Speech to Emotions Detection

Gayatri Ravjibhai Parmar
Computer Engineering, Silver Oak College of Engineering and Technology-Ahmedabad, Gujarat Technological University, Gujarat, India

Abstract **- Detection of mood and behavior by voice analysis which helps to detect the speaker's mood by the voice frequency. Here, I aim to present the mood like happy, sad and behavior detection device using machine learning and artificial intelligence which can be detect by voice analysis. Using this device, it detects the user's mood. Moreover, this device detects the frequency by trained model and algorithm. Algorithm is well trained to catch the frequency where it helps to identify the mood happy or sad of the speaker and behavior. On the other hand, behavior can be predicting in form, it can be either positive or negative. So, this device helps to prevent mental health issues and used in medical and gaming testing. Furthermore, it is easy to identify person's mood by their expression and by their actions on daily activities. But it is effective challenge to detect mood and behavior by voice frequency because rich environment affects the most. Thus, this device works as signal processing. Moreover, Machine learning, Data visualization and Data warehousing helps to discover the concepts and installation of database, while machine learning used to train the model and data visualization represent effective visualization of overall result from the data. Thus, these modules help to develop the basic and advanced concepts as well as research for this project.**

Index Terms **- Voice Detection, Signal Processing, Mood identification, Machine Learning, Visualization.**

I.INTRODUCTION

We can detect the mood of the person whom we meeting at the first time and also if we are not familiar with him or her using the analysing the tone of the voice of this particular person. The past research also gives us an information that how can we decide the sentiment or emotion of the person by using their enigmatic expression as well as by observing the eye of the respective person. Also, some research suggests that by defining the perception of the person's voice we are able to get more details about that person's state [1].

Previous studies already indicate that how can we retrieve the information from the voice. From this paper we are able to distinguish not only the perception of the person but also, we are able to get the information about type of perception like it is affirmative or the defeatist and based on that perception we are also able to predict the exact feeling behind the voice [2].

Sentiment analysis using the user's voice and by recognizing the user's voice able us to distinguish the speech signal in terms of how it is said and the semantic as well as non-lexical signal analysis is very helpful to achieve this task. By using the speech signal, we are able to retrieve the information which tells about the state of the user. Language is a well specified way to represent the affection of the user. By applying the conceptual knowledge to the language perception of the user's state become clearer. Past research shows us that by using the lexical as semantic analysis we are able to extract the voice signal of the user and are able to identify the state of the user. To bring out the emotion from the voice, vocal split of the language is implemented. Research also shows that in a civilized class, modification of the tone plays a crucial role to identify the state of the user [3].

Darvin proves that we can decide sentiment of the people by using their voice. As of now studies show that by using tone of the voice, we are able to judge the cerebral condition as well as fitness of the person. Communication way or method of the person gives indication about what he or she is delivering irrespective of their language. Many research works suggest that by using vocal analysis of the person we are able to identify the perception about the sense of the person [4]

II. CURRENT SITUATION OF THE RESEARCH

To start the research of my project first of all I go through the various research work which are already

done in this field and by briefly studying all the research I list out some major problems associated with the current research and try to solve as much as possible problems using modifying the current algorithms or by developing the new algorithms which are more efficient as well as accurate.

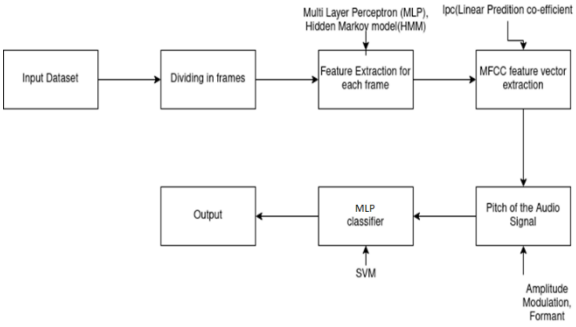

Fig: Block diagram of the system

This system covers mainly seven task which are given below.

1. Dataset selection
2. Converts data into the frames
3. Extraction of the feature from the frame
4. Feature vector extraction from the frame using MFCC
5. Pitch of the Audio signals and their property
6. MLP classifier for the classification
7. Output generation using python

Apart from the above seven modules some modules are remain same for most of the system like dataset, frame conversion, and the output displaying. So, I focused on the major modules which needs improvement and these are Feature extraction as well as classification and identification of the property for the audio signals. I figure out the changes needed in these modules and mentioned a comparison with the different research papers.

