SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

Prof. Hima Keerthi Sagiraju PhD Scholar, Department of Computer Science and System Engineering Andhra University Visakhapatnam, Andhra Pradesh :: keerthi.sagi2010@gmail.com

Pritam Sharma Andhra University College of Engineering For Women Visakhapatnam, Andhra Pradesh :: pritam99sharma@gmail.com

Priyamvada Akula Andhra University College of Engineering For Women Visakhapatnam, Andhra Pradesh :: priyamvadaakula1999@gmail.com

Sahithi Sahukari Andhra University College of Engineering For Women Visakhapatnam, Andhra Pradesh :: sahithisahukari@gmail.com

Chithajhallu Sai Siva Deekshita Andhra University College of Engineering For Women Visakhapatnam, Andhra Pradesh :: deekshitach31@gmail.com

Sri Sai Keerthana Pantula Andhra University College of Engineering For Women Visakhapatnam, Andhra Pradesh :: keerthana99pantula@gmail.com

ABSTRACT

Language has a very broad scope. Besides being used as a communication medium, language can also be interpreted to convey the expression of emotions and feelings of humans to humans, humans to animals, or humans with their surroundings. Language processing has 2 approaches, the first is semantic and the second is emotional. Emotions have an important role in everyday human interactions. *We can convey emotions using two ways either through facial expressions or through speech. Humans most natural way of expressing ourselves is speech. It plays a vital role in communication and expressing our emotions. So research's developed a Speech emotion recognition (SER) system which consist of collection of methodologies that process and classify speech signals to detect the embedded emotions. In this project we show how easy it the recognition of human emotions with the incorporation of technological capabilities using machine learning. The core 6 emotions identified were happiness, sadness, disgust, fear, surprise, and anger.*

Keywords: *Machine Learning, Human Emotion, feature extraction, Audio Extraction*

1. INTRODUCTION

Recognizing the emotion based on any factor like speech, facial expressions, text is an important challenging component in Human Computer Interaction. Language is a medium for communication to happen. Although there are some technologies which makes the machine understand the information based on content but assessing the emotion behind the content is difficult. Speech emotion recognition is used to detect the emotion using speech based on the pitch, tone. Feelings can be expressed through emotions.Emotions are one of the most difficult concepts to express. So, why is an emotion necessary and important to be defined? Emotions play a vital role in shaping an interaction between humans and humans or the surrounding environment .Amongst the numerous models used for categorization of these emotions, a discrete emotional approach is considered as one of the fundamental approaches. It can detect various emotions such as anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness. Emotions are not only understood by facial expressions but also with speech. Every speech of human being is associated with an emotion

In the past few decade Speech emotion recognition has been the source of research and has become one the fastest growing engineering technology. When SER was first developed, its main objective was for the needs of human psychology research. Another goal is to develop AI as a breakthrough from previous Speech Recognition (SE without involving emotions). Nowadays, in addition to psychological research needs, SER can also be applied in the areas of entertainment, security, banking, call centres, etc.In recent years, a large number of studies have focused on emotion detection using opinion mining on speech. Due to some intrinsic characteristics of the voice produced during calls, such as the loudness, voice quality and casual expression, emotion recognition on them is a challenging task. In this project u can learn how to recognise various emotions of a person that they are happy or sad or tensed using speech

Speech emotion processing and recognition system composes three parts, which were speech signal acquisition, feature extraction, and emotion recognition by using different classifiers in machine learning. There are three main types of features which are classified as Elicited features, Prosodic features and Spectral features. For spectral features different technologies are used like MFCC, LPCC and MEDC. For prosodic features technologies used are pitch, intensity, fundamental frequency, loudness, glottal parameters etc. Classification techniques for classifying emotions are used such as Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) etc.

This paper contains sections like complete description of concepts used in speech emotion recognition system and then follows the audio extraction techniques, classification techniques and different applications. Last section comprise conclusion.

2. CORE CONCEPTS

Speech emotion processing and recognition system was generally composed of three parts, which were speech signal acquisition, feature extraction, and emotion recognition by using different classifiers

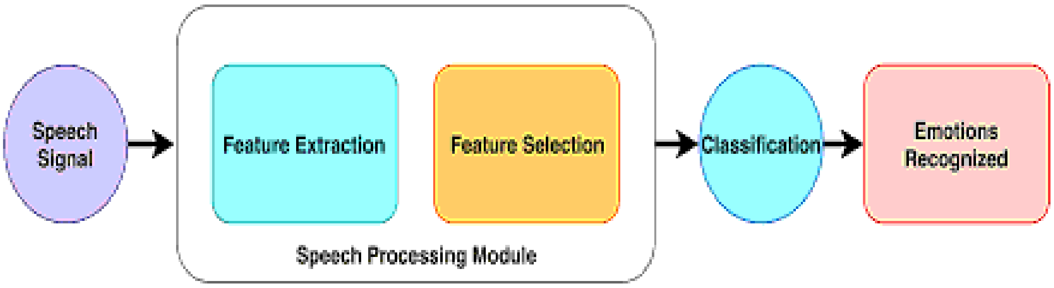


Figure-1: Speech Recognition Model

2.1 Emotion Models

2.1.1 Discrete Emotion Model

Discrete emotion theory is the claim that there is a small number of core emotions researchers concluded that there are nine basic The nine emotions are Shringara (love/beauty), Hasya (laughter), Karuna (sorrow), Raudra (anger), Veera (heroism/courage), Bhayanaka (terror/fear), Bibhatsya (disgust), Adbutha (surprise/wonder) and Shantha (peace or tranquillity).It explains that there are particular characteristics attached to each of these emotions, allowing them to be expressed in varying degrees.

2.1.2 Dimensional Emotion Model

For theoretical and practical reasons, researchers define emotions based on one or more dimensions. In "The Passions of the Soul", Descartes defines and investigates six main passions (wonder, love, hate, desire, joy, sadness, and sadness). Wilhelm Max Wundt, the father of modern psychology, proposed in 1897 that emotions could be described by three dimensions: "pleasant versus unpleasant", "arousing or subjugating" and "tension or relaxation".

2.2 Audio

Audio is the sound produced by the vibration's object. The sound produced by humans comes from the mechanism of action between organs such as the mouth, lungs, throat, vocal cords, and nose. The lungs press the air through the breath valve, then causes the vocal cords to vibrate, producing waves with pressure (impulses) through the air flow in the form of periodic quasi waves or what we call sound. Audio itself has 4 main attributes, namely:

2.3 Time/Duration

Duration is the length of waves in units of time. In the dataset we use, the average length of duration for each sound data is 3 seconds (3s).

2.4 Amplitude

Amplitude is the fluctuation in the change of sound waves. The shorter and the more frequent the waves, the higher the pitch or frequency.

