Human Emotion Identification from Speech using Neural Network

Bhoomi Rajdeep
M.Tech student – Computer Science Engineering Dept.
Rai School of engineering,
Rai University
Ahmedabad, India
Bhoomi9830@gmail.com

Hardik B. Patel
Assistant Professor
Computer Science Engineering Dept.
Rai School of Engineering,
Rai University
Ahmedabad, India

Sailesh Iyer
Professor and Dean
Computer Science Engineering Dept.
Rai School of Engineering,
Rai University
Ahmedabad, India

***Abstract*— Detection of mood and behavior by voice analysis which helps to detect the speaker's mood by the voice frequency. Here, I aim to present the mood like happy, and sad and behavior detection devices using machine learning and artificial intelligence which can be detected by voice analysis. Using this device, it detects the user's mood. Moreover, this device detects the frequency by trained model and algorithm.**

The algorithm is well trained to catch the frequency where it helps to identify the mood happy or sad of the speaker and behavior. On the other hand, behavior can be predicted in form, it can be either positive or negative. So, this device helps to prevent mental health issues and is used in medical and gaming testing.

Furthermore, it is easy to identify a person's mood by their expression and by their actions in daily activities. But it is effective and challenging to detect mood and behavior by voice frequency because a rich environment affects the most. Thus, this device works as a signal processing.

Keywords—Machine Learning, Speech reorganization, Human Emotion, MLP (Multilayer Perceptron), MFCC

I. INTRODUCTION

A. Overview

Various research work has been studied to identify the problems associated with the current research work and also solution been developed by modifying the existing algorithms as well as by developing a new method for the identification of the sentiment associated with the voice of the person. Block diagram of the system flow is given below to understand the proposed flow of the model.

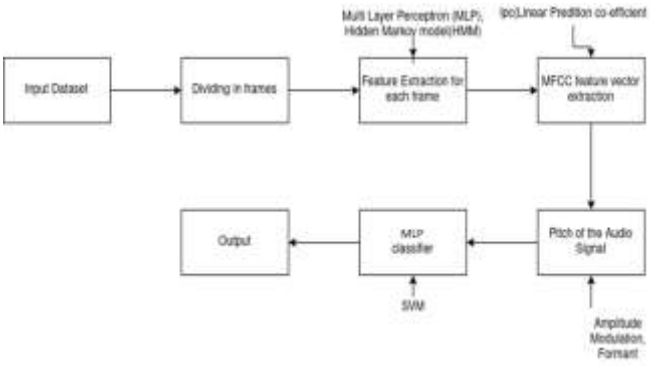

Fig 1.1.1: System flow diagram of the proposed model

Proposed system built with the help of the following modules.

1. Selection of the Dataset
2. Conversion of the data into the frames
3. Feature extraction process from the frame
4. MFCC feature vector extraction
5. Identification of the pitch from the Audio signal
6. Classification using MLP (Multilayer Perceptron)
7. Generation of the Output

In these seven modules, three modules are common in every research work such as, selection of the data, frame conversion and the generation of the output. Remaining four modules need more research to improve the accuracy of the model which are, feature extraction, MFCC feature

vector, identification of the pitch and the classification task. Following section describes the research work carried out related to these processes in a brief manner.

Feature Extraction

Most of the research work carried in the past used a Multi Layer Perceptron for the extraction of the features for the analysis of the features, and to identify the problem associated with the MLP feature extraction, various articles and research work is studied. Problems that are identified during the study of the research work are given below.

- In the Multilayer perceptron, when layers increase, number of the parameters required to train the model increases in a multiplicative order, which increases the complexity of the model.
- Information related to the geographical location is spatial in a nature and this kind of the information is hard to process.

During the study if the hidden Markov model from the various research work, following problems are found.

- well as the associated dependencies because of there are number of unstructured parameters are This method is unable to define a proper relationship as available in the model.

To resolve the issue, proposed model used a unique solution with the help of the Discrete Fourier Transform as well as use of the Mel Frequency Wrapping which are illustrated with the help of the following illustration.

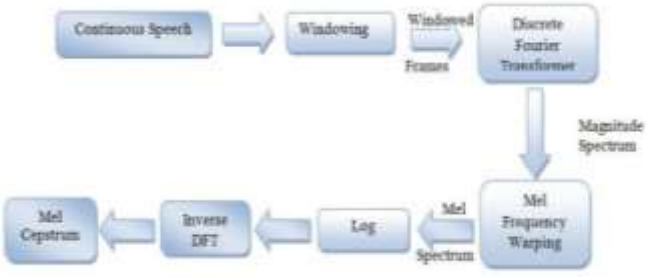

Fig 1.1.2: the Feature Extraction Process with the help of the Discrete Fourier Transform.

Various advantages of using Discrete Fourier Transform as well as the Mel Frequency Wrapping are given below.

With the help of the Fourier Transform, it is possible to perform loss less transformation, which means model can utilize all the parameters available in the signal. Also Fourier Transform provides an advantage of utilization of the various attributes related to the signal such as the amplitude of the signal, frequency domain, phase of the signal etc.

MFCC Feature Vector Extraction

For the feature vector extraction process, various models related to the coding of the speech has been studied which are used for the designing of the filters. For the selection of the coefficient of the filtering technique, various research work is taken for the reference and problems related to the work is identified. Various research work used a Linear method for the prediction which comes with the various disadvantages which are given below.

In birates of the signal, quality is reduced which ultimately reduce the quality of the signal. For the long distance transmission, loss-less compression is not possible with the help fo this technique and not suitable for the long distance transmission.

MFCC feature vector extraction process helps to overcome this disadvantage with the help of the Quantum Neural Network, which provides a parallel computation method to reduce the training time of the data. Also this technique used a IPSOQNN model which allows the fast recognition of the prediction of a signal. Also MFCC includes the characteristic for the identification of the phones available in the speech which helps to reduce the complexities of the computation.

Pitch of the Audio signal and their property

In general, amplitude modulation is used for the extraction of the various properties of an audio signal, but it suffers from the various problems such as amplitude modulation can not work with the lower bandwidth which means that bandwidth

required for the modulation process must be higher than the original signal. Another problem faced by the amplitude modulation is that it is very sensitive towards the noise, which means that if there is a higher noise is available in the signal then it is difficult for the model to recognise the orignial signal with the help of the amplitude modulation technique.

To identify the solution for this problem various research work has been studied and for the solution Pitch detection method is selected which provides a various advantages such as, this technique can easily identify the difference between the human voice and the instumental music. This technique is also capable of the identification of gender. This technique keeps a timestamp of the recording of a voice and based on the timestam, model can be able to identify that it is a morning, noon or the evening. With the help of the pitch, model can also able to identify the age of the person, who recorded the audio.

MLP classifier for the Classification

For the classification of the voice, various research work has been studied, which used Support Vector Machine for the classification. This technique has many advantages but still suffers from the disadvantages such as:

SVM algorithm is not suitable when the dataset is large in the size. Also there is a problem of the overlapping in the target class when noise is available in the signal. When there are more features are available in the dataset then it is difficult for the algorithm to provide a more accurate results.

