# Speech Emotion Recognition using Machine Learning

**R.ANUSHA[1], P.SUBHASHINI[2], DARELLI JYOTHI [3], POTTURI HARSHITHA [4], JANUMPALLY SUSHMA [5], NAMSAMGARI MUKESH[6]**

*Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad,*

anusharavankol@gmail.com[1], subhashinivalluru@gmail.com[2], jyothidarelli4@gmail.com[3], p.harshitha2000@gmail.com [4], jsushmareddy707@gmail.com[5], mukeshnamsamgari@gmail.com[6]

.

2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-1571-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOEI51242.2021.9453028

***Abstract*—Speech emotion recognition is an act of predicting human's emotion through their speech along with the accuracy of prediction. It creates a better human computer interaction. Though it is difficult to predict the emotion of a person as emotions are subjective and annotation audio is challenging, "*Speech Emotion Recognition*(SER)" makes this possible [1]. This is the same theory which is used by animals like dogs, elephants and horses etc do to be able to understand human emotion [1].There are various states to predict one's emotion, they are tone, pitch, expression, behavior etc. Among them, few states are considered to find the emotion through the speech. Few samples are used to train the classifiers to perform speech emotion recognition [2]. Thhis research work considers the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song dataset). Here, the three key features such as MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram and chroma are extracted.**

***Keywords*—Speech emotion recognition; machine learning; RAVDESS dataset; MFCC; Mel spectrogram; chroma; emotions; accuracy; predict; Mlp; Multi level percetron***

## I. Introduction

It is easy for a person to identify the emotion of other person irrespective of the semantics. Each emotion has each type of state. When we take example of a happy person, his/her face looks brightened with a look of glee, his/her voice will be filled with full of joy. Based on some observations like pitch, tone, behaviour etc. a person can quickly identify the state of the person whether he is happy, sad, angry, depressed, disgusted etc. [3]. As emotions are subjective, it is difficult for a machine to predict the emotion of the person. But Speech emotion recognition using machine learning can make this possible. Machine learning is used in my applications in our day to day lives. Using this for speech emotion recognition helps in various ways like saving person from depression etc. Let's see how this can be done using machine learning. Initially, let's discuss the literature survey (existing systems) with their conceptual explanation.

## II.LITERATURE SURVEY

Speech Emotion Recognition is a subject under research. Speech emotion recognition abbreviated as SER. It creates a natural Human Computer interaction. There are various kinds of methods used to identify the emotion from the speech, such as using support vector machine (SVM), Recurrent Neural Network, K-nearest neighbour, Hidden Markov Model(HMM).

### *A. Support Vector Machine (SVM)*

Support Vector Machine approach computes the audio parameters to identify the emotion and has high accuracy in predicting the emotion from the speech. But this approach can only classify the dataset into 2 classes only. That means we can only identify among the 2 emotions trained to the classifier. The other disadvantages of this approach are long processing time, background noise leading to error and it has low accuracy[4] .

### *B. K-nearest neighbour*

The other classifier is k nearest neighbour classifier. This is the simplest classification algorithm which identify the emotion of speech. This classifier uses pitch and energy of the audio to predict emotion and has accuracy of 64% for 4 emotions audio.

### *C. Hidden Markov Model (HMM)*

HMM models temporal sequencing from the audio. This modelling is useful in predicting the emotion from the speech. The main limitation of this classifier is feature selection process. As the features don't carry the complete information of the emotion of the speech. But it has good classification accuracy compared to other classifiers [5].

