Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals

Deepak Bharti (Author)
M.Tech (Scholar)
Department of CSE
Rayat Bahra University
Kharar, Punjab, India
Deepakbhartiofficial@outlook.com

Poonam Kukana (Co-Author)
Assistant Professor
Department of CSE
Rayat Bahra University
Kharar, Punjab, India
poonam.sihag@gmail.com

Abstract- **The** ***emotion recognition is the addressed communication that has been gaining more interest through the public. Speech emotion recognition has become a region of interest in the field of Human-Computer Interaction(HCI). Speech recognition (SR)*** **is the technology that is connected** ***with the methods and fields to identify the speech from the speech signals. Different technological; developments in the area of the SPMs (Signal Processing Methods), the recognition of the expression has become probable. Though there has been a massive growth in the field of voice recognition, there are various voice yields that have been implemented like Amazon Alex, Google, Home, and Apple Homepage that purposes basically on voice-based commands. SER (Speech emotion recognition) is a research area issue that tried to gather sentiments from the speech signals. Different surveys stated that the development in sentiment detection made a lot of the networks simpler and the world an appropriate location for living. Emotion recognition is a challenging issue so that the emotion may vary that depends on the situation, culture, person face -response that leads to ambiguous results; speech quantity is not adequate to precisely infer the emotion; lack of speech database in many languages. Moreover, SER has been used in various applications like interaction with robots, bank services, digital games, and so forth.In existing research, different speech emotions like Happy, Anger, and Sad, and were detected or recognized through the feature vectors. The various feature sets used were removed from the acoustic signals named such as Voice Pitch, MFCC (Mel Frequency cepstral coefficients), and STM (Short Term Energy). The various techniques have developed on the feature sets, and the influence of the increasing amount of the feature sets provide for the classifier. It presents the observation of the performance classification for India, Hindi, and Marathi speech. Moreover, the accuracy of the music or normal vocal-speech was 80%. In research work has designed an SER (Speech Emotion Recognition) model depends on the GFCC algorithm to citation the feature sets based on the DCT and High pass Filter method. After that, the ALO algorithm is using to select the instances with the help of coverage and Fitness function. The novel MSVM algorithm is using to classify the emotion-based on the feature set and evaluate the performance metric such as accuracy rate etc. In proposed work using the MATLAB simulation tool and evaluates the maximum accuracy rate and mitigate the error rates as compared with the existing parameters.***

Keywords:- Speech Emotional Recognition, GFCC, MSVM (Multi-Support Vector Machine, ALO (Ant Lion Optimization) and DCT (Discrete Cosine Transformation).

I. INTRODUCTION

Speech is conveying the data and context through pitch, emotions, speech and various features of the Human Vocal System (HVS). As the Human-Machine Interactions (HMI) progress, there is an essential to support the results of like communications by preparing the Machine Interfaces (MI) and Computer with the capacity to detect the Speech Emotion (SE) [1]. Nowadays, the huge quantity of efforts and resources are present in the growth of Artificial Intelligence (AI) and Smart Machines(SM) all for the main motive of shortening human life [2].

It is mainly a complicated speech signal that comprises the data nearby the message, speaker, linguistic and feelings. It is generated from the time changing vocal-tract scheme motivated by the time changing excitation resource [3]. On the other side, a person's mental stage takes place spontaneously instead of sensible effort. Several types of feelings are presented in the signal. The major challenge is to protect the hole among the data that is arrested by the microphone and related sentiment and to perfect the specified relationship The gap can be associated by contracting down the different sentiments in limited such as Anger, Gladness, Sorrow [4].

SER aimed to detect spontaneously the emotion stage of the social being from the voice of the male or female [3]. It is dependent on the in complexity analysis of the creation method of theSpeech Signal(SS), eliminating certain types that consist of expressive data from the voice of the utterer and receiving suitable pattern recognition techniques to detect the emotional stage. The speech recognition scheme of elements determined in Fig.1. The Speech Emotion Recognition (SER) scheme consists of four major phases which are described as Input sample, Valuable features, Detection/Recognition, and Emotion outcome.

