IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 2, April, 2022, pp. 271-278
Journal homepage: http://iieta.org/journals/ria

# Speech Emotion Recognition Using Machine Learning Techniques

Anil Kumar Pagidirayi*, Anuradha Bhuma

Department of ECE, S.V. University College of Engineering, Tirupati, Andhra Pradesh 517502, India

Corresponding Author Email: anilkumar.pgke@gmail.com

https://doi.org/10.18280/ria.360211

**Received:** 14 December 2021
**Accepted:** 24 February 2022

***Keywords:***
*Mel frequency cepstral coefficient, subspace-kNN (S-kNN), support vector machine, random subspace method, speech emotion recognition*

**ABSTRACT**

Mel Frequency Cepstral Coefficient (MFCC) method is a feature extraction technique used for speech signals. In machine learning systems, the Random Subspace Method (RSM) known as attribute bagging or bagged featuring used to classify the complete feature sets. In this paper, an innovative method is proposed which is a combination of RSM and kNN algorithm known as Subspace-kNN (S-kNN) classifier. The classifier selects the specific features extracted from MFCC are angry, sad, fear, disgust, calm, happiness, surprise, and neutral speech emotions in Speech Emotion Recognition (SER) system. Furthermore, in the proposed method the performance metrics of accuracy, Positive Predictive Values (PPV) rate, training time are evaluated on male and female voice signals when compared with previous classifiers like SVM and bagged trees.

## 1. INTRODUCTION

Natural Communication between humans for better interaction, emotions add colour to the language which acts as an essential component. Based on the emotion of the speaker the listener will acclimate the behaviour in accordance with different emotions conveyed by the speaker. Advancements in current technologies empowered human interaction with computers by non-traditional moods like signs, voice, facial movements etc. Modules of these emotional communication are still required [1]. For achieving a resistive human computer intellectual communication an essential computer human interaction is needed which is similar to two-way human conversations.

Speech acts as fastest and distinctive methodology for communication between humans. For the past fifty years tremendous research approaches have been done on speech emotions identification. The problem of human voice recognition still persists in machines, which indicates a lot of change in development of human dissertation over consortium of words. Despite of an outstanding development made in speech recognition [2], still we are far away in achieving specific communication between human and machine which are not realized by condition of speaker emotions. Extraction of speaker's state of emotion from his or her own voice then this method is known Speech Emotion Recognition (SER). This is overviewed briefly in current research approaches specially on SER.

The SER application is related in extensive range of applications in real world for example in call service centres helps to find the answering method of call associated to a client, in vehicles to identify the mental state of human while driving to prevent accidents, acts as diagnostic device for recognition of several syndromes in patient's medical emergencies, in E-tutorials and storytelling modules which accordingly synchronise to attitude of the audience mood, etc. In human speech there are numerous emotions based on several conditions. Basically, emotions are categorised into five primary emotions such as anger, calm, happiness, fear and disgust.

Emotions that are affected by the speech features are distinguished by some characteristics such as Quality, Spectral, steady features. The steady metric features of speech signals consisting of emotions contentment are affected by zero crossing rate, pitch and energy. Differentiation of speech signals are referred by pronunciation rate, energy, spectral data, and fundamental frequency (fo). A robustic relation exists between apparent emotion and speech quality. These are categorised as speech level, voice pitch, feature marginal composition and temporal. Spectral analysis of speech signals is represented as short timing features. Emotional content of speech pronunciations is depending on spectral energy distribution. High-stimulative emotions are noticed as happiness (or) anger by peak energies at higher frequencies similarly low-stimulated emotions are noticed as sadness are represented by low energy range. Detection of emotions by spectral analysis are carried out by MFCC and LPCC (Linear Predictive Cepstral Coefficients). Formation of speech by indeterminant flow of air in vocal tract system. Sound is produced in the vocal tract by the flow of air which gets affect by muscular tension of utterer due to stress. Speech detection in sound is important for unsteady speech features signals. Hence speech emotion detection plays a dynamic role in present applications like phycological condition of vehicle driver to avoid major accidents, Student's phycological state in classroom make teachers detect their mood by the way they answer the questions, so that teacher can change topic to make them actively involved in listening, children's mental state to develop better relationship with parents and friends, etc.

The sequence of this work is partitioned as follows. Section II portrays about literature analysis thoroughly on SER. Section III outlines the basic elements of SER system followed by feature extraction technique, Section IV the proposed SER system is presented and cross-validation with selection of feature. Section V provides experimental results with conclusion. The speech signals collected from the database are

preprocessed by removing noises using filters then the obtained data undergoes feature extraction to recognize various emotions then proposed classifier is applied to recognize exact emotion and accuracy rate has been summarized in the paper.

## 2. LITERATURE REVIEW

Many studies were proposed to identify the Automatic SER's which concentrated on emotion portraited methods, features extracted, reducing the dimensionality of features, classification of particular emotions, and regression algorithms.

