International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Speech based Emotion Recognition using Machine Learning

Ajay Gupta[1], Siddhesh Morye[2], Mukul Sitap[3], Supriya Chaudhary[4]

[1-3]*Student, Dept. of Information Technology, Vasantdada Patil Pratishthan's College Of Engineering, Mumbai, Maharashtra, India*
[4]*Professor, Dept. of Information Technology, Vasantdada Patil Pratishthan's College Of Engineering, Mumbai, Maharashtra, India*

Abstract – *The human voice can be characterized by several attributes such as pitch, timbre, loudness, and vocal tone. It has often been observed that humans express their emotions by varying different vocal attributes during speech generation. This paper presents an algorithmic approach for detection of human emotions with the help speech .The prime objective of this paper is to recognize emotions in speech and classify them in 6 emotion output classes namely angry, fear, disgust, happy, sad and neutral. The proposed approach is based upon the Mel Frequency Cepstral coefficients (MFCC) uses Crema-D database of emotional speech. Data Augmention is perfomed on input data audio file, such as Noise, High Speed, Low Speed etc. are added, thus more the varied data is available to the model better the model understands. Feature extraction is done using MFCC and then the extracted features are Normalized(for Independent Variable), Label Encoding(for Dependent Variable(for SVM,RF)),One Hot Encoding(for Dependent Variable(for CNN)) is done. After this the dataset is divided into Train, Test and given to different models such as Convolutional Neural Network(CNN),Support Vector Machine(SVM), Random Forest(RF) for Emotion prediction. We report accuracy, f-score, precision and recall for the different experiment settings we evaluated our models in. Convolutional Neural Network(CNN) was found to have the highest accuracy and predicted correct emotion 88.21% of the time.*

Hence, deduction of human emotions through speech analysis has a practical plausibility and could potentially be beneficial for improving human conversational and persuasion skills.

Key Words **: Speech Emotion Recognition, Support Vector Machine, Random Forest, *Convolutional Neural Network, Mel Frequency Cepstral coefficients.***

1. INTRODUCTION

The human voice is very versatile and carries a multitude of emotions. Emotion in speech carries extra insight about human actions. Human speech conveys information and context through speech, tone, pitch and many such characteristics of the human vocal system. As human-machine interactions evolve, there is a need to buttress the outcomes of such interactions by equipping the computer and machine interfaces with the ability to recognize the emotion of the speaker. Emotions play a vital role in human communication. In order to extend its role towards the human-machine interaction, it is desirable for the computers to have some built-in abilities for recognizing the different emotional states of the user [2,5].Today, a large amount of resources and efforts are being put into the development of artificial intelligence, and smart machines, all for the primary purpose of simplifying human life. Research studies have provided evidence that human emotions influence the decision making process to a certain extent [1-4]. If the machine is able to recognize the underlying emotion in human speech, it will result in both constructive response and communication.

In order to communicate effectively with people, the systems need to understand the emotions in speech. Therefore, there is a need to develop machines that can recognize the paralinguistic information like emotion to have effective clear communication like humans. One important data in paralinguistic information is Emotion, which is carried along with speech. A lot of machine learning algorithms have been developed and tested in order to classify these emotions carried by speech. The aim to develop machines to interpret paralinguistic data, like emotion, helps in human-machine interaction and it helps to make the interaction clearer and natural. In this study different classification models such as CNN,SVM,RF are used to predict in speech sample. The MFCC is used for the feature extraction .To train the model CREMA – D dataset was used along with Data Augmention.

In the next section 2, the previous related work on speech emotion recognition systems is explained. Subsequent, Section 3 provides an insight on the database that is used for implementing the system. Section 4 provides the methodology along with the approach used for feature extraction, classification algorithms used & the data flow diagram. In section 5, the experimental results are discussed along with emotion accuracy table & confusion matrix for the classification algorithms used. Section 6 describes the the various uses & applications of speech emotion recognition system. Section 7 concludes the paper along with the future scope.

2. RELATED WORK

Recognition of emotions in audio signals has been a field of study in the past. Previous work in this area included use of various classifiers like SVM, Neural Networks, Bayes Classifier etc. The number of emotions classified varied from study to study, they play an important aspect in evaluating the accuracy of the different classifiers. Reduction in the number of emotions used for recognition has generated more accurate results as depicted below.

