RESEARCH ARTICLE OPEN ACCESS

Speech Emotion Recognition Using Machine Learning

Dr. C sunitha ram [1], B. Varshini [2], B. Varsha [3]

[1] Assistant Professor, Department of Computer Science and Engineering, SCSVMV - Kanchipuram
[2] B.E Graduate (IV year), Department of Computer Science and Engineering, SCSVMV - Kanchipuram
[3] B.E Graduate (IV year), Department of Computer Science and Engineering, SCSVMV - Kanchipuram

ABSTRACT

Speech emotion recognition (SER) is an attempt to recognize emotional states related to human emotions in language. It is true that voices often express basic emotions through tone and pitch. The popularity of emotions has been a rapidly developing area of research in recent years. Unlike humans, machines cannot understand and express emotions. However, human-computer interaction can be improved by forcing automatic emotion recognition, reducing the need for human intervention. In this project, key emotions such as calmness, joy, fear, and disgust are analyzed from emotional audio signals. Use machine learning strategies such as the Multilayer Perceptron Classifier (MLP Classifier). It is used to classify given information into appropriate groups that can be nonlinearly separated. Mel Frequency Cepstrum Coefficient (MFCC), chroma, and Mel features are extracted from the audio signal and used to train the MLP classifier. To achieve this goal, use Python libraries such as Librosa, sklearn, pyaudio, NumPy, and sound report to examine voice modulation and understand emotions.

Keywords :- Speech Emotion Recognition, MFCC, MLP, Audio, Mel.

I. INTRODUCTION

In the natural human computer interaction (HCI), speech emotional Recognition (SER) has become one of the most important advertising strategies, in which the customers play an important role and recommend the perfect product or help him, so the demand for the product or company increases. Humans have the natural ability to use all of their senses to give maximum attention to the message received. Feeling emotions is natural for humans, but for machines, it is a very difficult business.

The Multilayer Perceptron (MLP)Conventional neural networks are increasingly being used for speech and also for various speech processing applications. Voice ubiquitylation is a method of changing a signal, received by a microphone, into a set of characters. They also can function the entrance to similarly processing to obtain speech understanding, a topic protected in section. As we know, speech recognition plays responsibilities that comparable with the human brain.

A. Objective:

The main objective of this project is to build a model to recognize emotion from speech using the Librosa and sklearn libraries and the RAVDESS data set. so why not have an emotion detector that will gauge emotion and, in the future, recommend you different things based on your mood this can be used by the different industries to offer different services like marketing companies suggesting you buy products based on your emotions, the automotive industry can detect the person emotions and adjust the speed of autonomous cars are required to keep away from any collisions.

B. Project scope:

Libraries. Librosa, soundfile and sklearn libraries (specifically) are used to build models using MLPClassifier.

From there, it then uses the speech emotion recognition API to translate speech into text. It then analyzes the audio signal. Then the audio files are masked and cleaned.

It then loads the data, extracts features from the data, and splits the data set into a training set and a test set. Then, we'll initialize an MLP Classifier and train the model. And finally, we will get the output as emotion from our live demo.

II. LITERATURE REVIEW

Emotion recognition from audio signals is a significant however tough element of human computer interaction (HCI). In the literature on SER, many techniques have been used to extract emotions from signals, including many established speech analysis and classification techniques. Recently, one deep learning method has been proposed. An alternative to the traditional SER method. This proposed framework provides an overview and description of deep learning methods. Recent literature uses these language-based emotion recognition techniques. The overview includes the database used. Extracted the sentiment, recognition and limitations related thereto.

TABLE I
LITERATURE SURVEY

S. no	**Author & Journal name**	**Methodologies**	**Advantages/ Disadvantages**
1	Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama (2020) [5]- Speech Emotion Recognition using Neural Network and MLP Classifier	In this paper the emotions in the speech are predicted using Neural Network,	The positive emotion recognition rate is higher than other approaches, but

		Multilayer Perceptron Classifier is used for the classification of emotions	neutral and negative emotions are often confused with each other
2	L J. At Dominguez Jimenez, KC Campo Landlines (2020) [6]- Biomedical signal processing and control	One class in one neural network	It is a complex process
3	Chaitanya Singla, Sukhdev Singh, Monika Pathak (2020) [7]- Automatic Audio Based Emotion Recognition System	They used a Commonsense effect-based approach (Real-world Knowledge concept models),	The System obtained an accuracy of around 70% for various audio files from the various voice sample
4	Navya Damodar, Vani H Y, Anusuya M A (2019) [8]- International Journal of Innovative Technology and Exploring Engineering	Voice Emotion Recognition using CNN and Decision Tree	this model can be improved by making the dataset times three times the original size to achieve the greater accuracy

