Implementation Paper on Speech Emotion Recognition & Accent Identification

Akanksha Gadikar[1], Omkar Gokhale[2], Subodh Wagh[3], Anjali Wankhede[4] Prof. Preeti Joshi[5]

[1234]Student, [5]Professor

[12345]Dept. of Information Technology Engineering, MMCOE, Pune, Maharashtra, India)

Abstract- In human machine interface application, emotion recognition from the speech signal has been research topic since many years. To identify the emotions from the speech signal, many systems have been developed. Speech has several characteristic features such as naturalness, pitch & tone, which makes it as attractive interface medium. It is possible to express emotions and attitudes through speech. Here in this paper, study has been carried out to recognize the human emotion through speech using the accent of the human. To recognize accent through speech various speech features were extracted. Based on these speech features Classification of the accent and emotion has been done using KNN. Here six emotion are considered like neutral, disgust, happy, sad, anger, and fear. The classification performance is based on extracted features using MFCC. Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.

Keywords—Emotion Recognition, Feature extraction, MFCC, KNN

I. INTRODUCTION

Human machine interaction are widely used nowadays in many applications. One of the medium of interaction is speech. Emotions are subjective experiences which play an important role in expressing mental and physical states of the humans and it is often associated with variety of feelings.

The Emotion Detection from Speech consists of two modules. First module is Speech to accent detection of user .This module identifies or gives result as accent of user which are living in particular region. This module is used to denote the region of user. Second module accent is considered as input. This module extract features of users speech to detect emotion of user.

In recent years, Speech Emotion Recognition has made great progress, especially after the utilization of deep learning. A typical SER system abstracts a collection of acoustic features or semantic features on top of automatic speech recognition (ASR) transcription, and then trains a multi-classifier classification model by machine learning methods such as SVM, decision tree and GMM.

The main characteristics of the proposed system are:

1. To develop model which identifies and extracts the features from the speech of the user.

2] To develop a model which will identify the regional aspects and detect the accent of the user.

3] To predict the emotional state of the user.

II. LITERATURE SURVEY

In [1] named “Speech Based Human Emotion Recognition Using MFCC”, IEEE WiSPNET year of 2017 conference proposed by authors named as M.S. Likitha , Sri Raksha R. Gupta.

A database consist of voices of 60 people with different emotions. Speech signal of speaker’s read using the function a wavread in MATLAB tool .MFCC method is used for detecting emotion from voice signals. Proposed work is based on feature extraction using MFCC and decision making using standard deviation. The speech signal made to undergo framing, after which it is passed through Hamming window for windowing process. Fast Fourier Transform was performed on the input signal. After which the Mel Frequency Cepstral Coefficients were obtained. The standard deviation for the mean value was found, and this value was passed through as if else a statement, where the obtained standard deviation of that particular emotion is compared with the optimized values of standard deviation for different emotions, and the corresponding emotion were displayed. It can predict the 3 basic emotions such as happy, sad, angry from MFCC waves.Some advantages of following proposed system are:

1) MFCC is simplest method for emotion detection.
2) Efficiency and performance remains constant even in noisy environment.

Hence this system can serve as noise robust emotion recognition system. Such efficiency in noisy environment extends the scope of the work wherein emotion recognition systems can be utilized in military.

In [2] named “Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural

Network Architecture", IEEE WiSPNET year of 2017 conference proposed by authors named as SaikatBasu, Jaybrata Chakraborty.

For the last two decades, several intelligent systems are proposed by re-searchers. These different systems also differs by the nature of features used for classification of speech signals. There are the widely used spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC).Gausian Mixture Model(GMM), Support Vector Machine (SVM) and Hidden Markov Model (HMM) are used by researchers for classification using a machine learning method. Only pitch and MFCC features are used for recognition of emotion.

In this article they get 13 MFCC (Mel Frequency Cepstral Coefficient) with 13 velocity and 13 acceleration component as feature extraction component and a CNN (Convolution Neural Network) and LSTM(Long Short Term Memory) based approach for classification. They chose Berlin Emotional Speech dataset (EmoDB) for classification purpose. It consists 535 utterances spoken by 10 different actors. All the recordings took place in the anechoic chamber of the Technical University Berlin, department of Technical Acoustics. RNN and Long Short Term Memory (LSTM) network can be used for training. In LSTM network, they have provided two hidden layer with 50 nodes in RST layer and 20 nodes in second layer. They have approximately 80 % of accuracy on test data. Some advantages are:

i) LSTM-RNN model was very effective and promising.

