|| Volume 6 || Issue 5 || May 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

SPEECH BASED EMOTION RECOGNITION

Amit Kumar Mishra[1], Karan Gupta[2], Sayam Bagrecha[3], Prof. Pravin Jangid[4]

Student, Dept of Computer Engineering, Shree L R Tiwari College of Engineering, Mira Road, Thane, Maharashtra, India [1 2 3]

Professor, Dept of Computer Engineering, Shree L R Tiwari College of Engineering, Mira Road, Thane, Maharashtra, India [4]

amitkumar.mishra@slrtce.in[1], krnxxy@gmail.com[2], bagrechasayam1998@gmail.com[3], pravin.jangid@slrtce.in[4]

---------- ***----------

Abstract: - Recognizing the emotions from the speech is essential for a natural connection between man and machine. In speech emotion recognition, emotional state of a speaker is extracted from his or her speech. A speech emotion recognition system is made to detect the emotion through the audio speech signals. The efficiency of emotion recognition system depends on type of features extracted and classifier used for detection of emotions. In this paper we have taken the audio features of MFCC, MEL, and Chroma. to recognize the emotion of speaker like sad, neutral, happy, angry. With the model proposed of MLP classifier we have achieved accuracy of 78% approximately.

Keywords: - *MLP Classifier, MFCC, Chroma, MEL.*

----------***----------

I INTRODUCTION

The importance of speech recognition is increasingly popular with improving user experience and engaging with Voice User Interfaces (VUIs). Personal speech is an effective means of communication that incorporates different aspects of language and measurement such as gender, age, language, highlighting, and mood. The sound waves that contain the human voice are different from all other sound-producing creatures because each wave has a different frequency. Voice-based gender identification has been a challenging task for voice and audio analysts.

There exist a set of features that is used for recognizing the voice. The human voice should be converted from the analogue to the digital form to extract useful features and then to construct classification models. Among the most common features utilized for voice recognition are mel-scaled power spectrogram (Mel), mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma). By getting the extracted features combined with the emotions label as a form of a training set, ML techniques are used to build a high-quality model for recognizing the voice emotions. The robustness and effectiveness of classifiers are determined by the quality of features that depend on a training set employing machine learning (ML) techniques.

1. FEATURE EXTRACTION

The acoustic characteristic of the speech signal is Feature. A small amount of data from the speech signal is extracted to analyse the signal without disturbing its acoustic properties [6][7][8]. This extracted signal is used for training and testing phases. We have used MFCC, Chroma and Mel Frequency Cepstrum as speech features by converting the raw wav form.

1.1 MFCC (Mel-Frequency Cepstral Coefficients):

Any sound produced by humans is determined by the structure of their vocal cords. If the structure is properly determined, any sound produced can be accurately represented. The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC (Mel-frequency cepstrum) accurately represents this envelope [4].

In MFCC the first 13 coefficients (the lower dimensions) of are taken as features as they represent the envelopes of spectra. And the eliminated higher dimensions express the spectral details. For different scenarios, envelopes are enough to represent the difference, so we can recognize various phonemes through MFCC. Below is the flow chart of MFCC in Figure (a).

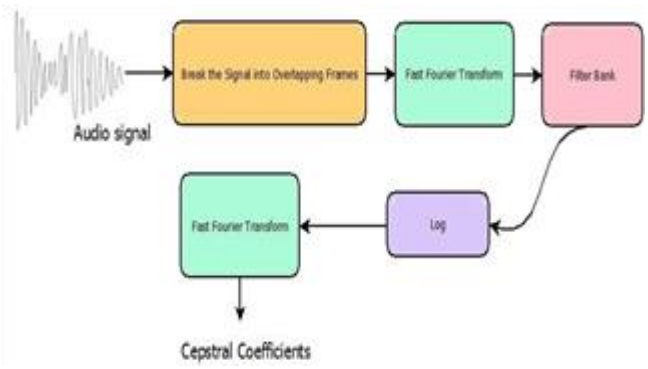

Figure (a)

1.2 Chroma Features:

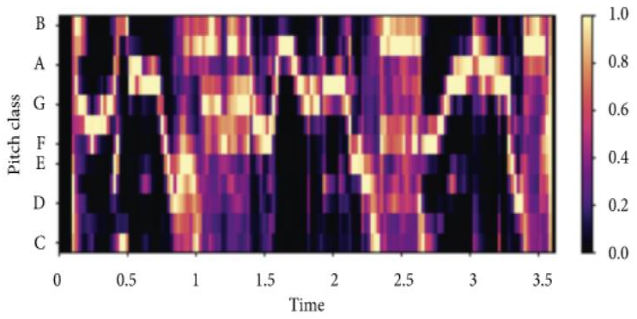

Figure (b)

Chroma-STFT incorporates a chromogram from a waveform or energy spectrogram, as shown in Figure (b). Chroma features are powerful representatives of the sound of music where the whole screen is displayed in 12 bins representing 12 different semitones (or Chroma) octave music.

