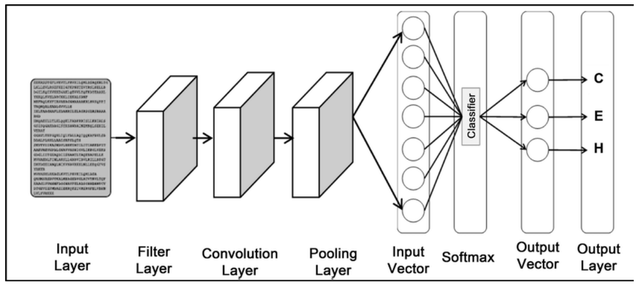


Figure 2:Graphical Workflow of a CNN Classifier

The mfcc feature is extracted as a vector for each audio clip. The array of these features is passed through various convolutional and pooling layers. The output here goes as input to a fully connected layer which then after mathematical calculations gives the output at the final layer

Inference

The parameter of interest: Mel-frequency cepstral coefficients(MFCC) and the relative performance of Convolutional Neural Networks(CNN) and Support Vector Machine(SVM).The CNN performs much better as compared to SVM in case of this dataset.

Experiments and analysis

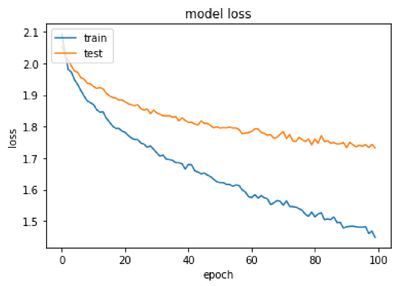
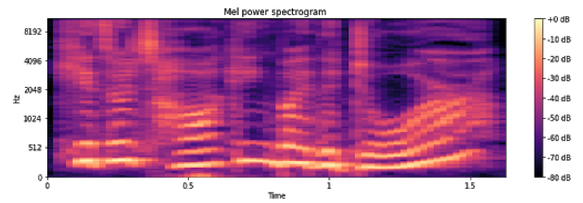


Figure 3:Loss function of CNN with number of epochs on classification into seven emotions

We compared and analysed the results of SVM and CNN on the given dataset by comparing their relative accuracy where CNN performed better. We then compared the results of classification into 7 categories without considering gender of the actor with 7 categories of only male actors.



Mel Power Spectogram of the female audio clip

References

- Intech Open Research Paperk
- Dataset
- Comparison of different classification methods
- Commonly used speech features and algorithms

Acknowledgements

A term project completed under the requirements of course CS 386: Artificial Intelligence (Instructor: Clint P. George)