To resolve the issues associated with the SVM algorithm, proposed solution used a Multi Layer Perceptron for the classification task which provides an efficient results even if the size of the dataset is large. It is possible to manage a large data with the help of MLP without deleting the any data from the dataset. This algorithm also provides an equal importance to every parameters available in the dataset and also utilise every veriable for the classification task.

B. Methodology

It is possible to create a speech-based emotion detection system with the help of a machine learning model and by training the neural network. The model can match the phases of the implementation with any other project that uses Machine Learning to make decisions. In the current proposal, the very first step is to collect data, after which I train the model and make the proper decision depending on the data I have collected. This process addresses the concerns of data representation and quality associated with the gathered data in the second stage, which is a collection of different machine learning tasks that are conducted over the collected data. The third step is feature extraction. The third stage is regarded as the model's heart, and it consists of methods for creating the model, learning from it, and training it over a given set of data, among other things. Finally, I display the appropriate output result in terms of emotion, which is based on the input data.

C. The Problem Areas Addressed by this Project

There are following areas that proposed solution resolves

- In various research work for the extraction of the features, MLP is used which provides a redundant result when the parameters grow. Proposed solution tries to resolve this issue with the help of the Discrete Fourier Transform [1].
- Amplitude modulation is used in the various research work for defining the property of the audio signal which requires the higher bandwidth than the original signal. Proposed solution used a pitch of the audio signal to resolve the issues associated with the amplitude modulation [2].

For the classification task various work used a SVM classifier which has a disadvantage of inaccurate results when dataset is large in the size. To resolve the issue of SVM proposed model used

MLP classifier which can operate in any size of the dataset [3].

D. *Objectives of the Research*

Following are the objectives of the proposed model which are described in a brief manner.

1. Voice Detection: With the help of the detection component, proposed model is able of detect the voice in more efficient manner. These components are helpful for the segmenting the voice signal into various frequency level which allows more flexible feature extraction process.
2. Feature Extraction: With the help of the Discrete Fourier Transform, feature extraction process is implemented which allows the loss-less feature extraction and utilize all the parameters available in the voice signal for the extraction of the feature.
3. Optimization Algorithm: With the help of the signal processor, frequency of the signal is optimized. With the help of the optimization algorithm, detection of the various moods available in the signal can be identified.
4. Database Creation: For the storage of the clips of the various voice signal, database is created in a secure fashion. Various types of the libraries are utilized for setting up the database which improves the performance of the various database related queries.
5. Classification of voice using Classifier: MLP classifier is used for the classification of the voice signal to improve the accuracy of the classification tasks. This algorithm is also useful for the extraction of the mood of a person at the different frequency level and based on the level of the frequency it provides an accurate result.
6. Analysis of the Speech: With the help of the speech signal analysis process, model is able to identify the mood of the speaker by utilizing the frequency of the speech signal. Feature algorithm is implemented to perform the analysis task and to detect the appropriate mood of the speaker.
7. Spectrum for the speech: This device is useful for collecting the data and the experiment performed and finds the actual frequency range associated with that data. After that it predict the mood which is closest to the actual values and finally discover the mood of the speaker.

II. Literature Review

A. *Speech Classification*

Speech helps to define the affection available in the sound and also provides an information related to the perceptual variation available in the speech. With the help of the various technical measures such as the amplitude of the signal as well as the wave length and tone of the signal, model is able to distinguish between the speech [4].

Text is a useful medium for the communication and also best way for the long-distance communication, model needs to convert the speech signal into the text format because signal consists more noise than the text format. Following section describes the speech and the text representation technique.

Speech and text Representation

To train the model in the domain of the voice, CMU US RMS ARCTIC speech database is very useful and consists many expressions which are used in the speech signal and signals available in this dataset are of 16 bit and each have the frequency of the 16KHz. Dataset consists 1600 various types of the sample each of the 16 bits. To understand the process of the text representation, one example is taken into the consideration which used a text representation of the sentence “we will ever forget it” and frequency associated with this text is illustrated in the following figure.

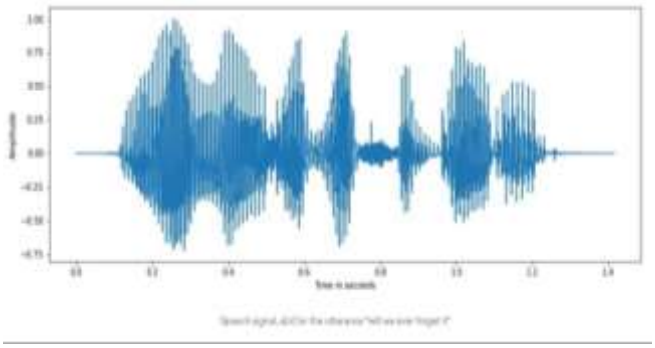

Fig 2.2.1 Representation of the sentence with the help of the speech signal

Above illustration represents the disparity of the amplitude and from the value of the highest amplitude available in the signal, we can divide the whole signal into the two different parts in which first part is known as the phonemes which is the fluent sound while the second part is known as the silent zone which indicates the absence of the speech in the given signal.

Database also consists the text equivalent to the given signal and it is defined by the various levels of the phonemes of the signal. Figure given below illustrates the various phonemes defined with in terms of the various time intervals.

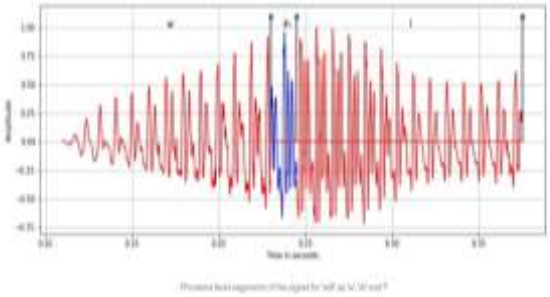

Fig 2.2.2: Illustration of the signal by different levels of the phonemes

In general, nature of the voice phonemes is quasi-periodic in nature where phonemes can be divided into two different categories which are vowels and semi-vowels. In the given example “ih” is the vowel and on the other end “w” and “l” are the example of the semi-vowels.

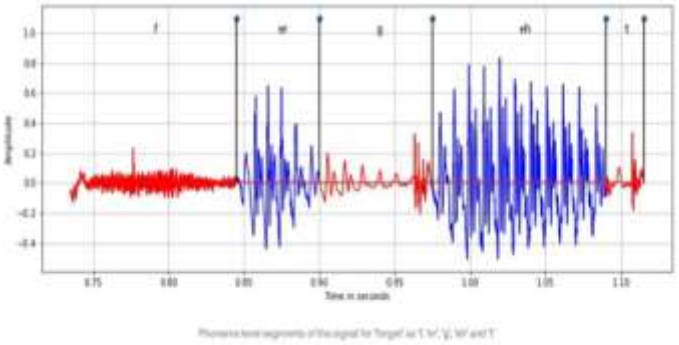

Fig 2.2.3: Representation of the different level of phonemes available in the signal

Figure given above illustrates the various types of the phonemes such as uttered one and uttered one or stop. In the above example uttered one phoneme are quasi-periodic in nature while uttered one or stop is clamorous in the nature and does not use for the construction of the vocal fold. Here we can see that “er” and “eh” are the uttered one while “f” is the unuttered one. Here “g” and “t” represent as a stop.