2.5 Sampling rate

Sampling rate is the number of audio samples per second, measured in Hz / KHz. Likes when we work with images, we use pixels as a measure of image qualities. The higher of pixel, means better the image produced. The highest sampling rate for an audio is 44100 KHz. However, the most frequent and commonly used sampling rate is 22050 or 22 KHz which is the upper limit of the sound that can be heard by humans.

2.6 Frequency

Frequency is another approach when we want to see the sound waves, but we can't see them directly based on the time domain. The frequency domain may not visually intuitive, but it will be useful when we go in to sound extraction section. In addition, technically the frequency domain requires less computing space for storage.

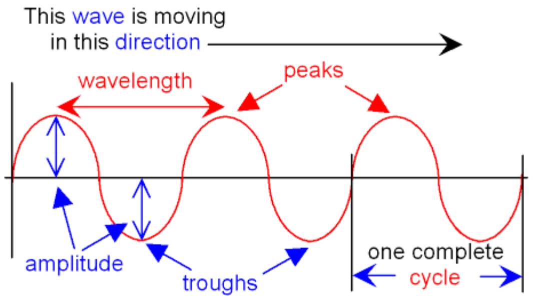


Figure-2: Frequency

3. PROPOSED MODEL

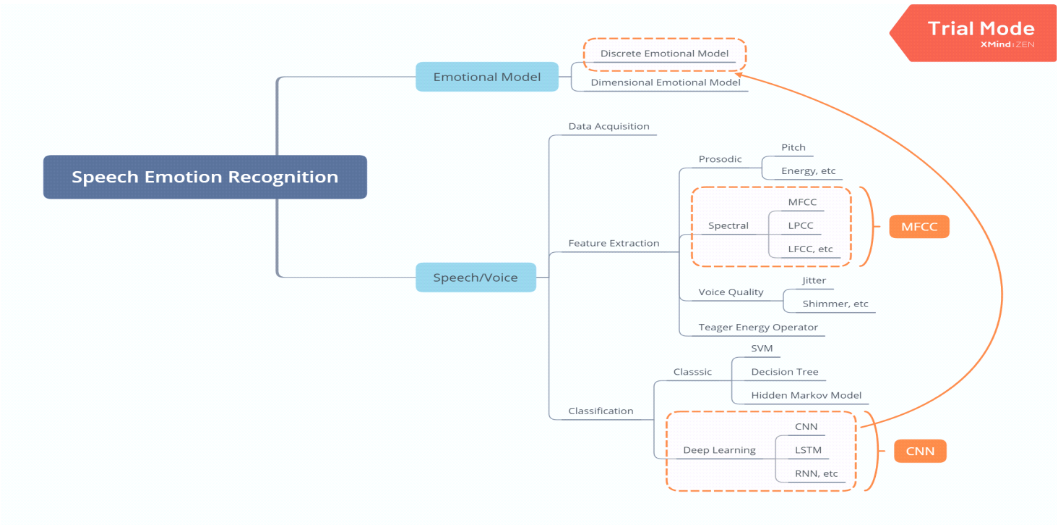


Figure-2: Flowchart of our proposed model

Human speech consists of many parameters which show theemotions comprise in it. As there is change in emotions this proper feature vector to identify the emotions. Features are categorized as excitation source features, spectral features and prosodic features. Excitation source features are achieved by suppressing characteristics of vocal tract (VT).Spectral features used for emotion recognition are Linear prediction coefficients (LPC), Perceptual linear prediction coefficients (PLPCs), Mel-frequency spectrum coefficients (MFCC), Linear prediction cestrum coefficients (LPCC),perceptual linear prediction (PLP).Prosodic features used for emotion recognition are pitch, energy, intensity. Statistical measurements are also used to distinguish emotions like minimum, maximum, standard deviation, range, mean, median, variance, skewness, kurtosis etc. of features.

Three features are extracted from the discourse signals given as information. The three features are, MFCC, Mel Spectrograph Frequency and Chrome.

3.1 Audio Extraction

- **MFCC**:

Mel Frequency Cepstral Coefficients (MFCC) is utilized to recover the sound from the given wav audio file by utilising distinct hop length and HTK-styles mel frequencies. Pitch of 1 kHz tone and 40 dB over the perceptual discernible edge is characterised as 1000 mels, utilized as reference point. The MFCC gives a Discrete Cosine Change in short DCT of a genuine logarithm of the transient vitality showed on the Mel recurrence scale. MFCC (Mel Frequency Cepstral Coefficients) was first recognized by Davis and Murmelstein in 1980. MFCC is one of the most famous methods for extracting features because its quite good ability to extract sound features. This method adapts the workings of human hearing. To get the MFCC coefficient we will divide the voice signal into several parts with the framing process. Then we will convert each part from the time domain to the frequency domain using Fourier Transform. From the results of the transform, we can calculate the energy at each frequency band using mel filter bank. This process will produce a mel spectrum. Then mel spectrum will we inverse again to get the MFCC coefficient value in the time domain. Below is a chart of the sound feature extraction process using MFCC(figure-2).will try to make it more concise by describing why each process needs to be done mel(f)=2595 x log 10 (1+f/700)

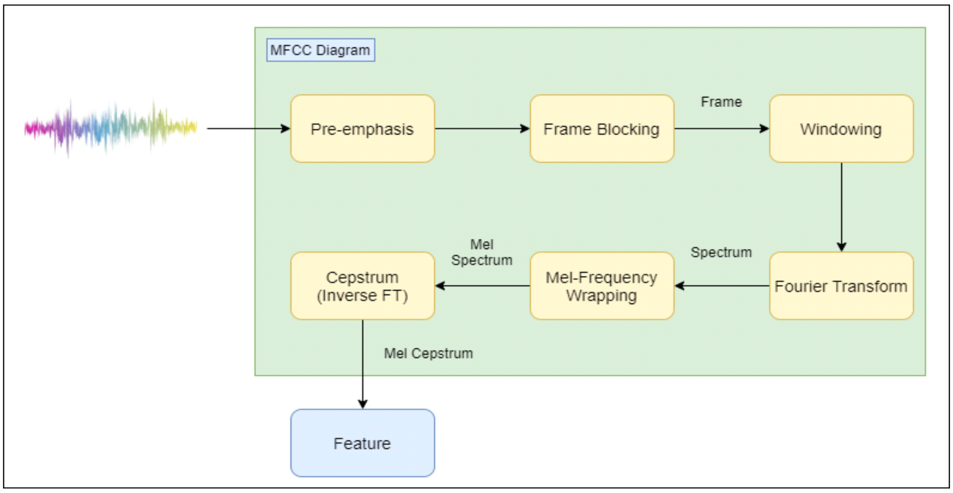
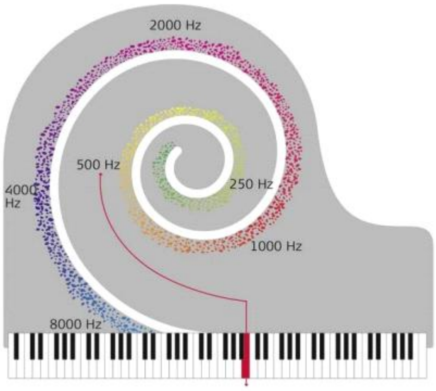


