Emotion Recognition from Speech using Acoustic Features and Linguistic Content

Monisri Rajendran
M.Sc Computing
18210193

***Abstract*—Emotion recognition from speech signals can provide a lot of information on the sentiment of the speaker. It can be useful in a number of scenarios where it is necessary to understand the emotional state of the speaker. This paper presents an approach to Emotion recognition by using both acoustic features and linguistic content of the speech. The speech signals (acoustic and Phonetic features) and the content of the Speech (text converted from audio) are analyzed separately and combined together to arrive at an output. Audio Classification is kept as the base model and Text classification model is used as a supporting model. The result shows that the combination model is better at predicting emotion than when the result from Audio Classification and content based classification is taken separately. RAVDESS (Ryerson University) Audio Emotional Data-set with 8 discrete emotions is used.**

***Index Terms*—Emotion Recognition, Speech signals, acoustic and Phonetic features, Audio Analytics**

I. Introduction

In this paper, we investigate a method to analyze the sentiment of a speaker from his or her speech. Emotions are part and parcel of human life and among other things, highly influence decision making [1]. Some of the basic emotions include joy, surprise, fear, anger, neutral etc., and also as two conventional classification positive or negative. Sentiment Analysis using Voice recognition can be applied in almost all areas where calls can be recorded like emergency calls, customer call centers, defense etc., An audio recording can be classified in to a set of predefined classes or emotions by supervised learning.

Real time audio analytics in a customer call center [2] might help an agent to understand the mood or emotion of the caller and motivate him to finish the call in a positive note when the caller is not happy. Similarly, also keep his own emotion under check when he is talking to the caller and tone down when he sounds angry to the customer. Audio analytics can be used in quality assurance to improve the efficiency of the calls and to check if the employees of the call center have performed to the highest standards and no inappropriate language were used and in Operational Intelligence to get the trend reports to find the types and quantities of calls occurring at a specific time. They can also be applied to areas like emergency call centers and defense to find out whether the call is a false call or if the caller is genuinely in trouble.

Two types of audio analytics can be performed including the speaker dependent and speaker independent speech recognition. In most researches it is found that the speaker independent analytics have a very low success rate. For speaker independent speech emotion recognition, Speaker Diarization can be used which clusters the speech segments based on the speaker [4]. [5] says that most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. In some cases, it is considered to be difficult even for humans to figure out the emotions from recorded speech. Other problems include low quality of the recorded speech, limited number of recordings, and the unavailability of phonetic transcriptions. In order to address this problem of low quality utterance and phonetic transcriptions, both audio and text analytics can be used in combination [6]. This can help increase the efficiency of the system. For text analytics, the speech audio can be converted to text using speech to text converters. The paper is organized as follows: section 3 contains the background and related works of the subject followed by section 4 containing the approach to the speech emotion recognition and different classification methods that are applied and some of the most suitable algorithms for the project. Section 5 mentions the results that are obtained.

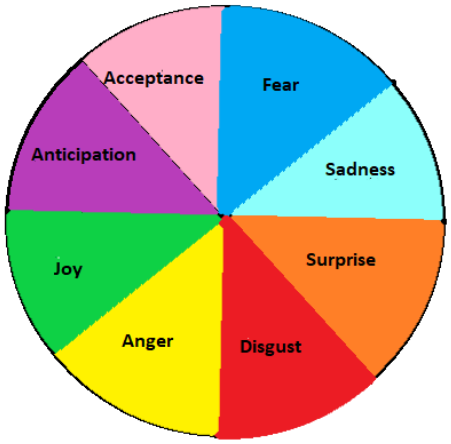

Fig. 1: Plutchik's wheel of Emotion [3]

II. Related work and background

The following section covers the different approaches for Emotion Recognition and their advantages, disadvantages and

usage. Similar research works where a combination system is used are discussed. Also contains the various feature extraction methods available for audio data and their applicability and suitability for different systems.

