JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

[1] Dr.Uttara Gogate , [2] Bhavika Ekilwale , [3] Siddhi Mahadik , [4] Manshi Jagdale
[1] Associate Professor , Dept. of Computer Engineering
[2,3,4] B.E. Students, Dept. of Computer Engineering
Shivajirao S. Jondhale College of Engineering, University of Mumbai

Abstract : *This project presents a comparative study of Speech Emotion Recognition (SER) systems. The speech signal is one of the most natural and fastest methods of communication between humans. This project has reviewed and compared the different classifiers that are used to discriminate emotions such as Happy , Sad , Neutral , Fearful , Surprise , Disgust ,Angry etc. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed. Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. After feature extraction, another important part is the classification of speech emotions. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task. A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to logistics regression (LR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals. There are number of datasets available for speech emotions, it's modelling and types that helps in knowing the type of speech. RAVDESS , Berlin and Spanish data- bases are used as the experimental data set.*

Key Words **- Speech Emotion Recognition, *Recurrent Neural Network (RNN), Support Vector Machines (SVM), Logistic Regression (LR), Modulation Spectral (MS) , Mel-Frequency Cepstrum Coefficients (MFCC),* Machine Learning**

I. INTRODUCTION

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion play in shaping human social. Emotional displays convey considerable information about the mental state of an individual has opened up a new research field called automatic emotion recognition, having basic goals to understand interaction and retrieve desired emotions.[1] In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions , speech , physiological signals, etc. SER aims to recognize the underlying emotional state of a speaker from her voice. [2]The area has received increasing research interest all through current years. For example, a teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learner's emotional state should be considered in the classroom.

II. LITERATURE REVIEW

The aim of any literature review is to summarize and synthesize the arguments and ideas of existing knowledge in a particular field without adding any new contributions. Being built on existing knowledge they help the researcher to even turn the wheels of the topic of research. It is possible only with profound knowledge of what is wrong in the existing findings in detail to overpower them.

EEG based affective models without labeled target data using transfer learning techniques approach was used. In this paper [1] has shown that by fusing EEG features and other features with bimodal deep autoencoders (BDAE), the shared representations are good features to discriminate different emotions. For the SEED dataset, compared with other feature merging strategies, the BDAE model is better than others with the best accuracy of 74.94%.

The core module of system in paper [2] was a hybrid network that combines Artificial neural network (ANN) and 3D convolutional networks (C3D) in a late-fusion fashion. ANN takes appearance features extracted by convolutional neural network (CNN) over individual video frames as input and encodes motion later, while C3D models appearance and motion of video simultaneously. Combined with an audio module, the system achieved a recognition accuracy of 59.02% without using any additional emotion-labeled video clips in training set, compared to 53.8% of the winner of Emoti W 2018.

In paper [**3**], they presented HoloNet, a well-designed Convolutional Neural Network (CNN) architecture regarding our submissions to the video based sub-challenge of the Emotion Recognition in the Wild (EmotiW) 2016 challenge. HoloNet has three critical considerations in network design:

1.To reduce redundant filters and enhance the non-saturated non-linearity in the lower convolutional layers, they used a modified Concatenated Rectified Linear Unit (CReLU) instead of ReLU.
2.To enjoy the accuracy gain from considerably increased network depth and maintain efficiency, they combined residual structure and CReLU to construct the middle layers.
3.To broaden network width and introduce multi-scale feature extraction property, the topper layers are designed as a variant of inception-residual structure.
4.They obtained a mean recognition rate of 57.84%, outperforming the baseline accuracy with an absolute margin of 17.37%, and yielding 4.04% absolute accuracy gain compared to the result of last year's winner team.

Author in paper [4], presented a novel set of harmony features for speech emotion recognition. These features are relying opsychoacoustic perception from music theory. First, beginning from predicted pitch of a speech signals, then computing spherical autocorrelation of pitch histogram. It calculate the incidence of dissimilar two-pitch duration, which cause a harmonic or in harmonic impression. In Classification step, Bayesian classifier plays an important rule with a Gaussian class-conditional likelihood. Experimental result in Berlin emotion database by using harmony features indicate an improvement in recognition performance. Recognition rate improved by 2% in average .

Authors in paper [5] ,proposed a segment based method for recognition of emotion in Mandarin speech. This approach is contain the following process. First, define the k parameter in weighted discrete k-NN classifier, the experimental testing of different k shows the best performance for k-NN is when k sets to 10. For selecting the foremost feature set, sequential forward selection (SFS) and sequential backward selection (SBS) are employed. SFS and SBS improves feature accuracy to 83% and 81% respectively. The highest accuracy in segment-based method achieves 80%. The experimental result build on private corpus by inviting 18 males and 16 females. It is essential to gather more expressive speech to explore the extensive emotional investigation, in the future.