In music, the word chroma feature or chromogram is closely related to the twelve distinct categories. Chroma-based features, also called "pitch class profiles", are a powerful tool for analysing music whose music can be logically categorized (usually twelve sections) and their tuning is about the same size. Important properties of chroma elements is that they capture harmonic and musical features, while being strong in transformation of timbre and instrumentation. The two main features of chroma are listed below:

(a) Chroma vector:

Representation of twelve spectral energy objects in which the bins represent 12 equal pitch classes of western genre music (space of the tones).

(b) Chroma Deviation:

The normal deviation of the 12 coefficients of chroma.

1.3 MEL

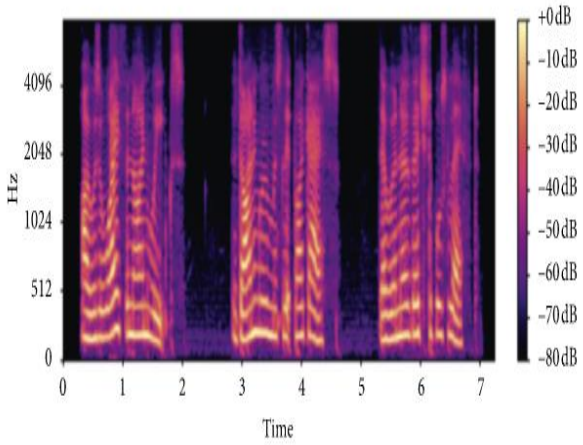

Figure (c)

The Mel Spectrogram is the result of the following pipeline:

1. Separate to windows: Sample input in windows size n_fft=2048, making hops of size hop_length=512 each time to sample the next window.
2. Compute FFT (Fast Fourier Transform) for each window to be converted from the time zone to the frequency domain.
3. Make a Mel scale: Take all the frequency, and divide it into n_mels = equally placed waves of 128.
4. Create a Spectrogram In each window, indicate the size of the signal in its parts, corresponding to the size of the mel scale.

II PROBLEM STATEMENT

As human speech it is one of the most natural ways of expressing oneself. We rely so much on it that we realize how important it is when we use other forms of communication such as emails and text messages where we often use emoji to express feelings associated with messages. Since emotions play an important role in communication, similar discovery and analysis are very important in today's digital world of remote communication. Emotional discovery is a challenging task, because emotions are inferior. There is no general consensus on how to measure or categorize oneself.

III PROBLEM OBJECTIVE

The purpose of this paper is to discover the feelings written by the speaker while he is speaking. For example, speech produced in a state of fear, anger, or excitement tends to be loud and swift, a high and wide range of pitch, and emotions such as sadness or fatigue create slurred and low speech. Discovering a person's emotions through a voice pattern and speech pattern analysis has many applications such as better facilitating human-machine communication. Basically, we present models of speech segregation that are expressed in terms of deep neural networks [1], Multilayer Perceptron Separation (MLP) based on acoustic factors such as Mel Frequency Cepstral Coefficient (MFCC), chroma and MEL.

IV LITERATURE REVIEW

Speech structures refers to the variability of speech pronunciation in time. When people express different emotions, the timing of the speech is different. Mainly in two aspects, one is the length of continuous pronunciation time, the other is the average rate of pronunciation. One is the length of continuous pronunciation time and the next one is the average rate of pronunciation.

The study of Z.Li [10] showed that the different pronunciations of emotions differ in the length of pronunciation and the speed of pronunciation. Compared with the length of the quiet call period, the call time for joy, anger and surprise is

greatly shortened. But compared to the quiet call time, the length of the sad call is longer. Compared with the silent call rate, the sad pronunciation rate is slower, while the excitement, anger, and surprise levels of surprise are relatively fast.

V IMPLEMENTED SYSTEM

5.1. Database Preparation: Here, we download and modify the database to be ready for release.

5.2. Dataset Upload: This process is about uploading data to Python which involves extracting audio features, such as getting various features such as power, pitch and vocal tract to stop in the speech signal, we will use the librosa library to do that.

5.3. Model Training: After editing and uploading the database, we simply train it in the appropriate sklearn model.

5.4. Model Testing: Measuring how well our model works.

In the model I did a grid search on the MLP Classifier to find the best hyper parameters. Thus, we find a single-layer neural network consisting of 300 units, a batch size of 256, 500 iterations and a variable learning rate [5].