A. *Emotion Recognition*

Humans express their emotions through their speech, voice and facial expressions. There has been various study on emotion recognition in human beings [7]. Emotions can be detected from speech audio by using acoustic and phonetic features. This is widely called as Speech Emotion Recognition (SER). Facial Emotion Recognition and text sentiment analysis are two other types of emotion recognition where FER is an image classification problem. Emotions can be detected by subtle changes in facial expressions. FER is considered to be one of the most difficult tasks in image classification. Content/text based sentiment analysis is a common and most widely used approach in Emotion recognition. Here, speech audio is converted to text and emotion is classified and detected based on the words used [2].

B. *Emotion Recognition from Acoustic and Phonetic Features*

Emotion can be recognized from audio by using some important acoustic variables including pitch (fundamental frequency), vocal energy, frequency spectral features, formants and temporal features (speech rate and pausing) [11].

C. *Text Based Emotion Recognition*

Linguistic content based systems [2] convert vocal speech in to content/text using speech to text converters and classify the words using bag of words or ranking method by labelling each word/phrase with a particular emotion. This is widely used in social media sentiment analysis like twitter sentiment analysis, movie review analysis etc.,

D. *Combination of different approaches*

All the three above said approaches can be combined for a better emotion recognition system. Recent studies show that these hybrid systems or combinations are better at recognizing emotions than other systems [7]. The results from each recognizer can be combined and arrived at a decision. The methods can be combined by keeping one method as the base system and the recognized emotion from the other system can be used for supporting the decision. One interesting method is to combine audio based classification and content based classification.

Speech emotion recognition is useful for a range of applications. It can be used to identify whether a person is under depression or suicidal risk. It is possible to differentiate a depressed and suicidal speech from a normal speech from the acoustic features of an audio signal [8]. In [8], the speech acoustics of separate male and female samples comprising normal individuals and individuals carrying diagnoses of depression and high-risk, near-term suicidality were analyzed and compared. It had a good classification performance for majorly depressed person and a poor classification performance for assigning Dysthymic patients.

Speech analytics can also be applied to call centers to analyze large chunks of audio data to measure the performance of the call center agent [9]. In [9], the audio signals are converted to text using google Speech API and the resulting text data was analyzed to measure the performance of the call center agents. [8] uses only acoustic and Phonetic audio analytics and [9] uses only text analytics on text converted data.

The efficiency of both the systems can be improved by combining both the acoustic and phonetic features and linguistic features. Thus, taking aspects like sarcasm, poor quality audio signals in to consideration.

Linguistic content of the spoken utterance is an important part of the conveyed emotion [10] [5]. The acoustic information can be classified using k-means, SVM, GMM classifiers and then fused with linguistic classification. It is said that the classification accuracy score for acoustic features alone is around 74.2%and linguistic features alone is around 59.6 and when both the features were fused, the accuracy is close to 83.1% and when fused using a MLP neural network, the accuracy is around 92.0% [5]. As similar to combining linguistic features, a lot of other features like facial recognition, video recordings of the facial expressions etc., to increase the efficiency further.

E. *Feature Extraction for Audio Analytics*

Acoustic features are obtained from speech signals. The speech signal is divided in to small chunks called frames. Some basic computations are applied to the signal including Fast Fourier Transform (FFT), Logarithm, Discrete Cosine Transform (DCT).

Some important feature extraction methods include the traditional methods[12] - Relative spectra filtering of log domain coefficients (RASTA) , Linear Predictive Coding (LPC), Energy Normalization, Perceptual Linear Prediction (PLP), wavelet based features, Mel-Scale Cepstral Analysis (MEL)

1) ***Linear Prediction based Methods*** *:* Linear Prediction based methods [12] include Linear prediction coding(LPC), Linear Prediction Cepstral Coefficients(LPCC). In this approach, the speech signal is approximated as a linear combination of the past speech samples.