In [3] named "Emotion Recognition from Speech using Emotional Statistical Parametric Speech Synthesis Using LSTM-RNN's", IEEE WiSPNET of 2017 conference year of 12 - 15 December 2017,proposed by authors named as Shumin An, Zhenhua Ling and Lirong Da.

Two modelling approaches, emotion dependent modelling and unified modelling with emotion codes, are implemented and compared by experiments. In the RST approach, LSTM-RNN-based acoustic models are built separately for each emotion type. A speaker-independent acoustic model estimated using the speech data from multi-speakers is adopted to initialize the emotion-dependent LSTM-RNN. Second approach builds a unified LSTM-RNN-based acoustic model using the training data of a variety of emotion types. Experimental results on an emotional speech synthesis database with four emotion types (neutral style, happiness, anger, and sadness).The emotion codes used by the unified modelling approach are capable of controlling the emotion type and intensity of synthetic speech effectively by interpolating and extrapolating the codes in the training set.

In [4] named "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks", IEEE WiSPNET of 2017 conference proposed by authors named as Wootaek Lim, Daeyoung Jang and Taejin Lee.

In this study, we investigate the result of the Speech Emotion Recognition (SER) algorithm based on CNNs and RNNs trained using an emotional speech database. The main goal of their work is to propose a SER method based on concatenated CNNs and RNNs without using any traditional hand-crafted features. In this work, they propose the CNNs, RNNs, and time distributed CNNs-based SER method that acquires signals on a 2D domain speech signal representation. In particular, they propose a method for analysing sequential audio data based on concatenated CNNs and RNNs.

Advantage:

i) By applying the proposed methods to an emotional speech database, the classification result was verified to have better accuracy than that achieved using conventional classification methods.

ii) Deep learning algorithms overcome limitations of handcrafted features.

Disadvantag**e:**

Deep learning algorithms are complex.

In [5] named "Speaker Accent Recognition by MFCC Using K-Nearest Neighbour Algorithm", IJAR in Computer Engineering , year of 2018 conference proposed by authors named as Munish Bhatia, Navpreet Singh, Amitpal Singh.

A K-Nearest Neighbour Algorithm involving Mel- Frequency Cepstral Coefficients (MFCCs) is provided to perform Speech signal feature extraction for the task of speaker accent recognition and speaker emotion detection. Mel-Frequency Cepstral Co- efficient is effectively used to perform the feature extraction of the input signal. For each input signal the mean of the MFCC matrix is used for pattern recognition . The K-nearest neighbour algorithm is based on evaluating minimum Euclidean distance measure from input data set to stored data set. Since large number of speakers of different accent are present, they can be grouped together depending upon their accent .Thus each signal coming from different group makes a distinct MFCC vector .

A speaker's accent distinguishes him/her as a member of a A speaker's accent distinguishes him/her as a member of a group. These groups have been classified on the basis of geographic areas, by social class . As human being we are sensitive to accent and can tell whether a speaker belongs to a group or not . For above scenario to be possible accent are pattern of speaking that distinguish a group from other group.

Advantages: 1)speaker accent recognition provides accuracy to speaker emotion recognition.

2) Speech Processing involves capturing of sound and

process its features by extracting it by using technique MFCC .

3) The K-nearest neighbour algorithm is based on evaluating minimum Euclidean distance measure from input data set to stored data set used for classification of accent and emotion of the speaker.

Disadvantages: Speech technology is still facing problem with pronunciation variation across different accent groups. In automated speech recognition, a variation in accent used in testing and training can lead increase in word error rate.

III. LGORITHMIC APPROACH K- NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

KNN Algorithm

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data

3.1 Calculate the distance between the query example and the current example from the data.

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries.
7. For classification, returns the mode of the K labels

Choosing the right value for K

1. As we decrease the value of K to 1, our predictions become less stable. Just think for a minute, imagine K=1 and we have a query point surrounded by several reds and one green. Point is most likely red, but because K=1, KNN incorrectly predicts that the query point is green.

2. Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point).