Table 1. Literature Review Table

SR .No	References and Year	Approach and method	Performance
1.	A. Revathi, N. Sasikaladevi ,R. Nagakrishnan , C. Jeyalakshmi (2020) [1]	EEG- based affective models without label target data using transfer learning techniques (TCA-based Subject Transfer)	Positive (85.01%) emotion Recognition rate is higher than other approaches but neutral (25.76%) and negative (10.24%)emotions are offered confused with each other.
2.	Leila Kerkeni ,Youssef Serrestou, Mohamed Mbarki , Kosai Raoof and Mohamed Ali Mahjoub. ,(2019) [2]	Semi Supervise Learning (SSL) technique	Delivers a stronger performance in the classification of high low emotional arousal (UAR =76.5%) and significance out performs traditional SSL methods by at least 5.0% (absolute gain).
3.	Patil , K.J. Zope,P.H..Suralkar.(2019) [3]	Video-based Emotional Recognition using CNN -RNN and C3D hybrid networks	Achieved accuracy 59.02% (without using any additional Emotional label video clip in training set) which is the best till now
4.	Martin, V. and Robert, V. (2021) [4]	HoloNet: towards robust emotional Recognition in the wild	Achieved mean Recognition rate of 57.84%
5.	Yelin Kim and Emily Mower Provos(2020) [5]	Data driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes	Achieved 65.60% UW accuracy, 1.90% higher than the baseline

III. OUTCOME OF LITERATURE SURVEY

The focus of existing system lies on accuracy rate approach, which combines content-based and collaborative based approaches. It shows that many of the disadvantages of existing system becomes difficult to catch. The problem statements have robust and automated speech recognition ,analysis of the captured audio of speech ,creating dataset for predicating and training. After studying many research paper we

observed that by using CNN model the accuracy rate of about 35.6% is achieved from the data model [1] [2], and by using Semi Supervised Learning (SSL) technique the accuracy of pitch was too low [3]. So to overcome that problem we decided to use RNN, SVM, LR algorithm and MFFC and MS features to increase the accuracy rate . To detect the emotions like Happy , Sad , Neutral , Fearful , Surprise , Disgust , Angry etc .Our system would be to understand a face and its characteristics and then make it weighted assumption of the identity of the person.

IV. METHODOLOGY

Emotion recognition can have interesting applications in human-robot interaction ,thus having the way for a scenario where human-robot interaction will normally take place in the real world. When a speaker expresses an emotion while adhering to an inconspicuous in to nation pattern, human listeners can never the less perceive the emotional information through the pitch and intensity of speech. On the other hand, our aim is to capture the diverse acoustic cues that are in the speech signal and to analyze their mutual relationship to the speaker's emotion .We propose a technique to recognize several basic emotions, namely Happy , Sad , Neutral , Fearful , Surprise , Disgust ,Angry ,based on the analysis of phonetic and acoustic properties in Figure 1.

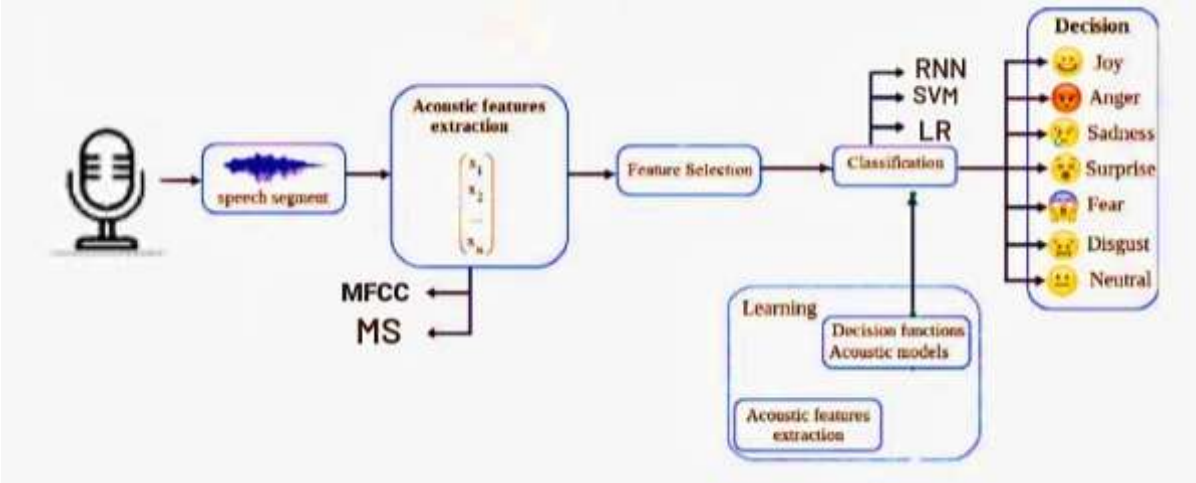

Figure 1. Block diagram

Emotional Speech Data:

