International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 08 | Aug - 2021 ISSN: 2582-3930

Speech Emotion Recognition System using MLPClassifier

Dr. A. Haritha[1], Lasya Sarada[2], Harsha Bharathi[3], Prabhakar[4], Preethi[5]
Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada.

***Abstract*— Speech is the maximum critical manner of expressing ourselves as people. Speech is not anything but the sound or a data uttered to express the emotions. People deliver facts via speech, tone, pitch and many different traits of the human vocal device. Emotion recognition plays a climatic role for a powerful verbal exchange. Human beings can hit upon the emotions without problems, while the machines cannot discover the emotions easily. Even though there are some technologies which makes the machine understand the records primarily based on content material however assessing the emotion in the back of the content material is hard. Speech is only natural then to increase this communication medium to computer applications. We define speech emotion recognition (SER) is a collection of methodologies that procedure and classify speech signals to detect the embedded emotions. SER isn't a brand-new area, it has been around for over a long time, and has regained attentions. Speech emotion recognition is a system that has a various audio files labeled into different emotions like sad, anger, happy, disgust, calm, neutral and surprised. Though there has been a significant boom inside the subject of speech recognition, there are numerous speech yields that have been performed in like amazon Alexa, google domestic, and apple Siri that functions basically on voice-primarily based instructions. Speech emotion recognition has determined growing programs in exercise in numerous fields like security, medicinal drug, entertainment, education. Speech emotion recognition (SER) is evolved based on Ryerson audio-visual database of emotional speech and song (RAVDESS) includes 7,356 documents. The database incorporates 24 professional actors (12 female, 12 male). And we also used MLP classifier which is a supervised machine learning algorithm which can be used for classification or regression problems. We consider speech transcriptions along with its corresponding speech features – Spectrogram and MFCC, which together provide a deep neural network both semantic relationships and the necessary low-level features required to distinguish among different emotions accurately. In proposed system using the Keras framework and evaluates the maximum accuracy rate and mitigate the error rates as compared with the existing parameters.*

Keywords--- Speech Emotion Recognition, RAVDESS, MFCC, Keras framework.

I. INTRODUCTION

Language is a medium to communicate with the men and women. Other than the use of the language as a medium for communication, it's also used to express the emotions and emotions to the opposite man or woman. Language is processed in two extraordinary ways, emotional and semantic. Emotional method is greater taken into consideration to recognize the emotions and the semantic technique is used to understand what the person is speaking about. The interplay occurs handiest when the emotion is identified because of these emotions play an important position inside the human conversations. There are structures that recognize the speech (speech recognition) however they do no longer expect the emotions. Emotions can make the verbal exchange valuable and efficient, so the researchers advanced the speech emotion recognition machine.

Emotions cannot be defined easily. This system is developed in order to build an AI machine that will be able to predict the emotions and help in person in a healthy conversation and also this is built as a breakthrough of the existing system called Speech Recognition. Emotions are the crucial one's when comes under the areas like medical field, military field, etc. Recognizing the emotions is a difficult task, since every individual has a unique tone, pitch and intonation of speech[1]. 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.

There are exceptional emotion models, discrete emotional model and the alternative is the dimensional emotional model. The discrete emotional model is used to expect the common emotions like unhappy, anger, glad, impartial, worry, disgust. All human beings are notion to have simple innate emotions that can be recognized by way of facial expressions and organic techniques. Theorists have carried out studies to decide which emotions are fundamental. A famous instance is Paul Ekman and his colleagues' move-cultural take a look at of 1992, wherein

they concluded that the six primary emotions are anger, disgust, worry, happiness, sadness, and marvel. Ekman explains that there are specific characteristics connected to each of those emotions, permitting them to be expressed in various stages. For theoretical and sensible reasons, researchers outline emotions based totally on one or greater dimensions. In "the passions of the soul", Descartes defines and investigates six principal passions (marvel, love, hate, preference, joy, unhappiness, and unhappiness). Wilhelm max Wundt, the father of modern psychology, proposed in 1897 that emotions will be defined by three dimensions: "pleasant versus unsightly", "arousing or subjugating" and "tension or relaxation".

Section II discusses the related work done in this area. Section III is about the proposed approach and Section IV shows the implementation part and results. Section V conclusion.