To understand the relation between the text and the phonemes, following example is taken which converts the whole sentence into the different phonemes.

*Let’s say we have a sentence such as “*will we ever forget it” then we can write respective phonetic sequence as ‘w’, ‘ih’, ‘l’, ‘w’, ‘iy’, ‘eh’, ‘v’, ‘er’, ‘f’, ‘er’, ‘g’, ‘eh’, ‘t’, ‘ih’, ‘t’ and we are able to depict that word “will” is depicted to the phonemes as ‘w’, ‘ih’, ‘l’

To understand the speech signal at a different level of the frequency, Short Term Fourier Transform is useful and based on the level of the frequency, spectrogram is generated for the visualization of the various frequencies available in the sound.

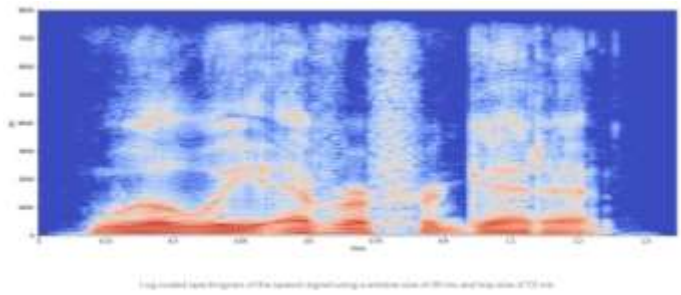

Fig 2.2.4 Spectrogram of the speech signal based on the Log scale

With the help of the spectrogram, mapping of the amplitude is possible in terms of a log scale. For example, if there is a frame size of the 30ms then it will provide 480 samples by multiplying 16KHz frequency with the 30ms frame size. If we reduce the frame size by let's say 5ms then spectrogram will provide a 80 samples. The main reason behind the reduction of the frame size is, there is a frequent variation in the signals and it is difficult to identify the variation while the sample size is large. By reducing the number of samples, model is able to get more accurate results. With the help of the frequency available in the signal we are able to analyze the variations in the amplitude of the signal in each interval of the time in an efficient manner.

There is a one drawback of using spectrogram which is when we reduce the size of the frame it creates a similar component for the various range of the frequency and it would be difficult to identify the difference between the two words with the help of the Fourier Transform.

Identification of the Noise in the signal

In general, Noise is an unwanted signal available in the original voice signal and it will create a huge impact on the effectiveness as well as accuracy of the output of a model. Identification and removal of that kind of the signal is necessary before applying signal into the model. Following section describes the removal process of the noise from the original signal [5].

In mathematical term signal is represented by s[n] and the respective noise is represented by w[n] then we can represent the outcome using the notation u[n] which is the addition of the amplitude of the original signal and the noise signal and described by the equation given below.

$$u[n]=s[n] + w[n]$$

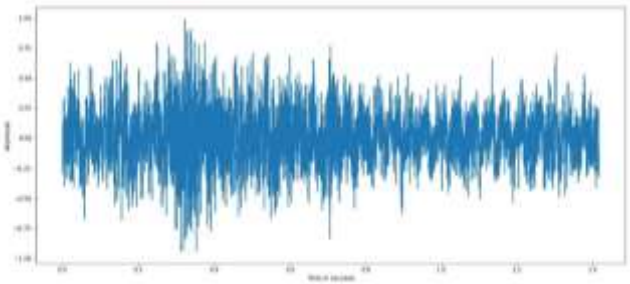

Fig 2.2.5: Example of Noise in NOISEX92 Database

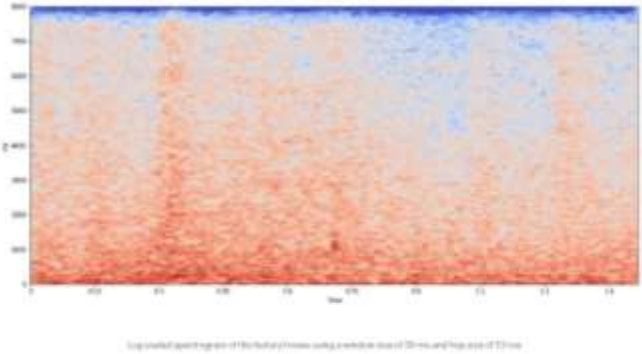

Fig 2.2.6: Log scaled spectrogram of the noise represent in the figure 2.2.5

In the above figure, rate of the sampling is 16KHz and it is same as the rate of original signal as mentioned earlier.

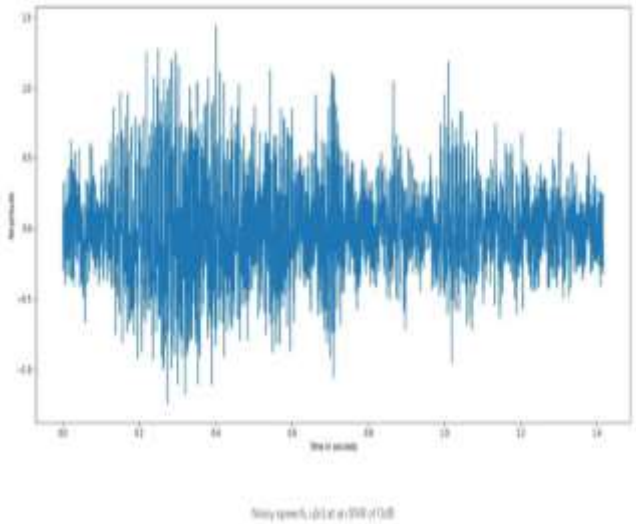

Fig 2.2.7: Noisy speech u[n] at an SNR of 0db

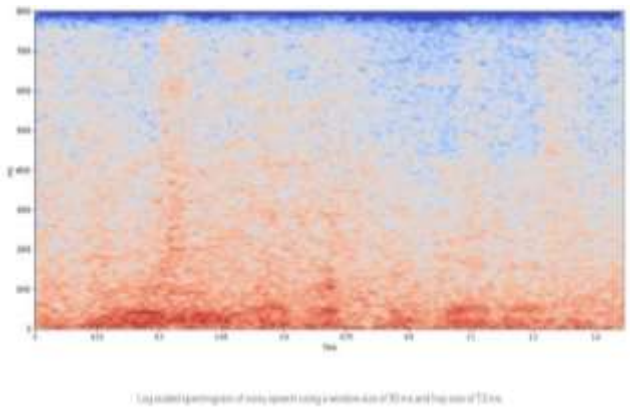

Fig 2.2.8: Log scaled spectrogram of noisy speech using a window size of 30 ms and hop size of 7.5 ms

Noise signal in the above illustration is represented with the help of time and the frequency domain and noise is adjusted with the original signal which represent the various division using the algorithm of an amplification. This algorithm helps to identify the noise available in the original signal and then help to improve the original signal by removing the Nosie.