In Emotional Speech data, we have explained the for every machine learning task, we need to have a training set of samples; SER is not different from the rest. The process of creating a training dataset for SER needs human agents to label the samples by hand, and different people perceive emotions differently. For example, one might tag an emotional voice as angry whilst the other perceives it as excited. There are some types of databases specifically designed for speech emotion recognition semi-natural and natural speech collections. Semi-natural collections are made by asking people or actors to read a scenario containing different emotions. Moreover, natural datasets are extracted from TV shows, YouTube videos, call centers, and such, and then labeled the emotions by human listeners. In further section, we have utilized 2 emotional speech databases in our experiments : Berlin Database and Spanish Database are standardized collections of emotions, which makes comparing results very easy. Earlier examples of databases for emotional speech used to contain a limited number of samples with a limited number of actors, but newer databases tend to create a larger number of samples and a wider range of speakers.

1. **Berlin Speech Emotional Database**

The standard overall performance and robustness of the recognition systems may be without troubles affected if it isn`t constantly well-knowledgeable with suitable database .Therefore, it is vital to have sufficient and suitable phrases with Inside the database to educate the emotion recognition tool and in the end look at its standard overall performance. The Berlin Database of Emotional Speech is one of the most widely used datasets for speech emotion recognition. It is a simulated dataset composed of 10 German sentences, five short sentences, and five long sentences. Ten speakers, five females, and five males were employed to create the dataset.

2. **Spanish Emotional Database**

We also use the most INTER 1SP Spanish emotional database contains utterances from two professional actors (one female and one male speakers).The spanish corpus that we have the right to access (free for academic and research use (Spa), was recorded twice in the 6 basic emotions plus neutral (anger , sadness , joy, fear , disgust, surprise, Neutral/normal). Four additional neutral variations (soft, loud, slow and fast) were recorded once and it contains more data (4528 utterances in total) .This paper has focused on only 7 main emotions from the Spanish Dataset in order to achieve a higher and more accurate rate of recognition and to make the comparison with the Berlin database detailed above.

Feature Extraction:

In feature extraction we are using the 2 extraction MFCC and MS , as a comparative analysis of the spectral features in isolation provides insights into the performance on a per feature basis. MFCC, Mel spectrogram (MS) were the highest performing individual spectral features across the datasets. When combined these features formed a vector of 155 data-points, and the highest performing permutation of spectral features

1. MFCC Features

Firstly, the results of the comparative analysis showed that the MFCC feature enabled the highest emotion recognition accuracy within each dataset. Demonstrating the importance of modelling for phonetic properties, found within the speech signal shape, in enabling accurate emotion classification across languages The Mel Frequency cepstrum coefficient is the maximum usage representation of the spectral value of a voice alert [9] .These are popular qualities of language because they explain human perception sensitivity by frequency evaluation . For each frame , the Fourier transform and electrical spectrum were predicted and mapped to the Mel frequency scale. The MFCC computing system is typically shown in Figure 2. In our study, we extract the main 12th order of the MFCC coefficient and sample voice alerts at 16KHz.Each MFCC function vector is 60 dimensions.

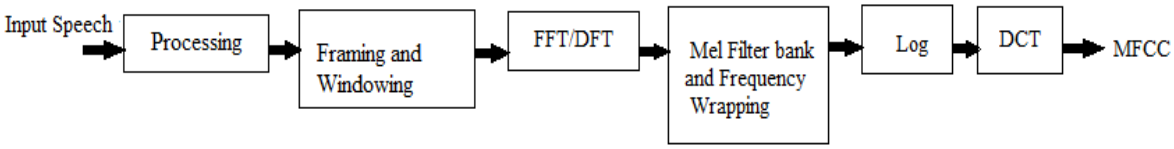

Figure 2. Schema of MFCC Extraction.

2. Mel-Spectrogram Features

Secondly, the performance of the Mel spectrogram feature varied across the datasets. Mel spectrogram performed highly. These functions are obtained by emulating the spectral time (ST) processing performed by the human auditory engine and taking into account the normal acoustic frequency as well as the modulation frequency. The procedure for calculating the ST representation is shown in Identification 2. To collect ST representations, phonetic symbols are first decomposed through an auditory filter bank. The Hilbert envelope of the critical tape output is calculated to form the modulation indicator .Modulation filter banks are implemented as well as Hilbert envelopes to perform frequency analysis .The spectral content of the modulation indica-tor is called the modulation spectrum, and the proposed function is called the modulation spectrum function (MSF). Finally, the ST plot is rendered by measuring the power of the decomposed envelope indicator.

Classification :

For modeling the emotional states, We are using three classification models such as support vector machine (SVM), Recurrent Neural Network (RNN), Logistic Regression (LR) to find best optimal model for our system by performing comparative analysis on this classifications.