Langari et al. [3] represents a PCA based innovative technique for extraction of speech features with the use of adaptive time-frequency coefficients to modulate and simulate the speech emotion recognition system using the Berlin Emotional Speech Database (EMO-DB) and SAVEE (80% accuracy) by application of Fractional Fourier Transforms and linking them with Cepstral featured coefficients which are not efficient.

Yu et al. [4] proposed a Support Vector Machine (SVM) based classifier in voice signals for recognizing the specific emotions based on spoken speech comparison based on linear discriminant classifiers (LDC), used to classify the emotional Chinese speech corpus using SVM of with accuracy up to 85%, but four emotions angry, neutral, sadness and happiness are only classified.

Harár et al. [5] proposed a SVM classifier-based SER with MFCC technique for extraction of speech features with TEO on 42 dimension 39 Features using RML (Ryerson Multimedia Laboratory) using for reduction of auto-encoder dimensionality which improves the identification rate but cannot applied larger data set need to be improved.

Lanjewar et al. [6] proposed a GMM and KNN based Speech emotion recognition method for Berlin emotion database on six emotions happy, angry, sad, fear, surprise and neutral. Spectral components using MFCC, wavelet features of speech which extract good accuracy only for angry but no other using these two algorithms.

Ezz-Eldin et al. [7] designed a Bagged trees (BOW) classifier-based Speech emotion recognition using RAVDESS database that results good accuracy based on 5min up to 88.1% but need to applied on other datasets to get good accuracy.

Zhang [8] proposed a binary weighted cuckoo search algorithm is used for SER and feature extraction by cross-validation that is used to certify the situations of various adapted features. A database of Open SMILE -2.3.0 used for classification of sentiments in speech signals by training 100 set of speech signals by accuracy of 83.2% an advanced intellectual dimension optimization reduction technique but further challenge arises to this model which has limitations to multi-dimensional platforms like bio-signals, facial expressions.

Prasanth et al. [9] represents the database Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) exclusion of three different features (MFCC, MEL and Chroma) from operational data resources by differentiating emotions and music with accuracy 85.12% using MLP technique but need to applicable for call service centres, voice-assistant remote aids.

Deriche [10] used to recognise speech emotion using HMM and ANN techniques on Berlin and Polish datasets to obtain better emotion recognition and gender identification combinedly up to 60% accuracy on six emotions without need of language identification.

## 3. ELEMENTARY SER SYSTEM

The general SER system consisting of three blocks such as speech pre-processing unit which produce the physical quantities of the speech signals. The second block consists of feature extraction block that extracts features from previous block. Here, K1, K2, … Kn are the speech featured signals extracted from physical quantities and applied to the third block which will classify the speech features. Finally, detection of specific emotion is done by using the classifier as shown in the Figure 1.

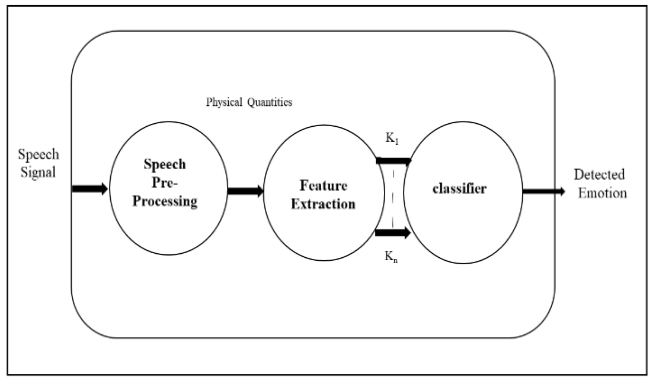


**Figure 1.** SER system [11]

An improved performance efficiency and accuracy are obtained in feature extraction process when the speech signal was pre-processed and applied to feature extraction block. The physical quantities obtained by pre-processing of speech signals contains pitch, phonetics and energy formants. In pre-processing stage, speech signals undergo three stages such as filtering, framing formats and windowing techniques. Filtering method provides the reduction of noise, formed due to the occurrence of disturbances in the surroundings or throughout the recordings of voice signals. Pre-emphasis filter will enhance the energy to higher frequencies in the speech signal because they may get attenuated in the process of speech signal production through vocal region.

The pre-emphasis filter's *z*-transform is represented as:

$$H(x) = 1 - bx^{-1} \quad (1)$$

where, *b* value ranges between 0.9 and 1.0.

Speech signal is a non-linear and non-stationary signal which is difficult for analyzation. So, framing technique analyses the speech signals by dividing them into equal numbered samples in independent manner. The size of the frame is decided by the feature extraction technique applied on it. Overlapping of frames take place to remove the barriers in between the frames. Partition of signal into frames will create some breakpoints at the edges of respective frames of input voice signal. Breakpoints between each frame is removed by applying frame through taped window. Window technique output y[m] is:

$$y[m] = x[m].w[m] \quad (2)$$

Here, the speech signal x[m] and window signal w[m] are at time 'm'. Among different windowing techniques applied in speech processing method, rectangular windowing technique is simplest one that shortly cuts off edges of the signal. Discontinuity is the major problem that occurred during Fourier analysis. So, continuity of the frame must be maintained between starting and ending of the edges. In order to evade this disjointedness a hamming window technique is applied. Hamming window avoids the discontinuity by shrinking the amplitude of signal near to zero of the window boundaries.

