# TABLE OF CONTENTS

LIST OF FIGURES vi

ABSTRACT vii

**CHAPTER-1 Introduction**  1

1.1 Origin of the Problem 1

1.2 Basic definitions and Background 1

1.3 Problem Statement 2

1.4 Real time Applications of Proposed work 2

**CHAPTER-2 Review of Literature** 3

2.1 Description of Existing Systems 3

2.2 Summary of Literature Study 4

**CHAPTER-3 Proposed Method** 5

3.1 Mel-Frequency Cepstral Coefficients (MFCC) 5

3.2 System Architecture Diagram 6

3.3 System Implementation 6

3.4 Training 7

**CHAPTER-4 Results and Observations** 8

4.1 Stepwise description of Results and Observations 8

**CHAPTER-5 Conclusion and Future study** 115.1 Conclusion 11

5.2 Future Study 11

**References** 12

**LIST OF FIGURES**

**Figures**

**Figure 3.2.1 Speech Emotion Recognition System** 6

**Figure 3.4.1 Training and Testing modules** 8

**Figure 4.1 Accuracy** 9

**Figure 4.2 depressed** 9

**Figure 4.3 angry** 10

**Figure 4.4 sad** 11

**ABSTRACT**

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. Emotion recognition is a rapidly growing research domain in recent years. Unlike humans, machines lack the abilities to perceive and show emotions. But human-computer interaction can be improved by implementing automated emotion recognition, thereby reducing the need of human intervention. In this project, basic emotions like calm, happy, fearful, disgust etc. are analyzed from emotional speech signals. We use machine learning techniques like Multilayer perceptron Classifier (MLP Classifier) which is used to categorize the given data into respective groups which are non linearly separated. Mel-frequency cepstrum coefficients (MFCC), chroma and mel features are extracted from the speech signals and used to train the MLP classifier. For achieving this objective, we use python libraries like Librosa, sklearn, pyaudio, numpy and soundfile to analyze the speech modulations and recognize the emotion.

**INDEX TERMS**: Speech emotion recognition, mel-frequency cepstral coefficient, artificial neural network, multilayer perceptron, mlp classifier.

**CHAPTER – 1**

**INTRODUCTION**

**1.1 Origin of the problem**

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. At present, speech emotion recognition is an emerging crossing field of artificial intelligence and artificial psychology; besides, it is a popular research topic of signal processing and pattern recognition. The research is widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on.

**1.2 Basic Definition and Problem**

**Speech emotion recognition:-**

Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions- including Neutral, Anger, Happiness, Sadness in which any intelligent system with finite computational resources can be trained to identify or synthesize as required.

**Mel – Frequency Cepstral Coefficient:-**

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

**Artificial neural network:-**

An artificial neural network (ANN) is the component of artificial intelligence that is meant to simulate the functioning of a human brain. Processing units make up ANNs, which in turn consist of inputs and outputs. The inputs are what the ANN learns from to produce the desired output.

**Multilayer perceptron:-**

A multilayer perceptron (MLP) is a class of [feedforward](https://en.wikipedia.org/wiki/Feedforward_neural_network) [artificial neural network](https://en.wikipedia.org/wiki/Artificial_neural_network) (ANN).

An MLP consists of at least three [layers](https://en.wikipedia.org/wiki/Layer_(deep_learning)) of nodes: an input [layer](https://en.wikipedia.org/wiki/Layer_(deep_learning)), a hidden [layer](https://en.wikipedia.org/wiki/Layer_(deep_learning)) and an output [layer](https://en.wikipedia.org/wiki/Layer_(deep_learning)). Except for the input nodes, each node is a neuron that uses a nonlinear [activation function](https://en.wikipedia.org/wiki/Activation_function).

**1.3 Problem Statement**

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and the associated affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. basic emotions like calm, happy, fearful, disgust etc. are analyzed from emotional speech signals.

**1.4 Real time Applications of Proposed work**

* **In call center**

Using speech emotion recognition, the employees recognize customers’ emotions from speech, so they can improve their service and convert more people.