1. Support Vector Machine

Support Vector Machines (SVM) is an optimal margin classifiers in machine learning. It can have a very good classification performance compared to other classifiers especially for limited training data .SVM is a supervised learning methods widely used for classification and regression with early practical implementation since 90's with high performance, simple and efficient computation of machine learning algorithms. or without SVM are 76.1% and 57.8% respectively in Figure 3 and Figure 4 we have demonstrated the accuracy and prediction of Gender and emotion .The Accuracy Rate using SVM was only 28%.

```
Confusion matrix :
 [[ 8  0  0  0  5  0  0  0]
 [ 0  0  0  0  3  0 10  0]
 [ 2  0  0  0  2  0  2  0]
 [ 2  0  0  0  5  0  3  0]
 [ 2  0  0  0  4  0  9  0]
 [ 0  0  0  0  0  0  3  0]
 [ 0  0  0  0  1  0  9  0]
 [ 2  0  0  0  3  0  0  0]]
```

```
[ ]  # classification report of model
     print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.50	0.62	0.55	13
1	0.00	0.00	0.00	13
2	0.00	0.00	0.00	6
3	0.00	0.00	0.00	10
4	0.17	0.27	0.21	15
5	0.00	0.00	0.00	3
6	0.25	0.90	0.39	10
7	0.00	0.00	0.00	5
accuracy			0.28	75
macro avg	0.12	0.22	0.14	75
weighted avg	0.15	0.28	0.19	75

Figure 3. Accuracy rate of emotion using SVM

```
[ ]  # classification report of the model
     print(classification_report(y_test,y_predictions))
```

	precision	recall	f1-score	support
female	0.96	0.86	0.91	329
male	0.86	0.96	0.91	305
accuracy			0.91	634
macro avg	0.91	0.91	0.91	634
weighted avg	0.91	0.91	0.91	634

```
[ ]  # confusion matrix
     print("Confusion matrix : \n", confusion_matrix(y_test,y_pre
```

```
Confusion matrix :
 [[282  47]
 [ 12 293]]
```

Figure 4 Accuracy rate of gender using SVM

2.Recurrent Neural Networks

Recurrent Neural Networks (RNN) are suitable for learning time series data. While RNN models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, LSTM (Long Short Term Memory) RNNs were proposed by Hochreiter et al (Seppand Jurgen, 1997) it uses memory cells to store

information so that it can exploit long range dependencies in the data .The comparision of predicted and actual lables as shown in Figure 5 it is shown the RNN output of labels for each emotion. And Figure 6 shows the accuracy rate using RNN was 67 %.

	Predicted Labels	Actual Labels
0	calm	calm
1	calm	calm
2	calm	calm
3	sad	happy
4	fear	fear
5	angry	angry
6	calm	calm
7	sad	sad
8	surprise	surprise
9	surprise	surprise

Figure 5. Comparision of predicted and actual lables in RNN

	precision	recall	f1-score	support
angry	1.00	1.00	1.00	13
calm	0.63	0.92	0.75	13
disgust	0.50	0.17	0.25	6
fear	0.50	0.40	0.44	10
happy	0.64	0.47	0.54	15
neutral	0.00	0.00	0.00	3
sad	0.59	1.00	0.74	10
surprise	0.60	0.60	0.60	5
accuracy			0.67	75
macro avg	0.56	0.57	0.54	75
weighted avg	0.64	0.67	0.63	75

Figure 6. Accuracy rate of emotion using RNN

3.Logistic Regression

Logistic Regression is a supervised classification algorithm which produces probability values of data belonging to different classes. There are three types of Logistic Regression algorithms, namely Binary class, Multi-class and Ordinal class logistic algorithms depending on the type of target class. The Wikipedia definition states that “Logistic regression computes the relationship between the target shown in Figure 7(dependent) variable and one or more independent variables using the estimated probability values through a logistic function” Figure 8. The Accuracy Rate using LR was only 75% .The logistic function, also known as a sigmoid function, maps predicted values to probability values.

	precision	recall	f1-score	support
0	0.82	0.69	0.75	13
1	0.91	0.77	0.83	13
2	0.75	0.50	0.60	6
3	0.71	0.50	0.59	10
4	0.92	0.80	0.86	15
5	0.60	1.00	0.75	3
6	0.60	0.90	0.72	10
7	0.44	0.80	0.57	5
accuracy			0.73	75
macro avg	0.72	0.75	0.71	75
weighted avg	0.77	0.73	0.74	75

Figure 7. Accuracy rate of emotion using RNN

```
[ ] # classification report of the model
    print(classification_report(y_test,y_predictions))
```

	precision	recall	f1-score	support
female	0.96	0.86	0.91	329
male	0.86	0.96	0.91	305
accuracy			0.91	634
macro avg	0.91	0.91	0.91	634
weighted avg	0.91	0.91	0.91	634

```
[ ] # confusion matrix
    print("Confusion matrix : \n", confusion_matrix(y_test,y_predictions
```