## III.PROPOSED SYSTEM

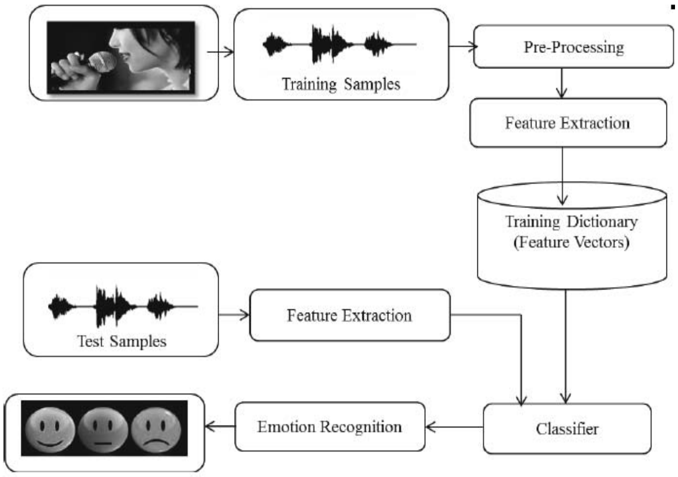
In proposed system, we are using RAVDESS dataset as a data for the system. The data present in the dataset is pre-processed to clean the audio and remove the disturbance from the audio to reduce the error in the output. The audio is divided into equal time intervals frames. Then the dataset is divided into 2 parts as training data and testing data. Training data is 80% of the dataset and testing data is 20% of the dataset.

The features are extracted from the audio and given to the classifier to predict the emotion.

The model is created by training data inputs to the classifier then this model is tested with the testing data inputs.

We get the accuracy by calculating the output of the model and the actual emotion in dataset.

*D. Architecture*



- The above figure represents the architecture of speech emotion recognition using machine learning.
- Based on the methodology, we divided the project into two modules. They are (1) Speech processing module and (2) Classification module.

*1) Speech Processsing*

Speech processing consist of two phases. The first one is (a) Preprocessing phase and the second one is (b) Feature extraction phase. Let's discuss these two phases.

*a) Preprocessing*

Preprocessing is the initial phase which is there after followed by feature extraction. Preprocessing includes:

SILENCE REMOVAL

There are many different types of silences present in a audio. To get clear audio of the person or the speaker, we need clear those silence present in the audio. So, this silent is removed by "silence removal". There are many methods to remove the silence, few of them are STE(short time energy)and ZCR(zero crossing rate). Below equation represents, short time energy method to remove the silence[6]:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

PRE-EMPHASIS

The most vital steps of preprocessing at high frequency is the pre-emphasis of the speech signal. Pre-emphasis is implemented by the following below equation[6].

$$x'(n) = x(n) - \propto x(n-1)$$

Where α is the pre-emphasis parameter whose value lies between (0.9) and 1.

NORMALIZATION

The signal sequence division of the highest value of the signal to ensure that the each sentence has a comparable volume level is done by normalization[6].

WINDOWING

$$y_1(n) = x_1(n)w(n), \quad 0 \le n \le N-1 \quad N-1$$

Hamming window is one among the most popular windowing techniques. The below formula describes hamming window function [6].

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

:

*b) Feature extraction*

Feature extraction is an important step in the whole process. The processing of the audio signal takes place here in feature extraction. In this module, extraction of feature vectors is carried out. Here, we have extracted three main features namely MFCC, Mel Spectrogram and chroma.

*2) Classification*

In our system, we are using MLP classifier for classification of the features. After completing the training process our model is created. After completion the testing, the accuracy of the model is calculated by the data given which is created by the classifier [7].

feed forward artificial neural network which is well known as ANN is super class of MLP which stands for multi-level perceptron. It has minimum of three layers which are one input layer, an output layer and other hidden layers[8].

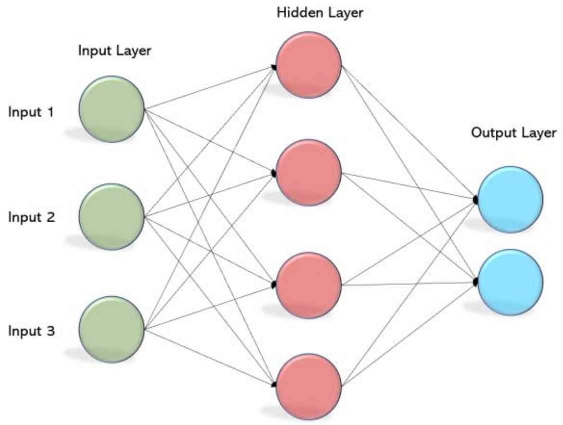
$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(\mathbf{w}^T\mathbf{x} + b)$$

Where, x represents input vector

W represents weights vector

B represents bias

Φ represents activation function.