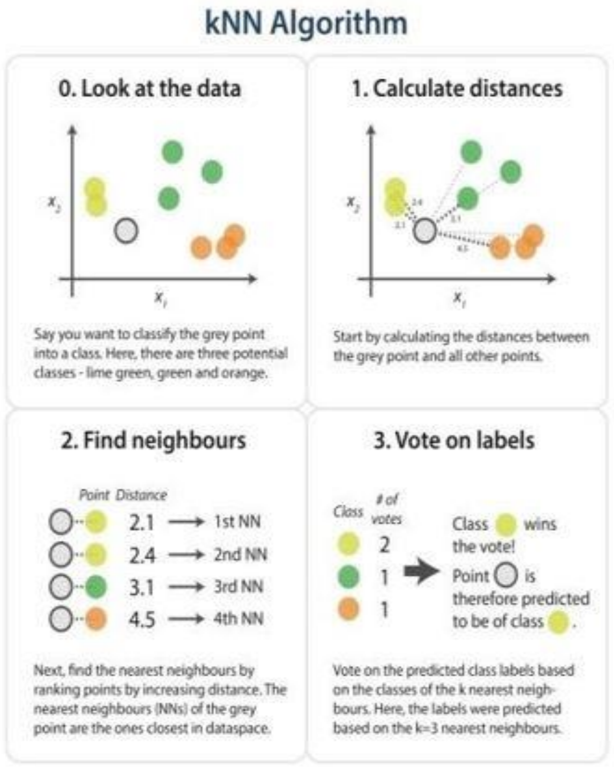

kNN Algorithm
0. Look at the data
x_2
x_1
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.
1. Calculate distances
x_2
x_1
Start by calculating the distances between the grey point and all other points.
2. Find neighbours
Point Distance
2.1 → 1st NN
2.4 → 2nd NN
3.1 → 3rd NN
4.5 → 4th NN
Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.
3. Vote on labels
Class # of votes
2
1
1
Class wins the vote!
Point is therefore predicted to be of class .
Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Fig.1: KNN Overview

IV. PROPOSED SYSTEM

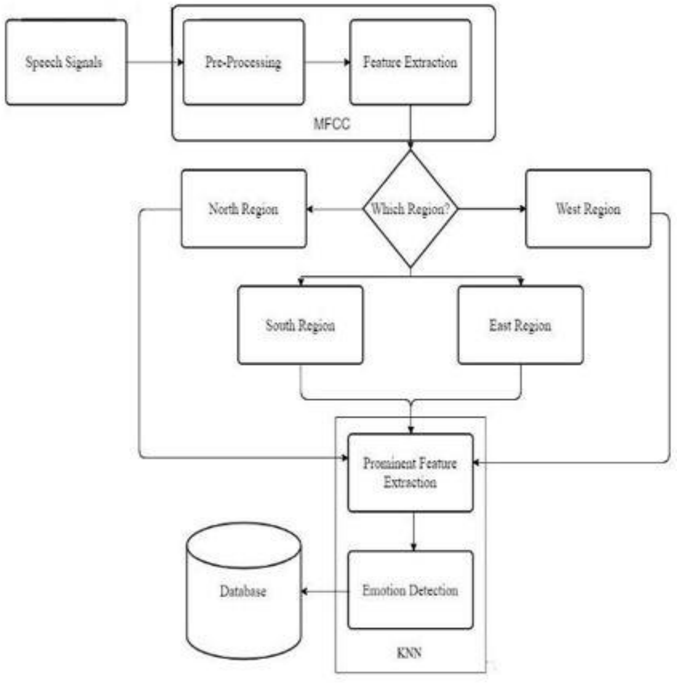


Fig.2: System Architecture

According to proposal, the modules are as follows :

MFCC Algorithm for State(Region) Identification and KNN for Emotion Detection.

I) Region Identification:
The acoustic characteristic of the speech signal is feature. Feature Extraction is a small amount of data from the speech signal is extracted to analyze the signal without disturbing its acoustic properties.

FCC Algorithm is used for feature extraction from Speech.

Following diagram defines the MFCC Flowchart:

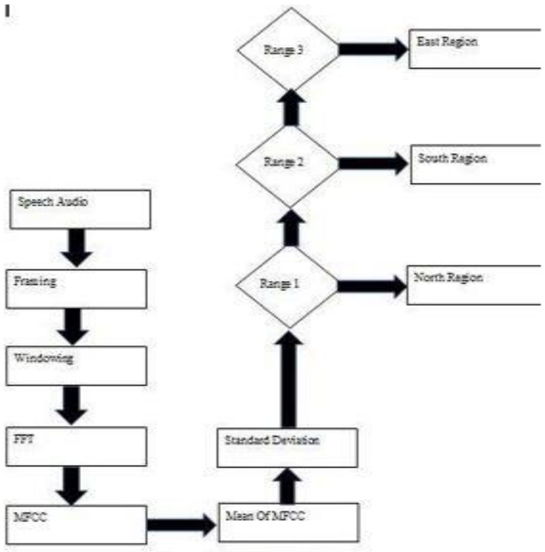


Figure 4.2: MFCC Flowchart

1] Pre-processing
Before using MFCC we make some preprocessing on the data set. All the speech les are with .wav extension, first we compute amplitude values of each le with a sample rate of 16000 sample per second. Then we take a weighted average according to the length of speech les and make them equal by adding zeros to the smaller le to make them equal to the average length le and crop all the larger le for the same purpose. After this process all les became of equal size.