There is always necessary to represent the noise in the form of the external signal. For example, wrong spelling of the word generates the wrong outcome and change the meaning of the entire sentence. This problem is illustrated in the following example.

A line like "Will we ever forget it?" would become "Will we never forget it?" if ever were altered to never, completely altering the original meaning. Instead of "Will we never forget it?" we may say "Will we never forget it?" which would change the meaning of the phrase and also make it illogical owing to incorrect spelling.

Speech Analysis Mechanism

This section explains the analysis technique related to the speech. In general, there are many regions are available in the signal of the speech when the speech is being recorded. To simplify the voice signal, it requires to divide the speech signal into the different small portions and then need to identify the silent regions available in the speech signal with the help of the filtering technique. Processed signals will be used for the classification as well as the recognition in the next phase [6].

To identify the silent regions available in the signal, filtering technique calculates the relative energy division in the available time frame. In the following illustration, frame size of the 20 ms is taken and energy signal is denoted by the s[n]. Energy signal is added to the square of the s[m] to that signal where range of the “m” is in between positive 10 ms to the negative 10 ms in a given sample size which is in terms of “n”. The reason behind the selection of a small frame size is that by reducing the size of the frame, it is possible to detect the differences associated with the different energy levels in the signal as well as it also helps to identify the silent regions of the signal.

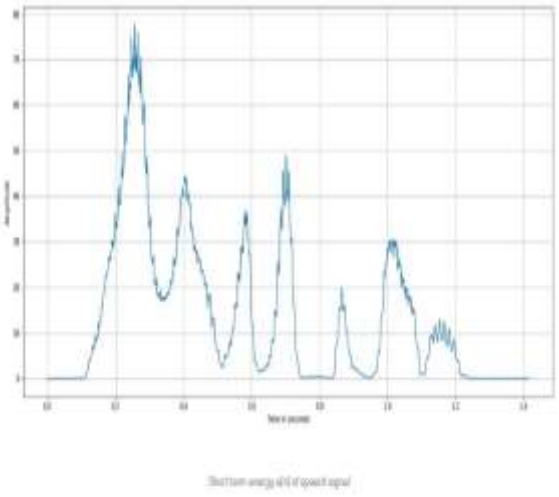

Fig 2.2.9: Speech signal in terms of the Short-Term Energy

Above illustration helps to identify the various energy levels of the voice signal. Above illustration also shows that the silent regions produce different energy levels than the original signal.

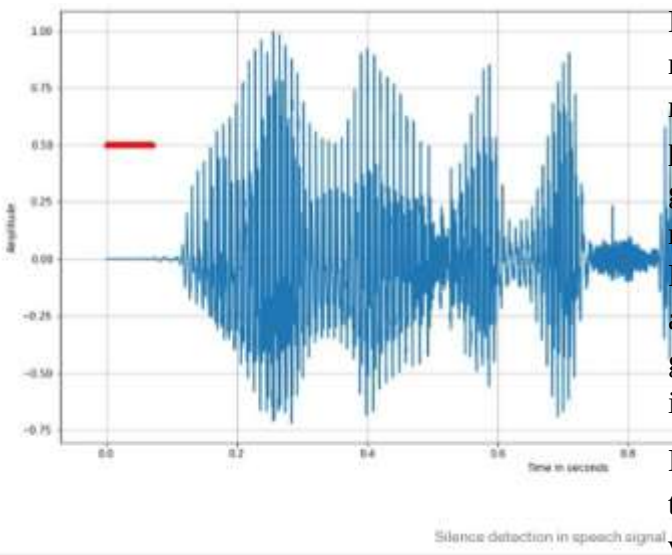

Fig 2.2.10: Illustration of the energy level of the silent region in the speech signal

In the above illustration, silent divisions of the signals are displayed with the red colour and the value of the threshold is near to 0.01 percentage in the respect of the total energy consisted by the actual signal and the selection of a threshold value is based on the observed fluctuations in the level of energy. The problem faced during the detection of the silent region is that it generates the high level of the energy signals and it is difficult task to identify that much high energy level signals. To resolve the issue, it is necessary to observe the fluctuation of the short-term energy signal with the help of the low-pass filtering technique which is useful to detect the high-level frequency noise signal.

B. Machine Learning Models

What is Machine Learning?

Machine Learning algorithms are a set of various functions which are used in the process of learning in terms of the supervised learning as well as the unsupervised learning. This process includes various phases such as train the algorithm with the help of the sample data and to predict the outcome by providing the new data to the model. In machine learning, selection of the algorithm is depending on the which type of the outcome is needed [7].

Data is faded to the algorithm in form of M*N matrix where M represent the column while the N represents the row of the matrix. For the input, particular database is selected and based on the given dataset, algorithm converts the data into the matrix format and then starts the training phase. For the new data, trained model tries to predict the appropriate result based on the historical data given into the training phase and also identify that in which the data belongs.

During the process of matrix conversion, we select two variables X and Y. Apart from these two variable X represents the attributes such as features of the data or it can be an independent variable. On the other hand, Y represent class label or we can say it is a dependent variable. X represent input variable while the Y represent an output variable. In below figure I illustrates how variable X and Y forms a matrix.

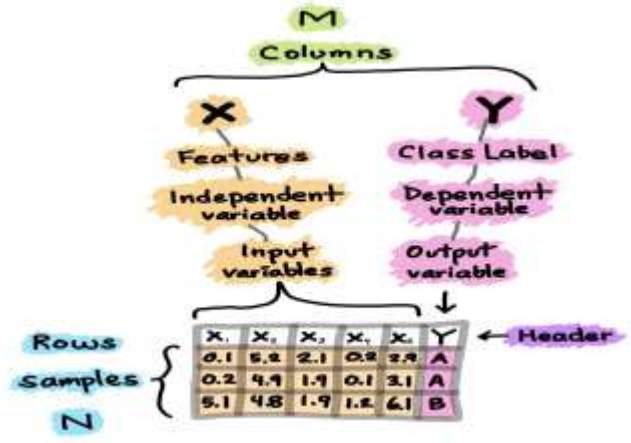

Fig 2.3.1: Representation of the Input and Output variable with the help of the Matrix

As mentioned earlier, machine learning algorithm uses, supervised as well as unsupervised approach for the training, these each approach has a basic difference in terms of performing the task of the classification and the regression. In the Supervised learning approach, algorithm utilize input as well as the output variables while in the unsupervised approach, algorithm utilize only input variable to perform the classification or the regression. In general, regression is used for the generation of the quantitative data while the classification is used for the generation of the qualitative data [8].

Exploratory Data Analysis

Exploratory data analysis is used for the building perception for the existing data with the help of the performing the following task on the given data [9].