2) ***Cepstral Based feature extraction methods*** *:* Ceptral based feature extraction methods [12] include LPCC, PLP (Perceptual Linear prediction) which is a combination of linear prediction and DFT (Discrete Fourier Transform) techniques, Mel frequency Cepstral coefficient method.

3) ***Wavelet Based feature extraction methods*** *:* Wavelet based feature extraction method [12] is another method. It is classified n to two types - Continuous wavelet Transform (CWT) and Discrete Wavelet Transform (DWT)

MFCC is considered to be the best feature extraction method that is widely used for many speech recognition systems as it

is able to mimic the human auditory system and it gives better performance rate.

F. MFCC feature extraction method

MFCC was introduced in 1980 by Davis and Mermelstein. After this, a lot of new approach on the same have been developed. MFCC is considered to be the most effective feature extraction method and is robust under many conditions [13].

Mel frequency cepstral coefficients are real numbers which corresponds to features of an audio signal which can be used as a feature. These coefficients are obtained form cepstrum obtained from the audio signal and applying Discrete Cosine Transform to the spectrum. Cepstrum is the rate of change in spectral bands. Using cepstrum we get the domain named quefrency from which we can get the values between each interval of the signal.The quefrency is a measure of time, where the signal is not present in the time domain. It is a time like domain and not time domain.

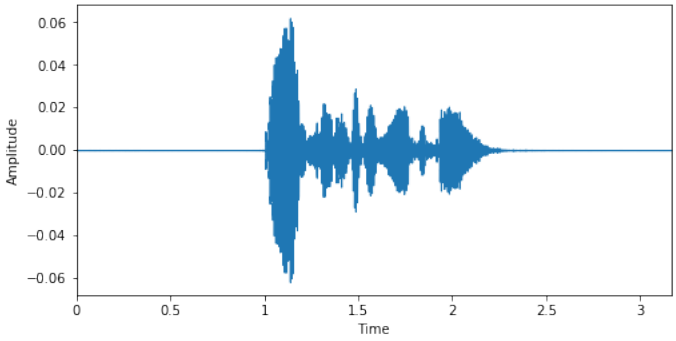

Fig. 2: Signal in Time Domain

The steps involved in MFCC feature extraction are as follows,

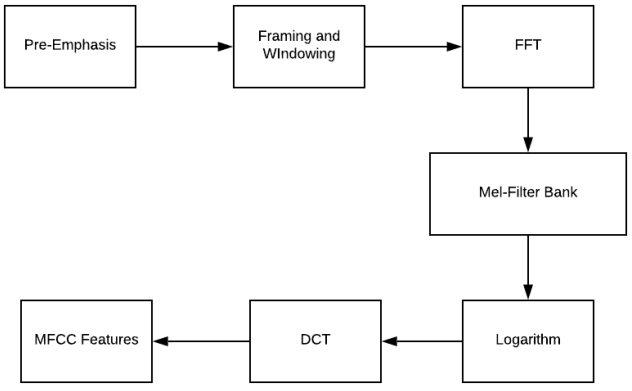

Fig. 3: MFCC Process

1) ***Pre-Emphasis****:* The noises are eliminated from the signal by passing it through High Pass and Low Pass filter

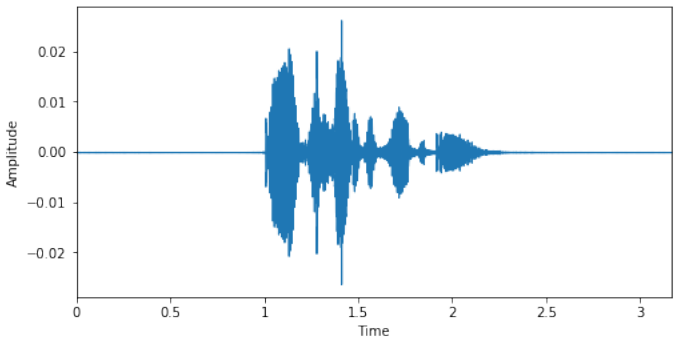

Fig. 4: Signal after Pre-emphasis

2) ***Frame Blocking****:* The signal is divided in to small chunks of data/frames