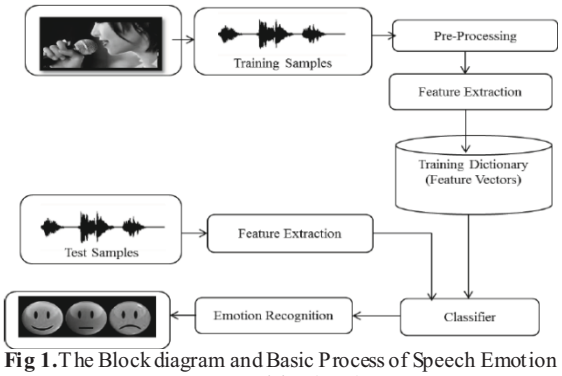

Fig 1. The Block diagram and Basic Process of Speech Emotion Recognition(SER)

However, a social being may not classify simply the usual emotions, it is not easy that the machines may offer an exact classification [5]. A typical group of emotions consists of anger, sadness, contentment, amazement and fear. Generally, the SER depends on the sincerity of the dataset. Various datasets are available namely: the Danish Emotional Speech corpus (DES) and EMO-DB, and four datasets from the Border scheme with French, Spanish, and English(US) emotional speech. These datasets worked as an emotional speech. There is an important proposed work on enhancing the recognition rate of the MSVM classification outcomes by developing various classification algorithms [6]. This research article mainly focuses on the feature sets and instance selection sets in instruction to enhance the recognition rate of the MSVM model. The research methodology can be developed in a previous speech emotion recognition system.

Fig 2 defines the complete SER system. The proposed system is separated into the information of the data set, signal pre-processing step, feature extraction in a signal, feature selection and classification (MSVM) in SER. The complete system is automated using the simulation tool MATLAB. The RAVDESS data set [7] which contained both gender signal samples was re-cycled to analyze examples was used to test analysis for three sentiments such as Anger, Happy, and Sad.

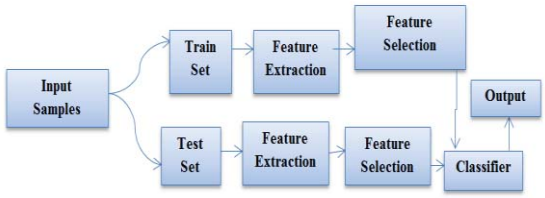

Fig 2. The Systematic Flow Chart

The research paper is arranged as follows, Section I gives a basic introduction of the speech ERS. Section II and explains the literature review of the emotion recognition system and Section III determines the research methodology advanced in feature extraction, selection and classifier training is elaborated. Section IV establishes the analysis of the simulation results.Finally, Section V explains the conclusion and further analysis based on the research.

II. RELATED WORK

It describes the various speech emotion recognition and detection methods. The classification method has been used to detect several emotions in the speech signals. The MFCC and GFCC utilizes the Feature Extraction(FE) methods used to excerpt the valuable feature set in the uploaded SS (Speech Signals). **Deshmukh, G et al., 2019 [8]** proposed research on the various opinions using the different feature sets like voice Pitch, MFCC (Mel Frequency cepstral co-efficient), STE (short term energy). These three features were extracted from the audio signals. The dataset is used for the preparation and testing of the acoustic data samples in girl's and boy's speech and proportion rate up to 4:1. Generally, the mean technique presented the maximum accuracy above the mode technique. Moreover, the mean of the group data values are taken as the consideration each rate presented more than the two rates achieves the maximum accuracy. The classified accuracy for the complete three opinions was searched that have increased up to 20% using the three features opposite to two features. Moreover, the annoyance and joy expressions classification accuracy rate was increased-up to fifteen to twenty percent using Short Term Energy feature vector. The opinion blues may not enhance the accuracy rate of the classification instead of the feature vectors. Blues is vulnerable to misclassification as happiness opinions. The happiness is not classified as angry because of the nearest values of the feature vector and sad, due to the low speech tone. Classification of the opinions for the region language named as Sanskrit and Marathi was developed. The classification recognition rate for the actual time audio signal regional language Hindi was achieved up to 100%.