VI DATA SET

Ryerson audio and video viewing database and Song (RAVDESS dataset) is a recorded database English sentences and songs in various emotional states [2] [3]. Experimental, we only used RAVDESS speech data. The recorded sentences were "Hi, nice to meet you", "Dogs sitting by the door". Twenty-four actors (12 males, 12 females) with North American accents participated in the recording. Total number of details composed of 1440 files. The RAVDESS database is divided into Eight categories of emotions: anger, disgust, fear, happiness, neutrality, sadness, surprise, and calmness

VII DEEP LEARNING

Scikit-learn [9] [11] is a Python component that integrates a variety of technological learning processes between intermediate and unregulated problems. This package focuses on bringing machine learning to non-professionals using language that is understandable for general purposes. Emphasis is placed on the ease of use, functionality, documentation, and API flexibility. It is less dependent and is distributed under a simplified BSD license, promoting its use in both academic and commercial applications.

VIII. CONCLUSIONS

In this paper, we have extracted the speech features like MFCC, Chroma, MEL that can be used for recognition of emotional states, and implemented the model that confirmed the test accuracy to be approximately 75%. In future we would like to implement more features on datasets in order to get the better speech emotion recognition rate.

IX RESULT

Following is the output between the actual values and the predicted values shown in figure 2.a

Out[75]:

	actualvalues	predictedvalues
58	male_fearful	male_happy
59	male_fearful	male_fearful
60	male_fearful	male_fearful
61	male_fearful	male_fearful
62	male_sad	male_sad
63	male_fearful	male_fearful
64	male_happy	male_happy
65	female_angry	female_angry
66	female_angry	female_fearful
67	male_angry	male_angry

Figure 2.a

The wave form of the wav audio signal is given below in figure 2.b

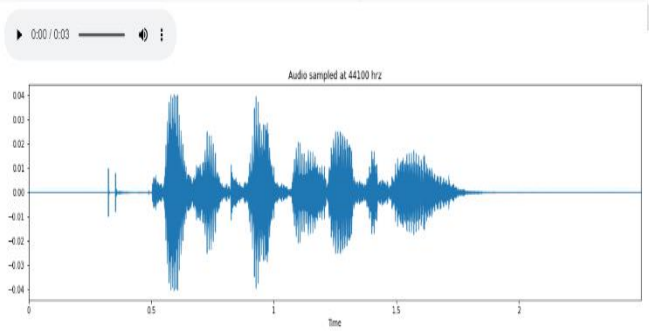

Figure 2.b

The extracted mfcc plot is shown in figure 2.c

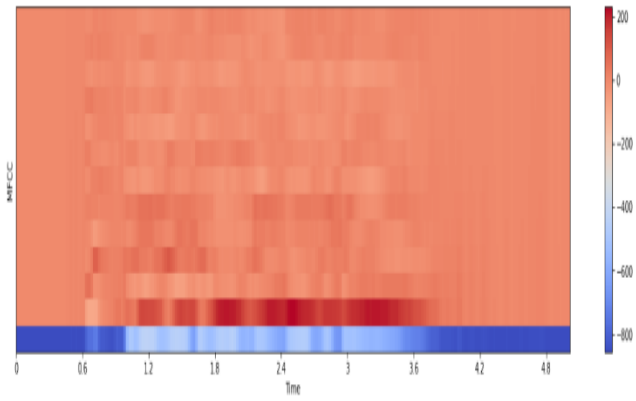

Figure 2.c

Following is the predicted result of the input voice, shown in figure 2.d

Figure 2.d

REFERENCES

[1] D Jian Wang, Zhiyan Han,"Research on Speech Emotion Recognition Technology based on Deep and Shallow Neural Network", 2019 Chinese Control Conference (CCC), 27-30 July 2019.

[2] Kyong Hee Lee; Hyun Kyun Choi; Byung Tae Jang; Do Hyun Kim," A Study on Speech Emotion Recognition Using a Deep Neural Network", 2019 International Conference on Information and Communication Technology Convergence (ICTC), 16-18 Oct. 2019.

[3] Panagiotis Tzirakis; Jiehao Zhang; Bjorn W. Schuller," End-to-End Speech Emotion Recognition Using Deep Neural Networks", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 15-20 April 2018.

[4] Utkarsh Garg; Sachin Agarwal; Shubham Gupta; Ravi Dutt; Dinesh Singh," Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma", 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), 25-26 Sept. 2020.

[5] Michael Neumann; Ngoc Thang Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019.

[6] M. S. Likitha; Sri Raksha R. Gupta; K. Hasitha; A. Upendra Raju," Speech based human emotion recognition using MFCC", 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 22-24 March 2017.

[7] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases.

[8] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li, "Speech emotion recognition using fourier parameters," IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69–75, January–March 2015

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, "Scikit-learn: Machine Learning in Python", 12(85):2825−2830, 2011.

[10] Z. Li, "A study on emotional feature analysis and recognition in speech signal," Journal of China Institute of Communications, vol. 21, no. 10, pp. 18–24, 2000.

[11] https://scikit-learn.org

****