SPEECH EMOTION RECOGNITION

Summer Project 2021

Brain and Cognitive Society (BCS), IIT Kanpur

Abstract

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. In this project, basic emotions like calm, happy, fearful, disgust etc. are analyzed from emotional speech signals. Using RAVDESS dataset which contains around 1500 audio file inputs from 24 different actors (12 male and 12 female) who recorded short audios in 8 different emotions, we will train a NLP-based model which will be able to detect among the 8 basic emotions as well as the gender of the speaker i.e. Male voice or Female voice.
After training we can deploy this model for predicting with live voices.

Objective

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. Speech emotion processing and recognition system is generally composed of three parts, the first being speech signal acquisition, then comes the feature extraction followed by emotion recognition.

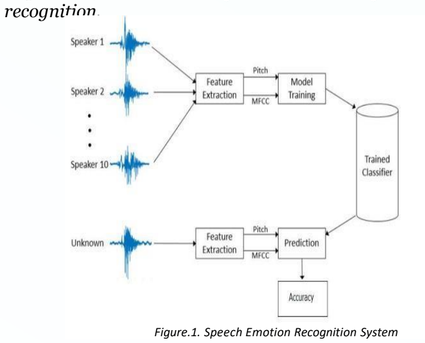


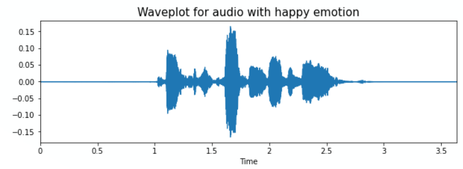
Figure.1. Speech Emotion Recognition System

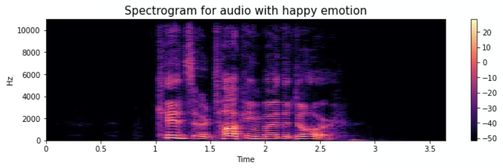
We extracted various features of the audio like ZCR, Spectral Centroid, MFCC and Chroma Frequency into an array and passed it as an input to our models.

We implemented 3 models, MLP, RNN-LSTM and CNN to detect the emotions of the input audio.

RAVDESS Dataset

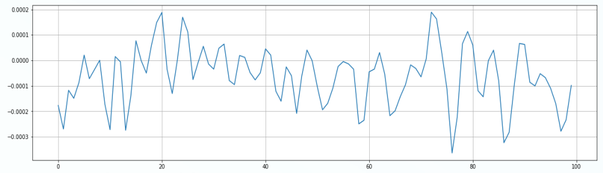
*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains **7356 files** (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.*



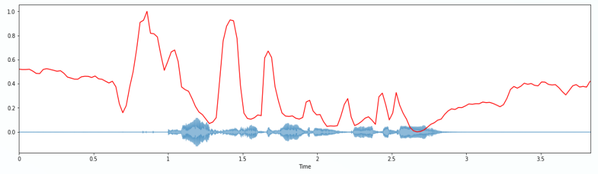


Feature Extraction

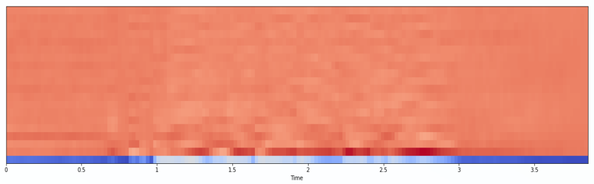
- ***Zero Crossing Rate (ZCR)***
  It is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.



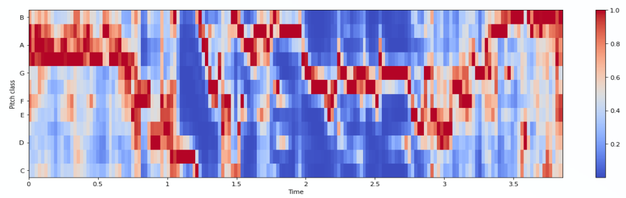
- ***Spectral Centroid***
  It is the center of 'gravity' of the spectrum. It is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located.
  Perceptually, it has a robust connection with the impression of 'brightness of a sound'.



- ***MFCC (Mel-Frequency Cepstral Coefficients)***
  In sound processing, it is a representation of the short term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency.