A. Extraction Of the Feature

I found that in most of the work which already done used a Multilayer perceptron model for the purpose of the feature extraction and to analyze the problems associated with this model I go through the various research work which are previously conducted and found some problems associated with this model which are given below.

- In MLP number of initial parameter increase in multiplicative order as we increases the layer of the model from one to two and so, on which generates highly redundant data and process becomes more complex at the execution level [5]
- In MLP models information regarding geographical data is in spatial form which is hard to process [6].

I also studied the problem occurred while using Hidden Markov model from the available research paper and one common problem in all research paper I found is given below.

- Hidden Markov Model uses numbers of parameters which are unstructured in the nature so, sytem is unable to define the associated dependencies between the different hidden states [7]

Finally, I reach at the conclusion that to overcome these problems I have to design a unique solution which resolve the associated problem with both model and for that purpose I applied Discrete Fourier Transform as well as Mel Frequency Wrapping and working of this system is illustrate in the figure given below.

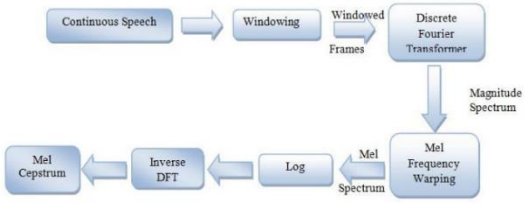

Fig: Discrete Fourier Transform for the Feature Extraction Process

There are some advantages of using Discrete Fourier Transform and Mel frequency wrapping which are given below. I listed out some advantages associated with using Discrete Fourier Transformation as well as Mel Frequency Wrapping which I given below

- The major advantage of using the Fourier transform is that it provides loss less transformation of the information means we can utilize whole signal during the process. The second advantage of using fourier transformation is that we can utilize all the information related to the signal like their amplitude, phase as well as the nature of the signal and we can utilize these information to transform the signal into the domain of the frequency [8].

B. Feature Vector Extraction Using MFCC

After the process of the feature extraction, I studied the various models used for the coding of the speech as well as filter designing. To select better coefficient for the filtering technique I studied the coefficient used for

the prediction in a Linear manner and during the study I found some major drawbacks associated with this technique which I listed below.

- Due to quality reduction in the bitrates of the speech signal, quality of the signal is drastically reduce [9].
- For the purpose of the long distance transmission this technique generates lossy compression which is not feasible when distance is longer [10].

To resolve these problems, I go through the technique called MFCC Feature vector extraction which comes with numerous advantages which I mentioned below.

- This technique uses a Quantum neural Network which provides a parallel computation methods so, that the training time for the data is reduced [11].
- This technique uses a IPSOQNN model which provides a very fast recognition for the signal [12] .
- When we train QNN using the IPSO the convergance rate is more higher and provides an accurate prediction by utilizing all the capabilities of global optimization [13].
- MFCC concentrates on the major characterisitcs of the phones available in the speech and it also reduces the complexities of the calculation [14]

C. Pitch Of the Audio Signal And Their Property

By using Amplitude modulation technique for the extracting the property of the audio signal there are some major problems while extracting the frequency which I listed below.

- Amplitude modulation is works on the higher bandwidth so, we require higher bandwidth then the original signal [15].
- Detector used for the amplitude modulation technique is very sensitive about the nose and when there is a higher noise available in the audio signal then it is much more difficult to recognize the original signal [16].

After researching on various technique used for the voice detection finally, I select the Pitch detection method which comes with various advantages over other techniques which are given below.

- This technique provides a distinguish between the human voice and the voice of the musical instrument [17]
- By this technique sytem is able to identify the gender of the person [18].
- By using this technique system is able to track the timestamp at when voice is recorded and it is easy to identify that it is morning or noon or the evening at the time of the recording [19].
- By using the pitch of the signal system can easily identify the approximate age of the person who recorded the audio [20].

D. MLP Classifier for The Classification

To study how the classification technique works I go through the various research papers and study the classifier named as SVM or Support Vector Machine and during my research work I found some major drawbacks of the SVM which I listed below.