Chaitanya Singla, Sukhdev Singh, Monika Pathak Proposed Automatic Audio Based Emotion Recognition System: They used Common Sense Affect based approach (Real-world Knowledge concept models), The System extracted various audio files from the various voice sample (2020).[5].
Navya Damodar, Vani H Y, Anusuya M a Proposed International Journal of Innovative Technology and Exploring Engineering, used Voice Emotion Recognition using CNN and Decision Tree, this model can be improved by making the dataset times three times the original size (2019) [6].
Jerry J, Aparna K, Shreya Ram, S. Rama proposed Speech Emotion Recognition using Neural Network and MLP Classifier, in this paper the emotions in the speech are predicted using Neural Network, Multilayer Perceptron Classifier is used for the classification of emotions, Emotions Recognition rate is higher than other approaches, but neutral emotions are often confused with each other (2020) [7].
L J. At Dominguez Jimenez, KC Campo Landlines proposed biomedical signal processing and control, they used a layer in a neural network 2020[8].

III. MODULE DESCRIPTION

A. Modules:

1. ***Data**:*

Data collection is one of the most important parts of building a machine learning model. Because no matter how well designed it won't learn anything useful if the training data is invalid.

2. ***RAVDESS***:

The dataset consists of approximately 150 audio file entries from 24 actors. 12 boys and 12 girls, where the actors made short recordings of different emotions. Each audio file is named so that the 7th letter matches the different emotional levels they represent.
Where these record short audios in emotions.
1=Neutral ,2=Calm, 3=Happy, 4=Sad, 5=Angry, 6=Fearful, 7=Disgust, 8=Surprise.

3. ***Data Pre-processing*:**

Data pre-processing in machine learning is an important step in improving data quality to aid in meaningful data extraction. insights from data. Data pre-processing refers to the technique of cleaning and organizing raw data to make it suitable for Building and train a machine learning model.

4. ***Speech to text translation using speech emotion recognition using API*:**

We initially tested the sound by translating it into text mode using the Speech Emotion API to find out what it was Application of machine learning techniques for speech emotion recognition

5. ***Analysing the file*:**

An audio signal can be analyzed in a variety of ways, depending on the type of information desired from the signal. Our step is to test the audio files by plotting waveforms and spectra to see sample audio files.

6. ***Masking and cleaning the audio file*:**

The next step is to clean the audio files by lowering the sample rate and removing the unwanted noise around the raw audio via masking.

B. *Feature Extraction*:

The following step includes extracting the capabilities from the audio documents to assist our mode and research among those audio documents. For function extraction, we employ the Librosa library in python which is one of the libraries used for audio evaluation additionally there are labels of feelings defined.
Also, there are labels of EMOTIONS defined, while the clean Dataset is loading with the calling of Feature Extraction process, each audio is grouped into the labels predefined

1. ***Librosa***:

Librosa is used for audio and music analysis, it is a python library. It features a flatter package layout, standardization of interfaces and names, backward compatibility, modular functionality, and readable code.

2. ***Extracted features*:**

The first step in any automatic speech recognition system is to extract a feature, i.e., identify the components of the audio signal. good at identifying linguistic content and removing all other information-carrying elements such as background noise, emotions, etc. Therefore, the extracted features can be:

1. MFCC: Frequency Factor Depression (MFCC) is a widely used feature in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980s and have been modern ever since.

2. Chroma: The chroma function is a descriptor representing the tonal content of a musical audio signal in a condensed form. Therefore, the chroma feature can be considered as an important prerequisite for high-level semantic analysis such as chord recognition or approximation of harmonic similarity. The better quality of the extracted chrominance function allows for much better results in these high-level tasks.

3. Mel: The Mel scale relates the perceived frequency or pitch of a pure sound to its actual measured frequency. Humans are much better to distinguish small changes in pitch at lower frequencies than at high frequencies. The integration of this scale makes our functions more in line with what people hear.

C. Building the model:

Since the challenge is a category problem, multi-layer perceptron appears the plain choice. we pick out this model to expect the proper emotions. This classifier connects to a neural network. Unlike different classification algorithms inclusive of support vectors or naive Bayes classifiers, the mlp classifier is based on an underlying neural network to carry out the task of classification.