II. Related Work

Abdul Ajij Ansari et al developed Speech Emotion Recognition using CNN. This paper presents the implementation of this function with the deep learning model of Convolutional Neural Networks (CNN). The architecture was an adaptation of an image processing CNN, programmed in Python using Keras model-level library and TensorFlow backend. Speech-to-Text (STT) technology leads to the faster innovations towards speech emotion recognition advancements can lead to advancements in various field like automatic translation systems, machine to human interaction, used in synthesizing speech from text. Convolution neural network deep learning model useful for the any recognition like speech, face, handwriting etc[2].

Deepak Bharti et al. developed A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals. Proposed work using the MATLAB simulation tool and evaluates the maximum accuracy rate and mitigate the error rates as compared with the existing parameters. SER aimed to detect spontaneously the emotion stage of the social being from the voice of the male or female. It is dependent on the in-complexity analysis of the creation method of the Speech Signal (SS), eliminating certain types that consist of expressive data from the voice of the utterer and receiving suitable pattern recognition techniques to detect the emotional stage[3].

Alif Bin Abdul Qayyum Convolutional Neural Network Based Speech Emotion Recognition. This is published in IEEE, in the year 2019. This paper presents a unique Convolutional Neural Network (CNN)based speech-emotion recognition system. A model is developed and fed with raw speech from specific dataset for training, classification and testing purposes with the help of high end GPu. SAVEE (Surrey Audio-Visual Expressed Emotion) database has been recorded from four native English male speakers, postgraduate students and researchers at the University of Surrey aged from 27 to 31 years[4].

Seunghyun Yoon et al. developed a multimodal speech emotion recognition using audio and text. This is published in IEEE, in the year 2018. Model encodes the information from audio and text sequences using dual recurrent neural networks (RNNs) and then combines the information from these sources to predict the emotion class. This architecture analyzes speech data from the signal level to the language level, and it thus utilizes the information within the data more comprehensively than models that focus on audio features. Deep learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, speech recognition, text-to-speech generation and other machine learning related areas. The speech emotion recognition task is one of the most important problems in the field of paralinguistics[5].

III. Proposed Approach

The proposed machine is designed to predict the emotions of the user's audio. The emotion can be predicted based upon the tone, pitch and intonation of the speech. This system is based at the discrete emotion model this is evolved to predict the maximum primary emotions such as happy, anger, sad, disgust, neutral and fear. There are 4 most important steps to predict the emotions. They are recording the audio, feature extraction, Classification and recognizing the emotion. The proposed system is a web app that is developed using the flask framework.

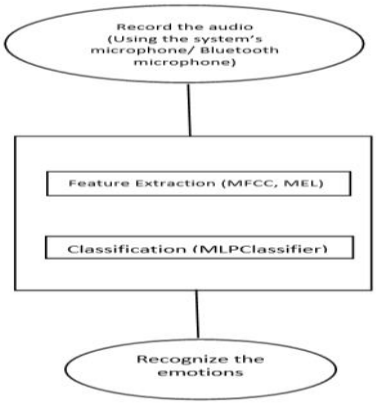


Figure 1: Proposed Speech emotion Recognition System

Record the Audio: The person will be capable of use the machine's microphone or bluetooth microphone to add the audio to the machine. This audio is saved in a "wav" format. This audio file is taken as input for the speech emotion recognition system.

Feature Extraction: There are two main feature extractors that are used in developing this project. MFCC and MEL are the features extractors used to extract the features of the audio file.

- **MFCC:** MFCC (Mel frequency cepstral coefficients) turned into first recognized via Davis and Mermelstein in 1980. MFCC is one of the maximum well-known strategies for extracting features due to its pretty correct potential to extract sound functions of a audio document. This MFCC adapts the workings of human hearing. To get the MFCC coefficient we will divide the voice sign into numerous elements with the framing procedure. Then we will convert each part from the time domain to the frequency domain the use of Fourier transform. From the results of the transform, we can calculate the strength at each frequency band using Mel filter bank. This method will produce a Mel spectrum. Then Mel spectrum do we inverse once more to get the MFCC coefficient value inside the time domain. Underneath is a chart of the sound feature extraction method using MFCC [6]

- **MEL:** Mel frequency warping is generally achieved with the use of filterbank. Filterbank is of filter that has the purpose to find out the strength size of certain frequency bands within the sound sign. For functions of MFCC, filters have to be carried out in the domain frequency. Over again, this approach is adapted from the manner the cochlea works. Within the cochlea, there are separate parts that modify the production of sounds primarily based on certain frequencies. Human belief of frequency in the signal sound does no longer comply with a linear scale. The frequency with which sincerely (in Hz) in a signal might be measured humans subjectively with use Mel scale. To calculate Mel scale we use:

$$Mel(f) = 1125\ln(1+f/700)$$

in which mel(f) is characteristic of mel scale and f is frequency.