### 3.1 Feature extraction techniques

Feature extraction techniques like Pitch, MFCC and LPCC (Linear Prediction Cepstral Coefficients) are used to extract features like zero crossing rate, phonetics formants and energy are extracted to find the features in SER systems.

3.1.1 Linear predictive cepstral coefficients

LPCC method is used to display the construction of human voice and performance well in fresh environment but hazy in noisy environment. LPCC is used to determine the parameters like spectra and pitch phonetics, used as significant feature extraction technique in audio and speech processing of voice signals [12-14]. LPC is designed as chronological approach to find the resonant construction of human vocal region which develops consequent sound. SER systems widely uses Pitch and MFCC feature extraction techniques and MFCC are applied in this work [15, 16].

Pitch. Pitch is defined as a fundamental frequency of speech signal and inverse of fundamental period gives pitch value. The pitch is varied with peak and low frequency of sound in the speech signal. Auto correlation function is used for estimation of the pitch from the waveforms applied.

$$\mathrm{R}(m) = \frac{1}{N}\sum_{i=0}^{N-l-1}(x[j+l] * x[j]); m \geq 1 \quad (3)$$

where, the applied speech signal is $x[j]$, discrete time signal is $j$ and delay applied by $l$. R(m) has a peak value, if $x[j]$ equals to $x[j+l]$.

3.1.2 Mel-frequency cepstral coefficients

MFCC of speech signals are calculated by following the steps in the shown Figure 2 and explanation each block as follows:

(1) Discrete Fourier Transform: Spectral data is extracted after pre-processing of speech data. Spectral data is extracted by DFT, a transformational method applied on discrete time signal in distinct frequency band.

DFT is computed by frequently used method known as Fast Fourier Transforms (FFT).

$$x(t) = \sum_{i=0}^{M-1}(x(j)e^{-j2\pi t\frac{i}{M}}) \quad (4)$$

where, $j$=0, 1, 2, ……, M-1.

DFT points get increased with increase in the frequency resolutions but the data get weaken by similar factor. Hence, solution is obtained by increasing the number of optimal points such that most of the speech data utilized in the frequency spectrum.

(2) Mel-scale filter bank: A wide range of frequency is required for obtaining FFT spectrum and the speech signal varies non-linearly on scale. *Mel* scale used to measure non-linearity and computed for specified frequency $f$ in Hertz:

$$\mathrm{Mel(f)} = 2595\log_{10}(1+\frac{f}{700}) \quad (5)$$

or

$$\mathrm{Mel(f)} = 1127\ln(1+\frac{f}{700}) \quad (6)$$

Mel-frequency is linearly proportionate to logarithmic function of linear frequency. It has frequencies spaced linearly with frequencies less than 1 KHz and logarithmic greater than 1 KHz. A triangular series band-pass filters are applied in a Mel-scale to determine the spectral components weighted sum of the filter then only output signal estimated to a Mel-scale.

(3) Discrete Cosine Transform (DCT) function: Features of MFCC are generated by transforming the Mel spectrum of a function to time domain shown in equation 5. And hence, series of acoustic vectors are transformed to specific input utterance. Here, logarithmic energy 'Ek' is generated by DCT and 'L' MFCCs are generated by the Triangular pass band filters application.

$$Cn = \sum_{t=1}^{N}(cos\, n * (t-\frac{1}{2})\frac{\pi}{N}) \quad (7)$$

where, $n$=1, 2… $L$, $N$ represents the log spectral coefficients and $L$ represents Mel-scale Cepstrum Coefficients generated.

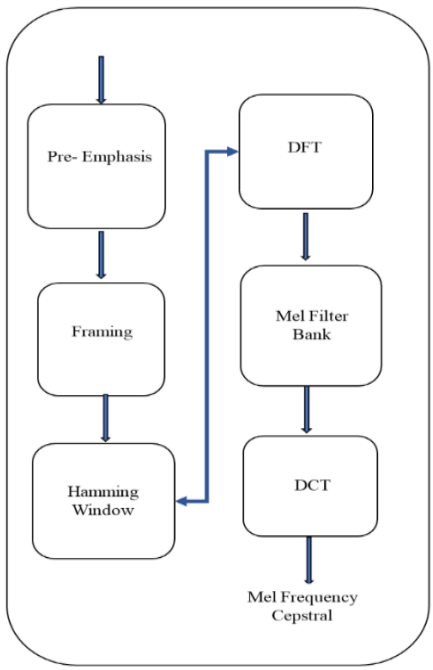


**Figure 2.** Feature extraction technique by MFCC [11]

### 3.2 Classifiers

3.2.1 Support vector machine

The main purpose of Support Vector Machine (SVM) is to determine the difference between two classes [17]. For

extending binary classifiers to multi-class classifier functions many proposals were suggested. In reality multi-classed SVMs are applied in numerous fields and tested their effectiveness in recognizing different categories of information given to it. SVM classifier is a part of supervised Machine learning method utilized for classification along with regression. Classification of data using SVM to find suitable subspace which can differentiate the peak margin that can separated from the training datasets implemented in the feature vector by a kernel function K, maximum applied kernel functions, such as linear, polynomial based training feature sets, new values are differentiated and calculated. Hence, SVM classifier can select good kernel functions and change the parameters for obtaining more identification rate but not feasible accuracy rate.

