SPEECH EMOTION RECOGNITION

Deep Learning on Python

Panda Projects

WeWork, Sector 15
Gurugram Haryana

Kunal Gehlot
Gehlotkunal4@outlook.com

Objective

To create an Emotion recognition engine to identify mood of a person through their speech recording.

Databases used

We used the **SAVEE** and **RAVDESS** databases:

- **SAVEE:** Surrey Audio-Visual Expressed Emotion (SAVEE) database has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. The data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled.
- **RAVDESS:** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness

Models used:

We first extracted features from the Audio by using LibRosa, which are MFCC (Mel-Frequency Cepstral Coefficients) with 39 Mel-Frequencies per audio, and then segregating the data into seven emotions, given: Anger, Fear, Disgust, Happy, Sad, Surprised, Neutral

Then we took two approaches, one of Machine Learning, with model and accuracy below mentioned:

Random Forest

Accuracy: 38.89%

KNeighborsClassifier(9 neighbors)

Accuracy: 35.93%

Concluding that machine learning would not work at all, we took the Deep Learning approach by using 2-dimensional CNN (Convolutional Neural Network)(Conv2D by Keras) and got our best yet result of 86% accuracy:

Conv2D

Accuracy: **86.43%**

Thus making the Conv2D the most useful Model to predict emotion.

Testing:

To test the data, a separate program to record audio is made and another one to load the saved models and test the sample audio against the model.