* **Telephone Interviews**

In the interviews interviewer can detect the emotion of the candidate using speech emotion recognition

**CHAPTER-2**

**REVIEW OF LITERATURE**

**2.1 Description of Existing Systems**

**Paper-1:-**

**Title: -** Speech emotion recognition using svm - 2018

**Authors:-** Peipei Shen**,** Zhou Changjun**,** Xiong Chen

Speech sample [2, 4, 6, 9] is first passed through a gender reference database which is maintained for

recognition of gender before it gets into the process. Statistical approach [5] is followed taking pitch as feature

for gender recognition [9]. A lower and upper bound pitch for both male and female samples could be found

using the reference database [14]. Input human voice sample was first broken down into frames of frame size

16 ms each. This was done for frame level classification in further steps.

For each frame MFCC(Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion

recognition. Reference database [14] is maintained which contains the MFCCs of emotions i.e. of Sad, Anger,

Neutral and Happy.

MFCC of the frames were compared with the MFCCs stored in reference database and the distance was

calculated between the comparable frames. Based on the distance of the analysis frame from the reference

database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional

frame count.

Speech sample [2, 4, 6, 9] is first passed through a gender reference database which is maintained for

recognition of gender before it gets into the process. Statistical approach [5] is followed taking pitch as feature

for gender recognition [9]. A lower and upper bound pitch for both male and female samples could be found

using the reference database [14]. Input human voice sample was first broken down into frames of frame size

16 ms each. This was done for frame level classification in further steps.

For each frame MFCC(Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion

recognition. Reference database [14] is maintained which contains the MFCCs of emotions i.e. of Sad, Anger,

Neutral and Happy.

MFCC of the frames were compared with the MFCCs stored in reference database and the distance was

calculated between the comparable frames. Based on the distance of the analysis frame from the reference

database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional

frame count.

Speech sample [2, 4, 6, 9] is first passed through a gender reference database which is maintained for

recognition of gender before it gets into the process. Statistical approach [5] is followed taking pitch as feature

for gender recognition [9]. A lower and upper bound pitch for both male and female samples could be found

using the reference database [14]. Input human voice sample was first broken down into frames of frame size

16 ms each. This was done for frame level classification in further steps.

For each frame MFCC(Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion

recognition. Reference database [14] is maintained which contains the MFCCs of emotions i.e. of Sad, Anger,

Neutral and Happy.

MFCC of the frames were compared with the MFCCs stored in reference database and the distance was

calculated between the comparable frames. Based on the distance of the analysis frame from the reference

database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional

frame count

Speech sample [2, 4, 6, 9] is first passed through a gender reference database which is maintained for

recognition of gender before it gets into the process. Statistical approach [5] is followed taking pitch as feature

for gender recognition [9]. A lower and upper bound pitch for both male and female samples could be found

using the reference database [14]. Input human voice sample was first broken down into frames of frame size

16 ms each. This was done for frame level classification in further steps.

For each frame MFCC(Mel Frequency Cepstral Coefficient)was calculated as the main feature for emotion

recognition. Reference database [14] is maintained which contains the MFCCs of emotions i.e. of Sad, Anger,

Neutral and Happy.

MFCC of the frames were compared with the MFCCs stored in reference database and the distance was

calculated between the comparable frames. Based on the distance of the analysis frame from the reference

database, one can classify the frame as anger, happy or normal. The output is displayed in terms of emotional

frame count

**Description**:- Speech sample is first passed through a gender reference database which is maintained for recognition of gender before it gets into the process. Statistical approach is followed taking pitch as feature for gender recognition . Input human voice sample was first broken down into frames of frame size 16 ms each. This was done for frame level classification using SVM in further steps.. Reference database is maintained which contains the MFCC of emotions i.e. of Sad, Anger, Neutral and HappyBased on the distance of the analysis frame from the reference database, one can classify the emotion as anger, happy or normal. The output is shown as the emotion detected.

**Paper-2** :-

**Title** :-Features and classifiers for emotion recognition from speech - 2020

**Authors**:- Theodoros Iliou, Ioannis Giannoukos

**Description** :- Survey from 2000 to 2011 describing various features (considering non-linguistic and linguistic information) and feature selection methods, and providing a comparison of classification performance of traditional classifiers, ANNs, and their combinations. The major shortcoming for direct comparison of SER systems is considered to be a lack of uniformity in the way the methods are evaluated and assessed.

