SPEECH EMOTION RECOGNITION
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.

Introduction

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. At present, speech emotion recognition is an emerging crossing field of artificial intelligence and artificial psychology; besides, it is a popular research topic of signal processing and pattern recognition. The research is widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on. 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.

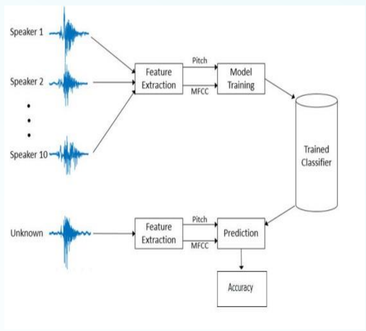
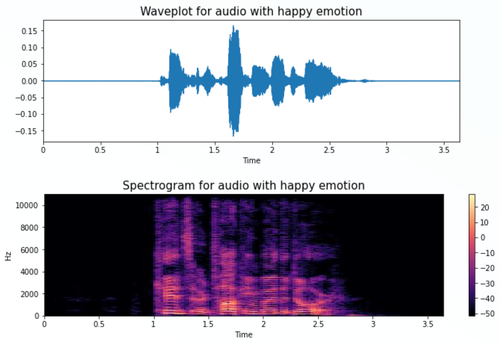


Figure.1. Speech Emotion Recognition System

Speech recognition is the process of converting an acoustic signal, captured by microphone or a telephone, to a set of characters. They can also serve as the input to further linguistic processing to achieve speech understanding, a subject covered in section. As we know, speech recognition performs tasks that similar with human brain.

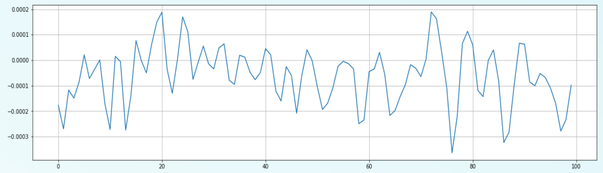
RAVDESS Dataset

*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.*

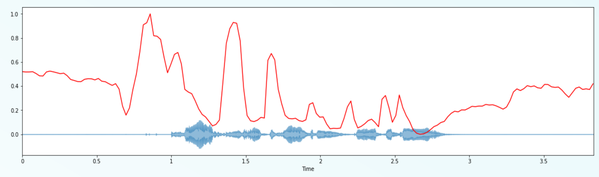


Feature Extraction and Model Implementation

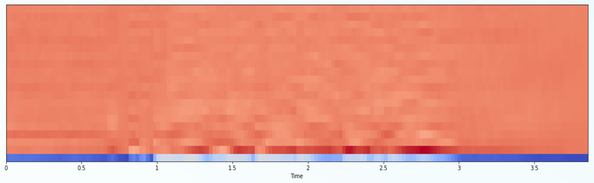
- ***Zero Crossing Rate (ZCR)***
  It is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.



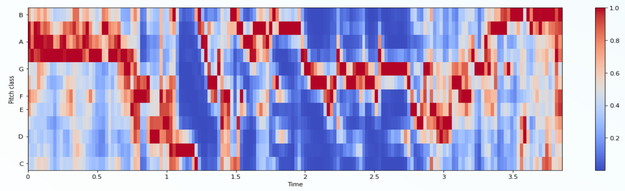
- ***Spectral Centroid***
  It is the center of 'gravity' of the spectrum. It is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located.
  Perceptually, it has a robust connection with the impression of 'brightness of a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



- ***MLP (Multi-Layer Perceptron) Model***
  For the training, we store the numerical values of emotions and their respective features correspondingly in different arrays. These arrays 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. In addition, Dropout regularization was used to prevent overfitting. Batch Normalization is applied over every layer which improves the performance, the stability of the learning process, and the activation function used is the SoftMax activation function.

- ***Convolutional Neural Network (CNN)***
  The architecture of CNN is originated from the visual cortex. The visual cortex has multiple layers, each of which can filter out irrelevant information. The preprocessing stage of Convolutional neural network is lower in comparison with other classification algorithms. 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. The input of the network is a list of 2D images constructed from 128 x 40 values of normalization

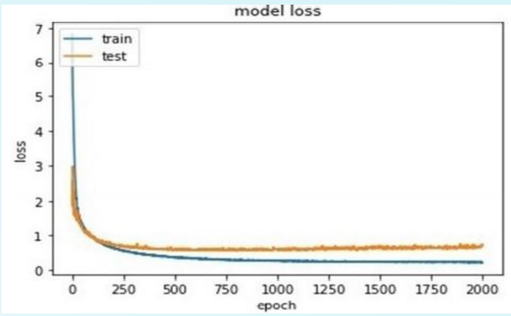
Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 40,128)	768
activation_1 (Activation)	(None, 40,128)	0
dropout_1 (Dropout)	(None, 40,128)	0
max_pooling1d_1 (MaxPooling1)	(None, 5128)	0
conv1d_2 (Conv1D)	(None, 5128)	82,048
activation_2 (Activation))	(None, 5128)	0
dropout_2 (Dropout)	(None,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

Total parameters: 87,944
Trainable parameters: 87,944
Non-trainable parameters: 0

Results

Classification on RAVDESS Dataset is performed with MLP, RNN-LSTM and CNN models.
- *MLP Model gave an accuracy of 63%*
- *RNN-LSTM Model gave an accuracy of 70%*
- *CNN Model gave an accuracy of 78%*

The CNN model was evaluated on 1000 samples, with 2000 epochs and trained with 87,944 parameters have shown an accuracy of 78.20%. The overall comparison of all models is depicted above in which the classification accuracy of CNN is better compared to other models.
The model loss is heavier with less data as the model is trained with more number of data, loss is found to be constant as shown in figure and with increasing number of data, the over fitting is reduced.



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 research on automatic SER is gaining momentum due to its improved ability on Human Computer Interaction. The accuracy can be improved by selecting relevant features. For improved results, mixed models of the approaches can be applied. The major challenge lies in recognizing the accent of a person and the usage of vocabulary. In future, it can be expanded more number of people and speech recognition can be done for regional languages.

Discussion and Future Work

The team would try and improve the models developed and deploy it for predicting the emotions of live voices.
A more detailed version can be seen in the documentation of the project.