Figure 3: Sound feature extraction process using MFCC

• Mel:

Individuals are incredibly improved at perceiving little or minute changes in pitch at low frequencies than they are at high frequencies. Mel Frequency Warping is generally done by using Filter bank. Filter bank is of filter that has the goal to find out the energy size of certain frequency bands in the sound signal. For purposes of MFCC, filters must be applied in the domain frequency. Once again, this method is adapted from the way the cochlea works. In the cochlea, there are separate parts that regulate the production of sounds based on certain frequencies.



parameters also gets changed. Hence it’s necessary to select

Figure-4:MEL SPECTRUM

Human perception of frequency in the signal sound does not follow a linear scale. The frequency with which actually (in Hz) in a signal will be measured humans subjectively with use

$$mel(f)=125\ln(1+f/100),$$

Where Mel scale, To calculate Mel scale we use, where mel(f) is function of mel scale and f is frequency. the most popular mel filter used is the triangular mel filter.

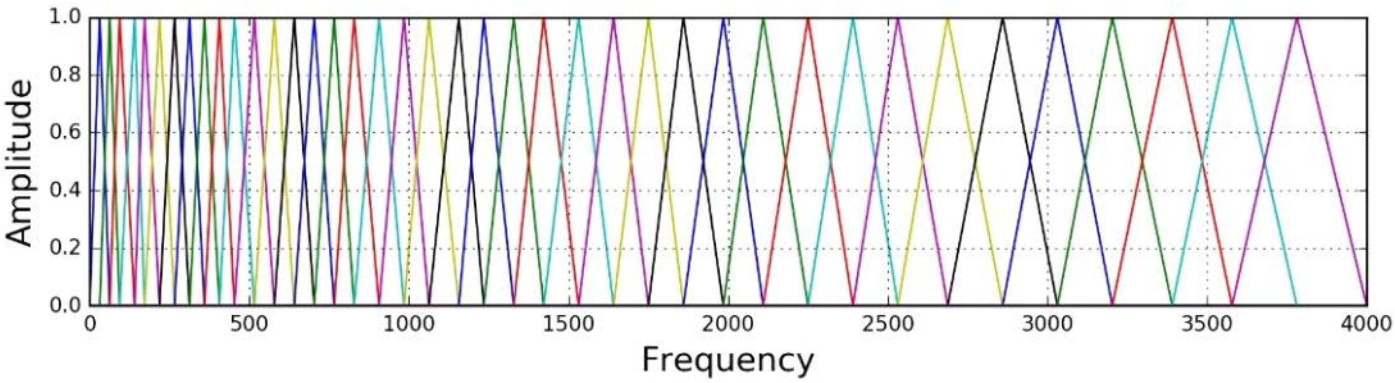


Figure-5: Mel Filter

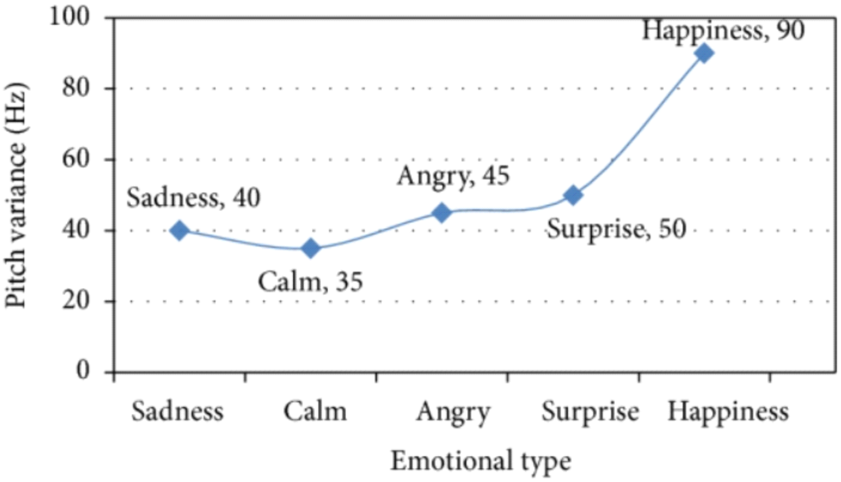
• **Chroma**:

A better nature of the extricated chroma include empowers which gives much better outcomes in these elevated level assignments. The chroma include the log-repeat size range across octaves.

$$Cf(b)=\sum Z-1z=0|Xlf(b+z\beta)||$$

3.2 Fundamental Frequency Construction

Banziger and Scherer proposed that, for the same sentence, if the emotions expressed were different, fundamental frequency curves were also different; besides the mean and variance of fundamental frequency were also different. If the speaker is in a state of sadness, than the fundamental frequency curve of the speech is bent downward generally. The below graph shows the curves of different emotional variance.



3.3Emotion Distribution of Gender

Regarding the distribution of gender, the number of female speakers was found to be slightly more than the male speakers, but the imbalance was not large enough to warrant any special attention.

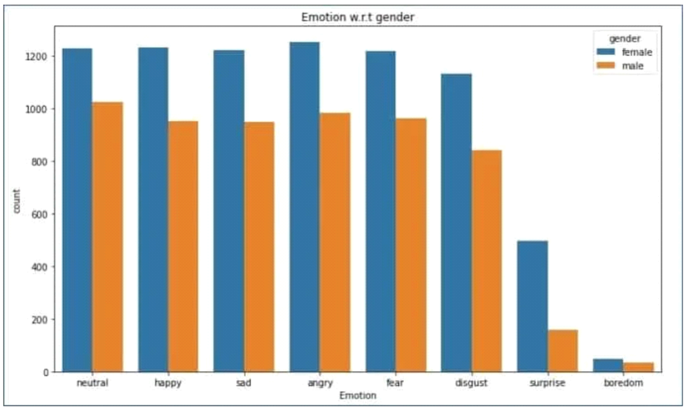
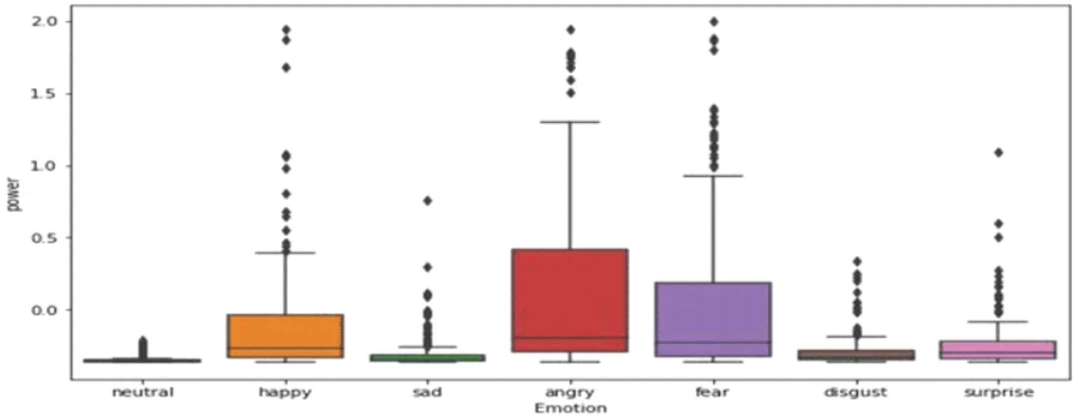


Figure: Variation in Energy Across Emotions

To ensure the uniformity in our study of energy across emotions variation as the audio clips in our dataset were of different lengths,and a power which is energy per unit time was found to be a more accurate measure. With respect to different emotions the metric was plotted. From the graph it is quite evident that the primary method of expression of anger or fear in people is a higher energy delivery. We also observe that both the disgust and sadness are closer to neutral with regards to energy although exceptions do exist.