```
Confusion matrix :
 [[282  47]
 [ 12 293]]
```

Figure 8 .Accuracy rate of gender using SVM

V. RESULT

In our Speech Emotion Recognition System all above results are predicted using Linear Regression as we did comparative analysis of three classification models so the accuracy rate of LR was more that SVM and RNN.

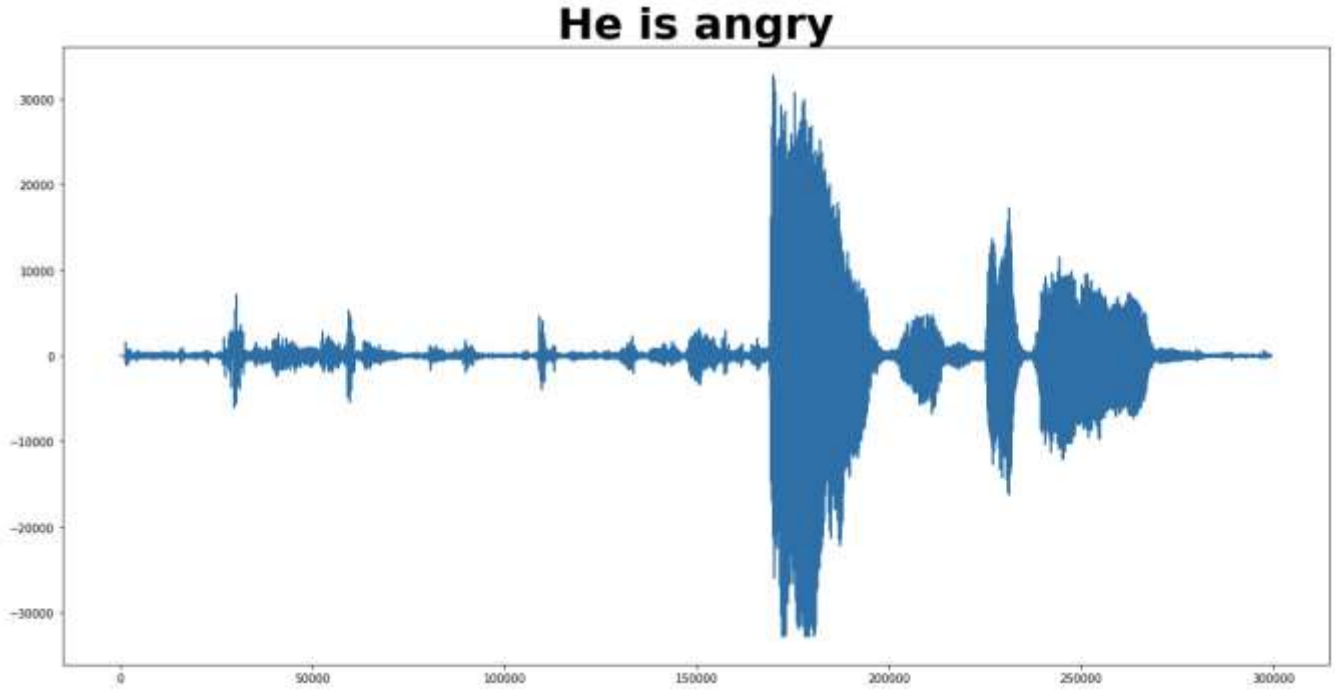

Figure 9. Result of predicted emotion from audio using Linear Regression

From this Figure 9 output ,we can visualize that the emotion predicted will be angry which is having more probability when compared to other emotions.

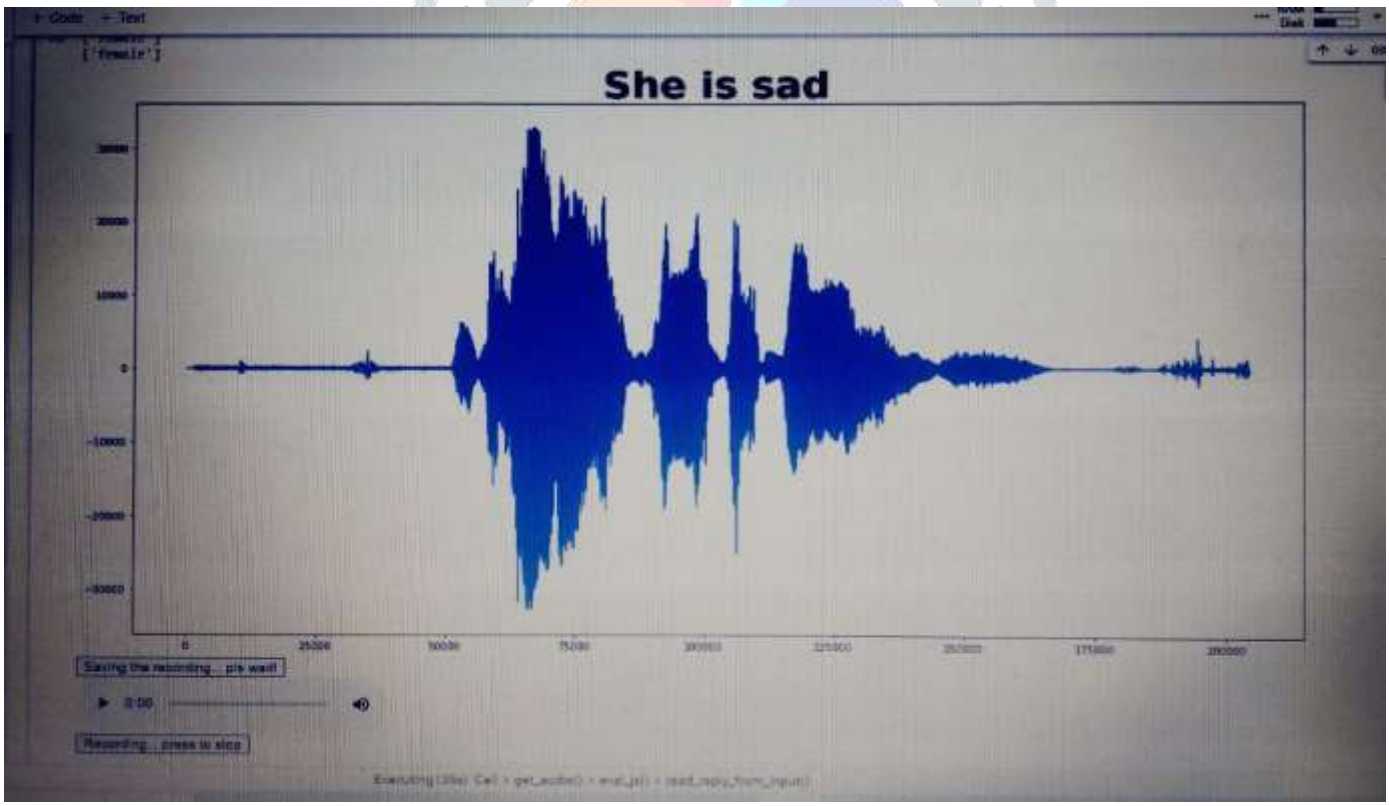

Figure 10. Result of predicted emotion from audio using Linear Regression

In Figure 10, From this output ,we can visualize that the emotion predicted will be sad which is having more probability when compared to other emotions.

Figure 11. Result of predicted emotion from audio using Linear Regression

In Figure 11,From this output ,we can visualize that the emotion predicted will be Disgust ,which is having more probability when compared to other emotions.

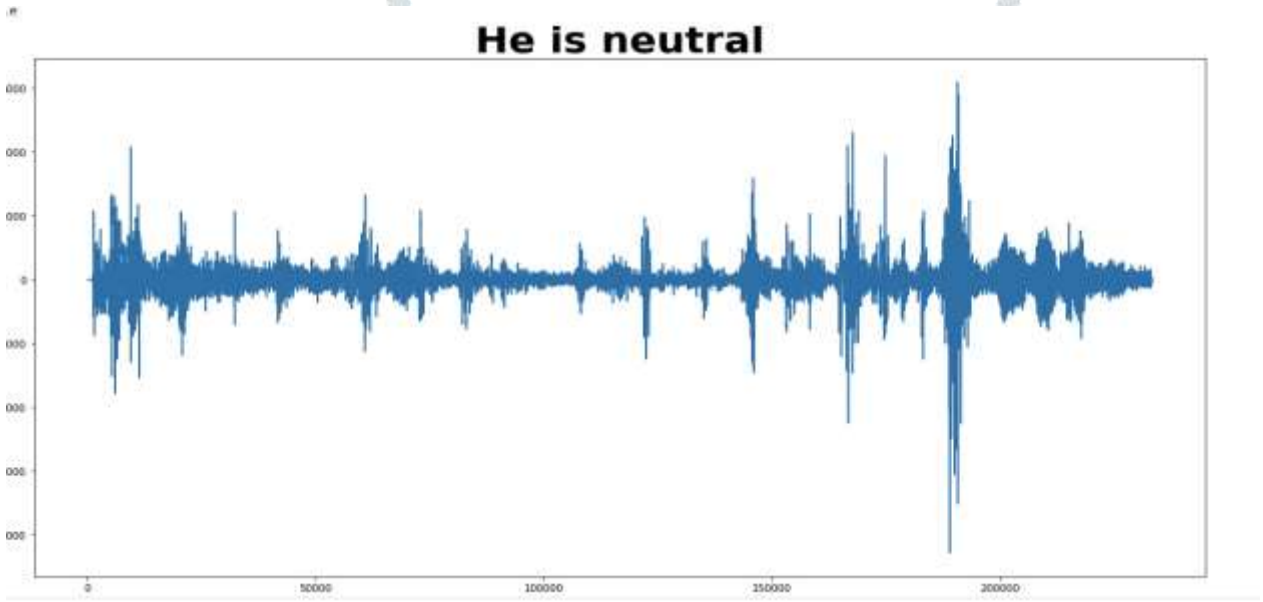

Figure 12. Result of predicted emotion from audio using Linear Regression