Jiang, W et al., 2008 [9] developed a Deep Neural (DN) structure to eliminate the data characteristic demonstrations from the different feature-sets that consist of terminated and unrelated data leads to low data recognition performance. After achieving the data features, the fusion system is trained to learn the demonstration of different features.Besides, a support vector machine was utilized as the last classifier for the recognition purpose. Experimental analysis was done on the IEMOCAP database represented that the planned structure enhances the recognition performance rate to achieve accuracy up to 64% compared to current state of art methods. **Khalil, R. A., et al., 2019[10]** provided research on the detailed study of the deep learning methods for Speech Emotional Recognition (SER). Deep learning (DL) methods like as Recurrent Neural Network, and Convolutional Neural Network have been the topic for more investigation in current years. Moreover,

DL techniques and the layer-wise structure are comprehensively described that depends on the classification of different natural opinion like as happy , sad, neutral, fear. These techniques offered a simple training model and also efficiently shared the weighs. Moreover, the restrictions of the deep learning methods include the high layer-wise structure , minimum efficiency for temporarily changing the input rate, and maximum learning at the time of the memorization of layer-wise data. The research methods create the base to compute the performance and restrictions of the present deep learning methods. **Sonmez, Y. Ü. et al., 2019 [11]**assumed the sub-space discriminant analysis and LTP, LBP, and NCA feature elimination on spectrogram using EMO-DB information , a success value up to 88% have been acquired. The success value was initially improved by incorporating other datasets. The novel methods were combined with the new methods that were utilized in both feature extraction and classification purpose to improve the accuracy rate. **Le, B. V. et al.,2014[12]** introduced classified methods that create the binary classification tree in an automated way and exploit numerous classifiers to identify different opinions. They proposed an outline that identifies the sentiments from Speech Signal (SS) with maximum accuracy rate and effectively compared to the approaches like Hidden Markov Model (HMM) and Support Vector Machine (SVM). This technique creates the binary (0,1) classification and tree decision automatically and optimized the classifier at every node of the tree so that the appreciation outcome may be attained with maximum accuracy rate and SNR rate. The detection stage is easy to develop on various mobile frameworks with minimum computation efforts as compared to other methods. Table I shows the comparative analysis with several methods such as MFCC, SVM, CNN, DBN, etc, the existing performance metrics and data set in speech emotion recognition system.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING TECHNIQUES

Author Name	Year	Methods	Metrics	Data set
Deshmukh et al., [8]	2019	MFCC SVM	Accuracy Rate (%) = 96%	Angry Happy Sad
Wenbing et al., [9]	2008	PoS Tagging Perceptron Model	Fmeasure (%)= 64%	MSR : Microsoft Research Corpus
Ruhul et al.,[10]	2019	DNN RNN DBN CNN DBM	-	-
Yesim et al., [11]	2019	MFCC FFT LPC PLP	Accuracy (%) rate = 88%	EMO-DB
Ba et al., [12]	2014	HMM SVM	TP,TN, FN, TP ,	EmoDataset
			Accuracy rate (%)	

Abbrevations:

MFCC: Mel Frequency Cepstral Co-efficient
SVM : Support Vector Machine
DNN: Deep Neural Network
RNN: Recurrent Neural Network
DFN: Deep Brief Network
FFT: Fast Fourier Transformation
LPC: Linear predicting Crop
PLP: Perceptual Linear Predictive
CNN: Convolutional Neural Network
DBM: Deep Boaltzman Machine
HMM: Hidden Markov Model
Tp: True Positive
Tn: True Negative
Fp: False Positive
Fn: False Negative

III. RESEARCH WORK

The proposed research work has implemented the GFCC, ALO with the MSVM classification model. It improves the SNR rate and accuracy rate. This method has increased the quality of the speech signal and reduce the noise level in the uploaded samples. The proposal model steps for emotion recognition procedure are elaborated as follows:

Step I: In the initialization phase, the load of the speech signal and input the speech signal which has to be checked. The uploaded input speech and the data sample speech should be time duration and frequency.

Step II: The interference has been eliminated. Here, the eliminating of noise only smooth the speech signal. No need to change the recorded or input speech sample data. In the research system, HPF (High Pass Filter) is used for eliminating the noise from the input speech signal.

Step III: The proposed system has developed the feature extraction method using the GFCC algorithm. It extracts the reliable feature sets. The efficiency allows a matric of 122*64 information.