4 CLASSIFIERS

Subsequent work with multilayer perceptron's has shown that they are capable of approximating an XOR operator as well as many other non-linear functions. They train on a set of both input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. Important issues in MLP design include number of units and specification of the number of hidden layers. 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.

Multi-layer Perceptron Classifier in short MLP Classifier relies on an underlying Neural Network to perform the classification. MLP Classifier implements a Multi-Layer Perceptron (MLP) algorithm and trains the Neural Network using Backpropagation. Building the MLP classifier involves the following steps:

- By defining and initiating the required parameters initialize the MLP Classifier.
- The data is given to the Neural Network to train it.
- Trained network is used to predict the output values.
- After predicting the output values now calculate the accuracy of the predictions.

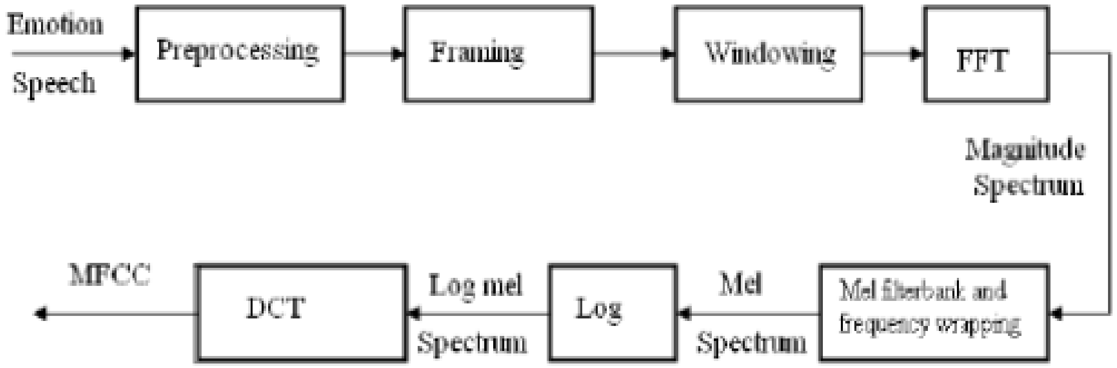


Figure -6 Amplitude Construction

Speech signal amplitude construction and speech emotional state also have a direct link.The volume of speech is generally low when the speaker is low/sad or depressed. Therefore, the analysis of speech emotion features of the amplitude construction is determined to be more meaningful.

The below graph is the comparison of emotional speech and calm speech, which is shown with the average amplitude difference. The graph shows that the amplitude of joy, anger, and surprise, three kinds of emotional speech, compared with calm voice signal is larger, while the sad speech amplitude is smaller.

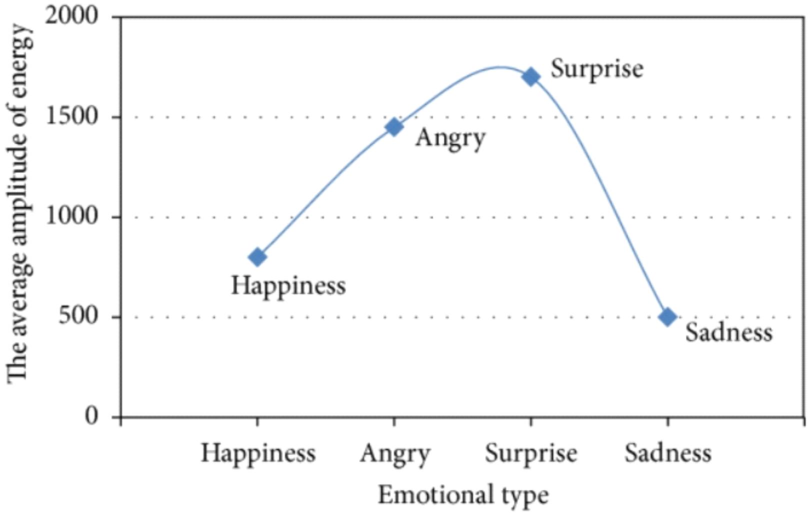


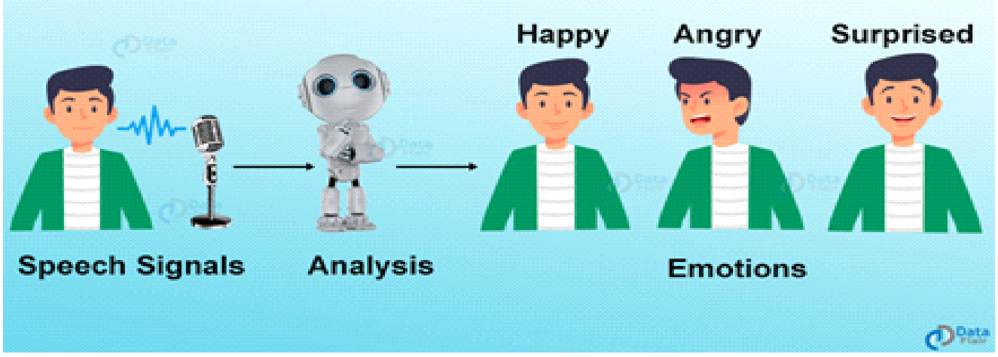
Figure-7: The distribution of emotional speech amplitude energy

5 APPLICATIONS:

There are maximum applications of emotion recognition system. Few of them are Security,Medicine,Entertainmnet,Education, Psychiatric diagnosis,Prosody in dialog system, Mobile based speech recognition system, Emotion recognition in call centre where emotions of customer can help to get better quality of service, Sorting of voice mail, Lie detection,Computer games.

Why is emotion detection important?

Human emotion recognition plays an key role in the world. Emotions are reflected from gestures of the body and through facial expressions. Hence understanding of emotion has a high importance of the interaction between human and machine communication.