- SVM algorithm is not suitable for the larger dataset [21].
- There is an overlapping in the target class when audio signal consist more noise [22].
- If there are more features are availble as compare to the training samples then this algorithm is unable to provide more accurate results [23].

So, I prefer to go with the MLP (Multi-Layer Perceptron) for the classification purpose which provides all the solution needed for the system and some advantages of using MLP is given below

- This algorithm is efficient even if the dataset is large in the size [24].
- This algorithm is able to manage thousands of data without deleting single data from the database [25].

This algorithm provides imporance parameter to each and every variable in the classification phase [26].

III. THE PROBLEM AREAS ADDRESSED BY THIS PROJECT

Here by studying various research work I found some major problems in the current methodology of mood detection and behaviour analysis by using voice, which I listed below.

- In most of the papers I found that authors used a MLP for the feature extraction which comes with the disadvantage that when number of parameters grow then it will increase the redundancy.
- In most of the papers, authors used amplitude modulation technique to define the property of the

audio signal but it consumes twice the bandwidth than the original signal which is a costly process and computation becomes more complex.

- To classify the data most of the authors used a SVM classifier which is suitable for the data with no or less noise. But when level of noise increases it gives an inappropriate result and efficiency of the algorithm is drastically reduced.

IV. OBJECTIVES OF THE RESEARCH

There are many objectives covered in the research and some major one is given below.

1. Detection of Voice
2. Feature Extraction of the voice
3. Designing an algorithm
4. Database creation
5. Classification of voice using classifier
6. Analysis of the speech
7. Defining spectrum for the speech
8. Reliability and the mobility of the application

I briefly explain all the objectives of the system to better understand the process in effective manner.

1. Detection of Voice: The mood of the person is detected by using detection components which are able to recognize the voice of the person more effectively. Using this components system is able to segment the speech into the different level of frequency and also allows the analysis of the feature extraction.
2. Feature Extraction of voice: After the detection of the voice this module extracts the features from the voice of the person and provides the different range of the frequency from that voice by using the extraction algorithm.
3. Optimization Algorithm: This algorithm is useful to optimize the frequency of the voice by using signal processor. By using various kinds of classification, it gives an accurate frequency of the voice. There are number of functions are used to detect the mood as well as frequency of the audio signal and by using this algorithm we are able to get actual values as compared to the predicted values.
4. Database Creation: Database is used to store the clips of the various voice tone and to perform this task we have to use different kinds of libraries as well as tools needed to setup the database and by setting up the database, we are able to take decision for the improvement in the execution.
5. Classification of voice using Classifier: To classify the various voice over the tone of the audio signal classification algorithms are helpful. By using this algorithm, we are able to extract the mood of the person at a various frequency level and based on that frequency we are also able to judge the behavior of the person.
6. Analysis of the Speech: Speech signal plays a vital role for the identification of the moods such as happy or sad where speech frequency is useful for the deriving the feature associated with that frequency and based on that feature algorithm help the system to select the appropriate mood of the speaker.
7. Spectrum for the speech: This device is useful for collecting the data and the experiment performed and finds the actual frequency range associated with that data. After that it predict the mood which is closest to the actual values and finally discover the mood of the speaker.

V. IMPLEMENTAION OF WORK

A. Data Set

In this research work I used a dataset named as RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) which consist approximately 7356 files which are rated by the individual on the basis of emotional validity, intensity as well as based on the genuineness.

B. Data Preparation

To prepare the data we have to create a function which is able to extract the emotion label as well as the gender label from the available files. For that purpose, we used a function called glob () which assign all the path names available in the file to the model.