D. MLP Classifier:

An MLP Layer consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. In the MLP classifier, an input layer is passed to some hidden layers which are used for the implementation of abstraction, and then result is a process and you can see predicted emotion. MLP classifier relies on an underlying Neural Network to perform classification. It can implement an MLP (Multilayer Perceptron) algorithm and trains the neural network using Backpropagation. Building the MLP (Multilayer Perceptron) Classifier involves the following steps:

1. Initialize the MLP classifier by setting and initializing the required parameters.
2. The data is passed to the neural network to train it.
3. The trained network is used to predict the output.
4. Then calculate the accuracy of the prediction.

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.

E. Predictions:

After fitting the model, test it by predicting emotions against the test data. After splitting of training and test data to save the model. The model is loaded again to predict the test data and stores its results in a CSV file with its labels to map the individual results to its wav filename.

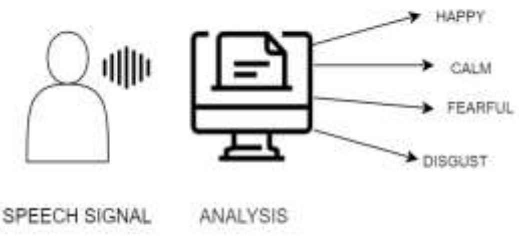
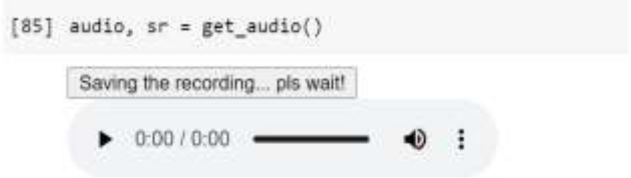


Fig. 1 Prediction of Emotions

F. Result Prediction:

Now, record the voice of any user via microphone and can then save it as a .wav file. Then later it can load our model to predict the result out of that .wav file.



```
[85] audio, sr = get_audio()
```

```
Saving the recording... pls wait!
```

Fig.2 Audio Recorder

G. System Architecture:

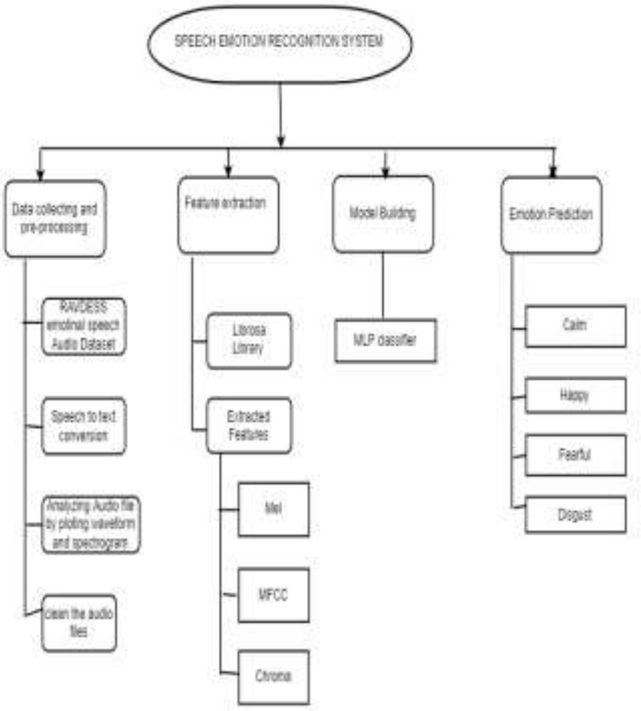
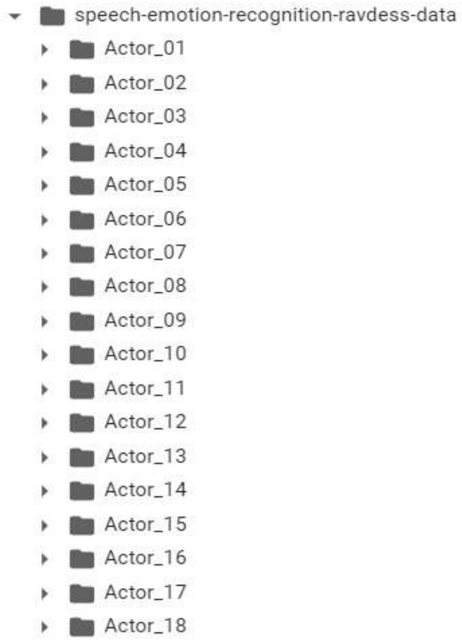