The following table summarises the previous study done on the topic.

Study	Algorithm Used	# of Emotion s	Accuracy (%)
[Kamran Soltani, Raja Noor Ainon,2007] [1]	Two layer Neural Network	6	77.1
[Li Wern Chew, Kah Phooi Seng, Li-Minn Ang, Vish Ramakonar, 2011] [2]	PCA, LDA and RBF	6 (divided into three independent classes)	81.67
[Taner Danisman, Adil Alpkocak, 2008] [3]	SVM	4/5	77.5/66.8
[Lugger and Yang, 2007] [4]	Bayes Classifier	6	74.4
[Yixiong Pan, Peipei Shen, Liping Shen, 2012] [5]	LSTM	3	95.1

Complete review on the speech emotion recognition is explained in [6] which reviews properties of dataset, speech emotion recognition study classifier choice. Various acoustic features of speech are investigated and some of the classifier methods are analyzed in [7] which is helpful in the further investigation of modern methods of emotion recognition. This paper [8] investigated the prediction of the next reactions from emotional vocal signals based on the recognition of emotions, using different categories of classifiers. Some of the classification algorithms like K-NN, Random Forest are used in [8] to classify emotion accordingly. Recurrent Neural network arises enormously which tries to solve many problems in the filed of data science. Deep RNN like LSTM, Bi-directional LSTM trained for acoustic features are used in [9]. Various range of CNN are being implemented and trained for speech emotion recognition are evaluated in [10]. Emotion is inferred from speech signals using filter banks and Deep CNN[11] which shows high accuracy rate which gives an inference that deep learning can also be used for emotion detection. Speech emotion recognition can be also performed using image spectrograms with deep convolutional networks which is implemented in [12].

3. DATA SET

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

This dataset is the sheer variety of data which helps train a model that can be generalised across new datasets. Many audio datasets use a limited number of speakers which leads to a lot of information leakage. CREMA-D has many speakers. For this fact, the CREMA-D is a very good dataset to use to ensure the model does not overfit.

4. METHODOLGY

4.1 Mfcc Feature Extraction

MFCC represents parts of the human speech production and perception. MFCC depicts the logarithmic perception of loudness and pitch of human auditory system.

MFCC are cepstral coefficients derived on a twisted frequency scale centerd on human auditory perception. In the computation of MFCC, the first thing is windowing the speech signal to split the speech signal into frames. Since the high frequency formants process reduced amplitude compared to the low frequency formants, high frequencies are emphasized to obtain similar amplitude for all the formants. After windowing, Fast Fourier Transform (FFT) is applied to find the power spectrum of each frame. Subsequently, the filter bank processing is carried out on the power spectrum, using mel-scale. The DCT is applied to the speech signal after translating the power spectrum to log domain in order to calculate MFCC coefficients [13].
The formula used to calculate the mels for any frequency is :

$$mel(f)=2595xlog10(1+f/700)$$

where mel(f) is the frequency (mels) and f is the frequency (Hz).

The MFCCs are calculated using this equation :

$$\hat{C}\ n=\sum_{k=1}^{n}(\log \hat{S}\ k)\cos[n(k-12)\pi k]$$

where k is the number of mel cepstrum coefficients, $\hat{S}\ k\hat{S}k$ is the output of filterbank and $\hat{C}\ n\hat{C}n$ is the final mfcc coefficients.

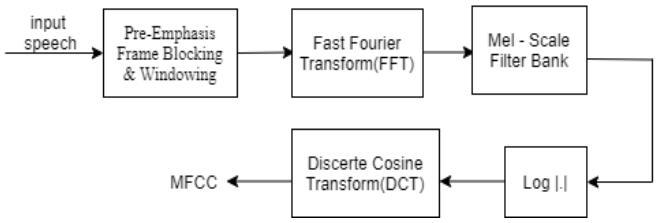


Fig1. MFCC Feature Extraction

The block diagram of the MFCC processor can be seen in Figure 1. It summarizes all the processes and steps taken to obtain the needed coefficients. MFCC can effectively denote the low frequency region better than the high frequency region, henceforth, it can compute formants that are in the low frequency range and describe the vocal tract resonances. It has been generally recognized as a front-end procedure for typical Speaker Identification applications, as it has reduced vulnerability to noise disturbance, with minute session inconsistency and easy to mine. Also, it is a perfect representation for sounds when the source characteristics are stable and consistent (music and speech).