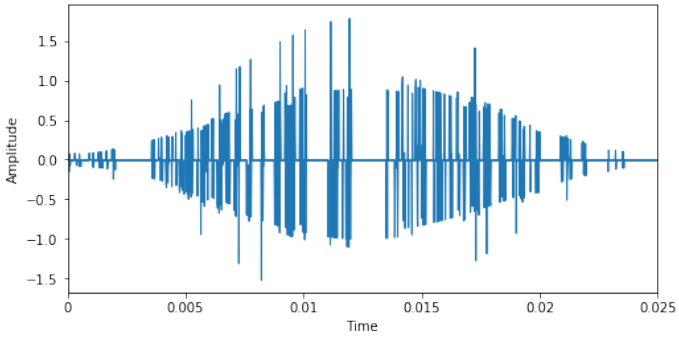

Fig. 5: Frame Blocking and Windowing

3) ***Windowing****:* We implement windowing to maintain the continuity between the first and the last point.

4) ***Applying Fast Fourier Transform****:* Fast Fourier Transform is applied to the windowed signal and the signal is converted to a frequency domain. This is called the Power Spectrum.

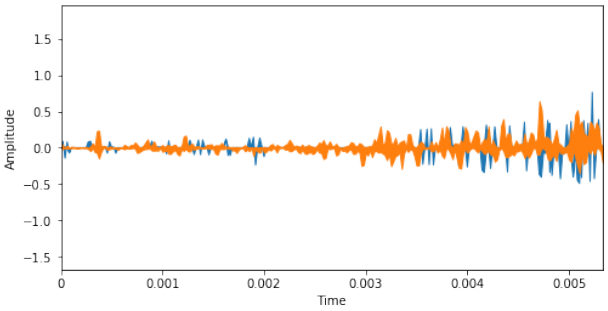

Fig. 6: Power Spectrum

5) ***Applying Triangular Band Pass Filter****:* Filter banks are computed by obtaining triangular filter to the power spectrum which obtains a periodogram.

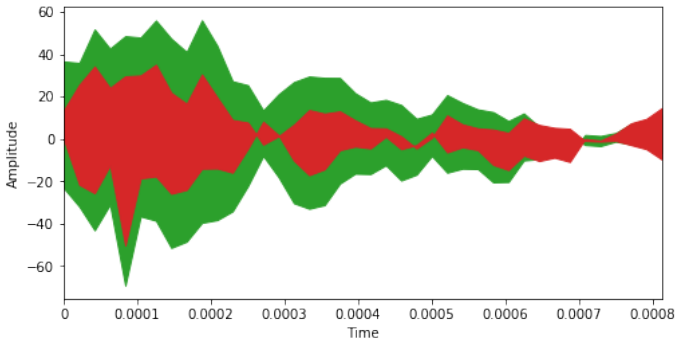

Fig. 7: Filter Banks

*6) **Discrete Cosine Transform**:* Discrete Cosine Transform is applied to the filter banks to obtain MFCC. These coefficients make up the mel-frequency cepstrum.

III. Method

A combination of both audio and speech text analytics was done by taking in to account the acoustic and phonetic features from audio and linguistic content from the speech converted text. The following section covers Data Collection, Data Preprocessing, Feature extraction, Feature selection and selection of classifier models for the method.

A. Data Collection

For audio analytics, RAVDESS [14] audio emotional dataset was used. The database contains 24 professional actors(12 female, 12 male), vocalizing two lexically matched statements. The speech includes calm, happy, sad, angry, fearful, surprise and disgust expressions. Each expression is produced at two levels of emotional intensity(normal, strong) with an additional neutral expression.The audio-only(16 bit, 48kHz .wav) file containing speech audio data with 60 trials per actor was used

B. Data Preprocessing and Feature extraction

There are various methods to extract features for audio analytics as discussed in the previous section. Out of which Mel-Frequency cepstral Coefficient method was used, as it is considered to be the most suitable method for emotion recognition. The data was loaded and each audio signal was applied with Pre-emphasis to remove any noise form the signal. The Pre-emphasized signal was then cut in to frames with a frame size of 0.025 and a frame stride of 0.01 such that the frames are overlapping and no data is lost.

