ISSN 2321 3361 © 2020 IJESC

Research Article Volume 10 Issue No.5

Speech Emotion Recognition using MLP Classifier

Sanjita. B. R[1], Nipunika. A[2], Rohita Desai[3]
Department of ECM
Sreenidhi Institute of Technology and Science, Telangana, India

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. Emotion recognition is a rapidly growing research domain in recent years. Unlike humans, machines lack the abilities to perceive and show emotions. But human-computer interaction can be improved by implementing automated emotion recognition, thereby reducing the need of human intervention. In this project, basic emotions like calm, happy, fearful, disgust etc. are analyzed from emotional speech signals. We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non linearly separated. Mel-frequency cepstrum coefficients (MFCC), chroma and mel features are extracted from the speech signals and used to train the MLP classifier. For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, numpy and soundfile to analyze the speech modulations and recognize the emotion.

Keywords: Speech emotion recognition, mel cepstral coefficient, artificial neural network, multilayer perceptrons, mlp classifier, python.

I. 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. The most propitious technique for speech recognition is the neural network based approach. Artificial Neural Networks, (ANN) are biologically inspired tools for information processing. Speech recognition modeling by artificial neural networks (ANN) doesn't require any prior knowledge of speech process and this technique quickly became an attractive substitute to HMM. RNN can learn the sublunary relationship of Speech – data & is capable of modeling time dependent phonemes. The conventional neural networks of Multi- Layer Perceptron (MLP) type have been increasingly in use for speech recognition and also for various other speech processing applications. 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.

II. PRESENT WORK

Traditional emotional feature extraction was based on the analysis and comparison of all kinds of emotion characteristic parameters, selecting all the emotional characteristics with high emotional resolution for the purpose of feature extraction. The traditional approach concentrates on the analysis of the features in the speech like time construction, amplitude construction, and frequency construction, etc. Speech time construction refers to the emotion speech pronunciation differences in time. Different emotions have different types of pronunciation time periods which can be recognized and analyzed by closely examining few datasets. Such variations can also be found in the frequency and amplitude of the parameters of respective audio signals. This method, however is the basic concept of categorizing emotions from speech, it also has many drawbacks like time taken is high, judging criteria may vary, and complex programming is required. There are also many models which were proposed earlier to improve the predicting accuracy of the SERS. For example, we have Support Vector Machine (SVM), which is a classifier that mathematically computes the parameters of the audio signal to be able to predict the emotion. This model has been very successful in the domain of SER. But the main disadvantage with SVM's is that it can only classify the data into two classes; either class 1 or 2. And other disadvantages include processing time, noise leading to errors in prediction and low accuracy.

III. PROPOSED SYSTEM

1. Neural networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be- it images, sound, text or time series, must be translated. It helps us cluster and classify. You can think of them as a clustering and classification layer on top of the raw data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. Neural networks emerged as an attractive acoustic modeling approach in ASR in the late 1980s. Since then, neural networks have been used in many aspects of speech recognition such as phoneme classification, isolated word recognition, audiovisual speech recognition,

audiovisual speaker recognition and speaker adaptation. Neural networks make fewer explicit assumptions about feature statistical properties than HMMs and have several qualities making them attractive recognition models for speech recognition.

1.1. Deep feedforward and recurrent neural networks

A deep feed-forward neural network is an artificial neural network with multiple hidden layers of units between the input and output layers. DNNs can model complex non-linear relationships. Its architectures generate compositional models, where extra layers enable composition of features from lower layers, giving a huge learning capacity and thus the potential of modeling complex patterns of speech data. One fundamental principle of deep learning is to do away with hand-crafted feature engineering and to use raw features. This principle was first explored successfully in the architecture of deep auto-encoder on the "raw" spectrogram or linear filter-bank features, showing its superiority over the Mel-Cepstral features which contain a few stages of fixed transformation from spectrograms. The true "raw" features of speech, waveforms, have more recently been shown to produce excellent larger-scale speech recognition results.