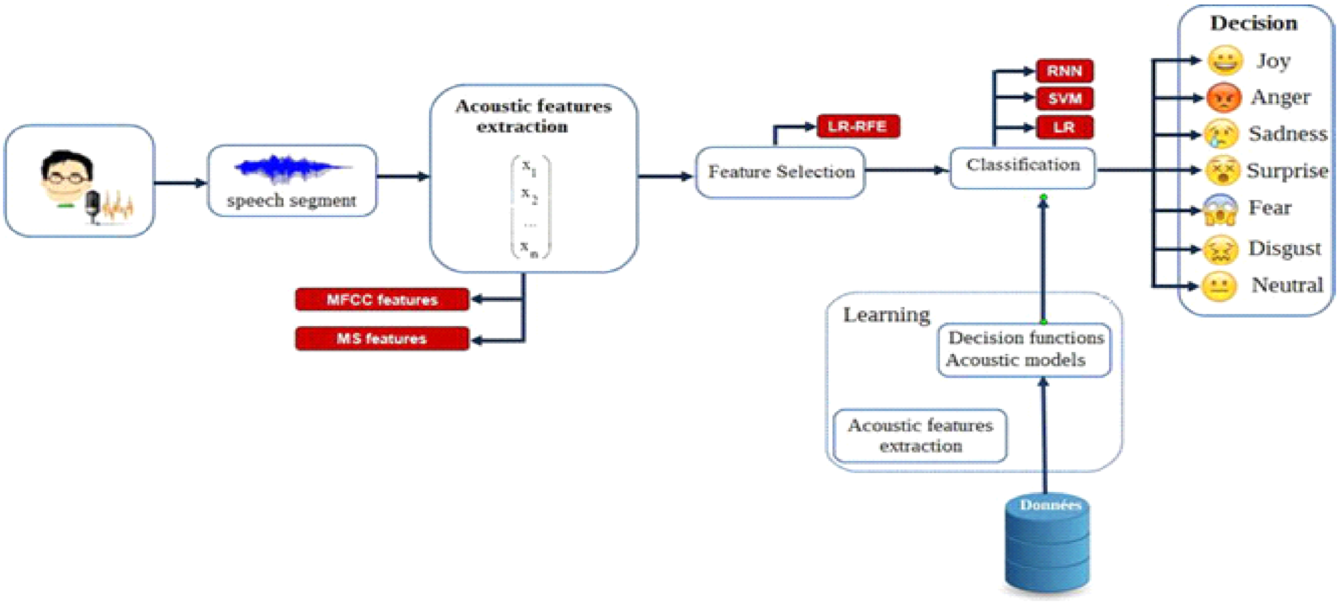


Algorithms is used for Emotion Recognition :

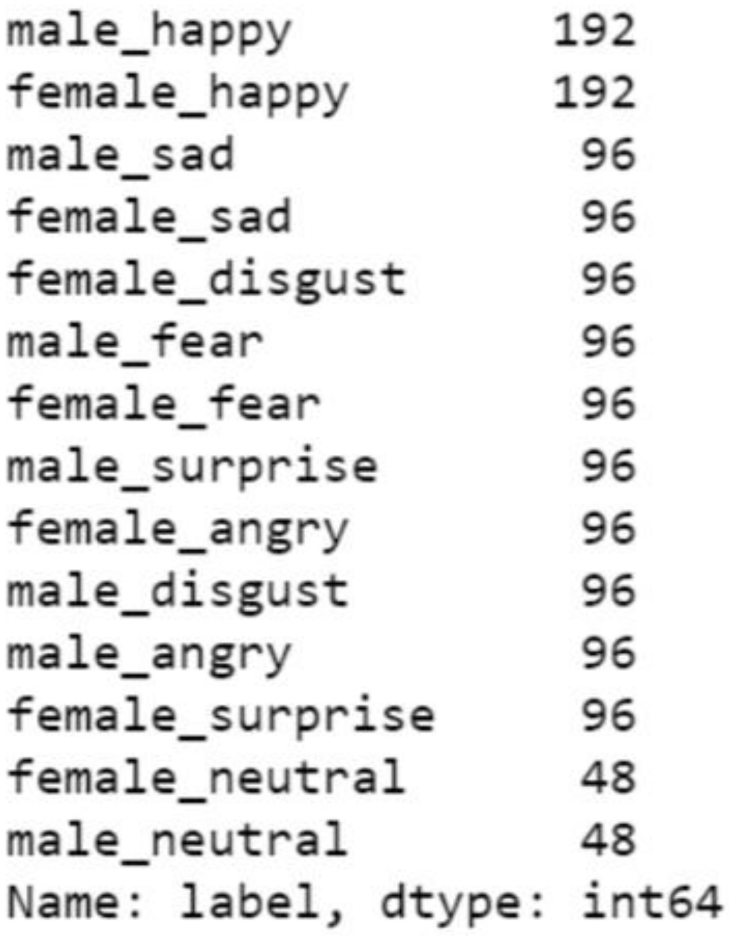
Three popular ML algorithms, SVM, RF were used for emotion intensity recognition. A comparative study and implementation of algorithms will measure face emotions and their intensities based on the different action units are presented

Can technology detect emotions:

As you answer the questions, an artificial intelligence (AI) system scans your face, scoring you for nervousness, empathy and dependability Emotion Recognition Technology (ERT) is in fact a burgeoning multi-billion-dollar industry that aims to use AI to detect emotions from facial expression.



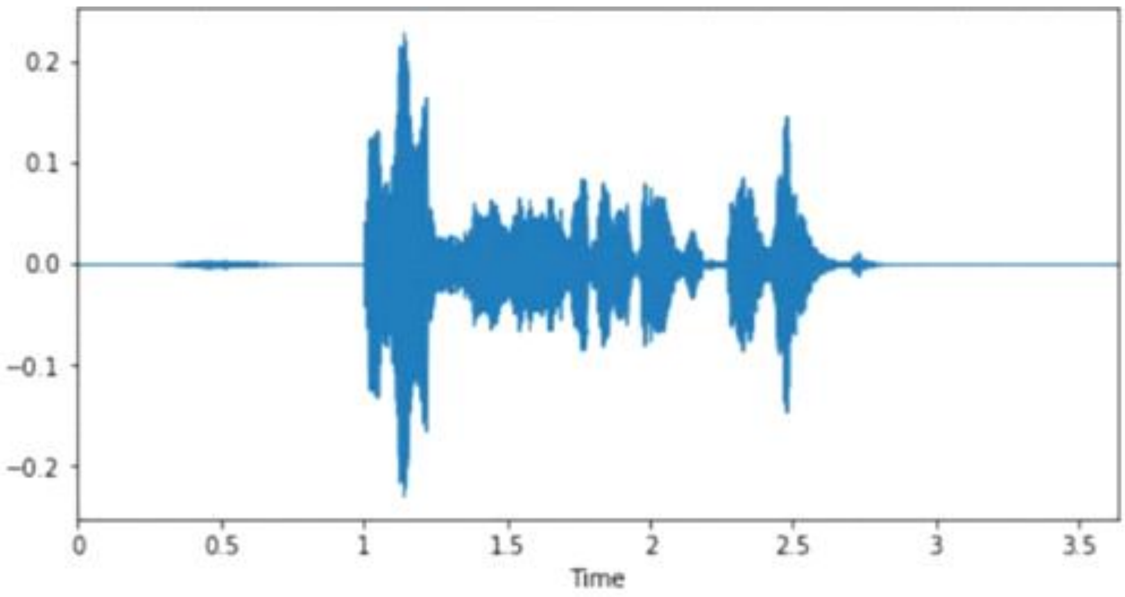
6 EXPERIMENTAL ANALYSIS:



```
male_happy          192
female_happy        192
male_sad             96
female_sad           96
female_disgust       96
male_fear            96
female_fear          96
male_surprise        96
female_angry         96
male_disgust         96
male_angry           96
female_surprise      96
female_neutral       48
male_neutral         48
Name: label, dtype: int64
```