This analysis technique uses a descriptive coefficient such as the mean, median, mode, as well as the standard deviation for the summarization of the data

Exploratory data analysis uses tools such as the whisker which is used for finding the difference between the available dataset. This analysis technique also uses the scatter diagram for defining the relationship between the different kinds of features.

This analysis method also uses data shaping for defining the relationship between the various instances of the data in terms of the parent-child relationship and also provides a task such as the pivoting of the given data as well as the filtration technique.

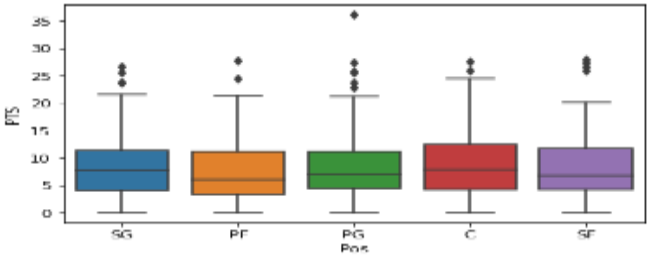

Fig 2.3.2: Illustration of the Box Plot

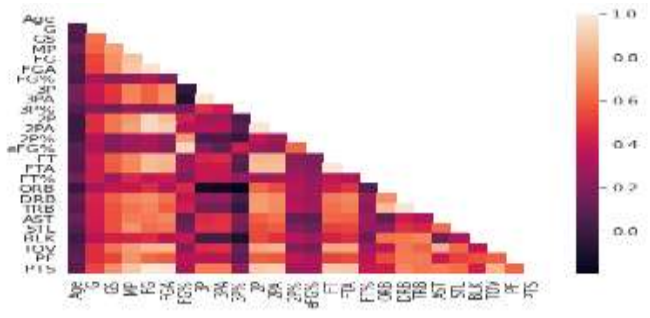

Fig 2.3.3: Illustration of the Heat Map

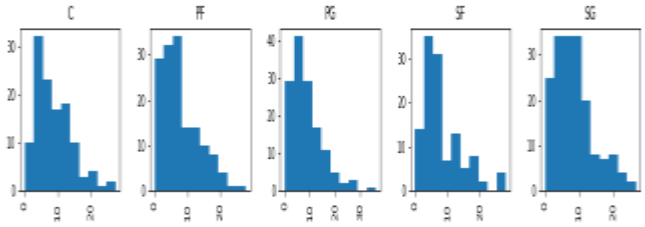

Fig 2.3.4: Representation of the Histogram.

Pre-Processing of Data

In the Pre-processing stage of the data, system applies various kinds of techniques to clean the data as well as to resolve the problems such as spelling check, normalization of data, standardization process to the various values of a data and approximate exact value of the data. It also applies techniques such as logarithmic transformation as well as altering techniques [10].

Splitting of a Data

We have to perform various kind of steps to split the data into different parts and description of each process is given below.

Train-Test Split Method

By using this method, we are able to identify the performance of the algorithm used in a machine learning as well as we are able to predict exact performance of the data by splitting the data into various parts and based on that data model reproduce the new or we can say unseen data [11].

Data splitting task is further divided into two different categories. Apart from these two categories one consists large set of data and used to train the data in the form of a subset with same characteristics while the other category consists smaller size of training data. This task is performed only once during the whole process.

By using the available training data, predictive model is designed which is able to test the remaining data available in the dataset. This method is sometimes known as the hyper-parameter optimization which forecast the new data and class of that data

The process of data splitting and training is demonstrate using figure given below.

Fig 2.3.5: Data Splitting on a Dataset

Train-Validation-Test Split

In this method rather than splitting the data into two parts, this method divides data into three parts namely Training, Validation and the Testing set In this method training set is used to develop the predictive model. Trained data are then validated by using the validation set which identifies the effectiveness of the trained data. In the last phase testing data is used to generate or to make prediction about the unseen or a new data. This process is illustrated in the diagram given below.

Fig 2.3.6: Train-validation Test Split

Cross-Validation

Cross validation technique divides data by n-folds, five-folds, ten-folds, and so on. This technique makes use of two folds: one for testing and the other for learning. [12].

For example, if we use 5-fold method than 1-fold is used for the testing while the remaining 4-fold is used as a training set in the model. Training process is an iterative process which provides a chance to every fold to become the testing data and based on the outcome select fold is used as a testing data. By this technique we are able to build a model which consist 5 different sub-models which consist a performance matrix in terms of the performance index. The process in which number of folds are same as the number of samples available in the data set then this technique is known as the Leave-one out cross-validation.

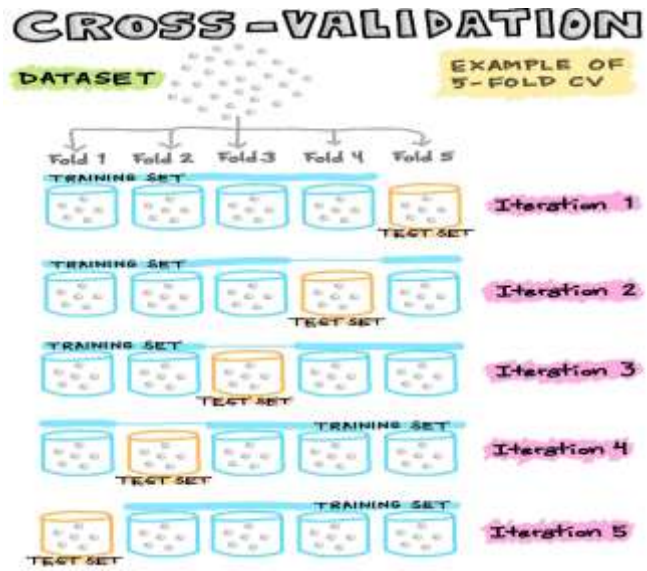

Fig 2.3.7: N- Fold Cross Validation

Model of a Process Building

In this section I try to figure out that how pre-processed data is further utilized to build a model. There are two kinds of data available in the dataset which are qualitative and quantitative data. To process these kinds of data I studied various kinds of classification and regression technique. I found that classification is suitable for the quantitative data while the regression is suitable for the qualitative data.

We can divide these algorithms into mainly two categories based on the approach of the learning. One is known as a supervised learning while the other one is known as an unsupervised learning. Supervised learning method provides a

mathematical support to connect two different variables, let's say X and Y in such a way that integration of these two variables generates a data called labeled data which are further utilized for the model building

Unsupervised learning algorithms on the other hand utilize only input variable from the given dataset and it tends to use unlabeled data in the process of the model building. In simple term we can say that it generates the new data from the available data [13].

The third kind of algorithm is based on the trial-and-error approach for the process of model building which is known as the Reinforcement learning

Hyperparameter Optimization

Using the hyperparameter optimization techniques machine is able to forecast the performance of the model and these parameters are used as a fundamental parameter which directly affect the learning process of the model It provides a flexible size of data block so it is easy to apply it on any kind of data size and universally applicable [14].