3.2.2 Ensemble bagged trees

Using MATLAB function implementation of classification is done by ensembled bagged tree classifier and boosted tree methods. Application of thirty decision trees is performed in bagged tree classifier. Analysis and fitting of simple data models for error detection using thirty decision trees are used in serial manner in early access [17]. So that, consecutive trees are fitted at every point in order to achieve solutions for net errors from the prior trees. Implementation of SVM in comparison and to find the performance of given classifier, a confusion matrix is simulated to evaluate necessary actions to be obtain better sensitivity (TPR), specific rate (TNR) and gives moderate accuracy.

## 4. PROPOSED METHOD

S-kNN classifier is the proposed method which classifies different speech emotions from speech signal by applying MFCC feature extraction technique. The Figure 3 represents the proposed SER system for identifying eight emotions such as anger, anxiety, disgust, fear, happiness, calm, neutral and sadness, taking neutral emotion as reference model. Speech features are extracted by using MFCC technique. Here the speech signal from the database gets pre-processed and generate the physical quantities to feature extraction block which modulates the energies of speech emotion components and then energies of peak frequency elements get modulated by pre-emphasis filter similar to pre-processing stage in MFCC technique. Further the speech signal is classified by S-kNN classifier by 5-fold cross validation which gives good accuracy. The explanation of the same is given below Figure 3.

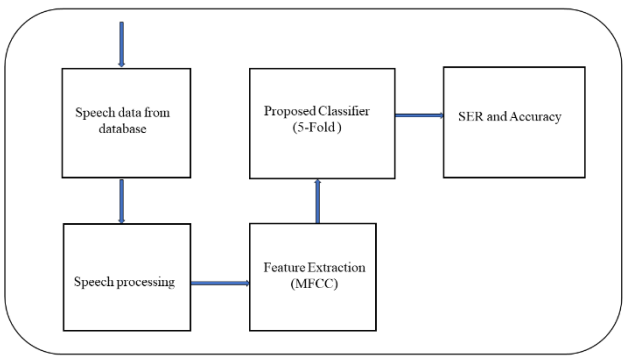


**Figure 3.** Proposed SER System

### 4.1 Ensemble subspace k NN classifier

Random selection of elements for a given featured vector is a stochastic method which constructs each classifier known as RSM. For kNN method a sample tested is compared with an example which selects specific features having nonzero contributions by the distance. Selection of subspace is geometrically portraying the similarity among the points and kNN are applied using the projected distances. Every time choosing random subspace combinedly with a new set of kNN are calculated. kNN chosen subspace are arranged based on majority voting classes for test sample. In S-kNN, similar training sample occurred more than once which happened more than single selected subspace among kNNs.

Formally, given a set of m points in an n-dimensional feature space $\{(x_1, x_2, \ldots, x_m)|x_j$ is real all $1\leq j\leq$m$\}$, We consider the n-dimensional subspaces $\{(x_1, x_2, \ldots, x_m)|x_j=1$ for j$\in$ I, $x_j=0$ for j$\notin$I$\}$, Where I is an n- element subset of $\{1, 2, .., m\}$, and $n<m$. In each path, selection of random subspace is taken from the choices $C(m, n)$ of the I subset. The subspace selected is predicted by all points. Projection of kNN points of every tested sample, where kNN$(1\leq k\leq N)$ points are initiated by using Euclidean distance among the training points that are projected. For selected k neighboured class labels $\{c_1, c_2, \ldots, c_k\}$ that are attached to list C. Allocation of the test point after passing through $p(p\geq 1)$ of the class has highest recurrent existence in the given list $C$.

S-kNN is derived from stochastic discrimination method consisting of nonlinear generated classifiers which combined appropriately with monotonic rise in accuracy. Specifically, each classifier don't have authoritative power but they makes things easy of unseen data for similar problems. So, freedom is introduced by stochastic method among many classifiers. Intensification of their decisions are made with the developed discriminative power. Similar approach is followed by RSM. Some dimensions of feature vector are ignored and consistency of classification is developed among samples that vary particularly in unnotified dimensions. Random set of dimensions opted are ignored by introducing certain independence among the module classifiers. In search of individual decisions, there will be increase in discriminative power similar to other classifiers combination. Taking individual components as kNN classifiers algorithmic difficulties are reduced in obtaining uniformity of the feature space that required in the previous SVM algorithm utilizing modest weak models.