DESCRIPTION :

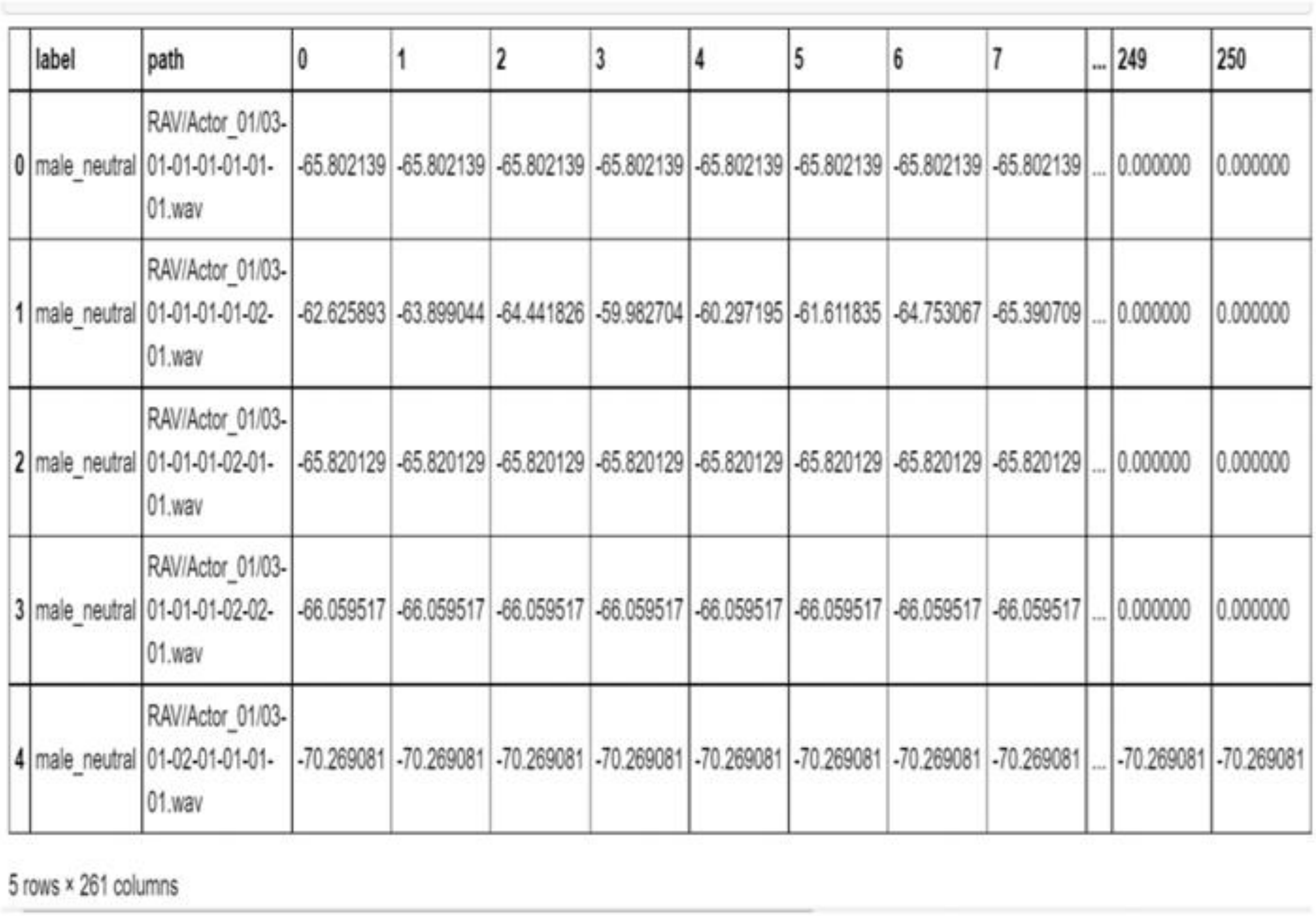
In the above diagram listed Emotional categories are Happy,Sad,Disgust,Surprise,Fear,Angry,Neutral... Gender categories are Male,Female.Here Numbers are storing in data type Int and names as Label



DESCRIPTION:

This is the wave plot of an audio file in the RAVDESS dataset.
And here according to recording of audio time will be calculated in Minutes and Seconds...
Also recognize the quality of audio in the terms of numbers like decimals etc...

