iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.41112

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Speech Emotion Recognition System Using Recurrent Neural Network in Deep Learning

Siddhant S. Patil[1], Shruti K. Patil[2], Ishwari S. Chankeshwara[3], Hrishikesh S. Rapatwar[4], Prof. Vidya. V. Waykule[5]
[1, 2, 3, 4, 5]*Computer Department, Savitribai Phule Pune University*

Abstract: ***In today's world, machine learning and deep learning together are enabling around 80% of the human interactions through the sheer ubiquity of the solutions provided by this domain. But one of the problems with the existing world is most of the people are not able to understand the actual emotional meaning and occurrence behind a person's speech. For instance, people having problem like Catatonia, etc. are not able to express themselves clearly or some industries which are considering some marketing strategy according to the customer mood, etc. can use this method. So, to bridge this gap between the people, it is important to develop a system that can assist them and then predict their emotional speech. This paper reviews the different approaches adopted to reduce the barrier of emotional communication which are already in existence and what methodology they used while doing so. In this context, we also present an approach of using the Recurrent Neural Network which is a part of Deep learning algorithms. The whole process of automated systems which continuously learn, adapt, and improve without much instruction is really fascinating. Our primary goal is to create a robust communication system through technologies that enable machines to respond correctly and reliably to human voices and provide useful and valuable services accordingly. In this review, an extensive report is made on the various approaches available for speech emotion recognition that has been done till now. All the model's and accuracy aspects are taken into consideration and are relayed according to it.***

Keywords: Deep Learning, Recurrent Neural Networks, Emotion Recognition, Speech Recognition, SER, RNN, Catatonia.

I. INTRODUCTION

Language is one of the most important methods for communication and speech is one of its main mediums. In human to machine interface, the speech signal is transformed into analogue and digital waveform which can be understood by the machine. Speech technologies are broadly used and seen to have unlimited uses. In many of the human-machine interface applications, emotion recognition from the speech signal is considered to be the research topic for many years. For this purpose, for the identification of the emotions from the speech signal, many systems have been developed until now. In this paper, speech emotion recognition based on the previous technologies which use different models and methods for the emotion recognition is reviewed and a new approach is suggested. They are used to differentiate emotions such as anger, happiness, neutral state, etc.

The intended system is going to be proposed such that it takes the input as speech both live and audio file and detects and recognizes the emotion behind that speech. After recognizing it, the output will be represented as the emotion in which the speech was spoken. There are various types of emotions included in this system such as happy, neutral, sad, etc. We have proposed to use the Recurrent Neural Network which is a part of Deep Learning Algorithms in order to increase our accuracy as compared to others models and methods which are in existence. In RNN, one data point or the current data depends upon the previous data point to perform an overall view. The model predicts the emotions based on the speech data provided during its execution.

II. LITERATURE SURVEY

A. Problems in the Current System

Speech Emotion Recognition is one of the most booming research areas around the world which is constantly growing its importance among research scientists around the world.

For the current system, there are few publicly available labelled datasets, and the lack of languages in which they are available is a major concern. A single dataset can contain uneven amount of data of the specified category. For example, in a speech dataset there can be 1000 files for the emotion angry and only 500 for happy. In such scenarios the model would be trained uneven and the predictions may not be accurate. The present systems and models used have comparatively lower accuracy and some have the problem of Negation Handling. Negation handling is when the overall meaning of the sentence is changed just because of the negated word added in the sentence somewhere. To address this problem some modern intelligent solution is required to improve its accuracy.

Apart from all these problems, one of the problems that arise is of the "context-dependency". There might be some words which are said in a different context or means, which have a different meaning all together. The frequency features for the word determines the actual emotional occurrence behind that word.

B. Present Work

1) Title*:* Emotion Recognition from Audio using Librosa and MLP classifier.

Prof. Guruprasad G[1], Mr. Sarthik Poojary[2], Ms. Simran Banu[3], Ms. Azmiya Alam[4], Mr. Harshith K R[5]

In this paper, the emotions in the speech are predicted using convolutional neural networks. Multi-Layer Perceptron Classifier (MLP Classifier) and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) are used for the Speech Emotion Recognition (SER) considering motive. The dataset contained 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Here there are three classes of features in speech which are examined namely, lexical, visual and acoustic features. Any of these combinations are considered here.

2) Title: Speech Emotion Recognition Using Fourier Parameters.

Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang

A new Fourier parameter model is used in this paper. Features such as pitch-related features, formants features, energy-related features, and timing features deliver important emotional cues. Also, time-dependent acoustic features, different spectral features similarly as linear predictor coefficients (LPC), linear predictor cepstral coefficients (LPCC), and Mel-frequency cepstral coefficients (MFCC) have a significant role to play in speech emotion recognition (SER). A FP model is formed to extract salient features from emotional speech signals. The FP features are considered effective in characterizing and recognizing emotions in speech signals. Moreover, it is possible to improve the performance of emotion recognition using more features.

3) Title: Speech Emotion Recognition Using Deep Learning Techniques: A Review.

RUHUL AMIN KHALIL[1], EDWARD JONES[2], MOHAMMAD INAYATULLAH BABAR, TARIQULLAH JAN, MOHAMMAD HASEEB ZAFAR[3], AND THAMER ALHUSSAIN[4]

This review overviews the different Deep Learning techniques used for Speech emotion recognition (SER). They have also discussed the datasets, limitations of the techniques, etc. Deep Neural Networks (DNNs) are derived upon feed-forward structures which comprised of one or more underlying hidden layers in between the inputs and outputs. The feed-forward architectures particularly as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) have a tendency to provide efficient results for the image and video processing. This review covers databases used, emotions extracted, and the contributions made towards speech emotion recognition and its limitations.

4) Title: An experimental study of speech emotion recognition based on deep convolutional neural networks.

*W. Q. Zheng, J. S. Yu, Y. X. Zou**

In this paper, an approach is taken to implement an Emotion recognition system based on deep convolution neural networks (DCNNs). To be specific, the log-spectrogram is computed and the principle component analysis (PCA) method is used for the reduction and the dimensionality and suppresses the interferences in it. After this, the PCA whitened spectrogram is split into non-overlapping segments. It also outperforms the SVM-based classification using the custom acoustics. But in this system, based on DCNNs (containing 2 convolution and 2 pooling layers) achieves only about 40% classification accuracy which can be increased.

5) Title: Hidden Markov model-based speech emotion recognition.

Bjorn Schuller, Gerhard Rigoll, and Manfred Lang

In this paper, the speech emotion recognition (SER) is done using the Hidden Markov models. In this paper, there are two different approaches proposed. The first proposed method is a global statistics framework of an utterance that is classified by Gaussian mixture models which made use of derived features of the raw pitch and energy contour of the speech signal. In the second method, it introduces increased temporal complexity applying continuously to the hidden Markov models while considering several states using low-level instantaneous features instead of global statistics.

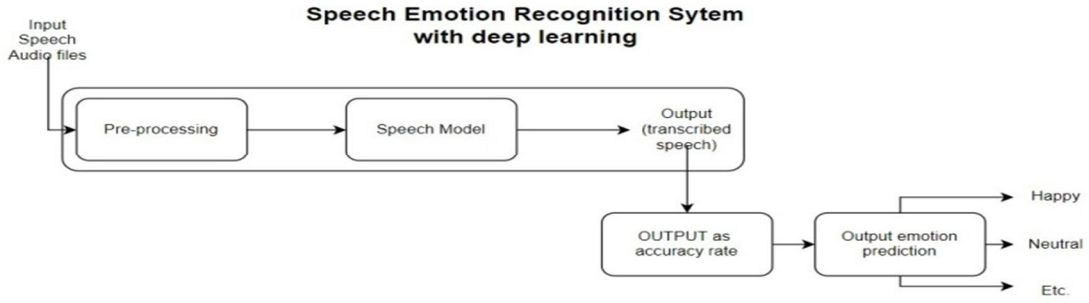
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

III. RESULT & DISCUSSION

A. *Proposed System*

1) *Problem Statement:* Emotions play an extremely important role in human mental life. Speech is a medium of expression of one's perspective and of one's mental state to others. Speech Emotion Recognition (SER) in terms can be defined as the extraction of the emotional state of the speaker from his or her speech signal and features in it. There are few emotions considered universally- including Neutral, Anger, Happiness, Sadness, etc. in which the intelligent systems with a number of finite computational resources can be trained to identify or synthesize as required. In this work, features of the audios are used for speech emotion recognition because these features contain the emotional information. *"Design and implement a model for speech emotion recognition system with Recurrent Neural Network in deep learning algorithms"*.

