Speaker Emotion Recognition System Project

Motasem Diab & Ezz Abu Asab (1162106 &1160730)

Department of Electrical and Computer Engineering, Birzeit University, Palestine

Abstract

The main aim behind this project was to be able to distinguish audio signals that are longitudinal waves which travel through the air and that consist of compressions and rarefactions [1] based on their attributes). To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed.

Mel-frequency cepstrum coefficients (MFCC) are extracted from the speech signals and used to train different classifiers.

Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task. [2]

Index Terms: Speech Synthesis, Emotion recognition and feature extraction.

1. Introduction

This project tackles the problem of Human Emotion Recognition using Audio Signal Processing and Classification using Machine Learning. Speech Processing has been developed as one of the vital provision regions of Digital Signal Processing. Speaker recognition is the methodology of immediately distinguishing who is talking dependent upon special aspects held in discourse waves. This strategy makes it conceivable to utilize the speaker's voice to check their character and control access to administrations. [3]. Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. This problem can come in handy when it comes to Emotion recognizing in the field of Psychology and Metal Health regarding emotional disorders. [4]

2. Background & Related Work

The following literature was used in the project.

- Automatic Speech Emotion Recognition Using Machine Learning by Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder.

  Some of the methodologies used to extract features from the sound files were used.

- Speaker Emotion Recognition Based on Speech Features and Classification Techniques by J. Sirisha Devi, Dr.Srinivas Yarramalle and Siva Prasad Nandyala.

  Some of the ideas presented in the FEATURE EXTRACTION AND MATCHING were tested in our project to get the better feature vector from the original.

3. Methodology

3.1. Data

The data presented to us included 6 Folders representing each one of the 6 target classes (happiness, sadness, anger, boredom, fear and neutral). A script was written to randomly split these files into training and testing files (75% Training and 25% testing). The training files were still sub-categorized into folders by the name of their class and the files themselves were named into more meaningful names (e.g. "anger{i}.wav).

3.2. Feature Extraction

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. Many spectrum features such as Mel-frequency cepstrum coefficients (MFCC) with delta and delta-delta "39-features", log Mel-filterbank energy, and Spectral subband Centroid (SSC) features were selected to extract the emotional features in our project. We used the built-in python functions for extract these features and we set a hamming window with size 25ms shifted by 10ms. For each sound file, the .wav was divided into frames then the features were extracted for each frame and appended to a big matrix of features with size [#of frames for all sounds: #of features], another one-dimensional array with the size of [#of frames] that has the labels for each corresponding frame.

3.3. Model Prediction

Many machine learning algorithms have been used for discrete emotion classification. The goal of these algorithms is to learn from the training samples and then use this learning to classify new observation. In our project, we decided to use three classification method: Gaussian-Naïve-Bayes (GNB), multi-layer perceptron (MLP), decision-tree (DT) and improved decision-tree (with random_state criterion), all four method work to make a model predict the class for specific feature vector. Our problem is that each sound file has a lot of frames, each frame is a feature vector, each feature vector has a predicted label, so every sound file has a number of labels that equals to the number of frames, to solve this issue we chose to predict the whole sound file based on the majority of labels, you can see the results for each method in the results section. [Experiments 1-4 in Experiments and Results]

3.4. Further Data Presentation

Following the initial predictions, there was a huge need to improve the accuracy levels, the topic of Feature Selection was heavily discussed as a solution to improve the accuracy and remove unnecessary features that made the model very heavy.

To do this the dataset (features and labels) were saved into a Comma Separated Values (CSV) file and a header was added to distinguish the concatenated columns of vectors. This resulted in "emot.csv".

3.5. Feature Selection (FS)

This CSV was imported back into our project and entered into a tree classifier which ran 100 iterations and gave a value of importance to features that were used in building the tree in each of the 100 iterations. Following that the values were plotted using the Panda Library data frame. As seen we can see in the following figure, these are the values of the importance of all the features (Figure 1).

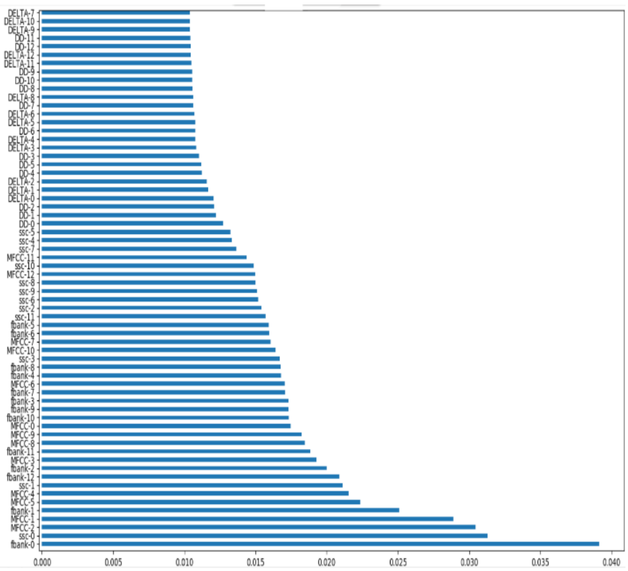

Figure 3.5.1 Feature Importance Plot

As obviously seen the features of Delta & Delta-Delta have the least importance, so they were removed from the CSV and the model.