C. Extraction Of the Feature

Here I used a Mel Spectrogram which is available in the library named as Librosa. It obtains the value in the form of Log-Mel Spectrogram for each audio file given to the model and then it averages all the values of the spectrogram. Finally, these averaged values are converted into the new data frames. Here, I put code used to generate the feature using the Librosa library.

```
def extract_feature(file_name, mfcc, chroma, mel):
```

```
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
            stft=np.abs(librosa.stft(X))
        result=np.array([])
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X,
sr=sample_rate, n_mfcc=40).T, axis=0)
            result=np.hstack((result, mfccs))
        if chroma:
chroma=np.mean(librosa.feature.chroma_stft(S=stft,
sr=sample_rate).T,axis=0)
            result=np.hstack((result, chroma))
if mel:

mel=np.mean(librosa.feature.melspectrogram(X,
sr=sample_rate).T,axis=0)
            result=np.hstack((result, mel))
return result
```

I calculate the absolute value from the element using the Librosa library. After that I applied MFCC to represent the information regarding the sound. Here I used 13 coefficients of MFCC to use it as a feature and it represented as an envelope of the spectra. I discard the higher dimensions of the feature to make the spectra simpler. MFCC is able to recognize the different kinds of phonemes using the difference between the features.

After that the outcome is transferred to the hstack () function which is an array of one-dimensional data to store the outcome in a single dimension.

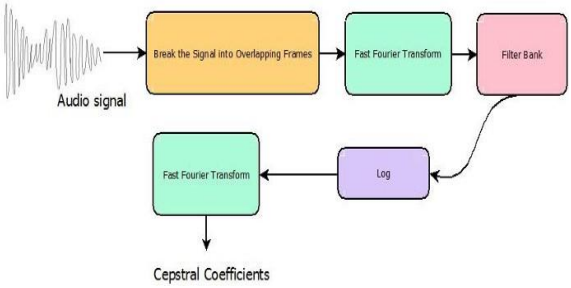

Fig: Cepstral Coefficient from the Audio Signal

D. Signals

Signal can be defined as a variation in a frequency with respect to the time. Here we have to collect various samples of an air pressure with respect to the time and have to measure rate of the sample data per second and it is commonly of 44.1 KHZ.

E. The Fourier Transform

Using Fourier Transform we are able to convert the various amplitudes collected at the time of sampling to the individual frequencies by using mathematical formulas. It means that using Fourier Transform we are able to convert signal from the frequency domain to the time domain and the output of this formulas is known as a spectrum.

F. Spectrogram

The Frequency of the signal is frequently changing over a time in the audio or a speech signal and these signals are known as a non-periodic signal. Using spectrum, we are able to represent the signals over a time and for this purpose a technique called Short-Time Fourier Transform is used. The main task of STFT is to compute several spectrums by performing the FTT on a different window segment of the signal. The output of this technique is known as the Spectrogram. By using spectrogram, we can be able to visualize the frequency of the signal over a time.

G. The Male Scale

Mel scale is defined as the unit of the pitch which represent the equal distances in pitch sounded equally distant to the listener. The reference scale between the scale and the normal frequency is selected as a 1000 mels to a 1000 Hz tone. Let us take an example to understand the Mel-scale. If any listener listens a frequency ranges from 600 Hz to the 650 Hz, then it is easy to tell the difference between two frequencies for him or her. But if he or she is listening the frequency of a scale like between 12000Hz to 12500 Hz then it is somewhat difficult for him or her to find the exact difference between the frequencies. The frequencies which are converted from frequencies to the Mel scale is known as the Mel-Spectrogram.

H. Pre-Processing of the Data

To pre-process the data following steps are performed which are given below.

1. Train and Test-Split the Data
2. Normalization of the data to improve the stability and the performance of the data
3. Perform Encoding on a target variable

In this process we split the data into various columns based on the category of that data and numbers of column are same as the number of categories present in the data. Each column is indicated by either zero or

the one based on the which corresponding column has been placed.

I. Classification Stage

Now MLPClassifier has an internal neural network for the purpose of classification. This is a feedforward ANN model.

In this stage I select the Multi-Layer Perceptron Classifier as an internal network to classify the data and this model is referred as a Feedforward ANN model

J. How To Initialize the Multi-Layer Perceptron Classifier?

Here I provided sample code by which we are able to initialize the MLP classifier.

model=MLPClassifier (alpha=0.01, batch_size=256, hidden_layer_sizes= (300,), learning_rate='adaptive')

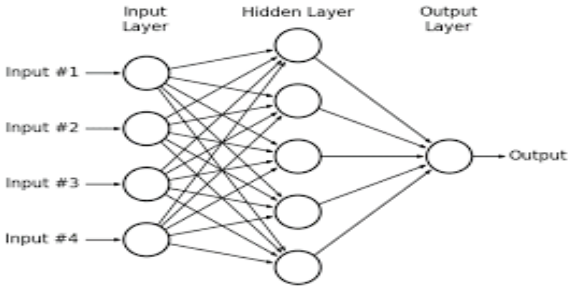

Fig : MLP with Hidden Layer

Here I calculate the accuracy_score () function to find out how accurate the model is and this function is imported from the SKLearn Library. Here I rounding the accuracy up-to two decimal points.