2) *Purpose:* A person is considered to be impassive if the person could not express himself/herself or showcase what his/her true emotions are or both. The count of such types of people is constantly increasing in today's world, and they are a closed, reserved society. They cannot communicate themselves effectively and efficiently. The progress of Information and Communication Technology has banged into all aspects of human life. It has changed the way we study, work, conduct business, travel, and communicate. The aim of the implementation is to provide people with a means to understand and communicate with each other effectively and efficiently. The system will provide various emotions depending upon the speech input provided through the user. The emotions such as happy, neutral, etc. will be estimated by the system to recognize the speech input. The user will be able to optimize his or her emotions of the speech through this system.



The system will make use of the Recurrent Neural networks (RNNs) and will use libraries such as:

1) Librosa
2) Scikit-learn (Sklearn)
3) Keras
4) Numpy
5) Pandas
6) PyAudio
7) Soundfile
8) Wave

The proposed system will be trained and tested on the two datasets combined for obtaining more accuracy and precision output.

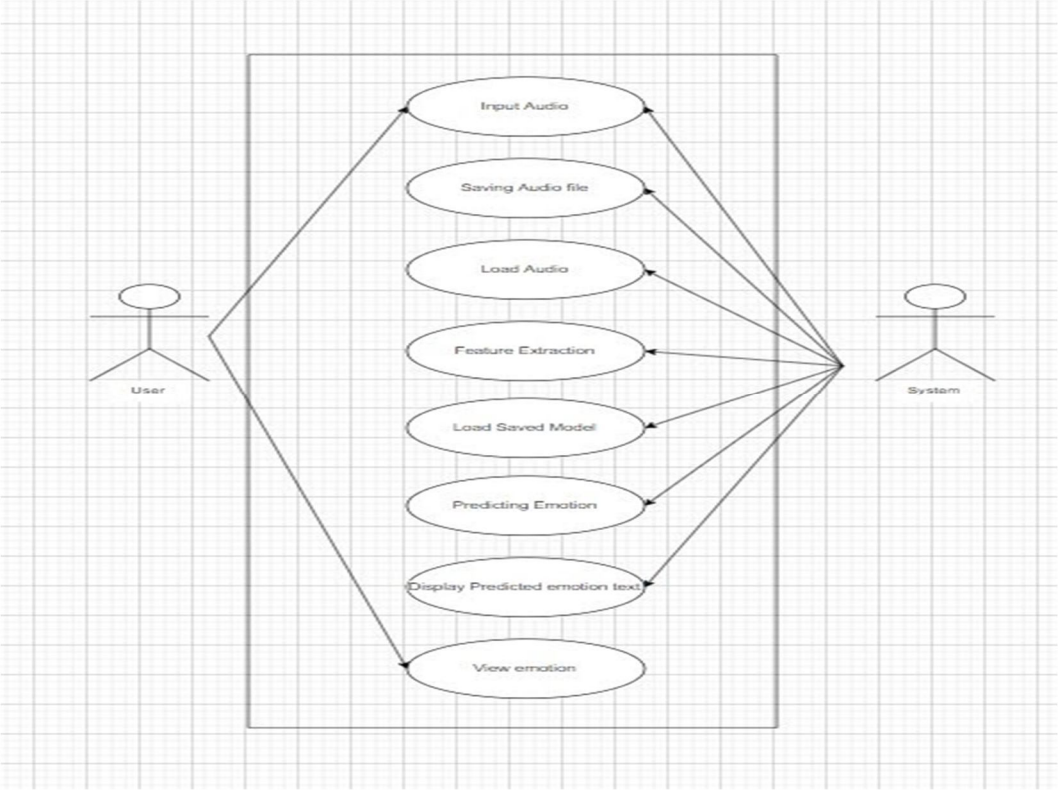
- *RAVDESS:* The **R**yson **A**udio-**V**isual **D**atabase of **E**motional **S**peech and **S**ong that contains 24 actors (12 male, 12 female), vocalizing two lexically-matched statements in a neutral North American accent.
- *TESS:* **T**oronto **E**motional **S**peech **S**et that was modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966). A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years).

The Deep learning network model Recurrent Neural Network is going to be used for the model for better precision and accuracy of the entire system.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

B. Methodology

1) *We have to Prepare Dataset:* In this step, we have to download and convert the dataset to be suited for extraction purposes.
2) *Loading the Dataset:* In this process, the dataset is loaded in Python which requires extraction of the audio features, for instance obtaining different features such as the pitch, power, and vocal tracts configurations from the speech signal, here the librosa library is used to do that.
3) *Model Training:* After the completion of the above two steps, we have to train it on the model.
4) *Testing the Model:* Final step for measuring how good and accurate the model is.