Each windowed signal was applied with Fast Fourier transform to obtain power spectrum. Triangular Band pass filters were applied to compute the filter banks. Finally Discrete cosine Transform was applied to the Filter Banks to obtain Mel-Frequency Cepstral Coefficients. The file names of the files from RAVDESS consists of 7 part numerical identifier. This identifier contains the stimulus characteristics including the emotion, emotional intensity, actor, statement etc.,

C. Feature Selection

The first 12 coefficients from MFCC and the stimulus characteristics were stored as separate records for further processing. Each audio file has about 298 frames from 3 seconds audio file.

From the 8 emotions, 4 important emotions lying in each quadrant from the valence-arousal circumplex model[15] was selected. In circumplex model, valence dimension describes the degree to which an emotion is pleasant or unpleasant. Arousal dimension describes the degree to which an emotion is associated with high or low energy.

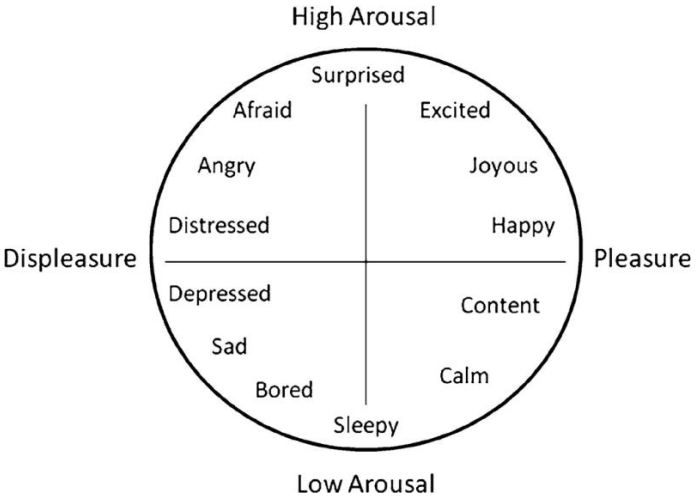

Fig. 8: Circumplex model of affect[15]

The 4 emotions include Neutral which belongs to pleasure and low arousal, happy belonging to pleasure and high arousal, Angry belonging to high arousal and displeasure and Sad belonging to low arousal and displeasure.

D. Classifier Models for Acoustic Features

The features extracted were then trained with classifier models. The audio based system performed classification using one of the multiple classifier algorithms that are available including K-nearest neighbors, Support vector machines(SVM), Artificial Neural Network(ANN) Random Forest classifier and GMM(Gaussian Mixture Model) were trained in both speaker dependent and speaker independent ways. Each classification model has its own merits and limitations.

*1) **Gaussian Mixture Model**:* GMM[16] are more appropriate for speech emotion recognition when only global features are to be extracted from the training utterances. GMM is considered to be the best classifier for speaker identification and verification

*2) **Artificial Neural Networks**:* Out of many types feed forward neural network used for speech emotion recognition, Multilayer perceptron layer neural networks are relatively common in speech emotion recognition as it is easy for implementation and it has well defined training algorithm [17].

*3) **k- Nearest Neighbor**:* k- Nearest Neighbor[18] classifies the emotions based on the classes using clusters.

*4) **Support Vector Machines**:* Support Vector Machines are also extensively used for Speech Emotion Recognition. SVM classifiers are mainly based on the use of kernel functions to non linearly map the original features to a high-dimensional space where data can be well classified using a linear classifier. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers.