SPEECH EMOTION RECOGNITION SYSTEM
Data collecting and pre-processing
Feature extraction
Model Building
Emotion Prediction
RAVDESS emotinal speech Audio Dataset
Speech to text conversion
Analyzing Audio file by ploting waveform and spectrogram
clean the audio files
Librosa Library
Extracted Features
Mel
MFCC
Chroma
MLP classifier
Calm
Happy
Fearful
Disgust

Fig. 3 System architecture for SER

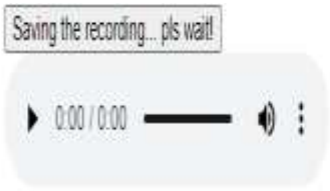
IV. RESULT AND DISCUSSION

The project is tested with real-time input. The output displays the emotion from the recorded audio is displayed correctly.



```
speech-emotion-recognition-ravdess-data
  Actor_01
  Actor_02
  Actor_03
  Actor_04
  Actor_05
  Actor_06
  Actor_07
  Actor_08
  Actor_09
  Actor_10
  Actor_11
  Actor_12
  Actor_13
  Actor_14
  Actor_15
  Actor_16
  Actor_17
  Actor_18
```

Fig.4 RAVDESS dataset



Saving the recording...pls wait!

Fig.5 Audio Recorder

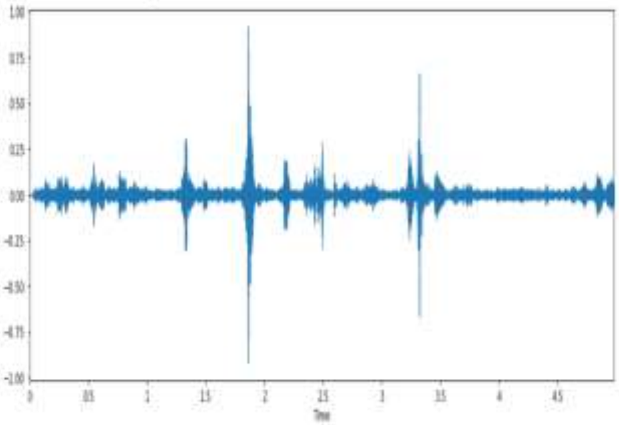
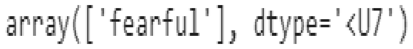


Fig.6 Wave plot of recorded audio(fearful)



```
array(['fearful'], dtype='<U7')
```

Fig.7 Result of emotion prediction

V. CONCLUSION

The emerging growth and development in the field of AI and machine learning have led to a new era of automation. Most of these automated device's work based on voice commands from the user.

Many advantages can be built over the existing systems if besides recognizing the words, the machines will comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automatic call center conversations, a diagnostic tool used for therapy, and an automatic translation system. Further enhancements to the model can be made so that it can function properly in real-time. To improve the accuracy of the model, we can increase the size of the dataset. The classifier can be integrated into the software or application so it can work in real-time. The more advances are made, the more different voices can be trained, and the dataset can be increased to implement a more realistic model.

VII. REFERENCES

[1] S. Casale, A. Russo, G. Scebba, "Speech Emotion Classification using Machine Learning Algorithms", 2008, IEEE International Conference on Semantic Computing.

[2] Vladimir Chernkh, Grigoriy Sreling, Pavel Prihodko, "ER From Speech With Recurrent Neural Networks", 2017.

[3] Rubi, C. Rana, "A Review: SE with Deep Learning Methods", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 5, May 2015, pg. 1017-1024.

[4] K. V. Krishna Kishore, P. Krishna Satish, "ER in speech Using MFCC and Wavelet Features", 3rd IEEE International Advance Computing Conference (IACC), 2013 .

[5] R. B. Pradeeba, K. Tarunika, Dr. P. Aruna, "Accuracy of SER through deep neural network and knearest", International Journal of Engineering Research in Computer Science and Engineering, Vol 5, Issue 2, February 2018.

[6] Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. (2020). A ML model for emotion recognition from physiological signals. Biomedical signal processing and control, 55, 101646.

[7] Singla, C., Singh, S., & Pathak, M. (2020, April). Automatic Audio Based ERS: Scope and Challenges. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

[8] Damodar, N., Vani, H. Y., & Anusuya, M. A. (2019). Voice emotion recognition using CNN and decision tree. Int J Innov Technol Expl Eng, 8(12), 4245-4249.