4.2 Classification Algorithms

1) SVM (Support Vector Machine)

SVM is a supervised learning algorithm used most widely for pattern recognition applications. The algorithm is simple to use and provides good results even when trained on limited size training dataset. More formally, it is an algorithm which constructs, in a high dimensional or infinite dimensional space, a hyperplane or set of hyperplanes which can be used for regression and classification tasks. It tries to learn a hyperplane that results in maximum separation between the data points belonging to different classes, leading to a better classifier. The data can be linearly separable or non-linearly separable. Non linearly separable data is classified by mapping it's feature space to a high dimensional space using a kernel function.

The data points are then linearly separable in this high dimensional space. The optimization problem in SVM reduces to

$$a = \min(\frac{||w||^2}{2}) \text{ subject to } \forall k,\ y_k(< w.x_k > +b) \geq 1$$

Where w is normal vector to the hyperplane and <w.xk> denote the inner product of w and x k for each data point (x k, y k) in the training set. Some of the widely used kernels are linear kernel and
radial bias kernel.
Linear kernel function is given as:

kernel(x, y) = (x, y)

Radial bias kernel is given as:

$$kernel(x, y) = e^{\frac{-||x-y||^2}{(2\sigma^2)}}$$

In this paper, SVM is implemented using random bias kernel.

2) RF (Random Forest)

Random Forests uses an ensemble of learning methods and is used for regression, classification and other tasks. Random Forests works by constructing a large number of decision trees at the time of training and it outputs the mean prediction or mode of the class of the individual trees. Random decision forests prevents decision trees from overfitting the training data.

Random Forests algorithm works as follows:

1. Random Record Selection: Each decision tree is trained on approximately 2/3rd of the training data.
2. Random Variable Selection: Only some of the variables used for prediction are chosen. This selection is done randomly and node is split according to the most optimal split of the node.
3. Trees grow to maximum extent without any pruning.

Random Forests algorithm is based on bagging. In bagging, a random sample is selected with replacements repeatedly from the training set and trees are then fit to these samples: By averaging the prediction values or by taking the majority vote from all the decision trees we predict the values of the unseen sample. In this paper, the random forests classifier was implemented using 100 forests.

3) CNN

Convolution layer
A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i.e., convolution operation and activation function.

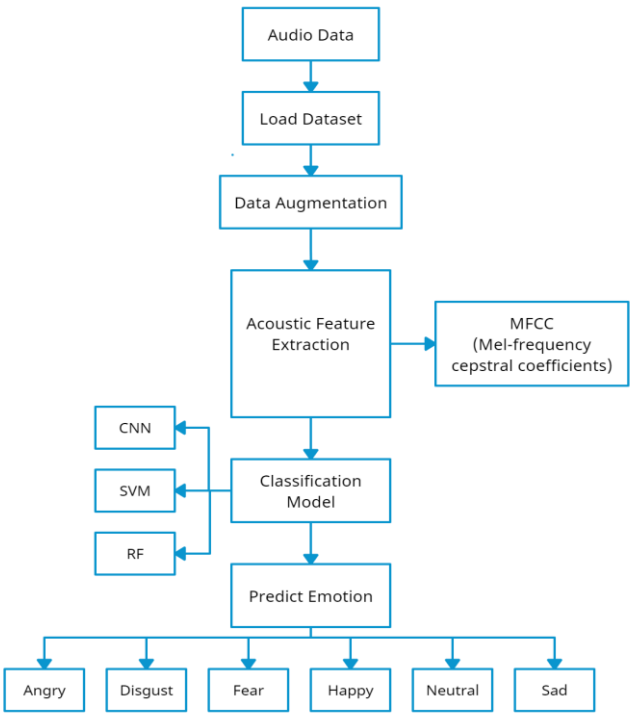
Nonlinear activation function
The outputs of a linear operation such as convolution are then passed through a nonlinear activation function.The most common nonlinear activation function used presently is the rectified linear unit (ReLU).

Pooling layer
A pooling layer provides a typical down sampling operation which reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions, and decrease the number of subsequent learnable parameters.