For example, in a random forest algorithm we have to optimize two parameters named as mtry and the ntree using the R package of an algorithm. Here mtry is used as a max_feature function while ntree uses n_estimator as a function. Two kinds of R package which are used in Both the RandomForestClassifier () and the RandomForestRegressor () from the scikit-learn library are used in this approach. The variables mtry(max feature) and ntree(n estimator) represent the number of variables we want to randomly sample at each division, and the number of trees growing in the model is represented by that of the factor ntree(n estimator).

In SVM algorithm we have to consider two different kinds of hyperparameter which are known as the C and the gamma parameter. Apart from these two parameters C is useful for the limiting the overfitting of the data while on the other hand gamma parameter is useful to control the width of the Radial Bias Function [15].

Selection of the Feature

By using this process model is able to select some subset of the feature from the collection of different features available. By using this process perception about the data is to be made. There are various kinds of feature selection algorithms are available like Pearson correlation, chi-square method, Lasso Method as well are Tree based methods are some of them

Machine Learning Tasks

Classification and the regression are performed by the machine learning algorithms and I described a brief overview of this task in below section.

Classification

Classification methods uses a various kind of qualitative and the quantitative data as an input and generates the data in a qualitative manner as an output.

The figure given below illustrates the process of the classification using the colored labels in which every color indicates a different dataset.

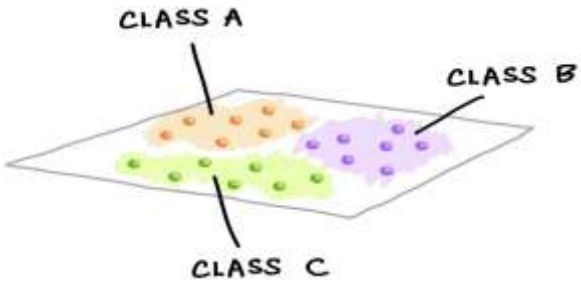

Fig: Example of a classification with different classes.

Performance metrics

For determining if a model's performance is beneficial or harmful, we use a matrices known as performance metrics, as well as some classification algorithms that comprise the following terminology, to determine whether the model's performance is good or bad.To understand the performance of the model we have to use

performance metric which identifies that performance of the model is good or bed by using some fundamental formulas which used with the classification algorithm Invalid source specified. and some of them are given below.

Accuracy of the model is calculated by the following equation.

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

sensitivity (Sn) of the model is calculated by the following equation

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

specificity (Sp) of the model is calculated using the formula given below

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Matthew's correlation coefficient (MCC) is calculated by the following formula.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

If the value of MCC is one then it indicates a best possible forecast while the negative one indicates a worst possible forecast. If this equation gives zero as an output that means it indicates a random forecast for the data.

Regression

By using the regression technique model is able to identify the strength of the relationship between the two variables Invalid source specified.. These two variables are Y represents the dependent variable while the X represents the independent variable in the relationship. This relationship can be defined as a simple function as a Y = f (X). Here X is an input variable and Y is a quantitative output generated by the regressor algorithm. Using scattered diagram, we can plot the outcome generated by the algorithm and this diagram plot the real data versus the forecasted output data

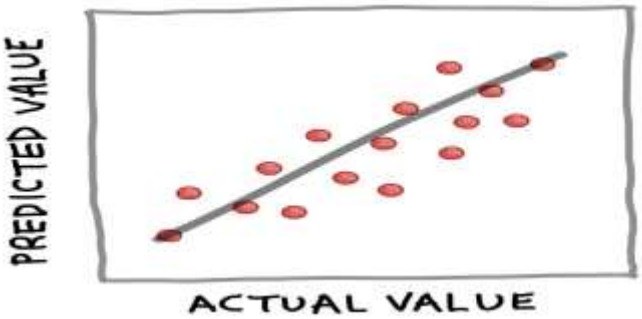

Fig: Scatter diagram of real values vs predicted values generated by the regressor

Performance metrics for the data

With the help of performance metrics, we can decide the precision of the model and how the forecasted data is fitted into the available dataInvalid source specified..

Coefficient of determination is useful for approximate the capabilities of the regression model and it is given by the following equation which is given below.

$$R^2 \equiv 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

If we obtain the result of R2 = 0.7, it shows that the model is capable of expressing 70% of the data, with the other 30% of the data leftover unchecked for (or untraceable for).

RMSE is a better method to estimate the error associated with the data and it is given by the following equation which is given below.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

III. PROPOSED WORK

A. *Releated work*

Human brain can easily detect the mood of the person, even if he or she is meeting first time with you and not familiar to you by analysing the tone

of the voice of a person. With the help of the past research work, we can get the information about how can we decide the sentiment or the emotion behind the voice of the person. This task is done with the help of the enigmatic expression as well as with the help of observing the human eye. Previous research suggests that with the help of detecting the perception of the voice we can identify the mental state of the person in a more accurate manner [16].

Previous research work indicates that how to retrieve the information with the help of the vocal information and model is not only distinguish the perception but also can get the helpful information about the nature and type of the perception of the voice such as affirmative or the defeatist. Based on the perception identified by the model we are able to predict the feeling attached with the voice [17].

Analysis of the sentiment with the help of the voice of the user and to distinguish the speech signal in terms of the semantic as well as non-lexical analysis helps to identify the mood of the person in a more accurate manner. With the help of the signal of the speech, model is able to retrieve the information related to the mental state of the user. With the help of the language, affection of the user can be identified and by applying the conceptual knowledge on that language provides the information related to the perception of the mental state of a user. Research work carried in the past also suggest that with the help of the lexical analysis as well as semantic analysis we can able to extract the information from the signal of the voice. For the extraction of the emotion from the voice signal, vocal split of the language is helpful [18].

It is proved by the Darvin that sentiment of the people can be decided with the help of the voice of a person. Past research suggests that the with the help of the tone of a voice we are able to identify the cerebral condition of the particular person. Many studies suggest that to identify the perception related to the sense of the person, vocal analysis is helpful [19].

[20] Proposed a research work to describe the perception based on the speed of the talking and also provides a proof related to the understanding of a perception related to the different state of the people which is affected by the speed of the talking. In general, speed of the talking increases when person is annoyed or freight. In same manner speed of the talking decreases when the person is in remorse or the dejection state.

B. Proposed work

Proposed flow of the system is illustrated with the following figure with the required modules for the design of the model.