E. Classifier Models for Linguistic features

The Linguistic content based system performs classification using the classifiers including Naive Bayes, Support vector machine and Maximum Entropy.

*1) **Support Vector Machines**:* Support Vector Machines [19] is considered to be one of the best classifier algorithm for text based classification. It is considered to give the highest prediction accuracy for such systems.

*2) **Naive Bayes Classifier**:* Naive Bayes is a fast learning algorithm. It has a straight forward approach. It uses Bayes theorem of theory of probability. It is suitable for text classification, sentiment analysis and spam filtering systems.

*3) **Maximum Entropy Classifier**:* Maximum Entropy classifier [20] as the name suggests selects the result which has maximum entropy. It is considered to be one of the best algorithm for text based classification

Any of the above classifiers can be used to classify the acoustic and the linguistic data and they can be ultimately fused together to provide better classifier accuracy.

For text-based sentiment analysis, a twitter emotion dataset with happy, anger,sad emotions was used. A bag of words approach was used for sentiment analysis.

IV. Results and Discussions

Multiple classification algorithms were trained for both audio classification and Text-based classification and the results were compared and the best of them was used for final combination approach.

In Audio based classification, k-NN classifier attained an accuracy of 44% for both speaker dependent and speaker independent classification.

Random Forest Classifier attained an accuracy of 55% for speaker dependent classification and an accuracy of 52% for speaker independent classification.

Gaussian Mixture Model attained an accuracy of 53% for speaker dependent classification and an accuracy of 49% for speaker independent classification

Artificial Neural Network attained an accuracy of 46% for speaker dependent classification and an accuracy of 41% for speaker independent classification.

Random Forest Classifier and Gaussian Mixture Model performed better for acoustic features and Support Vector Machine performed well for Linguistic content.

TABLE I: Speaker Independent Audio Classification

Classifier	Accuracy
k-nearest neighbour	0.44
Artificial Neural Network	0.41
Random Forest Classifier	0.52
Gaussian Mixture Model	0.495

TABLE II: Speaker Dependent Audio Classification

Classifier	Accuracy
k-nearest neighbour	0.44
Artificial Neural Network	0.46
Random Forest Classifier	0.55
Gaussian Mixture Model	0.53

In linguistic content based classification, Naive bayes gave an accuracy of 67% and Support Vector Machine gave an accuracy of 86% and Maximum entropy classifier gave an accuracy of 84%.

TABLE III: Linguistic Content Based Classification

Classifier	Accuracy
Naive Bayes Classifier	0.67
Support vector Machine	0.86
Maximum entropy Classifier	0.84

A. Combining the results

The acoustic and phonetic based classification was kept as the base classification method and tested with a new audio data-set. The same audio data-set was converted from speech to text using Sphinx audio converter and tested with text classifier. A threshold was set for the count of emotions predicted from each frame. The linguistic content based classification result was kept as a supporting predictor. A threshold was set for the count of emotions predicted from each frame. If the count is below the threshold, linguistic classification recognized emotion is used as the final prediction.

The combination system gave better prediction than when the audio classifier and text classifier was used separately. For text classifier, Support Vector Machine was used and combined with Gaussian Mixture Model and Random Forest Classifier separately. The combination system with Gaussian Mixture

TABLE IV: Combined classification

Audio Classifier	Text Classifier	Accuracy
Gaussian Mixture Model	Support Vector Machine	0.71
Random Forest Classifier	Support vector Machine	0.73

model gave an accuracy of 71% and and the system with Random Forest Classifier gave an accuracy of 73%.

V. Conclusions and Recommendations

Speech emotion recognition is quite challenging due to the shortage of available quality speech databases and poor recordings of audio signals. It is also challenging to identify the emotion when the speaker is undergoing physical or mental problems. In order to address all these shortcomings, it is suitable to combine both the acoustic features and linguistic features of the audio data. The above paper discusses the various approaches that can be used for classifying and segmenting the data and the best algorithms that can be used. It can be identified that the speaker dependent classification is more accurate than the speaker independent classification. For applications like call center speech analysis, Speaker dependent classification with a combination of acoustic, phonetic and linguistic classification is best suitable.