Since in this method random subclasses of elements of feature space are applied, it will be virtuous enough to the problems of relatively large numbered features. These problems are raised signal processing applications such as very large-scale data mining systems, speech processing and image recognition. For less features problem arises and it will be mandatory to enhance the feature vector with particular raw featured simple functions. Below algorithm helps to attain good emotion recognition in SER systems by application of S-kNN classifier followed by 4 steps [18].

**Algorithm for ESkNN**

Step1

For i=$l\rightarrow m$ do

i. Select $l$ random sampled features of size m from total d samples by replacing

| | | | | |
|---|---|---|---|---|
| | | learning part by bootstrap sample $B_i$ to construct with model out of *l* features. ii Instances left over are saved as OOB(i) sample of $B_i$ and kNN(k) function is called to construct model $C_i$. iii Find accuracy of $C_i$ using OOB(i) and store it as Acc(i). | | |
| | End for | | | |
| Step2 | For j= *l*→*m* do | | | |
| | | If Acc(j)>$Q_2$ then | | |
| | | | Select $C_j$, where $Q_2$ is the second quartile of the accuracies of all h models. else | |
| | | | | Drop $C_j$. |
| | | End if | | |
| | End for | | | |
| Step3 | i. From l selected features kNN is constructed using $B_i$ and returned. Best models are fused and arrange them say h, in descending order with reference to their accuracy. ii. Initialize q=1 and take the best model with the highest accuracy from the above selected model as the starting ensemble. | | | |
| Step4 | q=2→ *h* do | | | |
| | | If | | |
| | | | $BS^{(q)} < BS^{(q-1)}$ then Select the qth kNN model after application of validation data. | |
| | | Else | | |
| | | | Do not select the qth model for final ensemble. | |
| | | End if | | |
| | End for | | | |

An Ensembled subset of kNN classifiers are formed by following steps:

(1) A randomly sampled size of *l* samples chosen from total d samples. *l*<*d*, features without change in position from the feature space P, denoting feature vector P*l* are taken.

(2) Depending upon chosen random feature subcategory P*l*, a randomly sampled size m, is drawn from total d samples.

(3) Build the kNN classifier on random sample subspace.

(4) Compute the classifier accuracy by applying the similar feature set for its implementation.

(5) Repeat the above steps for *m* times and ranging of m classifiers based on their accuracies.

(6) Choose foremost *h* classifiers with peak accuracies.

Evaluation of selected classifiers as follows:

(1) Combination of second-best classifier with first best classifier is started in the ensemble, and performance evaluation of the classifier is calculated on the validation.

(2) Development of ensemble by addition of third best classifier is performed and the performance is calculated and applied to all the h classifiers.

$$BS^{(q)} < BS^{(q-1)} \tag{8}$$

Here $BS^{(q-1)}$ be the cleaver scored ensemble for best chosen kNN models excluding *r* th model and $BS^{(q)}$ be the cleaver scored ensemble of the finest models after considering the q th model, then selection of $q$ th model is shown in the above equation.

### 4.2 Cross-validation

Cross validation is the process of making the training more frequent, for training of models different training sets are applied, most frequently appeared data sets are chosen from different training groups. Here number of trainings are denoted as n train, and selected feature is threshold to n. Let us consider that for n train= 10 and n select=2 then 10 distinct data sets and test sets are chosen. Combination of classifier and feature extraction technique MFCC to choose the feature set, after that 5 sets of normalized feature subsets are generated. Comparing these with frequency more than 2 are finalized as feature subset. below equation represents the threshold value:

$$x_j^{(k+1)} = \begin{cases} 1, S(x_j^{(k+1)}) > \delta \\ 0, otherwise \end{cases} \tag{9}$$

where, $x_j^{(k+1)}$ is the new result and $S(x_j^{(k)})$ represents a sigmoid function and $\delta$ represents the threshold value. A suitable value of $\delta$ helps to select an accurate search direction. Cross validation of 5 in the proposed method helps to get preferrable accuracy.

## 5. RESULTS

The system is carried out using CSV files consisting of datasets combination of RAVDESS, CREMA-D, TESS and SAVEE extracted from audios, where the Dataset consists of female and male emotions based on 58 featured MFCCs values [19]. The database contains 49225 female and 35911 male variables are recorded at 44 KHz surrounding environment of emotions consisting of anger, calm, disgust, fear, happiness, neutral, sad, surprise are chosen. Further, the speech samples are downsampled to 16 KHz for Speech Emotion Recognition analysis using 58 featured MFCC feature extraction techniques. The confusion matrices and comparison with previous classifiers of SER system accuracy, Positive Predictive Values (PPV) rate, training time for features applied using MFCC and S-kNN classifier are shown in the below tables.

From the below tables, Tables 1 and 2 represent classification accuracy of female and male speech emotions, and the results generated will represent the accuracy of the SER system of the proposed classifier method. An average accuracy of 86.80% for the male, 92.40% for female speakers, PPV rate and Features training time are obtained. By this, we can say that the application of the S-kNN classifier with the MFCC feature extraction technique helps to obtain improved accuracy.

The above Tables 3 and 4 represent PPV rate comparison of female and male emotional speech database where Table 5 represents time taken for recognition.