In Figure 12 , From this output ,we can visualize that the emotion predicted will be Neutral ,which is having more probability when compared to other emotions.

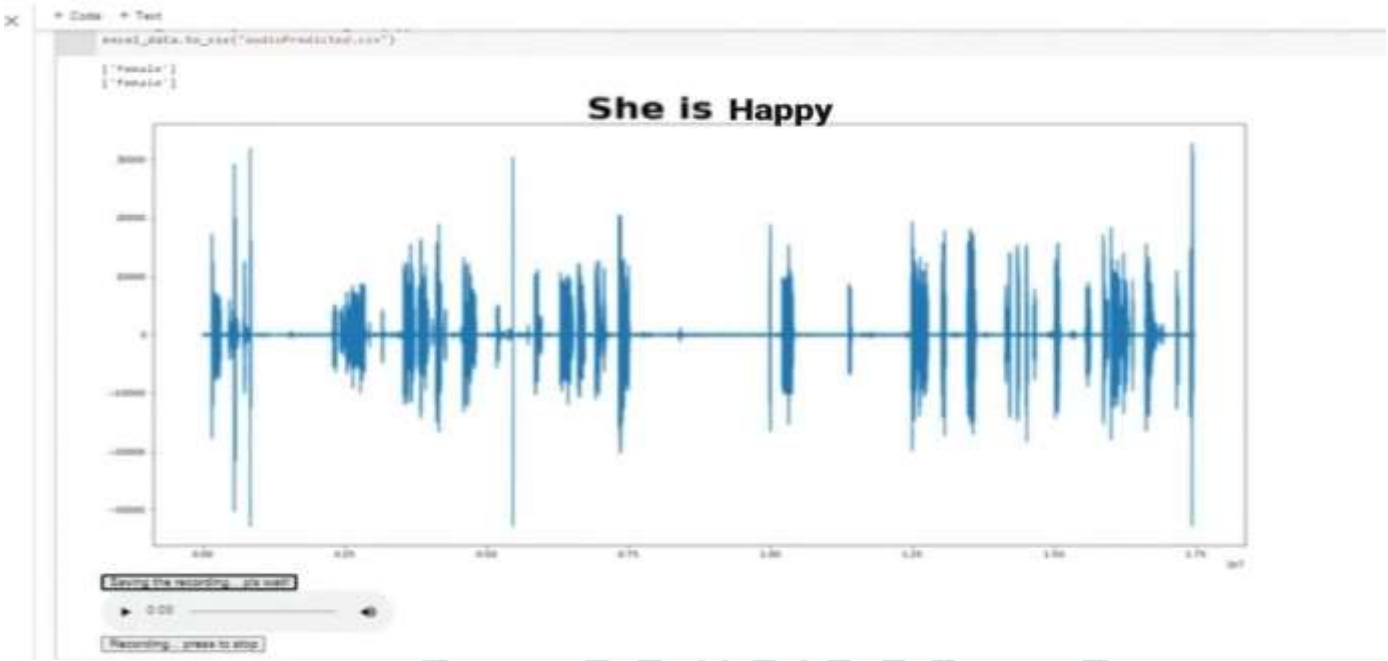

Figure 13 . Result of predicted emotion from audio using Linear Regression

In Figure 13 , From this output ,we can visualize that the emotion predicted will be Happy ,which is having more probability when compared to other emotions.

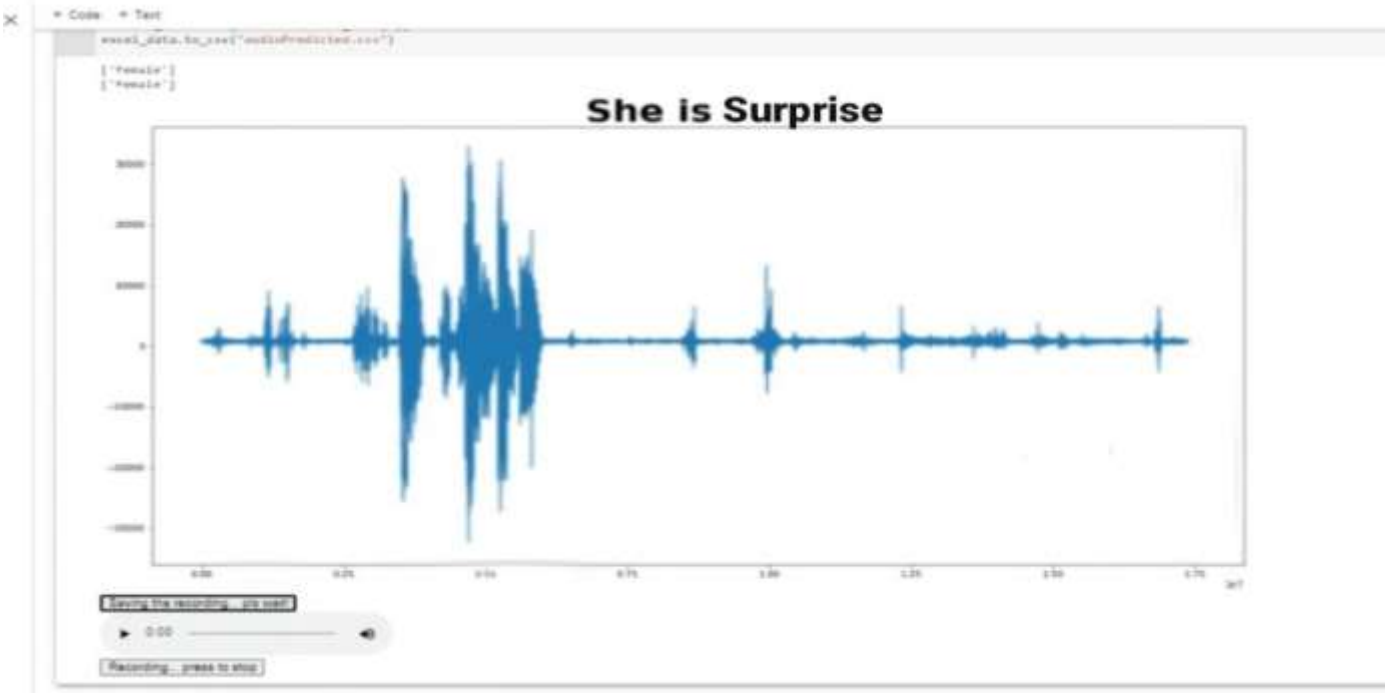

Figure.14 . Result of predicted emotion from audio using Linear Regression

In Figure 14 From this output ,we can visualize that the emotion predicted will be Surprise ,which is having more probability when compared to other emotions.

VI . CONCLUSION AND FUTURE SCOPE

After extensive research we came to conclusion that Logistic Regression and RNN will be best among all models as it provides more accuracy. We have got 89% accuracy by using LR model and for RNN we got 80% accuracy Happy , Sad , Neutral , Fearful , Surprise , Disgust ,Angry this are different 7 emotions we will give through this project. This speech based emotion recognition can be used in understanding the opinions/ sentiments they express regarding a product or political opinion etc.by giving the audio as the input to this