**Paper-3:**

**Title:-** Emotion detection from text and speech: a survey - 2017

**Authors**:- Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne , Reda Alhajj

**Description**:-The survey covers existing emotion detection research efforts, emotion models, datasets, detection techniques, their features, limitations, and some possible future directions. Emotion analysis from text is also thoroughly described. Review of the deep learning techniques for SER: RNN, recursive neural network, deep belief network (DBN), CNN, and auto encoder (AE).

**Paper-4:-**

**Title:-** Speech Emotion Recognition Using Deep Learning Techniques - 2016

**Authors:**- Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar

**Description**:- Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, contributions made toward speech emotion recognition and limitations related to it.

**2.2 Summary of Literature Study**

Considering our project, we started off with finding the already done researches and successful projects and papers over the internet and on the IEEE official website. We did find out the surveys made by Maisy Wieman, Andy Sun. Their research paper specifies the correct and on point information about the vocal pattern analysis to detect the emotions. We also tried to find out the algorithms and methods used to determine the features from the vocal pattern.

**CHAPTER-3**

**PROPOSED METHOD**

**3.1. Multilayer Perceptron Classifier (MLP Classifier)**

Subsequent work with multilayer perceptrons has shown that they are capable of approximating an XOR operator as well as many other non-linear functions. Multilayer perceptrons are often applied to supervised learning problems. They train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers. The number of hidden units to use is far from clear. As good a starting point as any is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units.

**Stepwise process**:-

1. Make the necessary imports

2. Define a function extract\_feature to extract the mfcc, chroma, and mel features from a sound file.

3. Define a python dictionary  to hold numbers and the emotions available in the RAVDESS dataset, and a list to hold those we want calm, happy, fearful, disgust.

4. Now, we load the data with a function load\_data() – this takes in the relative size of the test set as parameter. x and y are empty lists; we’ll use the glob() function from the glob module to get all the pathnames for the sound files in our dataset. The pattern we use for this is: “D:\\ravdess data\\Actor\_\*\\\*.wav”.

5. Time to split the dataset into training and testing sets! Let’s keep the test set 25% of everything and use the load\_data function for this.

6. Observe the shape of the training and testing datasets

7. And get the number of features extracted.

8. We use classifier, it optimizes the log-loss function using LBFGS or stochastic gradient descent. Unlike SVM or Naïve Bayes, the classifier has an internal neural network for the purpose of classification. This is a feedforward ANN model.

9.  Fit/train the model.

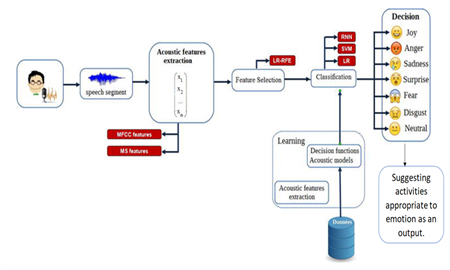
10. Let’s predict the values for the test set. This gives us y\_pred i.e., emotions predicted for the features in the test set.

11.  To calculate the accuracy of our model, we’ll call up the accuracy\_score() function we imported from sklearn. Finally, we’ll round the accuracy to 2 decimal places and print it out.

12. Based on the emotion detected the chatbot will give suggestions to the user.

**3.2 System Architecture Diagram**

Here we will take audio input from the user and we will calculate feautures like Mel-frequency cepstrum coefficients (MFCC), chroma and mel feautres and used to train the MLP classifier. So,we use libraries like Librosa, sklearn, pyaudio, numpy and soundfile to analyze the modulations and process the speech. Then we will obtain the emotion based on the audio input. Based on the emotion detected the chatbot will give suggestions to the user.



**Figure.3.2.1 Counselling using Speech Emotion Recognition System**

**3.3 System Implementation**

In the Speech Emotion Recognition System (SER), the audio files are given as the input. The data sets travels through a number of blocks of processes which makes it executable to help for the analysis of the speech parameters. The data is preprocessed to change it to the suitable format and the respective features from the audio files are extracted using various steps such as framing, hamming, windowing, etc. This process helps in breaking down the audio files into the numerical values which represents the frequency, time, amplitude or any other such parameters which can help in the analysis of the audio files. After the extraction of the required features from the audio files, the model is trained. We have used the RAVDESS dataset of audio files which has speeches of 24 people with variations in parameters. For the training, we store the numerical values of emotions and their respective features correspondingly in different arrays. These arrays are given as an input to the MLP Classifier that has been initialized. The Classifier identifies different categories in the datasets and classifies them into different emotions. The model will now be able to understand the ranges of values of the speech parameters that fall into specific emotions. For testing the performance of the model, if we enter the unknown test dataset as an input, it will retrieve the parameters and predict the emotion as per training dataset values. The accuracy of the system is displayed in the form of percentage which is the final result of our project.