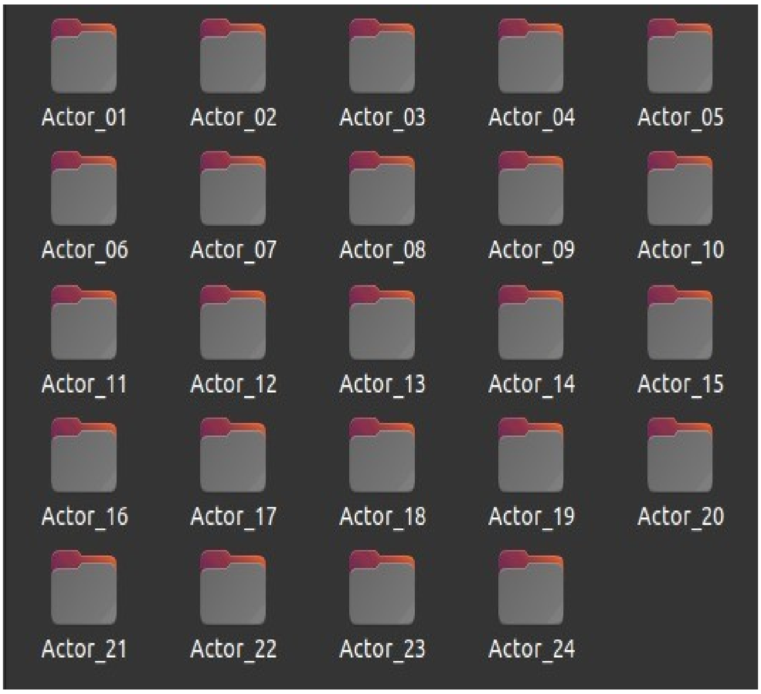
The classification is the main process in the system. We divided the extracted features into classes which predicts the emotion of the speech. In our system we are using the MLP (MultiLayer Perceptron) classifier [9].

We initialize the MLP classifier by defining the required parameters. After that we give the data to the network to train the model. The model is created to predict the emotion of the speech. The accuracy of the model is calculated by comparing the emotion predicted by the model and the actual emotion in the dataset [10].