Fully connected layer
The output feature maps of the final convolution or pooling layer is typically flattened, i.e., transformed into a one-dimensional (1D) array of numbers (or vector), and connected to one or more fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and down sampled by the pooling layers are created they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes.

Last layer activation function
The activation function applied to the last fully connected layer is usually different from the others. An activation function applied to the multiclass classification task is a softmax function which normalizes output real values from the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1.

4.3 Data Flow Diagram



In the first step the Data Set is loaded, from which one audio file is taken .Then that audio file is passed through Data Augmentation, where one without Augmentation, Noise ,High Speed, Low Speed, Stretch, Pitch are created .Then the Feature extraction is done from each data augmented audio file through MFCC.

This same process is being performed for all the audio files present in the dataset. Finally after the feature extraction from all the file is done . We randomly split our dataset into a train (80%) and test (20%) set. The same split is used for all the experiments to ensure a fair comparison. Here for Dependent variable (Label encoding (SVM,RF)) ,(One Hot Encoding (CNN)).

Classification Models are trained and their Evaluation metrics are been noted.

5. EXPERIMENTAL RESULTS

Three classification algorithms, namely Random Decision Forest, SVM and CNN classified an audio file into one of the 6 classes. Out of the three, CNN achieved the highest accuracy of 88.21%. Considering that many past papers achieved similar or less accuracy when trained on less than 6 emotional output classes, we consider our accuracy as a significant improvement. For all the three algorithms, we achieved highest classification accuracy for samples belonging to angry class while least accuracy was achieved for those belonging to happiness class for CNN, fear class for SVM & RF. The results for each algorithm is summarized in the following tables and graphs.