Step IV: After that feature extraction, introduced the ALO (Ant Lion Optimization) algorithm to select the feature set based on pitch, entropy, energy and auto correlation, etc.

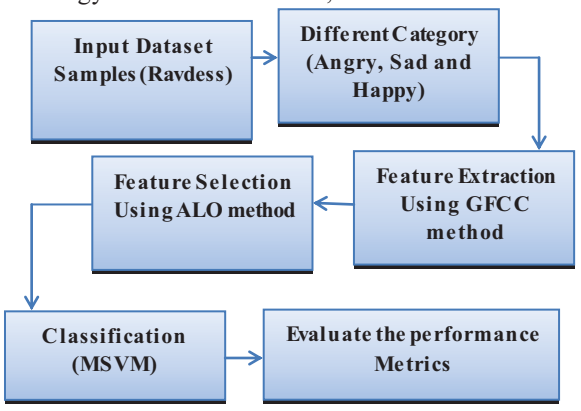

Fig 3. Proposal Flow Chart

Step V: The loaded speech data and input voice sample is trained with the help of the Multi-Class Support Vector Machine (MSVM) classification method.

Step VI: This proposed approach is implemented to verify the outputs from the simulated information. Create the SVM architecture model SS ← input samples, Class (C1 and C2). The system simulation is evaluated with the help of MSVM.

MSVM algorithm is a machine learning model. MSVM model has five steps such as :

1. Verify the set of speech samples. Inputs as well as target groups or classes.
2. Signal Pre-Processing such as noise removal, transform the data values and selection of the feature set.
3. Train model procedure uses the relationships among the input data and target data.
4. Testing model procedure to test the relationship and evaluate the accuracy rate.
5. Toemininate the overfitting, and validation, etc.

Step VII: If the output is equal to the train and test feature set then detect or recognize the emotion and show the speech emotion and also evaluate the performance metrics such as FAR, FRR, Accuracy, and Signal To Noise Ratio (SNR).

IV. Experiment Analysis

In this section, explains the **RAVDESS**[13] data set in the Speech Emotion Recognition (SER) system. This kind of data stored the 1-24 samples for the hero's emotions in audio records. It defines the proposed formula such as FRR, FAR, Accuracy, and Signal to Noise Ratio (SNR). The current research work is using 24 speech samples in each category. In proposed work are using three different samples such as Sad, Angry, and Happiness, or Joy. All the simulation are executed by MATLAB with GUI (Graphical User Interface) Tool.

RAVDESS data set description[13]:

The speech emotion dataset name is RAVDESS [36] full form Ryerson Audiovisual data set of expression voice and song comprises 7356 records (Size = 24.8 Gigabytes). The data set comprises twenty-four famous hero's twelve male and twelve female uttering 2 lexically coordinated declarations in Neutrak North American (NA) pronunciation. The language adds happy, sad, and angry emotions. Individual emotion is shaped at 2 different levels of expression intensity such as Normal and Strong, with an extra Impartial feeling. All situations are available in 2 different modality arrangements such as Audio only – 16 bit, 48kHz *.wav, Audio and Video – 720 H.264, aac 48kHz, *.mp4), and Video only – no sound.

Acoustic records all heros 1-24 are available as 2 division zip files size 200 MB each.

The video records are given division zip download for individual hero 1-24 size will be 500 Mb and are divided into separate speech and song. Proposed Parameters are defined as Accuracy Rate, FAR, FRR, SNR, and MSE.

