Speech Emotion Recognition

Renad Khaled , Raghda Alhkawaja, Shurouq Tanatra
1151356 , 1160831, 1152715
Department of Electrical and Computer Engineering, Birzeit University, Palestine
renad1997khaled@gamil.com,Raghoodaakhawaja@gmail.com,shurouqtanatra@gmail.com

Abstract

Humans have the natural ability to use all of their available senses to maximize awareness of the message received from the opposite person. Emotional detection is natural for humans but is very a difficult task for machines. The aim of the project is to build a model that helps discover human emotions from the way he talks with others, and this can be expressed through body language. Nowadays, people communicate with each other through feelings and emotional gestures determined by knowledge.

1. Introduction

A growing interest in the recognition and integration of users' emotions in the interaction with machines can be observed at the time. SER program is a series of methodologies that processes and categorizes speech signals to discover the emotions that are involved. There are diverse ways to formulate emotions, including commonly using separate models and dimensions. Whereas the SER method includes a description of emotions, a controlled learning structure; it will be trained in modern speech patterns to understand the emotions. Data must be processed until the classification method gets the requisite characteristics. In this project, we have used MFCC, Mel spectrogram, and Chorma-Stft. The solution is to determine which feature carries more knowledge and to incorporate such features in order to obtain a more reliable identification rate.

2. Background

Through our reading of several articles, we found that they are using several methods of speech analysis to reach the speaker's emotions:
In this paper [1], the focus was on finding new ways to recognize speech excitement; the methods used here are spectra and deep convolutional neural network (CNN). Seventy-five percent of the data was used for training, while the rest were used for testing phase. 61.75% accuracy for each spectrometer is considered and low accuracy is achieved.
Several experiments were conducted to evaluate the method used to identify the reactions:

- ❖ Fresh Trained CNN based SER
- ❖ Fine-tuned CNN based SER

In this paper [2], the basic idea was about introducing a new approach for automatic recognition the emotions of the speaker. First, a model for identifying feelings through vocal features is presented. Second, an approach to identify emotions is presented with an operative content using Belief Network, based on emotional key phrases. Finally, the two sources of information will be integrated into the easy decision merger using the neural network. Simple classifier K Means, Gaussian Mixture Models, Neural Nets or Support Vector Machines were used. Two-thirds of the data was used for training, and one-third for testing in three courses; a major gain was achieved in reducing error rates up to 8.0%.

3. Mythology

In this project, At first we Loaded the data and extract features for each sound file , after that we examine the feature extraction techniques, which is Mel-Frequency Cepstral Coefficients (MFCC), farther more we used Compute a Mel-scaled power spectrogram(mel spectrogram), finally we used chorma-stft , to extract features Speech to text Model in Python, And then provided the output to one form of the corresponding techniques, Then attaching both MLP Classifier and GaussianNpP. Finally, we attached SVM.

3.1. Feature Extraction:

To know any words that require us to extract its features, i.e. identify the components of a good audio signal to determine the linguistic content and get rid of all other things as shown in Figure 3.2, Among these features that were used:

3.1.1. Mel Spectrogram:

S = Mel Spectrogram (audio In, Js) returns a spectrum of the input sound in the Fs rate. Mel Spectrogram applies the frequency field to the audio signals that are framed at the appropriate time [3].

3.1.2. Mel Frequency Cepstral Coefficients (MFCCs):

A feature commonly used in automated speech recognition and voice recognition. Furthermore, the (MFCC) is the most widely utilized in speech recognition. It incorporates the benefits of cepstrum analysis with a crucial band-based perceptual frequency scale [4].
Stages of MFCC feature extraction for Audio Signals shown in figure 3.1

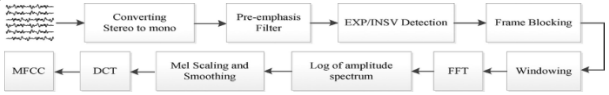
Converting Stereo to mono
Pre-emphasis Filter
EXP/INSV Detection
Frame Blocking
Windowing
FFT
Log of amplitude spectrum
Mel Scaling and Smoothing
DCT
MFCC

Figure 3.1:*Stages of MFCC Feature extraction for Audio Signals*

3.1.3. Chorma-stft:

The chroma attribute is a suffix, describing a simplified version of the tonal quality of a musical audio signal. Thus, chroma characteristics may be regarded as a significant precursor for high-level semitone studies, such as chord identification or harmonic similarity estimation. For such high-level operations, higher efficiency of the isolated chroma function makes for much better performance [5].

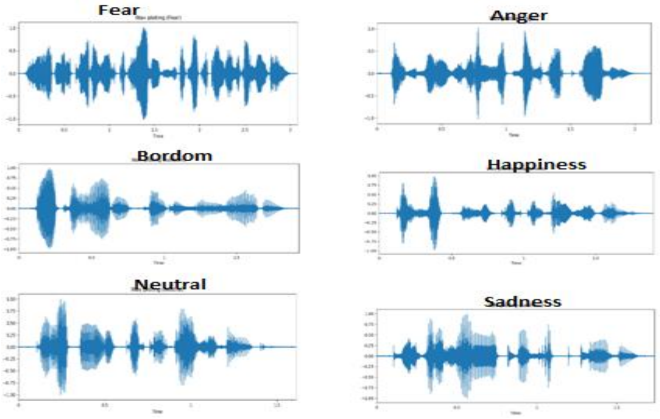

Figure 3.2:*Wav plotting for all emotions via the features*

3.2. Model Classification:

3.2.1. MLP Classifier:

It is a class of artificial feeding nerve networks (ANN). It consists of three layers: an input layer, a hidden layer, and an output layer. A supervised learning technique called reverse spreading is used for training [6]. The error can be presented as the following figure3.3.

$$e_j(n) = d_j(n) - y_j(n)$$

Figure 3.3:*Relation degree of error*

3.2.2. Gaussian NB:

The Gaussian Naive Bayes algorithm is an NB algorithm of a different kind. It is usually used where there are constant values in the app. It is therefore presumed that all devices obey the normal distribution of a Gaussian distribution [7].