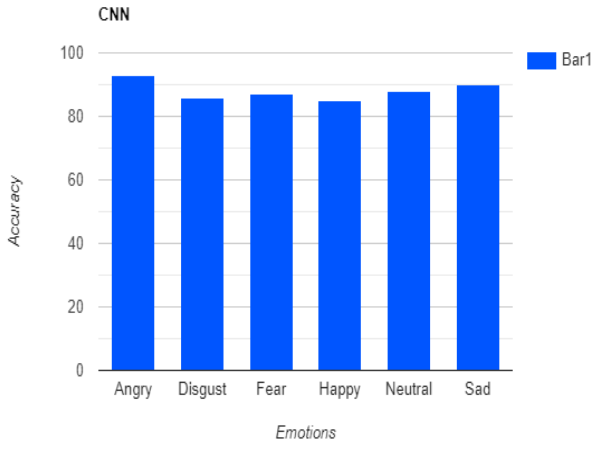


Fig.2 Emotion Accuracy Table for CNN

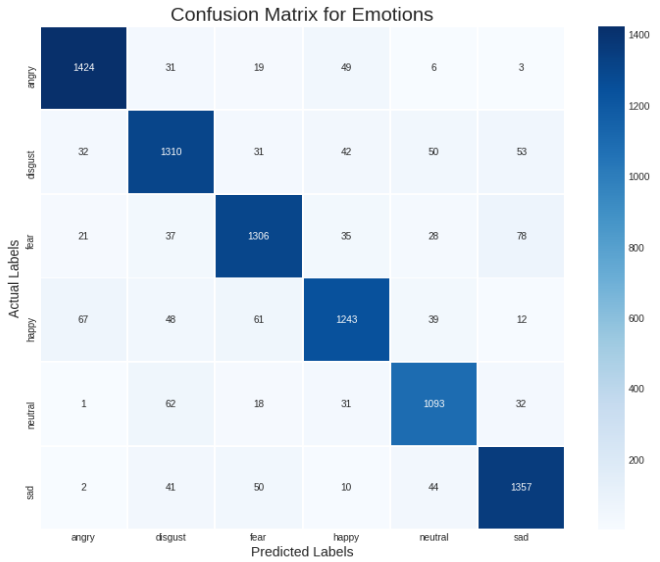


Fig.3 Confusion Matrix For CNN

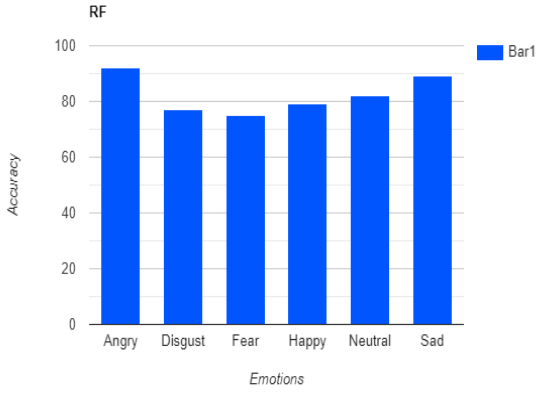


Fig.4 Emotion Accuracy Table for RF

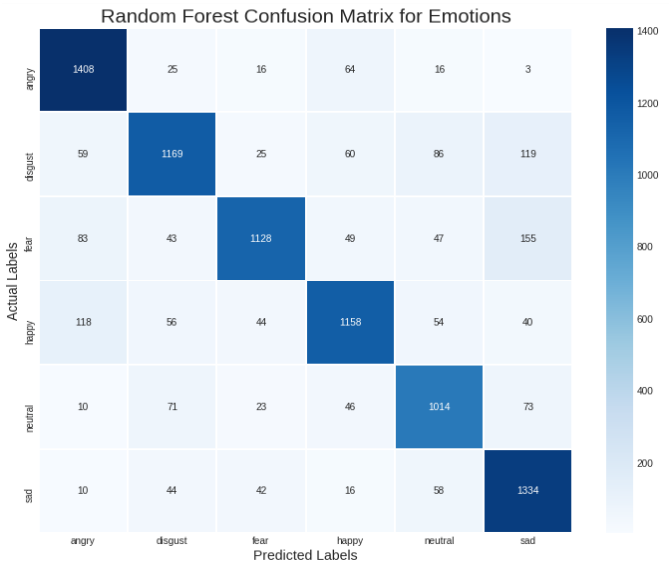


Fig.5 Confusion Matrix For RF

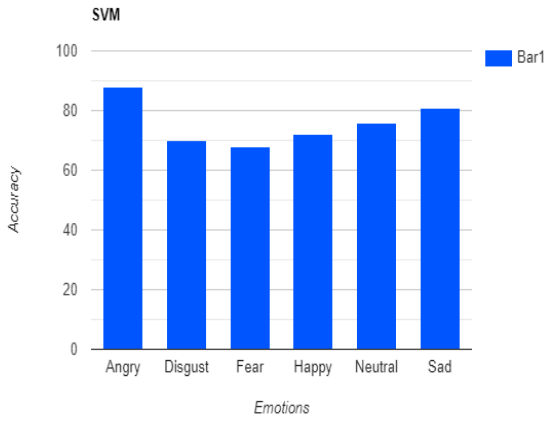


Fig.6 Emotion Accuracy Table for SVM

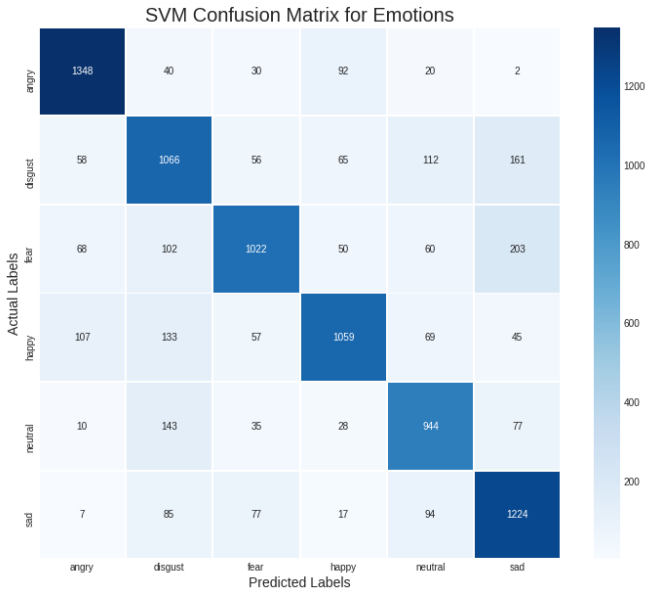


Fig.7 Confusion Matrix For SVM

6. USES & APPLICATIONS

Emotion Recognition is used in call center for classifying calls according to emotions[14]. Emotion Recognition serves as the performance parameter for conversational analysis[15] thus identifying the unsatisfied customer, customer satisfaction so on.