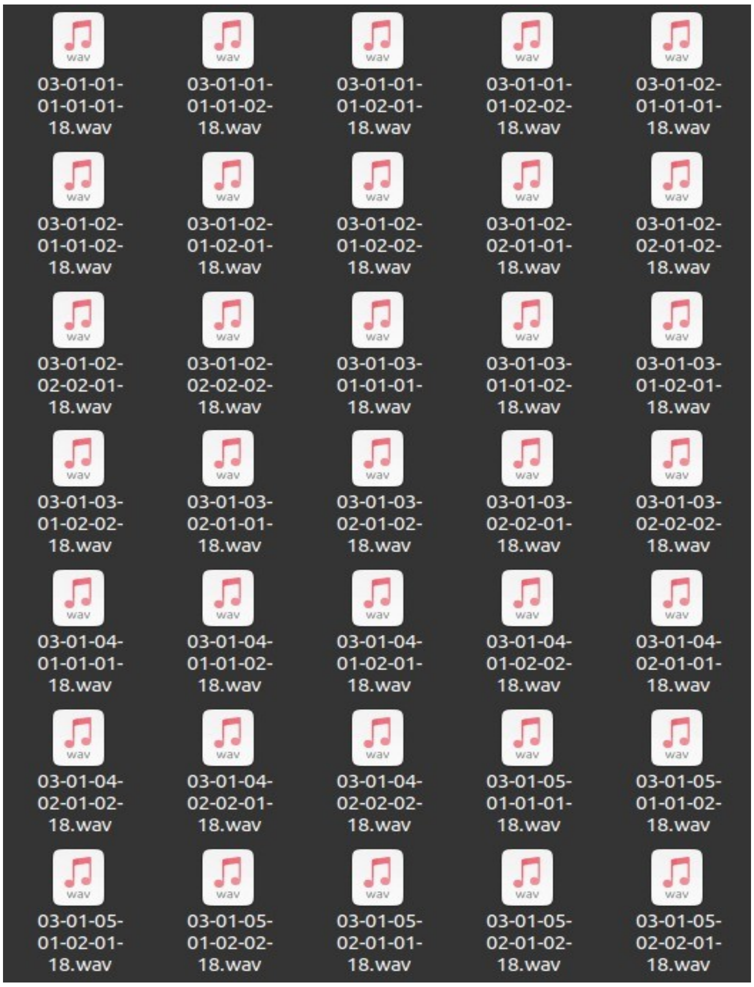
## II. EVALUATION

### A. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contain 7356 files whose total size is 24.8 GB. The database contains 24 professional actors where 12 of them are male actors and 12 of them are female actors. Happy, sad, calm, angry, fearful, surprise, and disgust expressions are included in speech, and calm, happy, sad, angry, and fearful emotions are included in song[11]. The below figure shows the 24 professional actor’s audios. The other figure shows the audios of one actor in different emotions.

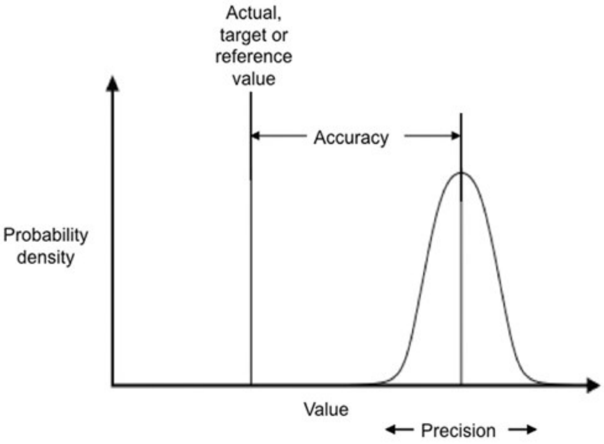


Actor_01 Actor_02 Actor_03 Actor_04 Actor_05
Actor_06 Actor_07 Actor_08 Actor_09 Actor_10
Actor_11 Actor_12 Actor_13 Actor_14 Actor_15
Actor_16 Actor_17 Actor_18 Actor_19 Actor_20
Actor_21 Actor_22 Actor_23 Actor_24



03-01-01-01-01-01-18.wav 03-01-01-01-01-02-18.wav 03-01-01-01-02-01-18.wav 03-01-01-01-02-02-18.wav 03-01-02-01-01-01-18.wav
03-01-02-01-01-02-18.wav 03-01-02-01-02-01-18.wav 03-01-02-01-02-02-18.wav 03-01-02-02-01-01-18.wav 03-01-02-02-01-02-18.wav
03-01-02-02-02-01-18.wav 03-01-02-02-02-02-18.wav 03-01-03-01-01-01-18.wav 03-01-03-01-01-02-18.wav 03-01-03-01-02-01-18.wav
03-01-03-01-02-02-18.wav 03-01-03-02-01-01-18.wav 03-01-03-02-01-02-18.wav 03-01-03-02-02-01-18.wav 03-01-03-02-02-02-18.wav
03-01-04-01-01-01-18.wav 03-01-04-01-01-02-18.wav 03-01-04-01-02-01-18.wav 03-01-04-01-02-02-18.wav 03-01-04-02-01-01-18.wav
03-01-04-02-01-02-18.wav 03-01-04-02-02-01-18.wav 03-01-04-02-02-02-18.wav 03-01-05-01-01-01-18.wav 03-01-05-01-01-02-18.wav
03-01-05-01-02-01-18.wav 03-01-05-01-02-02-18.wav 03-01-05-02-01-01-18.wav 03-01-05-02-01-02-18.wav 03-01-05-02-02-01-18.wav