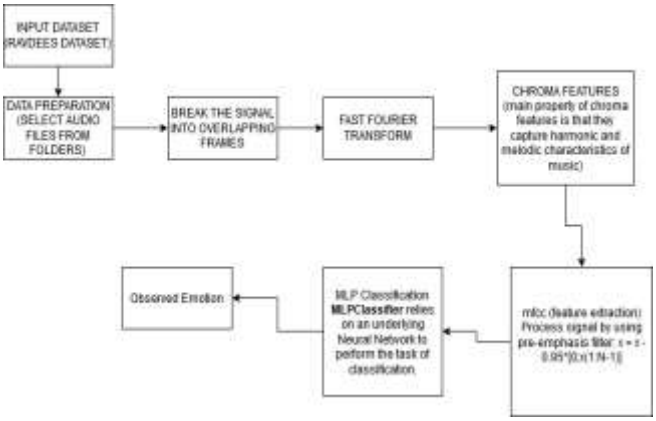

Fig 3.2.1: Flow of the System

The flow of the system given above defines main modules and algorithms used in the system which majorly includes the following things

1. RAVDEES DATASET
2. PREPARATION OF DATA
3. CONVERSION OF SIGNAL INTO FRAMES
4. FAST FOURIER TRANSFORM
5. CHROMA FEATURES
6. MFCC FEATURE EXTRACTION
7. CLASSIFICATION USING MLP
8. DISPLAY OF OUTCOME IN TERMS OF EMOTION

Data Set

Initial and foremost, data must be gathered from various audio files that contain distinct emotional categories and are accessible in the MP3 file

format. It's the first phase in the project's development. For the most part, we are employing the RAVADEES dataset for this purpose.

A numerical representation of these audio files is created as the next step in this process, allowing for further analysis to be performed on them. This process is known as feature extraction, and we will discuss it in further detail later in this section.

Data: It is the RAVDESS dataset, which stands for the Ryerson Audio-Visual Database of Emotional Speech and Song, that we will be working with. These 7356 files have been rated by 247 individuals ten times on the basis of their emotional validity, intensity, and sincerity.

Preparing Data

Preparing data: Developing a function to extract the emotion label and the gender label from each file that is uploaded (although we were only interested in classifying emotion, we also extracted the gender label in case we should decide to attempt to classify gender as well).

When we want to get all of the pathnames for the sound files in our dataset, we'll utilize the glob () method from the glob module.

Feature Extraction

For those who are unfamiliar with the concept of speech as a varying sound signal, people are capable of modifying the sound signal using their vocal tract or their tongue and teeth to pronounce it differently. (See also: Consequently, we must quantify the data and build better representations of the speech signal in order to retrieve the information contained within them. A feature extraction strategy is often used in signal processing, and in this section, I identify certain characteristics of a good signal. Some of the characteristics of an excellent feature are listed in the next section.

It is important to note that each trait is independent of the others and merely co-related to them. Consequently, we will be able to select the feature on an individual and independent basis.

Due to the nature of the feature, which is informative and more descriptive in nature, we are able to select emotional content from it for future investigation..

The feature must be present throughout all of the data samples, and features that are specific to a particular data sample must be avoided.

By using this attribute, we are able to progress with the value of the feature and we are able to remove outliers and missing values from that feature..

Feature Extraction: It was necessary to use librosa's Mel spectrogram in order to extract the log-Mel spectrogram values of each audio file, and then to take the average of those spectrogram values and put it into a new Data frame..

When this happens, the MFCC (which is nothing more than the coefficients that make up the Mel-frequency cepstrum) appropriately captures the information contained in the sound signal.

Aspects of MFCC that are generally considered features are the first 13 coefficients (the lower dimensions) of MFCC, which represent the envelope of spectra. Furthermore, the spectral features are expressed by the higher dimensions that were omitted. Envelopes are sufficient to capture the differences between phonemes, allowing us to distinguish between them using the MFCC system.

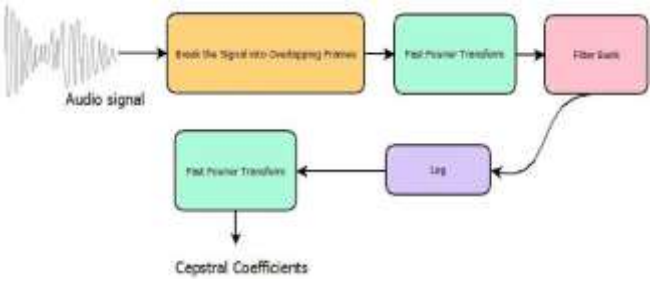

Fig 3.2.2: Flow of the Cepstral Coefficient

Signals

A signal is a change in a given quantity over a period of time. When it comes to audio, the quantity that changes is the air pressure. What method will we use to digitise this information? We can collect samples of the air pressure over a

given period of time. Even though the sampling rate might vary, it is most usually 44.1kHz (44,100 samples per second), which is the most common value. This is a waveform for the signal that can be understood, manipulated, and analysed using computer software, which we have done in this case

Fourier Transform

An audio signal is made up of a number of single-frequency sound waves that are combined together. When we take samples of a signal over time, we only record the amplitudes that result from the sampling process. In mathematics, the Fourier transform is a mathematical technique that allows us to break down a signal into its constituent frequencies and the amplitude of each frequency. In other words, it transfers a signal from the time domain to the frequency domain by converting its frequency component. The resulting structure is referred to as a spectrum.

Spectrogram

In signal processing, a spectrogram is a visual depiction of the frequency spectrum of a signal as it changes over time.

If our signal's frequency content changes over time, the fast Fourier transform can be a very useful tool for analyzing it. But what happens if the frequency content of our signal changes over time? This is true for the vast majority of auditory signals, including music and speech. Non-periodic signals are signals that do not repeat on a regular basis. A method for representing the spectrum of these signals as they change over time is required. The following may come to mind: Can't we compute various spectrums by running FFT on several windowed parts of the signal? Yes! A technique known as the short-time Fourier transform accomplishes exactly what is described above. In order to obtain the spectrogram, the FFT must be performed on overlapping windowed segments of the input signal.**3.2.7 The Mel Scale**

In order to detect emotions, we can make use of a subset of the features that falls under the category of Mel Frequency Cepstrum Coefficients (MFCC). When the term Mel is used, it refers to the scale that is utilised in the frequency measurement with respect to pitch measurement, and the formula for calculating the Mel scale is as follows.

$$m = 2595 \log_{10}(1 + (f/700))$$

The term Cepstrum refers to the Fourier Transform of the Log spectrum of the provided voice signal, which is defined in this context. Because humans are unable to hear frequencies on a linear scale, it is necessary to distinguish between lower frequencies and higher frequencies, according to some research. For example, we can easily distinguish between lower frequencies such as those between 100Hz and 150Hz, but it is difficult to distinguish between higher frequencies such as those between 10000Hz and 10050Hz, despite the fact that the real difference between the two is the same.

The Mel Scale is nothing more than a type of unit that defines an equal distance in pitch sounded as an equal distance to the listener as a unit of measurement.

MEL-SPECTOGRAM: When the frequencies are translated to the Mel scale, the result is a spectrogram called a Mel spectrogram.

Data Pre-Processing

The following steps are to be performed in this stage which are given below

1. Train, test, and divide the data.

2. Normalize Data — To improve model stability and performance.

3.One-hot encoding of target variable.

Specifically, it refers to the process of separating a column containing numerical categorical data into multiple columns based on the number of categories included in that column. Each column comprises the numbers "0" or "1" that correspond to the column in which it has been placed.**3.2.9 Classification**

4.Now MLPClassifier has an internal neural network for the purpose of classification. This is a feedforward ANN model.