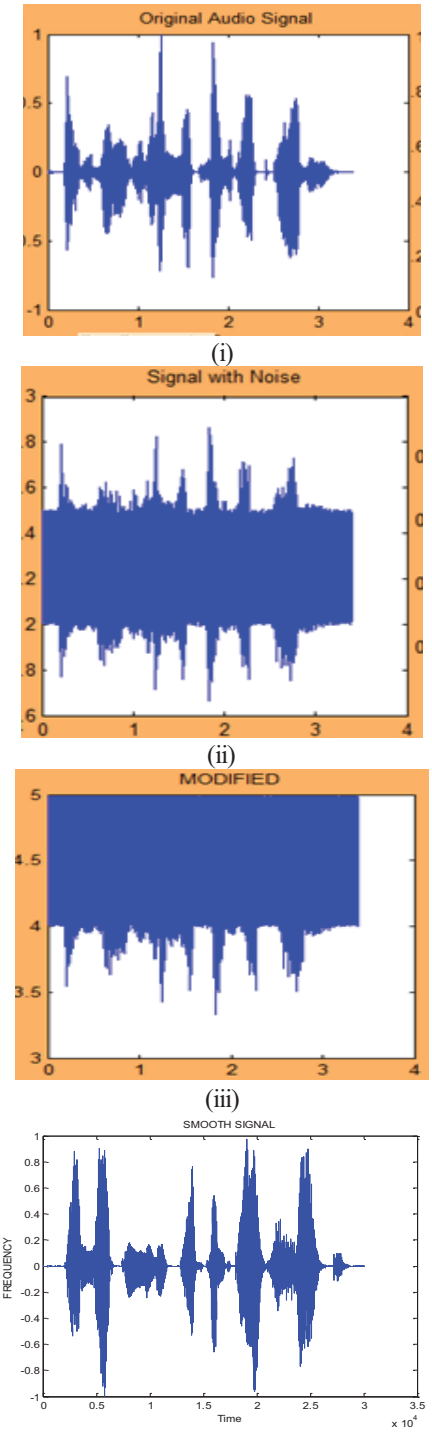

Fig 4. (i) Input Speech Sample (ii) Noise Speech Sample (iii) Modified or Decomposed Speech Sample and (iv) Smooth or Filter Signal

Fig 4(i) define the upload input sample (Sad, Angry, and Happy). It represents the input sample display in the axis tool. Fig 4(ii) presents that the noise level, and it has given the output speech signal. It has added artificial noise levels like 10,20,30, 40, and 50 percent. Then, the result shows the result in a noisy signal of the speech sample. Fig 4(iii) shows the decomposition speech signal. Fig 4(iv) defines that the filtered or smooth speech signal using the Gammatone filter approach.

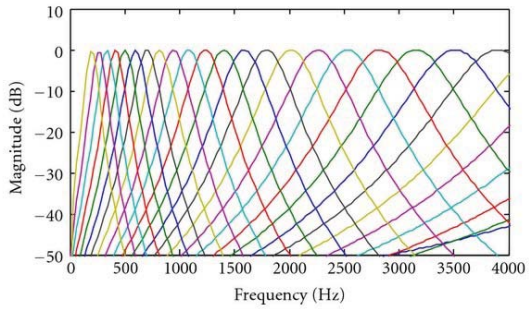

Fig 5. Feature Extraction

Fig 5 shows the GFCC feature extraction method has calculated the main properties of the GFCC. Normally, the GFCC evaluates the major phases such as Gamma tone filtration, downsampling, feature set, discrete cosine transformation, and GFCC sets. It shows the magnitude with different-2 number of frequency rate in Hz.

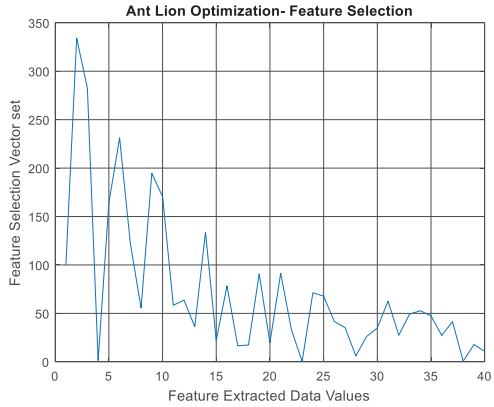

Fig 6. Feature Selection Process

Fig 6 depicts the feature selection process by the Ant Lion Optimization algorithm. This algorithm is used to select the feature sets and saved inside the database. It found the best valuable feature set in the various emotion types and optimized the vector sets. ALO algorithm has improved the feature set values and optimized the speech signal feature sets.

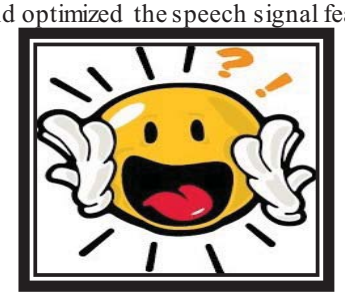

Fig 7. Recognize the category of the speech emotion

The above Fig 7 shows that the MSVM classification performance is detected or recognized by the speech emotion (Happy Case) and expression in the image format.