VI. CONCLUSION

By go through the various research work and finding the research gap associated with the past work, I am able to recognize the current problem in the system and also able to design the solution for my research work. By changing the existing algorithm and by improving the performance of the algorithm I am able to get the desired result with better accuracy.

REFERENCES

[1] P. E. G. B. Pascal Belin, "Understanding Voice Perception," British Journal of Psychology, vol. 102, no. 4, pp. 711-725, 2011.

[2] J. Lundergren, "The psychological impact of a person's voice," The Heart of Tech, 2015.

[3] S. V. C. P. M. Kalamani, "Feature selection algorithms for automatic speech recognition," 2014.

[4] G. Priest, "Charles Darwin's Theory of Moral Sentiments," Journal of the History of Ideas, vol. 78, no. 4, pp. 571-593, 2017.

[5] B. A. Hamid Moghaddasi, "Study on the Efficiency of a Multi-layer Perceptron Neural Network Based on the Number of Hidden Layers and Nodes for Diagnosing Coronary- Artery Disease," Jentashapir Journal of Health Research, pp. 4-5, 2017.

[6] HMühlenbein, "Limitations of multi-layer perceptron networks - steps towards genetic neural networks," ELSEVIER, vol. 14, no. 3, pp. 249-260, 1990.

[7] P. T. Chandralika Chakraborty, "Issues and Limitations of HMM in Speech Processing: A Survey," International Journal of Computer Applications, vol. 141, no. 7, pp. 13-17, 2016.

[8] H. X. C. L. Jiufei Luo, "A comparative study of interpolation discrete fourier transform algorithms under strong noise," IEEE, p. 158, 2016.

[9] B. S. Atal, "The history of linear prediction," IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 154-161, 2006.

[10] B. S. Atal, "The history of linear prediction," IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 154-161, 2006.

[11] S. M. Md. Afzal Hossan, "A novel approach for MFCC feature extraction," 2011.

[12] F. G. E.A. Grimaldi, "PSO as an effective learning algorithm for neural network applications," 2004.

[13] F. G. E.A. Grimaldi, "PSO as an effective learning algorithm for neural network applications," 2004.

[14] C.-f. C. Wei Han, "An efficient MFCC extraction method in speech recognition," 2006.

[15] N. Paraouty, "Interactions between amplitude modulation and frequency modulation processing: Effects of age and hearing loss," The Journal of the Acoustical Society of America, vol. 140, p. 121, 2016.

[16] S. A. S. Alkadhim, "AMPLITUDE MODULATED SIGNALS: Generation Methods," ResearchGate, p. 17, 2020.

[17] D. Gerhard, "Pitch Extraction and Fundamental Frequency: History and Current Techniques," ResearchGate, p. 7, 2003.
[18] M. J. Hernández, "A tutorial to extract the pitch in speech signals using autocorrelation," ResearchGate, pp. 10-11, 2016.
[19] W. J. Hess, "Pitch Determination of Speech Signals — A Survey," 2016.
[20] I. A. L. A. S. Kolokolov, "Measuring the Pitch of a Speech Signal Using the Autocorrelation Function," Springer, vol. 80, pp. 317-323, 2019.
[21] S. M. A. Sasan Karamizadeh, "Advantage and Drawback of Support Vector Machine Functionality," 2014.
[22] Y. S. Yingjie Tian, "Recent advances on support vector machines research," 2012.
[23] A. D. A. S. Sourish Ghosh, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019.
[24] V. E. B. Popescu Marius, "Multilayer perceptron and neural networks," 2009.
[25] A. S. R. Arti Rana, "Application of Multi-Layer (Perceptron) Artificial Neural Network in the Diagnosis System: A Systematic Review," IEEE Xplore, p. 216, 2018.
[26] A. F. A. Ahad, "Speech recognition using multilayer perceptron," 2002.