- ***Chroma Frequency***
  Chroma frequency is an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.



Model Implementation

- ***MLP (Multi-Layer Perceptron) Model***
  The arrays containing features of the audios are given as an input to the MLP Classifier that has been initialized. The Classifier identifies different categories in the datasets and classifies them into different emotions.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 120)               19320
_________________________________________________________________
dense_6 (Dense)              (None, 80)                9680
_________________________________________________________________
dense_7 (Dense)              (None, 50)                4050
_________________________________________________________________
dense_8 (Dense)              (None, 20)                1020
_________________________________________________________________
dense_9 (Dense)              (None, 2)                 42
=================================================================
Total params: 34,112
Trainable params: 34,112
Non-trainable params: 0
```

- ***RNN-LSTM Model***
  We used RMSProp optimizer to train the RNN-LSTM model, all the experiments were carried with a fixed learning rate of 0.1. The batch size utilized was 32 with an epoch size of 200. Batch Normalization is applied over every layer and the activation function used is the SoftMax activation function.

- ***Convolutional Neural Network (CNN)***
  The activation layer called as the RELU layer is followed by the pooling layer. The specificity of the CNN layer is learnt from the functions of the activation layer.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 40,128)	768
activation_1 (Activation)	(None, 40,128)	0
dropout_1 (Dropout)	(None, 40,128)	0
max_pooling1d_1 (MaxPooling1)	(None, 5128)	0
conv1d_2 (Conv1D)	(None, 5128)	82,048
activation_2 (Activation))	(None, 5128)	0
dropout_2 (Dropout)	(None, 5128)	0
flatten_1 (Flatten)	(None, 640)	0
dense_1 (Dense)	(None, 8)	5128
activation_3 (Activation)	(None, 8)	0

Conclusion

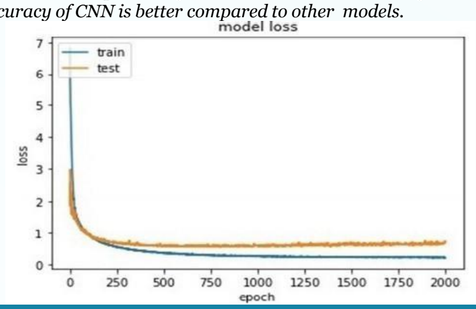
The significant part of SER are the signal processing unit in which relevant features are extracted from speech signal and classified in order to bring out the emotion to the particular class. It is shown that CNN being the best classifier compared with machine learning techniques. The accuracy can be improved by selecting relevant features. For improved results, mixed models of the approaches can be applied.

Results

Classification on RAVDESS Dataset is performed with MLP, RNN-LSTM and CNN models. The accuracies obtained by different teams are as follows:

- **Team 1**
  MLP Model: 63%, CNN Model:73%, LSTM Model: 72%
- **Team 2**
  MLP Model: 60%, CNN Model:70%, LSTM Model: 60%
- **Team 3**
  MLP Model: 62%, CNN Model:71%, LSTM Model: 68%
- **Team 4**
  MLP Model: 71%, CNN Model:78%, LSTM Model: 75%
- **Team 5**
  MLP Model: 66%, CNN Model:72%, LSTM Model: 66% (Emotion recognition)
  MLP Model: 99%, CNN Model:99.5%, LSTM Model: 97% (Gender recognition)

The overall comparison of all models showed that classification accuracy of CNN is better compared to other models.



Documentation

- ***GitHub Repositories***
  https://github.com/pushpanshu0501/Speech-Emotion-Recognition--1
  (Team 1)
  https://github.com/aarjavjain824/BCS_Speech_emotion_recognition.git
  (Team 2)
  https://github.com/Nitesh2k19/Speech_Emotion_Recognition.git
  (Team 3)
  https://github.com/anmolag190153/BCS_summer_project_SER
  (Team 4)
  https://github.com/Gaukmo1/BCS_SER_project
  (Team 5)

- ***Documentation***
  https://www.overleaf.com/project/60cc4cd9e461496da6c8ce1c

- ***Demo***
  https://drive.google.com/drive/folders/1toToX54XUQ5zpvyOXhztZbLY6y9wdqnN?usp=sharing