### *B. Results*



*1) Accuracy::* Percentage of records which are exactly classified is termed as Acurracy. Classification models are evaluated by accuracy rates. There is a formula for calculating accuracy. The formula is [12]:

$$\mathrm{Acc} = \mathrm{TP} + \mathrm{TN} \,/\, \mathrm{Total\ samples}$$

*2) Precision and Recall:*

Performance of classification systems is evaluated by both precision and recall[13]. The fraction of related instances within all retrieved instances is called as precision. Recall is also known as sensitivity. The fraction of retrieved instances with in all the related instances is called recall or sensitivity [13].

$$\text{precision} = \frac{tp}{tp + fp} = \frac{\text{retrieved and relevant documents}}{\text{all retrieved documents}}$$

*Mathematical definition of precision*

$$\text{recall} = \frac{tp}{tp + fn} = \frac{\text{retrieved and relevant documents}}{\text{all relevant documents}}$$

*Mathematical definition of recall*

*3)F1 Score :*

The measure of combining both the precision and the recall is termed as F1 score. It is calucated as the harmonic mean of the precision and recall [14].

$$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

## III. Conclusion

By implementing this project we can identify the emotion of the speech using Machine Learning which can be used in improving the human computer interaction. This system can be used in improving the virtual voice-based assistants which can understand the emotion of the human and can respond accordingly and in marketing, improving customer service in call centers etc. In this system we get an accuracy of about 80% approximately.

## IV. Future works

Some of the drawbacks can be resolved in the future to make the model more accurate and efficient.

- ❖ We can improve the model by training the model a variety of datasets which increases the accuracy of the model.
- ❖ Removing the disturbance from the input audio which deviates from correct prediction.
- ❖ Adding more emotions to the system since this system can identify only 8 emotions.
- ❖ Extracting more features from the speech to improve the classification process.

## References

[1] https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/

[2] Mehmet Berkehan Akçay, Kaya Oğuz, 2000. www.sciencedirect.com/science/article/abs/pii/S0167639319302262

[3] Akalpita Das,Laba Kr. Thakuria,Purnendu Acharjee,Prof. P.H. Talukdarhttps://www.ijser.org/researchpaper/A-Brief-Study-on-Speech-Emotion-Recognition.pdf

[4]YashpalsingChavhan,ManikraoDhorehttps://www.researchgate.net/publication/43785303_Speech_Emotion_Recognition_Using_Support_Vector_Machines

[5] Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576, Singaporehttps://www.sciencedirect.com/science/article/abs/pii/S0167639303000992

[6] Bashar M. Nema, Ahmed A. Abdul-Kareem Department of Computer Science, College of Science, Mustansiriyah University, IRAQ

https://www.iasj.net/iasj/download/62b38cc626857948

[7] Prabhakar Reddy G,Arun Dalton, Sai Prasad K "Fuzzy logics associated with neural networks in intelligent control for better world". International Journal of Reasoning based Intelligent Systems.

[8] https://en.wikipedia.org/wiki/Multilayer_perceptron

[9]R. Anusha, Boggula. Lakshmi,Spruthi kankala ,T. Mounika, Deep Stock Prediction" International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[10]R. Anusha , T. Nirmala, N.Thulasichitra "System for smart protection of crop" .International Journal of Grid and Distributed Computing Vol. 13, No. 2, (2020), pp. 1707-1715

[11] https://smartlaboratory.org/ravdess/

[12] Subhashini Peneti (MLR Institute of Technology), Hemalatha E(Jawaharlal Nehru Technological University)-DDOS Attack Identification using Machine Learning Techniques

[13] ThomasWood,https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall

[14] TeemuKanstrén, https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

[15] Pushpa Rani, K., Lakshmi, L., Teegala, N., Dhanalaxmi,"An application for the development of smart library as an academic initiative "International Journal of Advanced Trends in Computer Science and Engineering Volume No 8(1.3): PP:55-58 · July 2019

[16] Pushpa Rani, K., Reddy, M., Anjaneyulu, B., Hari Chandana, B."A formal assessment paper on EDUSET-MBA universities custom search "International Journal of Innovative Technology and Exploring Engineering Volume-9 Issue-1, November 2019 ISSN: 2278-3075