The most popular mel filter used is the triangular mel filter.

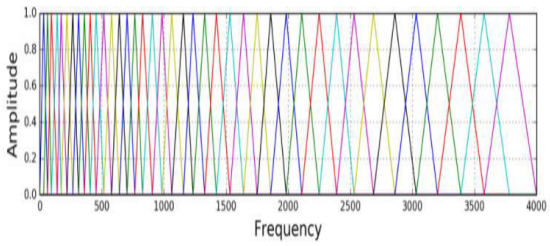


Figure 2: MEL Frequency Spectrum

Classification: The classification used to increase this machine is multi-layer percepton classifier also referred to as MLPCassifier in short. MLP is a network made up of a percepton. This mlpclassifier depends upon the neural networks for the classification. Multi-layer perceptons are able to performing many non-linear functions. Multi-layer perceptron (MLP) is a supervised getting to know set of rules that learns a function $f(\cdot):R^m \rightarrow R^0$ by training on a dataset, where m is the wide variety of dimensions for input and o is the quantity of dimensions for output. Given a set of features $X=x_1,x_2,...,x_m$ and a goal y, it is able to examine a non-linear function approximator for either classification or regression. It is one of a kind from logistic regression, in that between the input and the output layer, there may be one or extra non-linear layers, called hidden layers.

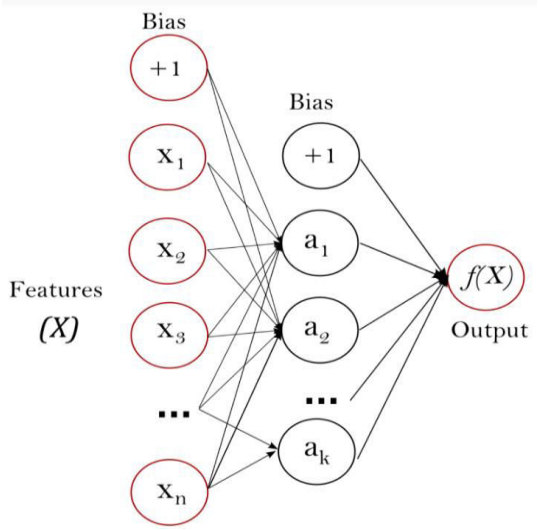


Figure 3: MLP Classifier

The multi-layer perceptron functionality to learn non-linear models and capability to analyze models in actual-time (on line studying) the use of partial fit[8]. MLP is a feed-forward neural network, that maps the inputs to an appropriate set of outputs. Usually, the network consists of an input layer, a hidden layer(s), and an output layer as proven. Given an input node xi, the output of the hidden node hj is given as

$$Hj = \varphi 1 + n \sum i=1 \, wij + \theta j \, ,$$

Where wi,j represents the weight among the ith input and jth hidden node, and θj represents the bias value.

In contrast, the output could be given as

$$Output = \varphi 2 + n \sum j=1 \, wjk + \theta k$$

The mapping of inputs to outputs is an iterative process, in which in every iteration, weights wi,j are updated. One of the typically used set of rules is the Back propagation algorithm, which updates the weights the usage of

$$Wji(t + 1) = wji(t) - e \, \partial e \, f \, wji$$

The mistake between computed and preferred output is used to update the weights.[9]

This technique allows in breaking down the audio files into the numerical values which represents the frequency, time, amplitude or another such parameters that can help within the evaluation of the audio files. After the extraction of the required features from the audio files, the version is trained. For the training, we save the numerical values of emotions and their respective functions correspondingly in one-of-a-kind arrays. Those arrays are given as an input to the MLP classifier that has been initialized. The classifier identifies special classes within the datasets and classifies them into exceptional emotions. The model will now be able to apprehend the tiers of values of the speech parameters that fall into precise emotions. For testing the performance of the version, if we enter the unknown test dataset as an input, it's going to retrieve the parameters and expect the emotion as consistent with education dataset values. The accuracy of the system is displayed inside the shape of percentage that is the final end result of our challenge.