1.2. Mel-Frequency Cepstral Coefficients (MFCC)

The Mel-frequency cepstral coefficients (MFCC) is one of the most popular audio feature. It is a representation of the speech signals where a feature called the cepstrum of a windowed short-time signal is derived from the FFT of that signal. Afterwards the signal goes to the frequency axis of the mel-frequency scale using a log based transform, and then de-correlated using a modified Discrete Cosine Transform. The steps to extract MFCC features are including pre-emphasis, frame blocking and windowing, FFT magnitude, Mel filterbank, log energy, and DCT. MFCC utilizes the mel-scale, which is tuned to the human's ear frequency response. Due to this, MFCC has been proven to be invaluable in the speech recognition field and has been attempted to be integrated with emotion recognition. According to Spectral audio features such as MFCC is best suited for a N-way classifier.

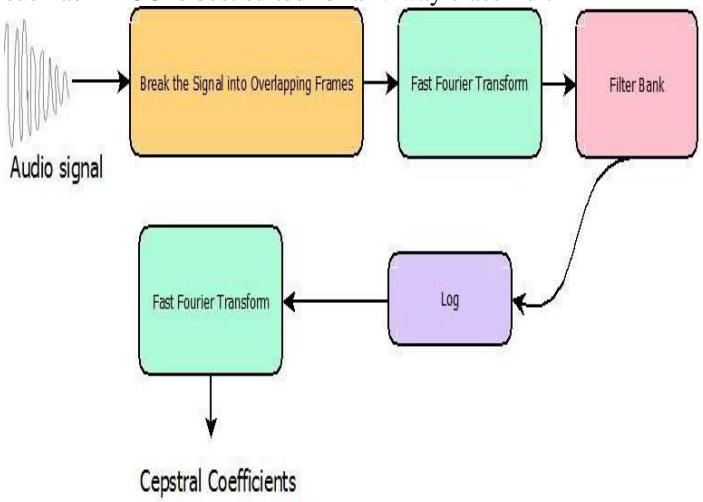

Figure.1. MFCC

1.3. Multilayer Perceptrons Classifier (MLP Classifier)

Subsequent work with multilayer perceptrons has shown that they are capable of approximating an XOR operator as well as many other non-linear functions. Multilayer perceptrons are often applied to supervised learning problems. They train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers. The number of hidden units to use is far from clear. As good a starting point as any is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units

IV. SPEECH EMOTION RECOGNITION USING MLP CLASSIFIER

In the Speech Emotion Recognition System (SER), the audio files are given as the input. The data sets travels through a number of blocks of processes which makes it executable to help for the analysis of the speech parameters. The data is preprocessed to change it to the suitable format and the respective features from the audio files are extracted using various steps such as framing, hamming, windowing, etc. This process helps in breaking down the audio files into the numerical values which represents the frequency, time, amplitude or any other such parameters which can help in the analysis of the audio files. After the extraction of the required features from the audio files, the model is trained. We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. 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. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing the performance of the model, if we enter the unknown test dataset as an input, it will retrieve the parameters and predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage which is the final result of our project.

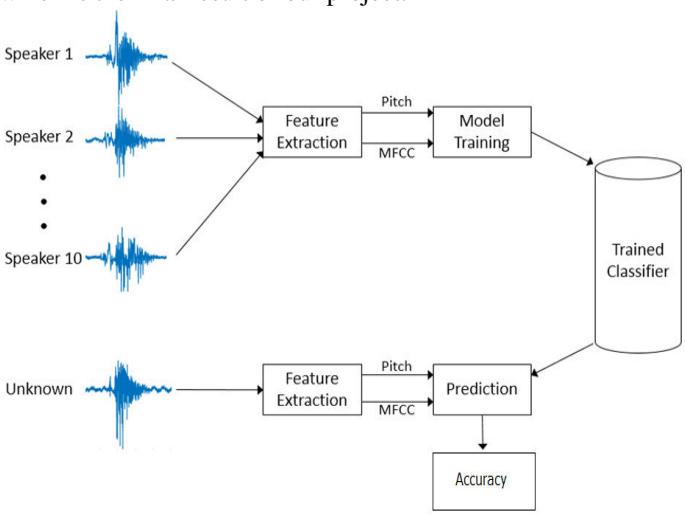

Figure.2. Speech Emotion Recognition System

V. TRAINING

Once configured, the neural network needs to be trained on your dataset.