2] Pre-emphasis
Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increase the energy of signal. This increment of energy level gives more information.

3] Framing
In this process, speech sample is segmented into 20-40 ms frames. The length of human voice may vary, so for fixing the size of speech this processes is necessary. Although the speech signal is non-stationary in nature frequency can be changed over the time period), but for a short duration of time, signal behave like a stationary signal.
Overlapping of frames is done to smoothen the transitions between frames by a size. The first frame consists of the first N = 256 (typical value) samples. The second frame begins at M= 100 (typical value) samples after the first frame, and overlaps it by N-M samples and so on.
This process continues until the entries speech signal is accounted.

4] Windowing
After frame blocking, the signal is pass through a windowing function to minimize discontinuities and spectral distortion at the extremes of each frame. The result of windowing a signal x(n) is given by Y(n)= x(n) w(n) , 0 & lt;= n & lt;= N-1where w(n) is the window function, 0 &lt;= n &lt;=N - 1, and N is the number of samples in each frame.
Windowing function reduce the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a10ms span. That means each frame contain some overlapping portion of previous frame.

5] Fast Fourier Transform (FFT)
FFT is used to generate the frequency spectrum of each frame. Each sample of each frame converted from time domain to frequency domain by the FFT. FFT is used to find all frequencies present in the particular frame. FFT is performed on the windowing signal to it into frequency domain. The discrete Fourier Transform (DFT) over discrete signal x(n) of N samples, converts each frame of N samples into the frequency domain from its time domain. The FFT is the fast algorithm to implement DFT.

6] Mel Frequency Warping
The Mel-frequency scale is linearly spaced for frequencies below 1000 Hz and logarithmically spaced above 1000 Hz. A pitch of 1000 Hz tone, which is equivalent to 1000mels when it is above the perceptual hearing threshold level by 40 dB.formula can be used to compute the mels for a given frequency f in Hz. Mel (f) =2595 * log 10(1+f/700)After getting the filter bank energy of each frame, log function is applied to them. It is also inspired by human hearing perception. A human does not listen loud volume on a linear scale. If the volume of the sound is high, human ear cannot recognize large variations in energy. Log energy computation gives those features for which human can listen clearly.

7] Discrete Cosine Transformation(DCT)
In the final step DCT is calculated of the log filter bank energies. We have used 25ms frames with 10ms of sliding. We have also used 26 band pass filters. From each frame we

computed 13 MFCC features. We have also calculated energy within a frame. After getting 13 MFCC features, we also computed 13 velocity components and 13 acceleration components by calculating time derivatives of energy and MFCC.

II) Emotion Detection Module

Speech prominent features are extracted and used for emotion detection.T he proposed system results six basic emotions which include Happy, Sad, Angry, Disgust, Fear and Neutral. Using MFCC for feature extraction and KNN for classifying the emotion and region.