Dataset: Processing and classifying voice statistics is a tough as it may be a piece heavy at the labeling part. This is due to the fact every person has exceptional views on emotions. Here we use the RAVDESS dataset (the Ryerson audio-visible database of emotional speech and song) that is a voice recording dataset of 24 actors (male and female). This voice recording voiced two statements in an impartial North American accent. Speech sounds encompass expressions of calm, happiness, disappointment, anger, worry, wonder, and disgust. Each expression is produced at two stages of emotional depth (regular, robust), with additional neutral expressions. RAVDESS consists of 1440 files: 60 trials in keeping with actor x 24 actors = 1440 information.

Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only)
- Vocal channel (01 = speech, 02 = song)
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised)
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door")
- Repetition (01 = 1st repetition, 02 = 2nd repetition)
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female)

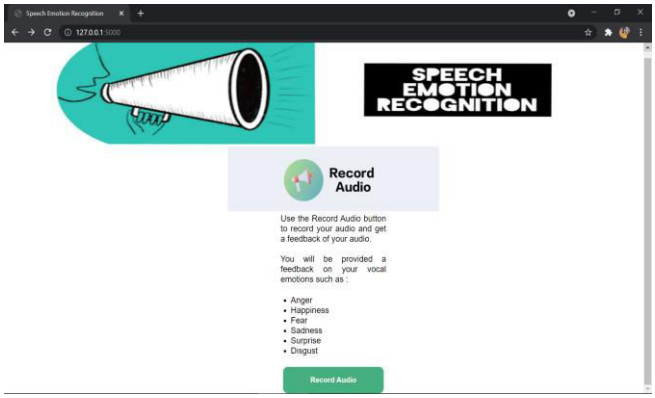
Filename example: 03-01-08-02-02-02-14.wav

- Audio-only (03)
- Speech (01)
- Surprised (08)
- String intensity (02)

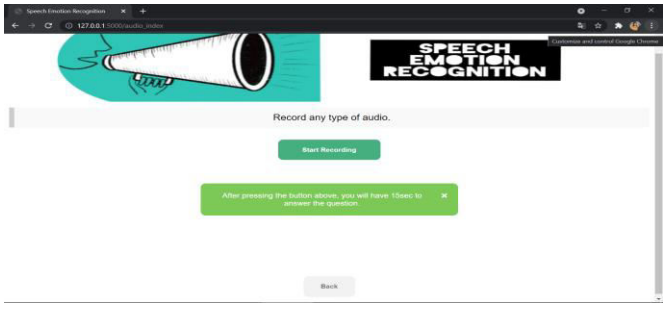
- Statement "dogs" (02)
- 2nd Repetition (02)
- 14th Actor (14) Female, as the actor ID number is even.

Emotion Recognition: In this step the user will get the output of the emotion recognized by the system. All the emotions that are present in the audio file are represented to the user in the percentage. The highest percent emotion is the emotion that is shown to the user as the output.

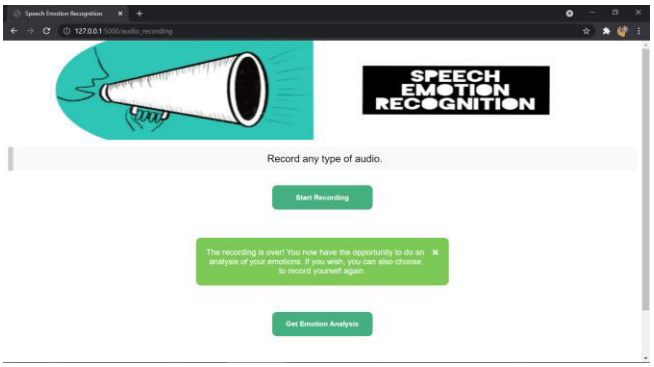
V. RESULTS



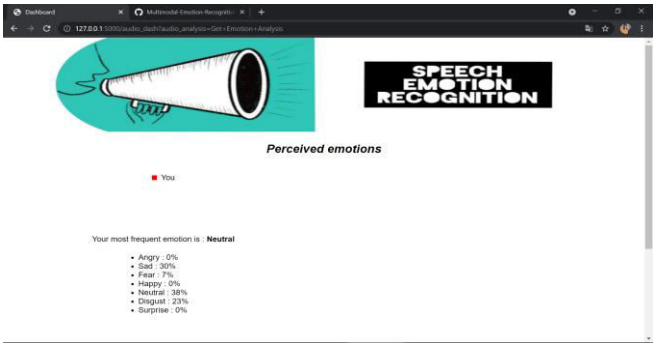
Description: This is the home page for the speech emotion recognition system. The user can use the Record audio button to go to the recording the audio page.