1.Data Preparation

You must first prepare your data for training on a neural network. Data must be numerical, most common example

being real values. If you have categorical data, such as a sex attribute with the values “male” and “female”, emotion attributes such as “happy”, “sad”, “angry” etc. you can convert it to a real-valued representation which is called a one hot encoding.

2.Training

The input to the model should be the features extracted along with the emotion category that it belongs to, stored correspondingly into respective arrays so that, classifier will be able to identify the patterns, correlations and then classify the data. This training helps the model to understand, which emotions have what range of the respective features. So, when an unseen data is given as an input, it will be able to correlate and predict the emotion.

3. Prediction

Once a neural network has been trained it can be used to make various predictions. You can make predictions on test data in order to estimate the skill of the model on unseen data. You can also deploy it operationally and use it to make predictions continuously.

IV. RESULTS

Accuracy was calculated for one emotion at a time.

```
# Calculate the accuracy of our model.
accuracy = accuracy_score(y_true=y_test, y_pred=y_pred)
# Print the accuracy
print("Accuracy: {:.2f}%".format(accuracy*100))
```

Accuracy: 100.00%

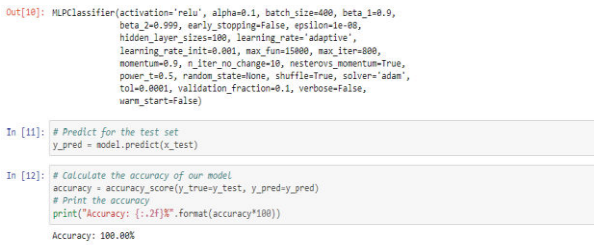

Figure.3. Results

V. ADVANTAGES OF USING MLP FOR SER

1. Provides the flexibility to work with nonlinear values
2. Less number of parameters required
3. Higher performance compared to previous systems
4. Better classification of parameters is shown.
5. Can handle missing values, model complex relationships and support multiple inputs

VI. DISADVANTAGES OF USING MLP FOR SER

1. MLPs always need fixed number of inputs to be provided for fixed number of outputs, there is a fixed mapping function between the inputs and the outputs in these feed-forward neural networks that pose a problem when a sequence of inputs is provided to the model.
2. Network must be retrained when a new emotion is added to the system

VII. CONCLUSION

This paper shows that MLPs are very powerful in classifying speech signals. Even with simplified models, a limited set of characters can be easily identified. We have obtained higher accuracies as compared to other approaches for individual emotions. The performance of a module is highly dependent on the quality of pre-processing. Mel Frequency Cepstrum Coefficients are very dependable. Every human emotion has been thoroughly studied, analyzed and the accuracy has been checked.The results obtained in this study demonstrate that speech recognition is feasible, and that MLPs can be used for any task concerning recognizing of speech and demonstrating the accuracy of each emotion present in the speech.

VIII. REFERENCES

[1].HTTP://PRACTICALCRYPTOGRAPHY.COM/MISCELLANEOUS/MACHINE-LEARNING/GUIDE-MEL-FREQUENCY-CEPSTRAL-COEFFICIENTS-MFCCS/

[2].https://www.stat.purdue.edu/~vishy/papers/VisMur02b.pdf

[3].https://www.sciencedirect.com/science/article/pii/S1110866512000345

[4].https://pdfs.semanticscholar.org/d945/29ae612b76287c7dcdccd5bf09f3c5e772af.pdf

[5].https://en.wikipedia.org/wiki/Deep_learning