V. RESULT

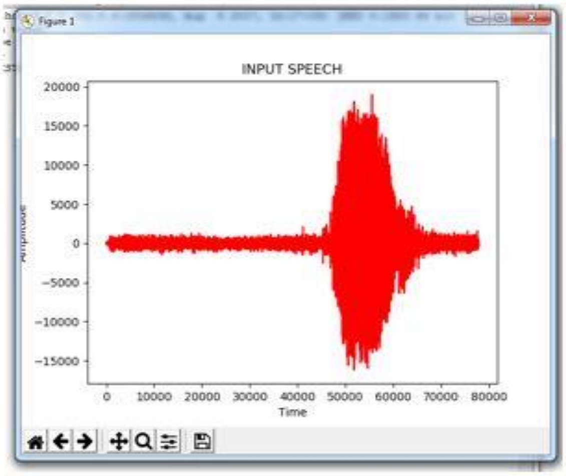


Fig.3: Output 1

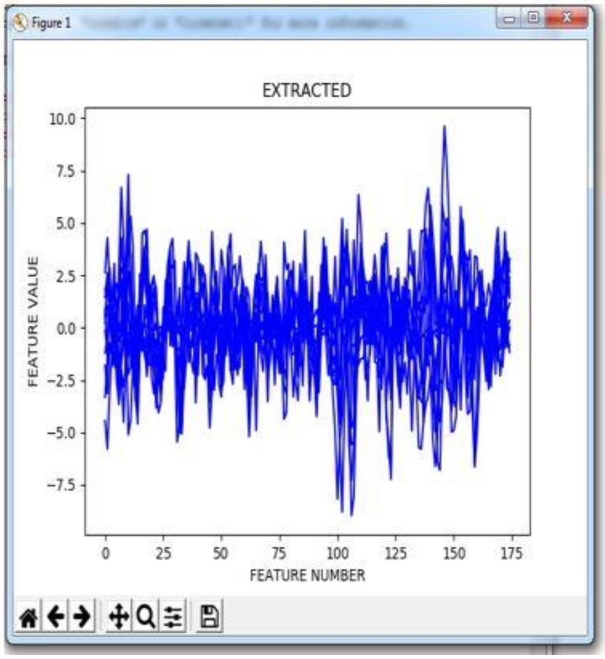
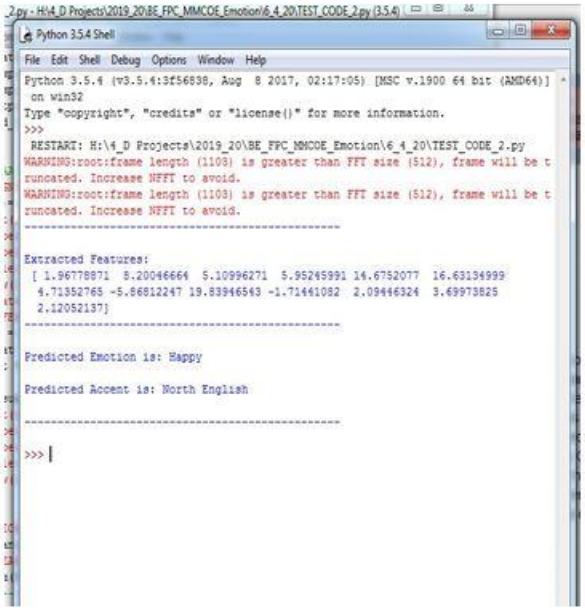


Fig.4: Output 2


_2.py - H:\4_D Projects\2019_20\BE_FPC_MMCOE_Emotion\6_4_20\TEST_CODE_2.py (3.5.4)
Python 3.5.4 Shell
File
Edit
Shell
Debug
Options
Window
Help

```
Python 3.5.4 (v3.5.4:3f56838, Aug  8 2017, 02:17:05) [MSC v.1900 64 bit (AMD64)]
 on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: H:\4_D Projects\2019_20\BE_FPC_MMCOE_Emotion\6_4_20\TEST_CODE_2.py
WARNING:root:frame length (1103) is greater than FFT size (512), frame will be t
runcated. Increase NFFT to avoid.
WARNING:root:frame length (1103) is greater than FFT size (512), frame will be t
runcated. Increase NFFT to avoid.
-----------------------------------------------

Extracted Features:
 [ 1.96778871  8.20046664  5.10996271  5.95245991 14.6752077  16.63134999
  4.71352765 -5.86812247 19.83946543 -1.71441082  2.09446324  3.69973825
  2.12052137]
-----------------------------------------------

Predicted Emotion is: Happy

Predicted Accent is: North English

-----------------------------------------------

>>> |
```

Fig.5: Output 3

VI. CONCLUSION

Although none of the approaches proved to be good enough for practical purposes with the present extent of development, they were good enough to prove that translating speech into in a feature space works for recognition purposes. Speech emotion could be useful in speech understanding, recommendation, retrieval and some other related applications. In this project, focus on challenging issue of recognizing speech emotions such as happy, sad, anger, fear, and neutral. Speech database collected from linguistics laboratory. Mel Frequency Cepstral Coefficient (MFCC) features are extracted from the speech signal. The neural network model classifier used to recognize the emotion from the features taken. The rate of recognizing the emotion from the speech signal is about 88.3%. Experimental results show neural network classifier is better which offers a new efficient way of solving problems. Finally, the new approach developed for training the neural network's architecture proved to be simple and very efficient. It reduced considerably the amount of calculations needed for finding the correct set of parameters. If the traditional approach had been used instead, the amount of calculations would have been higher.