Description: This is the Recording audio page for the speech emotion recognition system. The user an use the Start Recording button to record the audio of 15 sec and this audio is saved in voice_recording.wav.



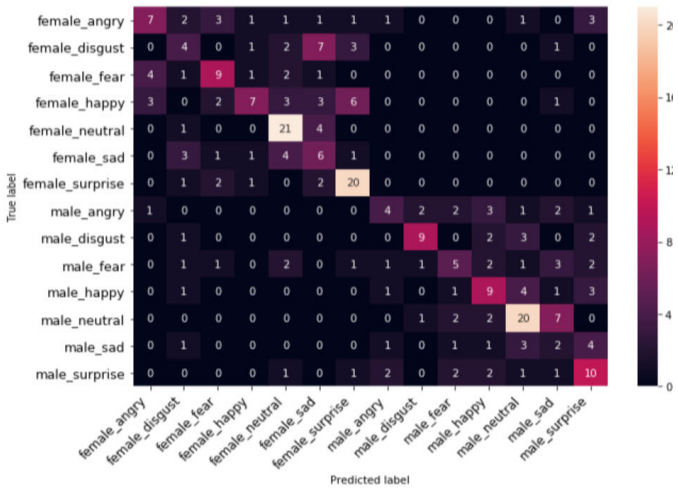
Description: This page is seen after recording the audio for 15secs. The user can use the Get emotion analysis button to get the predicted emotions by the system.



Description: This is the Audio dashboard page for the speech emotion recognition system. The results are shown in this page.

CLASS	PRECISION	RECALL	F1 - SCORE
ANGER	0.88	0.80	0.84
SAD	0.89	0.82	0.86
FEAR	0.89	0.83	0.86
HAPPY	0.91	0.90	0.91
DISGUST	0.83	0.79	0.81
NEUTRAL	0.89	0.87	0.88
TOTAL	0.88	0.84	0.86

Table 1: Recognition Data for the Speech emotion Recognition



Description: Confusion Matrix

VI. Conclusion and Future Development

On this paper a discrete emotional version of speech emotion popularity is discussed. This system is evolved as a web app with the use of the flask framework. This model takes the audio as input in wav format. The device will process the input through extracting the features of the audio report the use of the MFCC and Mel feature extractors. The next step is type that's achieved the use of the MLPClassifier. This MLP classifier is the efficient classifier for classification. The results suggests that the six one-of-a-kind emotions can be categorized with a great accuracy core. Due to the fact that is a web app the user may be able to use this device successfully and quite simply. This device may be applied as a multimodal method for the future improvement.

References

[1] Speech Emotion Recognition using Neural Network and MLP Classifier Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama Volume 10 Issue No.4

[2] Speech Emotion Recognition using CNN Abdul Ajij Ansari, Ayush Kumar Singh, Ashutosh Singh International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 06 | June 2020

[3] A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals Deepak Bharti (Author) Poonam Kukana (Co-Author), Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)

[4] Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. Alif Bin Abdul Qayyum, Asiful Arefeen, Celia Shahnaz, 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON), 28-30 November,2019, Dhaka, Bangladesh

[5] Multimodal speech emotion recognition using audio and text, Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, IEEE, 2018.

[6] Kerkeni, Leilaand Serrestou, Youssef and Raoof, Kosaiand CIMer, Catherine and Mahjoub, Mohamed andMbarki, Mohamed,"Automatic Speech Emotion Recognition Using Machine Learning;' March,2019.

[7] X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001

[8] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. ACM SIGKDD Explor. Newsl. 2009, 11, 10–18.

[9] Abdul Rehman Javed, Muhammad Usman Sarwar, Suleman Khan, Celestine Iwendi, Mohit Mittal and Neeraj Kuma, Analyzing the Effectiveness and Contribution of Each Axis of Tri-Axial Accelerometer Sensor for Accurate Activity Recognition, Sensors · April 2020.