model. .Various other emotions can be added lot of researchers are exploring this field due to its wide importance. Exploring the vast literature review it has been found that maximum research was done in field of text emotion analysis. Still the audio emotion analysis needs much work and improvement in terms of accuracy. We are made our system more accurate in audio. Integrated the system with various

platforms. Learning from continuous emotions or combining the audio modality with other modalities such as facial expressions to improve performance**.** The major challenges we face is in making increasing accuracy rate .We have accomplished system module accuracy ,still our system has many limitation .There is further scope for modification accurate such as Various other emotions can be added ,Integrating the system with different platforms ,Learning from continuous emotions or combining the audio modality with other modalities such as facial expressions to improve performance**.**

REFERENCES

[1] A. Revathi, N. Sasikaladevi, R. Nagakrishnan, C. Jeyalakshmi. ,"Robust emotion recognition from speech" Gamma tone features and models International Journal of Speech Technology, may 2020,pp.20-24.

[2] Patil, K.J. Zope, P.H.; Suralkar, S.R. "Emotion Detection From Speech Using Mfcc and Gmm". Int. J. Eng. Res. Technol. (IJERT) 2018,1, 9.

[3] Hsu, C.W.; Chang, C.C.; Lin," C.J. A Practical Guide to Support Vector Classification." 2019; pp. 1396–1400.

[4] S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., Mar. 2020, pp. 293–303

[5] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2017

[6] Asia-Pacific. Martin, V. and Robert, V. (2019). Recognition of Emotions in German Speech Using Gaussian Mixture Models

[7] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2016.

[8] S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., Mar. 2019, pp. 293–303

[9] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[10] Kalsum, Tehmina, et al. "Emotion recognition using hybrid feature descriptors." IET Image Processing 12.6 (2018): 1004-1012

[11] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, May 2020.

[12] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[13] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 20118

[14] L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2016

[15] Leila Kerkeni ,Youssef Serrestou , Mohamed Mbarki , Kosai Raoof and Mohamed Ali Mahjoub. , 2020," Speech Emotion Recognition: Methods and Cases Study,2019,pp.130-144.