**Table 1.** Comparison of FEMALE SER system

| Female SER system Classification accuracy (%) | | | |
|---|---|---|---|
| **Female emotions** | **SVM** | **Bagged Trees** | **Proposed** |
| Angry | 70.40% | 93% | 96.1% |
| Calm | 74.40% | 92% | 95.1% |
| Disgust | 59% | 86.30% | 91.2% |
| Fear | 57.70% | 86.80% | 91.2% |
| Happy | 55.90% | 86% | 91.3% |
| Neutral | 63.50% | 87.70% | 91.0% |
| Sad | 71.90% | 90.20% | 92% |
| Surprise | 85.50% | 96.50% | 98.40% |
| Average Accuracy | 64.80% | 89% | 92.4% |

**Table 2.** Comparison of Male SER system

| Male SER system Classification accuracy (%) | | | |
|---|---|---|---|
| **Male emotions** | **SVM** | **Bagged Trees** | **Proposed** |
| Angry | 85.20% | 90.10% | 92.40% |
| Calm | 63.50% | 91.50% | 94.20% |
| Disgust | 70.40% | 77.50% | 84.40% |
| Fear | 70.60% | 76.00% | 83.00% |
| Happy | 72.00% | 77.70% | 85.40% |
| Neutral | 69.50% | 80.80% | 85.80% |
| Sad | 72.90% | 83.40% | 87.90% |
| Surprise | 61.50% | 87.00% | 93.10% |
| Average Accuracy | 73.00% | 81.00% | 86.80% |

**Table 3.** Comparison of Female PPV rate of SER systems

| Female SER system Positive Predictive Values (PPV) Rate (%) | | | |
|---|---|---|---|
| **Female emotions** | **SVM** | **Bagged Trees** | **Proposed** |
| Angry | 72.70% | 88.50% | 94.30% |
| Calm | 62.00% | 88.70% | 92.00% |
| Disgust | 56.70% | 86.50% | 88.40% |
| Fear | 63.60% | 90.40% | 93.50% |
| Happy | 61.00% | 90.60% | 93.80% |
| Neutral | 66.40% | 89.00% | 91.50% |
| Sad | 61.20% | 86.80% | 90.90% |
| Surprise | 83.10% | 97.20% | 98.40% |
| Average PPV rate | 65.80% | 88.03% | 92.85% |

**Table 4.** Comparison of Male PPV rate of SER systems

| Male SER system Positive Predictive Values (PPV) Rate (%) | | | |
|---|---|---|---|
| **Male emotions** | **SVM** | **Bagged Trees** | **Proposed** |
| Angry | 86.70% | 83.10% | 91.70% |
| Calm | 88.01% | 89.20% | 93.20% |
| Disgust | 71.60% | 78.30% | 83.30% |
| Fear | 77.00% | 82.40% | 87.40% |
| Happy | 76.20% | 80.90% | 87.90% |
| Neutral | 73.20% | 79.10% | 85.20% |
| Sad | 69.31% | 80.10% | 84.00% |
| Surprise | 89.50% | 91.90% | 91.60% |
| Average PPV rate | 78.91% | 83.125% | 88.03% |

**Table 5.** Comparison of Classifiers training time of SER systems

| S. No | Classifiers | Female SER Training Time (Sec) | Male SER Training Time (Sec) |
|---|---|---|---|
| 1. | SVM | 337.4 | 805.5 |
| 2. | Bagged Trees | 76 | 57 |
| 3. | S-kNN | 112.5 | 119.4 |