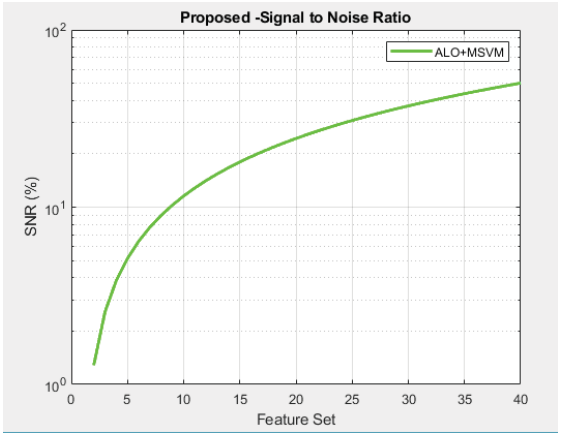

Fig 8. Proposed –SNR Parameter

Fig 8 shows the Signal to Noise Ratio (SNR) parameter and represented the speech quality. SNR performance metric shows the improvement of the signal speech quality in wave format. Proposed work, has improved the speech quality signal and reduce the noises in the uploaded speech samples.

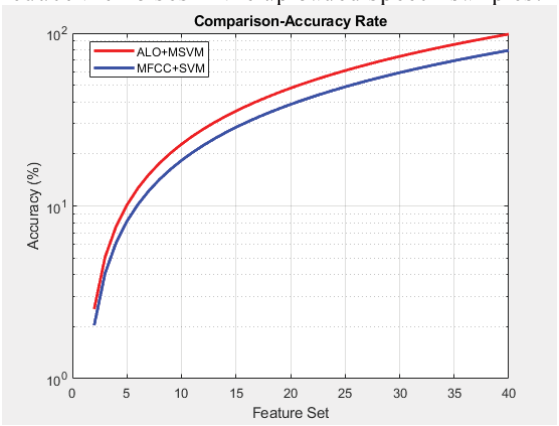

Fig 9. Comparison Analysis with Proposed and Existing Methods

The above Fig 9 explains the comparative analysis with proposed GFCC, ALO, and MSVM algorithms and Existing using MFCC and SVM classification models. In proposed model has achieved a high accuracy rate as compared with existing methods. The ALO+MSVM classification model has 97 percent and existing model has an accuracy rate 79.48 percent.

TABLE II. PROPOSED METRICS

Proposed Parameters	Values
Accuracy	97 %
SNR	50.1 %
FAR	0.0799
FRR	0.00066
MSE	0.7439

Table II shows the proposed parameters (GFCC, ALO and MSVM) algroithms such as accuracy rate 97 per cent, snr value is 50.1 per cent, error types are FAR value 0.0799, FRR value 0.00066and MSE value 0.7439.

TABLE III. COMPARATIVE ANALYSIS

Algorithm	ALO+MSVM	MFCC+SVM
Accuracy Rate (%)	97 %	79.48 %

Table III defines the comparative analysis with proposed and existing methods in the speech emotion recognition system. In proposed GFCC, ALO, and MSVM algorithm have achieved the accuracy value is 97 percent, and the existing MFCC[14] and SVM[15] algorithm accuracy value is 79.48 percent. In proposed algorithm has improved the system performance by up to 19 percent.