3.2.3. *SVM:*

The benefits of vector supporting devices are: Efficient in high dimensional spaces. In situations when the number of measurements is larger than the number of tests, they remain effective. It also includes a subset of decision-making teaching points, and it is also memory efficient. Finally, versatile: various roles for the decision feature may be defined in the kernel [8].

3.2.4. *Logistic Regression:*

It is a mathematical technique that uses a logistic equation to predict a binary dependent variable in its simplest form, though there are also more nuanced extensions. Farther more logistical regression analyzes predict the parameters of a logistical model. The logistical model is used to measure the likelihood of a certain form or occurrence e that happens such as pass/fail, win/lose, alive/dead, or healthy/sick [9].And the samples were presented by tow ways:

- ❖ RPM: A model that allows a probability distribution of a value of either zero or one to make learning more powerful.
- ❖ Raw pixel: A model works by inserting images to identify the features.

3.2.5. *Voting Classifier:*

Classifier Voting allows two forms of voting [10]:

- ❖ Hard: Predicting the majority of votes and a recurring prediction is chosen between models.
- ❖ Soft: prediction based on average probability, using all model predictions.

4. Experiments and results

The project's aim is to create a model that can be conveyed by body language, trying to explore people's emotions through the way he speaks to others. After that, we read the sounds from the attached files and found the sample rate for each sound, after which they used three types of features as a following (MFCC, Mel spectrogram, chroma-stft).

Speech data for this project are given and are already divided into two subsets:

- ❖ Training (80% of data)
- ❖ Testing (20% of data).

We use this knowledge to train and test the program using the implementation details. The accuracy was changed each time we run the code, and the highest accuracy was 83% as shown in table 4.1

The models that we used to obtain the accuracy:

- ❖ MLP Classifier:

  in this type we used six parameters the first parameter alpha about float value equals 0.01, the second parameter batch_size about integer value equals 256, after that epsilon value for numerical stability in adam equal 1e-05, farther more hidden_layer_sizes is the number of neurons in hidden layer equal 300, after this learning_rate equals 'adaptive' keeps the learning rate constant to 'learning_rate_init' as long as training loss keeps decreasing, and max_iter Maximum number of iterations and equal 500.

  During the experiment, we obtained an accuracy of 78.75%, and in the following figure4.2 we represented all the emotions of this model, and as the figure shows when emotion for happiness, he got two false predictions, one at anger and the second at fear.

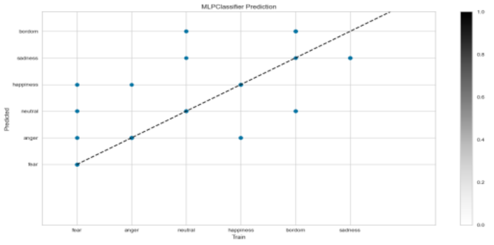

Figure 4.1:*MLP Classifier prediction*

- ❖ Gaussian NB and SVM

  In this type we used two methods as a following:
  1. Fit Gaussian Naive Bayes/SVM according to X, y
  2. Predict: Perform classification on an array of test vectors X.

- ❖ Voting Classifier

  Combine the predictions of several base estimators:
  1. Linear SVC: It is implemented by liblinear should scale better to large numbers of samples.
  2. K Neighbors Classifier: Classifier implementing the k-nearest neighbor's votes.
  3. Random Forest Classifier: Meta estimator is used to improving accuracy and control over-fitting.

- ❖ Logistic Regression

  Use two ways of analyzing samples:
  1. RPM: Parameters are estimated using SML. The first parameter is random_state about integer value and use to define the state of the random permutations' generator. Then use verbose parameter is default equals zero and means silent mode. And use the learning_rate parameter for weight updates and equals 0.06. Finally use n_iter parameter is the Number of iterations over the training dataset to perform during training and set equals 10.
  2. Raw pixel: using fit and predicts method.

Models	Best Accuracy
MLP Classifier	**83.75%**
Gaussian NB	**58.75%**
SVM	**56.25%**
Voting Classifier	**73.75%**
Logistic regression using raw pixel	**72.50%**
Logistic regression using RPM	**31.25%**

Table 4-1:*The accuracy for different models*

5. Conclusion

The aim of this project is to build a model that allows the detection of human feelings from the way it communicates with others and was implemented using the Python language by the Pycharm program. This development shows important results by looking at the integration of audio and linguistic information by defining emotion of speech as a solid framework and a valuable feature derived from that accuracy to 85%, and finally, different classification methods were compared. And the best model MLP classifier and the accuracy was 83.75%.

6. Preferences

[1] (Badshah, Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W., 2017) access in 15/5/2020
[2] (Björn Schuller, 2004) access on 15/5/2020
[3] (melspectrogram.html, 2020)
[4])Mel-frequency_cepstrum(2020 ،
[5] (Shah, 2020)
[6] (Multilayer_perceptron, 2020)
[7] (gaussian-naive-bayes-classifier-implementation-python, 2020)
[8] (svm.html, 2020)
[9] (Logistic_regression, 2020)
[10 (ensemble-learning-using-the-voting-classifier-a28d450be64d, 2020)