C. Deep Learning

Deep Learning is a machine learning method based on the neural networks which can imitate the human way of processing. It uses layers to extract high-level features from raw input by gaining knowledge as humans do. Deep Leaning can clarify complex feature abstraction by building a hierarchy in which each level of abstraction is created with information gained from the preceding layer. Deep learning algorithms make predictions repeatedly from each layer of the network. This iteration continues several times increasing the accuracy of the result. The number of processing layers is the reason for adopting the name 'deep.'

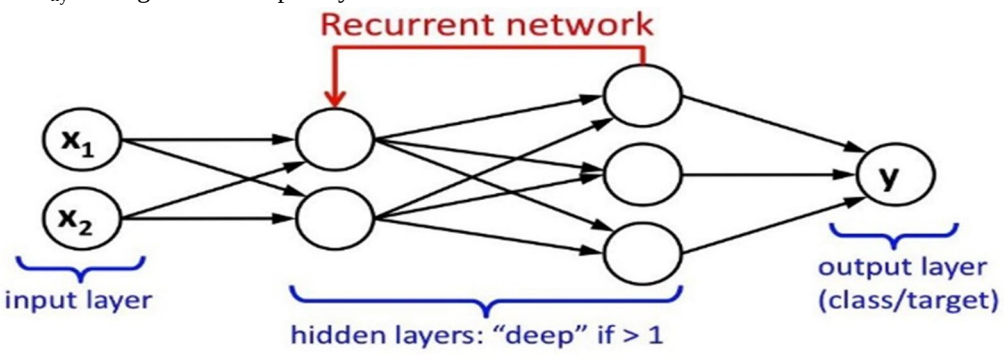
Deep learning algorithms are useful for predictions from learning large datasets. With deep learning, the machine learns from the data (training set) it is given and applies that knowledge to a new set, and it gets better as it identifies more features and adds them to the teaching set, to increase accuracy.

D. RNN

A Recurrent Neural Network (RNN) is a type of neural network where the output from the previous step is fed as input to the current step. These networks allow the reusing of the outputs as the inputs while having a hidden state. RNNs are mainly used for sentiment classification, video classification, parts of speech tagging, entity labeling, etc. In short, recurrent neural network processes sequences while retaining a memory (output – called as a state) of the current element of a sequence which is then given to the next element. This element not only considers current input but also a memory of the preceding element. The memory allows the network to understand and take all the context to make accurate predictions.

To calculate current state, $h_t = f(h_{t-1}, x_t)$

where h_t is the current state, h_{t-1} is the previous state, and x_t is the input state.
To calculate output, $y_t = W_{hy}h_t$
where y_t is output and W_{hy} is weight at the output layer.



In the figure, the hidden layers take input(s) from the input layer which actually contains the processed output, and are carried forward towards the output layer. This process is recurrent i.e., it is repeated until the final prediction is made, and hence called a recurrent neural network. RNN is basically designed to predict the output as humans predict it. Humans consider the entire statement instead of considering separate words of a statement to predict the final output. For example, *"The weather was bad at first but, the sunlight lighted up the mood."* A machine learning model that considers separate words to predict sentiment, would predict that this sentence is negative. But, RNN will consider the words like 'but' and 'lighted up' and predict that the sentence turns from negative to positive and hence, will give positive output.