**3.4 Training**

Once configured, the neural network needs to be trained on your dataset.

1. **Data Preparation**

You must first prepare your data for training on a neural network. Data must be like a

numerical, most common example being real values. If you have categorical data, such

as a sex attribute with the values “male” and “female”, emotion attributes such as the

“happy”, “sad”, “angry” etc. you can convert it to a real-valued representation which is

called a one hot encoding.

**2. Training**

The input to the model should be the features extracted along with the emotion as a

category that it belongs to, stored correspondingly into respective arrays so that, the

classifier will be able to identify the patterns, correlations and then classify the data.

This training helps the model to understand, which emotions have what range of the

respective features. So, when an unseen data is given as an input, it will be able to

correlate and predict the emotion**.**

**3. Prediction**

Once a neural network has been trained it can be used to make various predictions.

You can make predictions on test data in order to estimate the skill of the model on

unseen data. You can also deploy it operationally and use it to make predictions

continuously.

Validation Set (20%)

Testing Set (20%)

Training Set (60%)

To Avoid Over fitting

To Test Accuracy

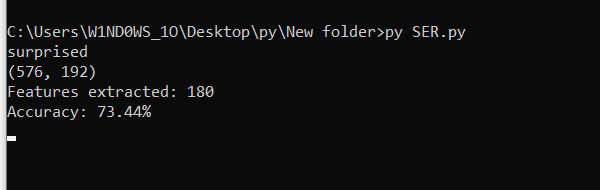
To Train model

**Figure 3.4.1 Training and Testing modules**

**CHAPTER-4**

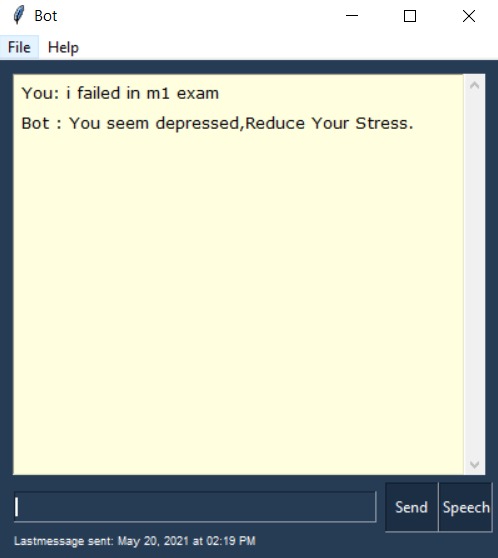
**Results and Observations**

Accuracy was calculated for one emotion at a time. # Calculate the accuracy of our model. accuracy = accuracy score (y\_true=y\_test, y\_pred=y\_pred) # Print the accuracy print("Accuracy: {:.2f}%".format(accuracy\*100)) Accuracy: 100.00%



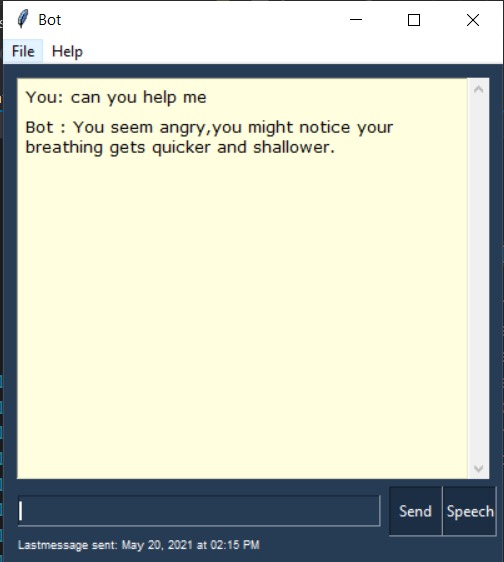
**Figure.4.1 Accuracy**

The accuracy obtained is nearly 73.5%.