One point to note is that, the accuracy of the prediction goes down as we increase the number of classes or emotion as some emotions are similar and fall under the same arousal and valence dimension. MFCC feature extraction method is the best feature extraction technique and predicts well for audio based analytics.

As a future work, all the three ways of Emotion recognition including the Facial Emotion Recognition can be combined to detect emotions in human beings.

References

[1] V. Nanavare and S.K.Jagtap, "Recognition of human emotions from speech processing - sciencedirect." Procedia Computer Science,Volume 49(2015),Pages 24-32, Apr, 2015.

[2] S. Ezzat, N. E. Gayar, and M. M. Ghanem, "Emsys: An emotion monitoring system for call center agents." International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 4 pp. 619 -627, 2012.

[3] R. Plutchik and H. Kellerman., "Emotion: Theory, research and experience. volume 1. theories of emotion.." (Pp. 399, Volume 1, Academic Press: London, Psychological Medicine, Cambridge Core, 1980.

[4] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis,." PLOS ONE, vol 7, issue 1,, 2015.

[5] M. E. Ayadi, M. S.Kamel, and FakhriKarrayb, "Survey on speech emotion recognition: Features, classification schemes, and databases - sciencedirect." Pattern Recognition, Volume 44, Issue 3, 572-587, Mar, 2011.

[6] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture - ieee conference publication." IEEE International Conference on Acoustics, Speech, and Signal Processing.

[7] Q. Yao, "Multi-sensory emotion recognition with speech and facial expression." https://aquila.usm.edu, Dissertations,710, 2016.

[8] D. France, R. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk - ieee journals & magazine." IEEE Trans on Biomedical engineering, Vol 47, issue 7, July, 2000.

[9] B. KARAKUS and G. Aydin, "Call center performance evaluation using big data analytics." DOI: 10.1109/ISNCC.2016.7746116 Conference: 2016 International Symposium on Networks, Computers and Communications (ISNCC).

[10] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogs." Proceedings of the International Conference on Multimedia and expo, 549-552, 2013.

[11] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods - sciencedirect." Speech Communication, Volume 48, Issue 9, Sep, 2006.

[12] U. Sharma, S. Maheshkar, and A. N. Mishra, "Study of robust feature extraction techniques for speech recognition system - ieee conference publication." In: International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015.

[13] S. Gupta, J. Jaafar, W. Fatimah, wan Ahmad, and A. Bansal, "Feature extraction using mfcc." Signal Image Processing : An International Journal (SIPIJ) Vol.4, No.4,, Aug, 2003.

[14] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english." PLOS ONE, May, 2018.

[15] A. Tseng, R. Bansal, and et al, "Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders." J Autism Dev Disord 2014 Jun;44(6):1332-1346.

[16] D. A.Reynolds, "Speaker identification and verification using gaussian mixture speaker models - sciencedirect." Speech Communication,Volume 17, Issues 12, Pages 91-108, Aug 1995.

[17] R. P. Lippmann, "Review of neural networks for speech recognition." Neural Computation, MIT Press Journals, 1989.

[18] T. LakshmiPriya, N.R.Raajan, N. P.Preethi, and S.Mathini, "Speech and non-speech identification and classification using knn algorithm - sciencedirect." Procedia Engineering,Volume 38, Pages 952-958, 2012.

[19] M. Ahmad, S. Aftab, M. Salman, and N. Hameed, "Sentiment analysis using svm: A systematic literature review." International Journal of Advanced Computer Science and Applications, Jan 2018.

[20] H. Htet and Y. Y. Myint, "Social media (twitter) data analysis using maximum entropy classifier on big data processing framework (case study: Analysis of health condition, education status, states of business." Journal of Pharmacognosy and Phytochemistry; SP1: 695-700), 2018.