V. CONCLUSION AND FUTURE SCOPE

A speech emotion recognition system is defined using the Machine learning method using MSVM to recognize the various types of expression. Thus, the feature extraction method using Gammatone Frequency Cepstral Co-efficient (GFCC) algorithm has been extracted from the RAVDESS data set, and these features have been defined. It analyses how the classification and feature extraction performance in the recognition rate of emotions in speech (Sad, Happy, and Angry). It has selected the high discriminant feature sets. The Ant Lion Optimization (ALO) feature selection method presents that more data is not always better in Machine Learning (ML) uses. The ML using the MSVM method model has trained and calculated to recognize emotional categories (Sad, Happy, and Joy) from these feature sets. SERS evaluated the high accuracy rate of 97 percent on the RAVDESS data set using MSVM classifier with feature extraction (GFCC) and feature selection (ALO). For existing data sets, all the classifiers get an accuracy of 79.48 percent, when the feature extraction with MFCC was applying to the feature sets. The research work has concluded that the MSVM, GFCC, and ALO models have achieved high accuracy rate and high signal to noise ratio in comparison with SVM and MFCC algorithms. The research work has used the GFCC algorithm to extract the feature sets in the uploaded sample speech. GFCC algorithm has various steps like Gammatone Filter Bank, Downsampling, Discrete Cosine Transformation (DCT), and Feature extracted set. After that, the ALO algorithm has developed to select the feature set using the Fitness Function. It is calculated by the fit value based on the ALO algorithm. MSVM classification algorithm is implemented to classify the emotion of the speech samples. The simulation tool used by MATLAB is to evaluate the performance of the parameters like AUC, FAR, FRR, MSE, and SNR and compared with the existing algorithms (MFCC and SVM).

The future scope of the proposed work is to implement an efficient feature set fusion approaches to enhance the multi-model system performance. And also to develop to hybrid optimize the speech model and hoping to achieve high speech emotion feature sets. Finally to consider other models in the multi-modal like Speech Emotion Recognition System (SERS).

REFERENCES

[1] Pao T, Wang C. and Li Y., "A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition." *2012 FifthInternational Symposium on Parallel Architectures,Algorithms and Programming*, Taipei, 2, no.3, (2012): 157-162.

[2] LikithaM. S., GuptaS. R. R., HasithaK. and RajuA. U., "Speech based human emotion recognition using MFCC." *2017 International Conference on WirelessCommunications, Signal Processing and Networking(WiSPNET)*, Chennai, 4, no. 5, (2017): 2257-2260.

[3] Joshi, D. D., and Zalte M. B, "Speech emotion recognition: a review." *IOSR J. Electron. Commun. Eng.(IOSR-JECE)* 4, no. 4 (2013): 34-37.

[4] Basu, S., Chakraborty, J., Bag, A., and Aftabuddin, M. "A review on emotion recognition using speech." In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, 2, no.7, (2017):109-114.

[5] Ingale, A. B.,and Chaudhari, D. S. "Speech emotion recognition." *International Journal of Soft Computing and Engineering (IJSCE)* 2, no. 1 (2012): 235-238.

[6] Lin, Y. L., and Wei, G. "Speech emotion recognition based on HMM and SVM." In *2005 international conference on machine learning and cybernetics*, 8, no. 5, (2005): 4898-4901.

[7] Gharavian, D., Sheikhan, M., Nazerieh, A., and Garoucy, S. "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network." *Neural Computing and Applications* 21, no. 8 (2012): 2115-2126.

[8]Deshmukh, G., Gaonkar, A., Golwalkar, G., and Kulkarni, S. "Speech based Emotion Recognition using Machine Learning." In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 4 , no 4(2019): 812-817.

[9] Jiang, W., Huang, L., Liu, Q., andLü, Y. "A cascaded linear model for joint chinese word segmentation and part-of-speech tagging." In *Proceedings of ACL-08: HLT*, pp. 897-904. 2008.

[10] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. "Speech emotion recognition using deep learning techniques: A review." *IEEE Access* 2, no. 7 (2019): 117327-117345.

[11] Sonmez, Y. Ü., and Varol, A. "New Trends in Speech Emotion Recognition." In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*,6 , no.8(2019): 1-7.

[12] Le, B. V., and Lee, S."Adaptive hierarchical emotion recognition from speech signal for human-robot communication." In *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 7, no.6 (2014): 807-810.

[13] LivingstoneR., Steven. 2020. "RAVDESS Emotional SpeechAudio". *Kaggle.Com*. https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio.

[14]Han, W., Chan, C. F., Choy, C. S., and Pun, K. P. "An efficient MFCC extraction method in speech recognition." In *2006 IEEE international symposium on circuits and systems*, 5, no. 2 (2006): 4.

[15] Vishwanathan, S. V. M., and Murty, M. N.," "SSVM: a simple SVM algorithm." In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, 3, no.2, (2002): 2393-2398.