Such techniques could also be applied in observing companies' interactions with customers through call centres. Currently to analyse emotions in such conversations, a human specialist with limited capacities has to be included. But if one employs machines to do the task, then it will be much cheaper and output will be more consistent.

Public services could also benefit from such an approach. Namely, it is possible to analyse emotions in the voices, i.e. speeches, of parliament members. Such information could be

of high interest and value for the society, as attitudes and honesty of politicians could be investigated.

Also, emotion recognition could be used in different NGOs dealing with civil society issues. For example, by using emotion recognition in speech it is possible to track emotional states and behaviour of different social groups. Also, in the academia such techniques could be used in order to achieve higher granularity, especially in social science research. For example, one could detect emotions in speech while conducting interviews.

SER is used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen[16]. This can also help the salesperson or the entrepreneur of a particular commodity whether the customer of him is satisfied with its service or not. Thus, this process will provide a lively and most usefulness to the environment.

7. CONCLUSION & FUTURE SCOPE

In this paper, an approach to emotion recognition in speech based upon Random Decision Forest, SVM and CNN classifiers was presented. CREMA – D dataset was used and a total of six in 6 emotion output classes namely angry, fear, disgust, happy, sad and neutral. Feature extraction was done using MFCC, total of 58 features were extracted for emotion prediction. Data Augmention was performed on the audio samples to increase the variety of dataset so that the model could perform better. For SVM classification model RBF kernel was used .For RF classification model 100 n-estimators was used. For CNN we tried multiple epochs, and the loss function used is categorical cross entropy, the optimizers used was adam. For all the models accuracy, precision, recall and F1 score was noted.

Hence after numerous trails and tests it was found that CNN performed the best among the other models with an accuracy of 88.21%.

The paper presents only the prediction of six human emotions using speech. It can be expanded to predict more human emotions. The CNN classification algorithms wrongly predicted some of the samples belonging to happy class ,and in SVM,RF belonging to fear class. This can be rectified by extracting more features to better distinguish between these two class.The plan to further make the Speech Emotion Recogniton system more roboust & real time analysis would be done.

	SVM	RF	CNN
Accuracy Score	0.760096	0.822610	0.882158
Precision Score	0.763155	0.825476	0.882146
Recall Score	0.760096	0.822610	0.882158
F1 Score	0.759643	0.821769	0.882051

Fig.8 Accuracy Table for Classification Algorithm

8. REFERENCES

[1] K.V .Krishna Kishore, P.Krishna Satish, "Emotion Recognition in Speech Using MFCC and Wavelet Features", 3rd IEEE International Advance Computing Conference (IACC) , 2013.

[2] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine ", International Journal of Smart Home, 2012

[3] Ashish B. Ingale and Dr.D.S.Chaudhari,, "Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine", International Journal of Advanced Engineering Research and Studies,Vol. 1,Issue 3 , 2012.

[4] Davood Gharavian, Mansour Sheikhan, Alireza Nazerieh, Sahar Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network", NeuralComputing and Applications , Volume 21, Issue 8 , pp 2115–2126, 2011.

[5] Li Wern Chew, Kah Phooi Seng, Li-Minn Ang, Vish Ramakonar, Amalan Gnanasegaran, "Audio-Emotion Recognition System using Parallel Classifiers and Audio Feature Analyzer", Third International Conference on Computational Intelligence, Modelling & Simulation, 2011.

[6] S. G. Koolagudi and S. R. Krothapalli, "Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features," Int. J. Speech Technol., vol. 15, no. 4, pp. 495–511, 2012.

[7] J. Rong, G. Li, and Y. P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009.

[8] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017.

[9] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[10] C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, 2017.

[11] 14. H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[12] 15. A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.

[13] Some Commonly Used Speech Feature Extraction Algorithms By Sabur Ajibola Alim and Nahrul Khair Alang Rashid Submitted: October 4th 2017.

[14] F. Dipl and T. Vogt, "Real-time automatic emotion recognition from speech," 2010.

[15] 7. S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," 2016 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2016 - Proc., no. November 2017, pp. 1278–1283, 2016.

[16] 8. 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," Acoust. Speech, Signal Process., vol. 1, pp. 577–580, 2004.