The removal of these features meant that we need to run the extraction and prediction for the model on the dataset [Experiment 5 & 6 in Experiments and Results].

3.6. Other Implementations

After the new testing dataset was released, we proceeded to turn the previous dataset into 100% Training. Following that the code was slightly modified to print onto the output.txt.

4. Experiments and Results

The project's experiments and results can be seen in the following two tables (Table 1 & 2).

Experiment	Result
1 (Regular DT Before FS)	Anger accuracy = 100% Bordom accuracy = 68.75% Fear accuracy = 53.33% Happiness accuracy = 26.66% Sadness accuracy = 92.307% Neutral accuracy = 43.75 % Average accuracy = 67.32673%
2 (MLPC Before FS)	Anger accuracy = 100% Bordom accuracy = 75% Fear accuracy = 46.666% Happiness accuracy = 0.0% Sadness accuracy = 92.387% Neutral accuracy = 6.25 % Average accuracy = 57.425 %
3 (Gaussian NB) (Before FS)	Anger accuracy = 73.07% Bordom accuracy = 81.25% Fear accuracy = 46.66% Happiness accuracy = 60.0% Sadness accuracy = 100% Neutral accuracy = 0.0 % Average accuracy = 60.396 %
4 (Improved DT Before FS)	Anger accuracy = 100% Bordom accuracy = 75% Fear accuracy = 60.0% Happiness accuracy = 33.33% Sadness accuracy = 100% Neutral accuracy = 56.25 % Average accuracy = 73.2673 %
5 (Regular DT After FS)	Anger accuracy = 100% Bordom accuracy = 75% Fear accuracy = 53.3% Happiness accuracy = 33.33% Sadness accuracy = 92.3% Neutral accuracy = 50.0 % Average accuracy = 70.29 %
6 (Improved Decision Tree After Feature Selection)	Anger accuracy = 100% Bordom accuracy = 81.25% Fear accuracy = 66.666% Happiness accuracy = 33.33% Sadness accuracy = 92.307% Neutral accuracy = 62.5 % Average accuracy = 75.2475 %

Table 1 Experiments

EXP#	Ang.	Bor.	Fear	Happ.	Sad.	Neu.	AVG
1	100	68.7	53.3	26.6	92.3	43.7	67.3%
2	100	75.0	46.6	0.0	92.3	6.25	57.4%
3	73	81.2	46.66	60.0	100	0.0	60.3%
4	100	75	60.0	33.3	100	56.2	73.26
5	100	75	53.3	33.3	92.3	50.0	70.29
6	100	81.2	66.6	33.3	92.3	62.5	75.24

Table 2 Accuracy Percentage for Experiments

5. Conclusions

In this project, we managed to distinguish the 6 aforementioned emotions to very good accuracy. It was concluded that this was possible using MFCC, FBANK and SSC features, the features of Delta and Delta-delta were used at the start but eventually removed to improve the accuracy due to their lack of importance when tested in a TreeClassifer.

Other Implementations such as SVM and K-Means were tried but were not used in the final project.

The features of Chroma and MS were at some point discussed to be used, however, we faced problems while using them so they weren't added to the dataset.

We studied how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly different features is selected. Feature selection techniques display that extra data is not always good in machine learning uses.

6. References

[1] "Audio Signal Processing" Oct. 29, 2019. Accessed on: May. 20, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Audio_signal_processing

[2] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Clede., "Automatic Speech Emotion Recognition Using Machine Learning". Accessed on: May. 20, 2020. [Online].

[3] J.Sirisha Devi, Dr.Srinivas Yarramalle, Siva Pasad Nandayla., "Speaker Emotion Recognition Based on Speech Features and Classification Techniques" Accessed on May. 20, 2020. [Online].

[4] Ali H, Hariharan M, Yaacob S, Adom AH. Facial emotion recognition using empirical mode decomposition. Expert Systems with Applications. 2015;42(3):1261-1277 . Accessed on: May. 20, 2020.