**Figure 4.** Confusion matrix of Female SER system using SVM

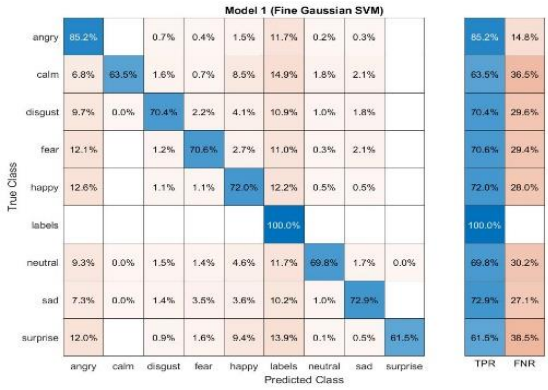


**Figure 5.** Confusion matrix of Male SER systemusing SVM

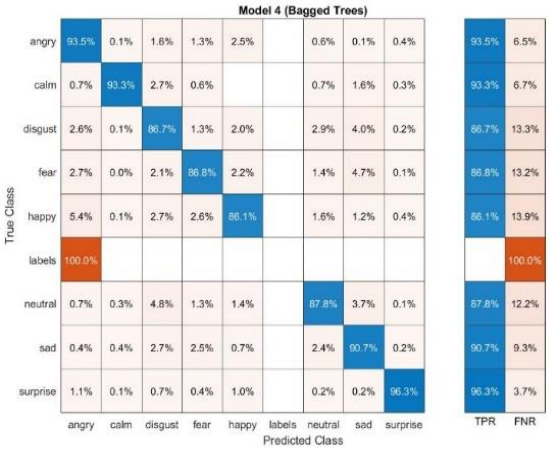


**Figure 6.** Confusion matrix representation of Female SER system using Bagged Trees

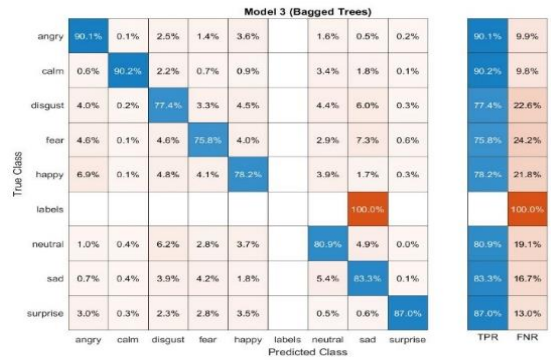


**Figure 7.** Confusion matrix representation ofMale SER system using Bagged Trees


**Figure 8.** S-kNN applied Female SER system


**Figure 9.** S-kNN applied Male SER system

The above Figures 4-9 represent the confusion matrices obtained by performing different classifier techniques consisting of true class positive models, and graphs, obtained by performing simulation of taken data from speech data base.

Figures 10-13 are graphical representation of numerical data taken from Tables 1-4.

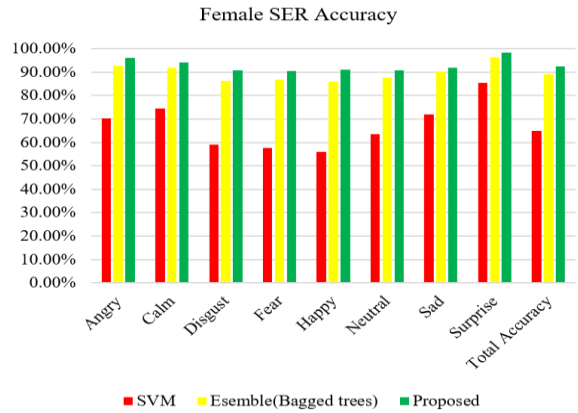


**Figure 10.** Comparison Bar graph of Female SER system using SVM, Bagged Trees Subspace kNN

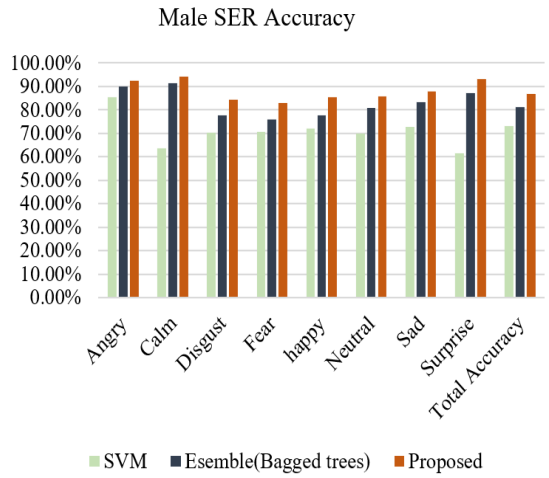


**Figure 11.** Comparison bar graph of Male SERsystem using SVM, Bagged Trees Subspace kNN

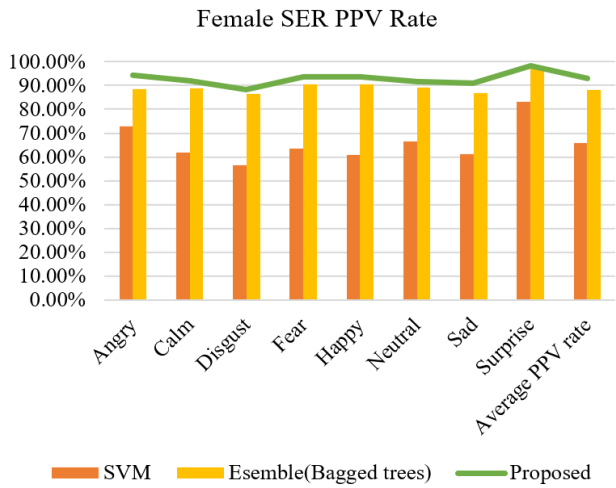


**Figure 12.** Comparison Bar graph of PPV rate of Female SER system

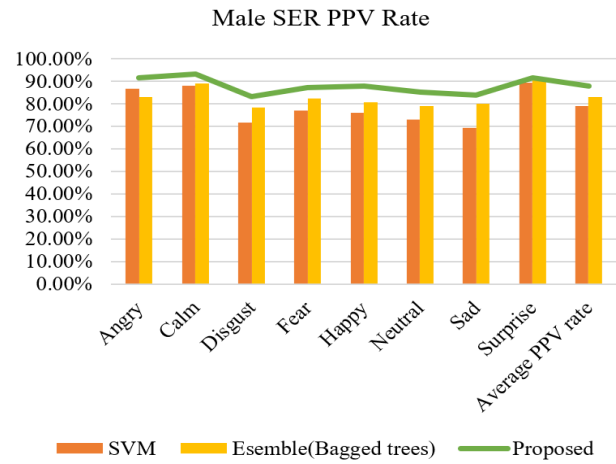


**Figure 13.** Comparison Bar graph of PPV rate of Male SER system

## 6. CONCLUSION

The proposed classifier S-kNN used in the SER system classifies different emotions like anger, calm, disgust, fear, happiness, sad, surprise with reference to neutral speech. CSV