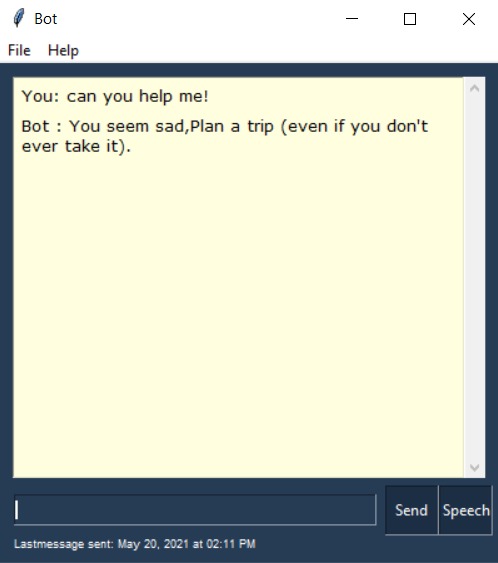
**Figure 4.2 Depressed**

we open the chatbot and we will provide it with an audio as input then the ML model will recognize the emotion depending on the different features processed from the audio file. The bot will then provide counselling, in this case a suggestion. As we can see in the above figure 3.2 depressed is the emotion classified by the model so it will suggest ways to reduce stress



**Figure 4.3 Angry**

This particular test case provides a better understand of the working ML model. As we can see the input query is similar to the input query in the previous test case but depending on the features obtained from the audio file provided as an input the emotion to be recognized is different, in this case the emotion is recognized to be anger and suggestions are provided accordingly in dealing with it. The suggestion provided is to control your breathing.

****

**Figure 4.4 Sad**

As the emotion is recognized to be sad our model will suggest us do activities which could help in changing our state of mind. In this scenario go on a trip is the activity suggested to a sad person which provides a change of scenery and frequent interaction with friends bound to help a person dealing with sadness.

**CHAPTER – 5**

**CONCLUSION AND FUTURE STUDY**

**5.1 Conclusion**

Hence our project presents a new way to give the ability to machine to determine the emotion with the help of the human voice. It will give the machine the ability to have a better approach towards having a better conversation and seamless conversation like human does.

**5.2 Future Study**

We will try to integrating an app that will suggest counsel statements to any one who uses the app according to the voice recognization software and reply accordingly.

**REFERENCES**

[1] L. Rabiner, B-H Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, 1993

[2] B. Schuller, S. Steidl, A. Batliner, "The INTERSPEECH 2009 Emotion Challenge"

INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, pp. 312-315, 2009.

[3] Laurence Devillers, Laurence Vidrascu, "Real-Life Emotion Recognition in Speech, Speaker Classification II: Selected Projects. Lecture Notes in Computer Science, vol. 4441, pp. 34-42, 2007.

[4] E. M. Albornoz, D. H. Milone, H. L. Rufiner, "Spoken Emotion Recognition using Hierarchical Classifiers," Computer Speech and Language, vol.25, pp. 556-570, 2011.

[5] J. K. Joy, Aparna; Ram, Shreya; Rama, S, "Speech Emotion Recognition using Neural Network and MLP Classifier," IJESC, vol. 10, no. 4, 2020.

[6]N. Amir, "Classifying emotions in speech: a comparison of methods", Eurospeech 2001, Poster Proceedings, Scandinavia, 2001, pp. 127-130.

[7] J. P. Pinto, “Multilayer perceptron based hierarchical acoustic modeling for automatic speech recognition,” Ph.D. dissertation, EPFL, Lausanne,Switzerland, 2010.

[8].H.K. Palo, Mihir Narayana Mohanty and Mahesh Chandra. Use of different features for Emotion Recognition using MLP network. Springer India 2015, Computational Vision and Robotics, Advances in Intelligent Systems and Computing

[9] Xiao, Z., E. Dellandrea, Dou W.,Chen L., “Features extraction and selection for emotional speech classification”. 2005 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp.411- 416, Sept 2005.

[10] J. K. Joy, Aparna; Ram, Shreya; Rama, S, "Speech Emotion Recognition using Neural Network and MLP Classifier," IJESC, vol. 10, no. 4, 2020