6.1 Feature extraction:

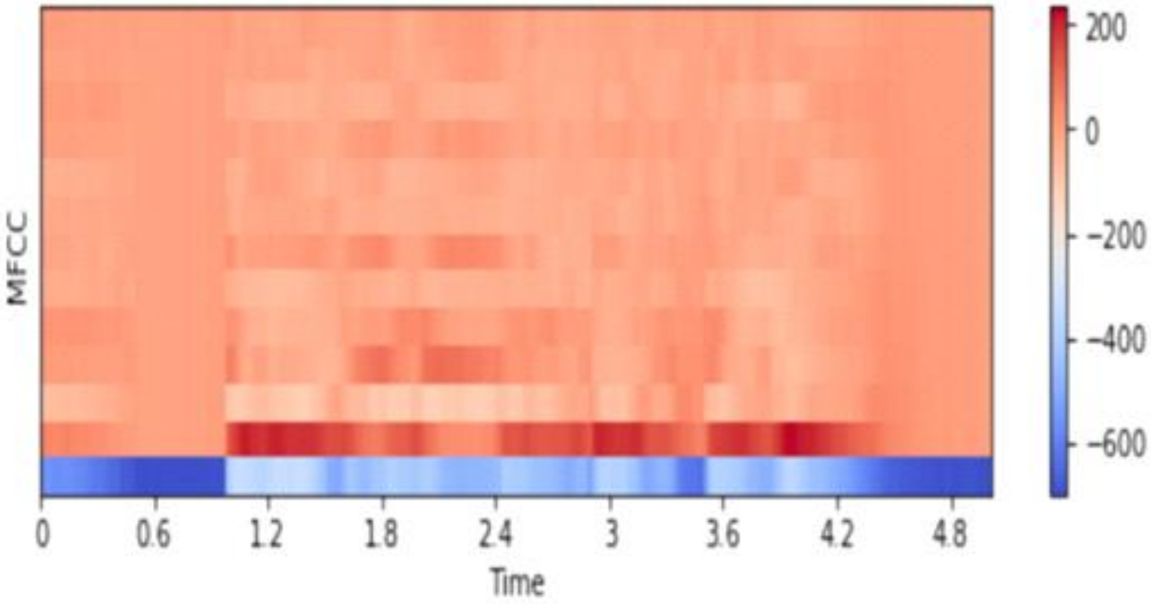


	label	path	0	1	2	3	4	5	6	7	...	249	250
0	male_neutral	RAV/Actor_01/03-01-01-01-01-01-01.wav	-65.802139	-65.802139	-65.802139	-65.802139	-65.802139	-65.802139	-65.802139	-65.802139	...	0.000000	0.000000
1	male_neutral	RAV/Actor_01/03-01-01-01-01-02-01.wav	-62.625893	-63.899044	-64.441826	-59.982704	-60.297195	-61.611835	-64.753067	-65.390709	...	0.000000	0.000000
2	male_neutral	RAV/Actor_01/03-01-01-01-02-01-01.wav	-65.820129	-65.820129	-65.820129	-65.820129	-65.820129	-65.820129	-65.820129	-65.820129	...	0.000000	0.000000
3	male_neutral	RAV/Actor_01/03-01-01-01-02-02-01.wav	-66.059517	-66.059517	-66.059517	-66.059517	-66.059517	-66.059517	-66.059517	-66.059517	...	0.000000	0.000000
4	male_neutral	RAV/Actor_01/03-01-02-01-01-01-01.wav	-70.269081	-70.269081	-70.269081	-70.269081	-70.269081	-70.269081	-70.269081	-70.269081	...	-70.269081	-70.269081

5 rows × 261 columns

DESCRIPTION :

Feature extraction is one of the determining parts in the success of the classification model later. At this stage, we use MFCC to extract voice features in our data. The MFCC coefficient to be used is 13.
In Label it cagerorized only Male category and only Neutral Emotion.
It also displayed path and level of voice recognisation level.

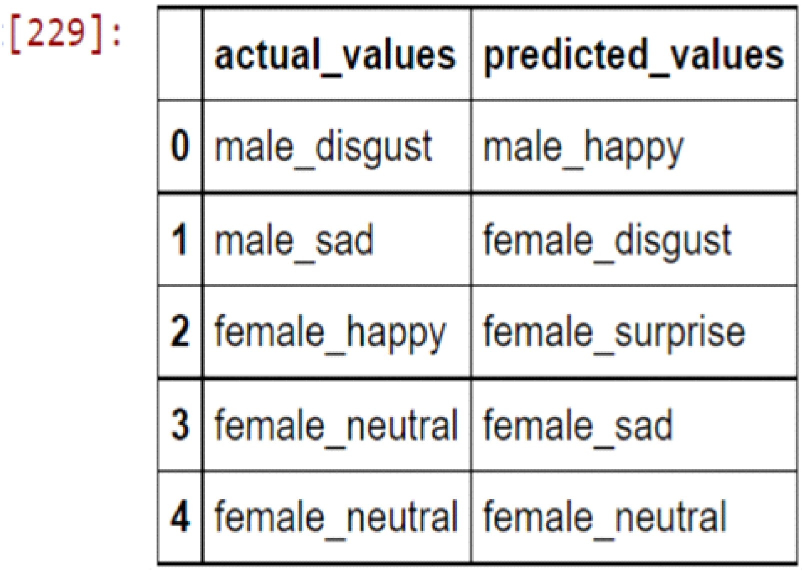


DESCRIPTION :

If a cepstral coefficient has a positive value, the majority of the spectral energy is concentrated in the low-frequency regions.

On the other hand, if a cepstral coefficient has a negative value, it represents that most of the spectral energy is concentrated at high frequencies.It is calculated in the categories of MFCC and Time in the terms of integers

6.2 Building model :



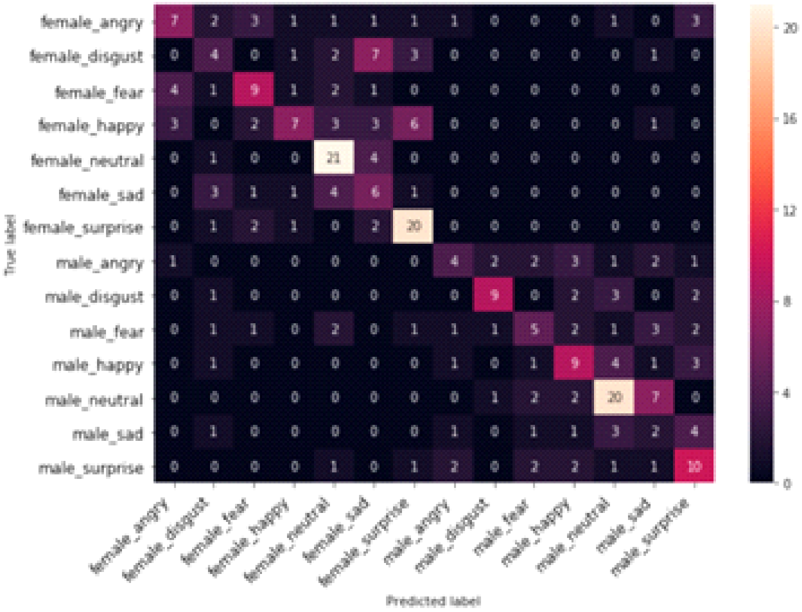
[229]:

	actual_values	predicted_values
0	male_disgust	male_happy
1	male_sad	female_disgust
2	female_happy	female_surprise
3	female_neutral	female_sad
4	female_neutral	female_neutral

DESCRIPTION:

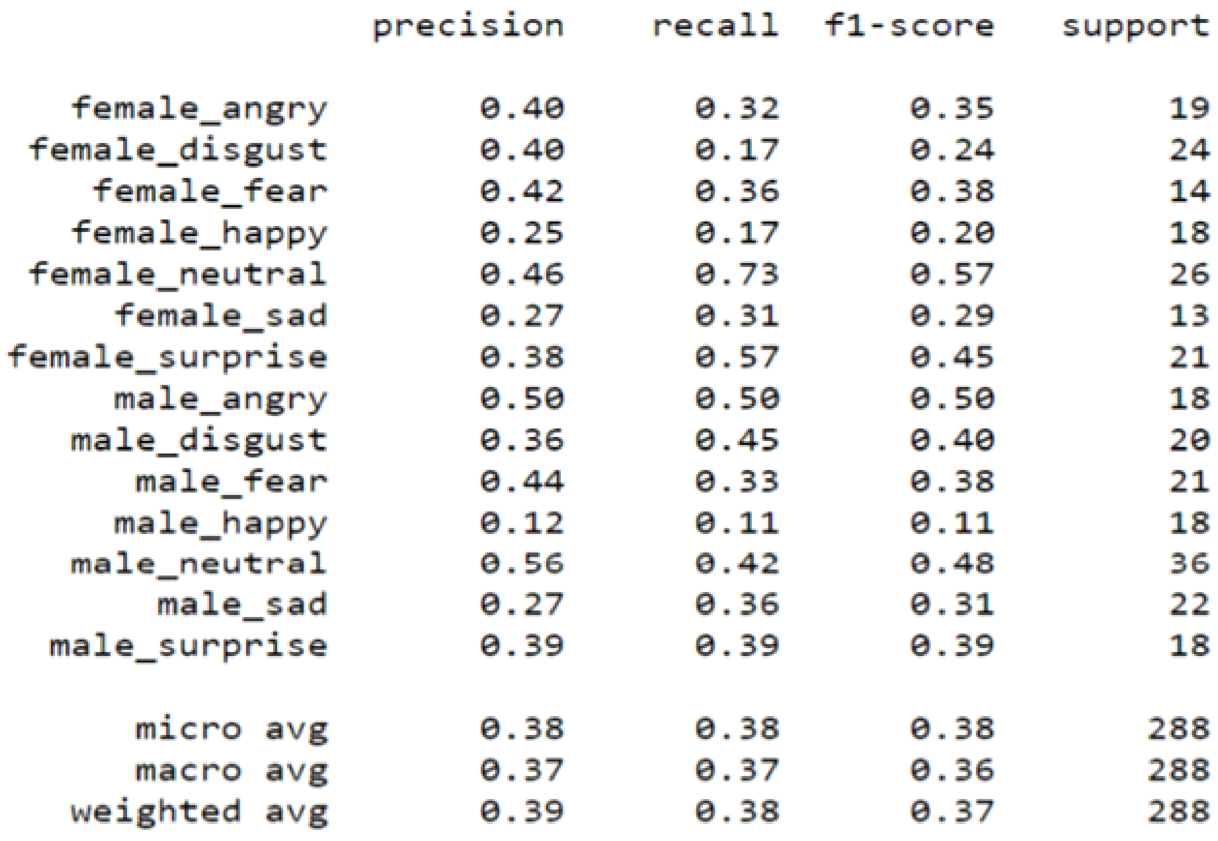
Comparison of actual values and the predicted values.

Here both Male and Female are indicated and Emotions are Happy,Sad,Neutral,Disgust,Surprise.



	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	7	2	3	1	1	1	1	1	0	0	0	1	0	3
female_disgust	0	4	0	1	2	7	3	0	0	0	0	0	1	0
female_fear	4	1	9	1	2	1	0	0	0	0	0	0	0	0
female_happy	3	0	2	7	3	3	6	0	0	0	0	0	1	0
female_neutral	0	1	0	0	21	4	0	0	0	0	0	0	0	0
female_sad	0	3	1	1	4	6	1	0	0	0	0	0	0	0
female_surprise	0	1	2	1	0	2	20	0	0	0	0	0	0	0
male_angry	1	0	0	0	0	0	0	4	2	2	3	1	2	1
male_disgust	0	1	0	0	0	0	0	0	9	0	2	3	0	2
male_fear	0	1	1	0	2	0	1	1	1	5	2	1	3	2
male_happy	0	1	0	0	0	0	0	1	0	1	9	4	1	3
male_neutral	0	0	0	0	0	0	0	0	1	2	2	20	7	0
male_sad	0	1	0	0	0	0	0	1	0	1	1	3	2	4
male_surprise	0	0	0	0	1	0	1	2	0	2	2	1	1	10

Here in the above diagram indicated Emotions in the form of Matrix between Actual Values and Prediction Values
It is a Confused and Mixed Matrix



```
                 precision    recall  f1-score   support

   female_angry       0.40      0.32      0.35        19
 female_disgust       0.40      0.17      0.24        24
    female_fear       0.42      0.36      0.38        14
   female_happy       0.25      0.17      0.20        18
 female_neutral       0.46      0.73      0.57        26
     female_sad       0.27      0.31      0.29        13
female_surprise       0.38      0.57      0.45        21
     male_angry       0.50      0.50      0.50        18
   male_disgust       0.36      0.45      0.40        20
      male_fear       0.44      0.33      0.38        21
     male_happy       0.12      0.11      0.11        18
   male_neutral       0.56      0.42      0.48        36
       male_sad       0.27      0.36      0.31        22
  male_surprise       0.39      0.39      0.39        18

      micro avg       0.38      0.38      0.38       288
      macro avg       0.37      0.37      0.36       288
   weighted avg       0.39      0.38      0.37       288
```

Description:

Here is the Complete result of Emotion Speech Recognisation
In the Categories of Precision,Recall,F1-Score,Support...

All Categories expressed the level od Emotion in the terms of Decimal Numbers

7. CONCLUSION AND FUTURE SCOPE :

Through this project, we have displayed how we can leverage Machine learning to obtain the under emotion from speech audio data on the human expression of emotion through voice.

Application-Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

A few possible steps that can be implemented to make the models more robust and accurate are the following

1) An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.

2) Figuring out a way to clear random silence from the audio clip.

Exploring other features of audio data to check their applicability in the domain of emotional speech recognition.

3)These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.

4)Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of system due to in some cases the expression of emotion is vocal.

5)Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

8. REFERENCE :

[1]vol. 26, no. 1, pp. 72–75, 2005.

[2] A. Krizhevs[1] C. Peng, “Research emotional and recognition in speech signal,” Journal of Jing jang University, ky, I. Sutskever, and G. E. Hinton, “Image net classification with deep neural networks,” in Proceedings of the 26th Annual Conference on Neutral Information Processing Systems (NIPS '12), pp. 1097–1105, Lake New, UK, December 2012.

[3]L. Zhao, C. Jiang, C. Zou, and Z. Wu, “Study on emotional recognisation analysis in speech,” Acta Electronica volve. 32, no. 4, pp. 606–609, 2004.

[4]G. Penguan “Research on emotional recognition,” Application , vol. 24, no. 10, pp. 101–103, 2007.

[5]L. Zhao, X. Qian, C. Zhou, and Z. Wu, “Study on emotional feature discover from speech ,” Journal of Data and Processing, vol. 15, no. 1, pp. 120–123, 2000.

[6]P. Guo, Research of the Method of Speech Emotion Extraction and the Emotion feature, Northwestern University, 2007.

[7]Y. Kim, H. Lee, and E. M. Provost, "Machine learning for robust feature generation in audio emotion recognition,” in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.

[8]G. E. Hinton, and Y. Techical, “ machine learning algorithm for Machine learning beliefs ,” Computation, vol. 18, no. 7, pp., 2006.

[9] A.P. Wanare, S.N. Dandare, "Human Emotion From Speech", Int. Journal of Research and Applications, vol. 4, no. 7, pp. 74-78, July 2014