files consisting of datasets of RAVDESS, CREMA-D, TESS, and SAVEE database are applied to the proposed classifier using the MFCC feature extraction technique for analysis of the present work. An accuracy of 92.4% for female, 86.80% for male speakers, 92.85% female PPV rate and 88.03% male PPV rate are obtained by using the S-kNN classifier, which is relatively better than ensemble bagged trees and SVM classifiers. Hence S-kNN classifier helps to classify the speech emotions more accurately when compared with other classifiers. But when compared with the feature prediction training time there is need for decrease in training time compared to other classifiers. We can achieve better accuracy and training time of features in the detection of speech emotions by combining multiple classifiers or feature extraction techniques.

## REFERENCES


[1] Khanna, P., Sasikumar, M. (2011). Recognizing emotions from human speech. In Thinkquest~ 2010, pp. 219-223. https://doi.org/10.1007/978-81-8489-989-4_40

[2] Umamaheswari, J., Akila, A. (2019). An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pp. 177-183. https://doi.org/10.1109/COMITCon.2019.8862221

[3] Langari, S., Marvi, H., Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. Informatics in Medicine Unlocked, 20: 100424. https://doi.org/10.1016/j.imu.2020.100424

[4] Yu, C., Tian, Q., Cheng, F., Zhang, S. (2011). Speech emotion recognition using support vector machines. In International Conference on Computer Science and Information Engineering, pp. 215-220. https://doi.org/10.1007/978-3-642-21402-8_35

[5] Harár, P., Burget, R., Dutta, M.K. (2017). Speech emotion recognition with deep learning. In 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, pp. 137-140. https://doi.org/10.1109/SPIN.2017.8049931

[6] Lanjewar, R.B., Mathurkar, S., Patel, N. (2015). Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. Procedia Computer Science, 49: 50-57. https://doi.org/10.1016/j.procs.2015.04.226

[7] Ezz-Eldin, M., Hamed, H., Khalaf, A. (2020). Bag-of-words from image to speech a multi-classifier emotions recognition system. International Journal of Engineering & Technology, 9(3): 770-778.

[8] Zhang, Z. (2021). Speech feature selection and emotion recognition based on weighted binary cuckoo search. Alexandria Engineering Journal, 60(1): 1499-1507. https://doi.org/10.1016/j.aej.2020.11.004

[9] Prasanth, S., Thanka, M.R., Edwin, E.., Nagaraj, V. (2021). Speech emotion recognition based on machine learning tactics and algorithms. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.12.207

[10] Deriche, M. (2017). A two-stage hierarchical bilingual emotion recognition system using a hidden Markov model and neural networks. Arabian Journal for Science and Engineering, 42(12): 5231-5249. https://doi.org/10.1007/s13369-017-2742-5

[11] Bandela, S.R., Kumar, T.K. (2017). Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, pp. 1-5. https://doi.org/10.1109/ICCCNT.2017.8204149

[12] Gupta, H., Gupta, D. (2016). LPC and LPCC method of feature extraction in Speech Recognition System. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, pp. 498-502. https://doi.org/10.1109/CONFLUENCE.2016.7508171

[13] Chamoli, A., Semwal, A., Saikia, N. (2017). Detection of emotion in analysis of speech using linear predictive coding techniques (LPC). In 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 1-4. https://doi.org/10.1109/ICISC.2017.8068642

[14] Ram, R., Palo, H.K., Mohanty, M.N. (2013). Emotion recognition with speech for call centres using LPC and spectral analysis. International Journal of Advanced Computer Research, 3(3): 189.

[15] Langari, S., Marvi, H., Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. Informatics in Medicine Unlocked, 20: 100424. https://doi.org/10.1016/j.imu.2020.100424

[16] Subhashree R., Rathna G.N. (2016). Speech emotion recognition: Performance analysis based on fused algorithms and GMM modelling. Indian Journal of Science and Technology, 9(11): 1-8. https://doi.org/10.17485/ijst/2016/v9i11/88460

[17] Yu, C., Tian, Q., Cheng, F., Zhang, S. (2011). Speech emotion recognition using support vector machines. In International Conference on Computer Science and Information Engineering, pp. 215-220.

[18] Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., Lausen, B. (2018). Ensemble of a subset of kNN classifiers. Advances in Data Analysis and Classification, 12(4): 827-840. https://doi.org/10.1007/s11634-015-0227-5

[19] [RAVDESS, CREMA-D, TESS and SAVEE] MFCCs for SER (2021). Your Machine Learning and Data Science Community/Datasets (www.kaggle.com).