- Initialize the Multi-Layer Perceptron Classifier

model=MLPClassifier (alpha=0.01, batch_size=256, hidden_layer_sizes= (300,), learning_rate='adaptive')

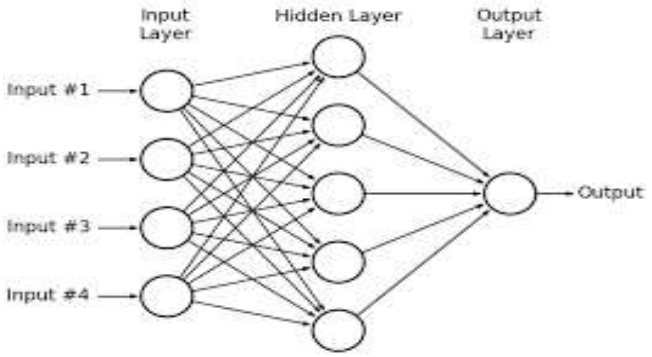

Fig 3.2.3: Multilayer Perceptron Classifier

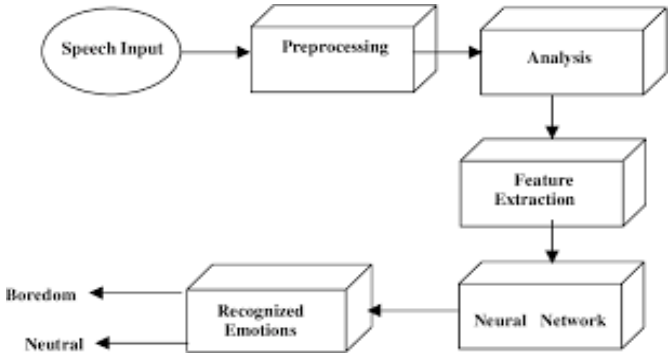

Fig 3.2.4: Flow of the system to calculate accuracy

This method, which is imported from sklearn and is used to calculate model correctness, is called accuracy score (). At the end of the process, accuracy is rounded up to two decimal places before printing it.

CONCLUSION

The research required for my project includes everything from data selection to data pre-processing, pitch detection to feature selection, Mel Frequency identification to classification. Throughout this journey, I have studied various research papers, which has enabled me to identify the research gap between the existing system and the proposed system. At the outset of this project, I set out to improve the accuracy of speech-based emotion detection by modifying the currently existing algorithms to meet each criterion. I also attempted to design some unique algorithms that would improve the accuracy further. This project is suitable for a variety of corporate categories, as well as for social media platforms where the emotions of the client are extremely important.

currently, I am only able to cover four emotions: happy, sad, and fearful, all of which are due to the limited time available to me. In addition, I am able to incorporate additional emotions into my project and to construct a neural network that is capable of learning from a variety of datasets and that will be able to handle external speech signals in the future.

REFERENCES

[1]. Shaheen, "Impact of Automatic Feature Extraction in Deep Learning Architecture," Gold Coast, QLD, Australia, 2016.

[2]. Moritz, "An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* , vol. 23, no. 11, pp. 1926-1937, 2015.

[3]. Evgeniou, "Support Vector Machines: Theory and Applications," 2001.

[4]. Shoshan, "Speech and Music Classification and Separation: A Review," *Journal of King Saud University - Engineering Sciences*, vol. 19, no. 1, pp. 95-132, 2006.

[5]. Chavan, "Speech recognition in noisy environment, issues and challenges: A review," 2015.

[6]. Haridas, "A critical review and analysis on techniques of speech recognition: The road ahead," *International Journal of Knowledge-Based and Intelligent Engineering Systems* , vol. 22, no. 1, pp. 39-57, 2018.

[7]. Çelik, "A Research on Machine Learning Methods and Its Applications," *Journal of Educational Technology and Online Learning*, pp. 1-8, 2018.

[8]. S. Arya et al. , Journal of the Electrochemical Society, 10, 023002(2021). DOI: 10.1149/2162-8777/abe095

[9]. S. Arya et al. Journal of The Electrochemical Society,168,

027505(2021). DOI: 10.1149/1945-7111/abdee8

[10]. HB Patel, SS Iyer, Comparative Study of Multimedia Question Answering System Models, 2022, ECS Transactions 107 (1), 2033

[11]. Hardik Patel, Anuradha Bharti, Classification of Brain Signals of User in Gaming Using Artificial Intelligence, Studies in Indian Place Names (SIPN) [ISSN 2394-3114] vol-40 (9), 70-73.

[12]. Hardik Patel, Indrjeet Rajput, Prevention Of Information Leakage From Indexing In Cloud, International Journal For Technological Research In Engineering vol-01, issue-10, page no. 1158-1161.

[13]. Hardik Patel, Indrjeet Rajput, Indexing Based Various Search techniques in Cloud Computing, MonTeC-2014, 27 feb,2014, (ISBN: 978-81-929173-0-6) page no-189-192.

[14]. Chowdhury, "Classification using Supervised Machine Learning Techniques," Orem, UT, USA, 2020.

[15]. C. H. Yu, "Exploratory data analysis in the context of data mining and resampling," *International Journal of Psychological Research* , vol. 3, no. 1, pp. 1-11, 2010.

[16]. Kolokolov, "Signal Preprocessing for Speech Recognition," *Automation and Remote Control,* vol. 63, no. 3, pp. 494-501, 2002.

[17]. Browne, "Cross-Validation Methods," *Journal of Mathematical Psychology,* vol. 44, no. 1, pp. 108-132, 2000.

[18]. Alloghani, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, Springer, 2020, pp. 3-21.

[19]. Diaz, "An effective algorithm for hyperparameter optimization of neural networks," *Ibm Journal of Research and Development,* vol. 61, no. 4, pp. 1-13, 2017.

[20]. Lessmann, "Optimizing Hyperparameters of Support Vector Machines by Genetic Algorithms.," in *International Conference on Artificial Intelligence, ICAI*, Las Vegas, Nevada, USA, 2005.

[21]. Vashisht, "Speech Recognition using Machine Learning," *IEIE Transactions on Smart Processing and Computing,* vol. 10, no. 3, pp. 233-239, 2021.

[22]. Tandel, "Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey," in *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020.

[23]. Murarka, "Sentiment Analysis of Speech," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 6, no. 11, pp. 1-5, 2017.

[24]. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Philos Trans R Soc Lond B Biol Sci.,* vol. 364, pp. 3449-3451, 2009.

[25]. S.S. Iyer, K.I. Kamaljit, Practical evaluation and comparative study of text steganography algorithms. Int. J. Innov. Res. Comput. Commun. Eng. 5(3), 74–77 (2016). ISSN (Online) 2278-1021 ISSN (Print) 2319-5940

[26]. S.S. Iyer, K.I. Kamaljit, Practical evaluation and comparative study of big data analytical tools, in Int. J. Innov. Res. Comput. Commun. Eng. 5(2), 57–64 (2017). ISSN (Online): 2320-9801 ISSN (Print): 2320-9798.