VII. REFERENCES

[1]. M.S. Likitha,1 Sri Raksha R. Gupta. "Speech Based Human Emotion Recognition Using MFCC", IEEE WiSPNET 2017 conference, pp. 2257-2260, 2017 IEEE.

[2]. SaikatBasu, Jaybrata Chakraborty. "Emotion Recog- nition from Speech using Convolutional Neural Network with RecurrentNeural Network Architecture",IEEE Xplore

Compliant - Part Number:CFP17AWO-ART, ISBN:978-1-5090-5013, pp. 333-336, 2017 IEEE.

[3]. Shumin An, Zhenhua Ling and Lirong Dai. "Emotional Statistical Parametric Speech Synthesis Using LSTM- RNNs", Proceed-ings of APSIPA Annual Summit and Conference 2017, pp. 1613-1616, 12 - 15 December 2017.

[4]. Panagiotis Tzirakis, George Trigeorgis . "End-to-EndMultimodal Emotion Recognition Using Deep Neural Networks",IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESS-ING, VOL. 11, NO. 8,pp. 1301-1309, DECEMBER 2017.

[5]. Wootaek Lim, Daeyoung Jang and Taejin Lee. "Speech Emotion Recognition using Convolutional and Recurrent Neural Net-works", Audio and Acoustics Research Section, ETRI.

[6]. Pengyu Cong, Chaomin Wang."UnsatisdZed Customer Call Detection with Deep Learning",2016 IEEE.

[7]. S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh."Speech Emotion Recognition", 2014 IEEE.

[8]. George Trigeorgis, Fabien Ringeval, Raymond Brueck- ner, Erik Marchi."END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NET-WORK", pp.5200-5204, 2016 IEEE.

[9]. JianweiNiu, Yanmin Qian, Kai Yu. "Acoustic Emotion Recognition using Deep Neural Network", pp.128-132, 2014 IEEE.

[10]. Mohsin Y Ahmed, Zeya Chen, Emma Fass and John Stankovic. "Real Time Distant Speech Emotion Recognition in Indoor Environments", November 2017.

[11]. Ma Xiaoxi, Lin Weisi, Huang Dongyan, Dong Minghui, Haizhou Li. “Facial Emotion Recognition”, IEEE 2nd International Conference on Signal and Image Processing , 2017 IEEE.

[12]. Aditya Ambekar, Gaurav Deshmukh, Parikshit Aswarmol, Piyush Dave, “Speech Recognition using Recurrent NeuralNetworks” , 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.

[13]. Devamanyu Hazarika, SoujanyaPoria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, Rada Mihalcea, “CASCADE: Contextual Sarcasm Detection in Online Discussion Forums”, arXiv:1805.06413v1, May 2018.

[14]. Anthony Hu, Seth Flaxman , “Multimodal Sentimental Analysis to Explore the structure of Emotion”, Applied Data Science Track Paper, KDD 2018.

[15]. Jonathan Herzig, Michal Shmueli-Scheuer, David Konopnicki, “Emotion Detection from Text via Ensemble Classification Using Word Embedding”, ICTIR 17 , 2017.

[16]. Emad Barsoum, Cha Zhang, Cristian Canton Ferrer and Zhengyou Zhang, “Training Deep Networks for Facial Expression Recognitionwith Crowd-Sourced Label Distribution”, Microsoft Research, September 2016.

[17]. Rajesh K M, Naveenkumar M, “A Robust Method for Face Recognition and Face Emotion Detection System using Support VectorMachines”, 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016.

[18]. Asif Salekin, Zeya Chen,Mohsin y Ahmed,John Lach,Donna Metz,Kayla De La Haye,Brooke Bell,John A. Stankovic, “Distant Emotion Recognition”, ACM 2017.

[19]. Safaa Azzakhnini, LahoucineBallihi, DrissAboutajdine, “Combining Facial Parts For Learning Gender, Ethnicity,and Emotional State Based on RGB-D Information”, ACM Trans. Multimedia Comput. Commun. Appl., Vol. 14, No. 1s, Article 19. Publication date: March 2018.

[20]. LigangZhang, Brijesh Verma, Dian Tjondronegoro, Vinod Chandran, “Facial Expression Analysis under Partial Occlusion: A Survey”, ACM Computing Surveys, Vol. 51, No. 2, Article 25. Publication date: April 2018.