E. Advantages

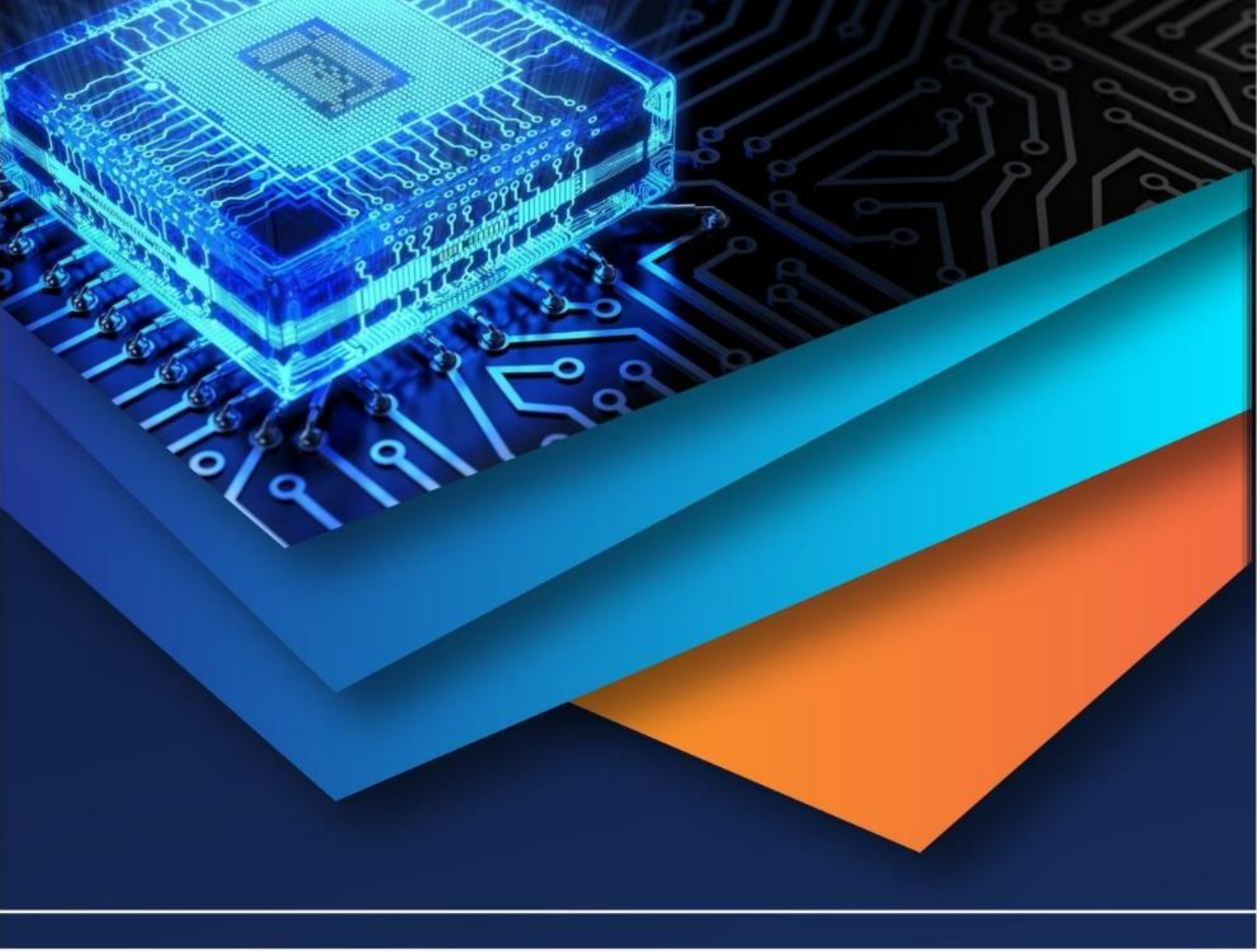
1) Possible to process the input of any length.
2) It remembers each and every piece of information through time called Long Short Term Memory (LSTM).
3) It has hidden layers and further dense layers for training the model
4) Weights in this method are shared across time.

IV. CONCLUSION

In this review, using the Recurrent Neural Network (RNN) one of the deep neural networks is used to extract the emotional characteristic parameter from emotional speech using the datasets. A live speech input or an audio file is provided to the model for prediction. This system can be employed in a variety of setups like for disabled people (Catatonia) for expressing themselves, in medical sciences to understand the patient, in linguistics speech research, and many more.
In future work, we will continue to further study speech emotion recognition based on Deep Learning Algorithms and further expand our understanding of the topic. Our ultimate aim is to study how to improve the recognition accuracy of speech emotion recognition.

REFERENCES

[1] Prof. Guruprasad G[1], Mr. Sarthik Poojary[2], Ms. Simran Banu[3], Ms. Azmiya Alam[4], Mr. Harshith K R, "Emotion recognition from audio using librosa and mlp classifier" – IRJET 2021.
[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
[3] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 827-831, doi: 10.1109/ACII.2015.7344669.
[4] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.
[5] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), 2003, pp. I-401, doi: 10.1109/ICME.2003.122093.
[6] Sattar, Rusul & Sadkhan, Eng. Sattar B.. (2020). Emotion Detection Problem: Current Status, Challenges and Future Trends.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY