



## Multi-Attribute Feature Extraction and Selection for Emotion Recognition from Speech Through Machine Learning



Kummari Ramyasree<sup>1,2</sup>, Chennupati Sumanth Kumar<sup>1\*</sup>

<sup>1</sup> Department of E&ECE, GITAM Deemed to be University, Visakhapatnam 530045, AP, India

<sup>2</sup> Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad 501506, Telangana, India

Corresponding Author Email: [schennup@gitam.edu](mailto:schennup@gitam.edu)

<https://doi.org/10.18280/ts.400126>

### ABSTRACT

**Received:** 2 August 2022

**Accepted:** 10 December 2022

#### Keywords:

*speech emotion recognition, spectral, wavelet, prosodic, correlation analysis, fisher criterion, SVM, recognition rate*

Speech-based emotion recognition is still challenging due to its complexity despite being widely used in applications relating to emotions. In this paper, we developed a framework by considering three features: Prosodic features, Wavelet, and Spectral features. Under Prosodic, pitch and energy are considered, while under wavelet features, the approximation and detailed sub-bands at various scales are considered. Mel-Frequency Cepstral Coefficients (MFCC), Formants, and Long-Term Average Spectrum (LTAS) are all measured from speech signals as part of spectral features. Further, the significant features are selected based on nonlinear statistics, and dimensionality reduction is accomplished through Fisher Criterion. Spearman Rank Correlation is employed to find the nonlinear statistics under correlation analysis. For categorization, a Support Vector Machine and Decision Tree are used. The proposed method is simulated over RAVDESS, SAVEE, EMOVO, and URDU databases, and the observed recognition rates are approximately 79.66%, 88.99%, 87.68%, and 95.78%, respectively.

## 1. INTRODUCTION

Humans exhibit their emotions, such as fear, happiness, surprise, sadness, contempt, etc., in everyday life in response to the various events they encounter. Emotions have a direct relation to our mental health, and they have a significant effect on decision-making. Furthermore, researchers in different multi-disciplinary domains (ex. Cognitive science, neurology, and psychology) have put considerable interest in understanding, studying, and analyzing human emotions [1]. Further, Human emotions are the major indicators in the analysis of human behavior. They can be applied in different applications like Human Behavior Analysis, Human-Computer Interaction, Lie detection in Crime analysis, etc. In all these intelligent applications, emotion identification is very important. For example, in HCI, if the computer is able to identify the emotion of the operator properly, then it can process the next consecutive operations perfectly. Similarly, for lie detection in crime analysis, emotion has a major role and can be diagnosed by proper emotion analysis. Hence, emotion analysis is very important for several real-time applications.

The rapid advances in artificial intelligence technology have led to an increased research interest in systems that can recognize the feelings of human beings. There are widespread applications for such kinds of techniques that can identify the emotions of human beings. Some applications include Biomedical, Engineering [2], Human-computer interaction [3], etc. The cognitive appraisal theory claims that people's interpretations of specific events and decisions about the corresponding situations, whether positive or negative, can affect how well they accomplish their objectives or goals, which determines their emotional responses to those situations

[4].

Nonetheless, the general emotional states happen in conjunction with several psychological changes in the body's functions like voice, facial expressions, hormone levels, skin temperature, respiration, brain signals, breathing rate, heart rate, etc. Hence, emotions can be regarded as complex mental states linked with body reactions. Physiological signals, including Electroencephalogram (EEG), Electromyography (EMG), Electrocardiography (ECG), Respiration Rate (R.R., Galvanic Skin Response (GSR) as well, and facial expressions, can be used to analyze the emotional state of human beings [5]. Among these signals, speech and facial expression have been utilized mainly by researchers for emotional state identification. Furthermore, compared to the image, the speech signal is complex free and deals with more compact dynamic information. Further, a person's speech can be recorded much more quickly than other signals requiring additional equipment [6]. Due to its wide range of applications, including call centers, Smart T.V.s, computer games, robot interactions, criminal investigations [7-9], and psychological medical diagnostics [10, 11], Speech Emotion Recognition (SER) has attracted a lot of attention.

### 1.1 Motivation

The main motivation behind this research is to lessen the burden on the manual emotion analysis methods, which take more time for analysis. Furthermore, traditional methods also suffer from huge manual complexity. Hence, we were motivated to develop an automatic emotion detection mechanism based on speech signals. Even though several methods were developed earlier for Speech-based Emotion recognition, they consider a limited set of features to describe

the emotion through speech signals. Most of the methods employed the basic prosodic and statistical features, which have limited recognition performance. Hence, we consider three sets of features from three different aspects such that the recognition system can get more knowledge about emotion attributes and recognize the emotions more accurately.

In this paper, we propose a new speech emotion recognition system by considering the multiple attributes of emotions from speech signals as features. At feature extraction, we presented to extract three sets of features: prosodic, spectral, and Wavelet features. We used a novel correlation analysis-based method for feature selection, using the Fisher criterion for dimensionality reduction. For classification, we employed ensemble learning by combining two supervised learning algorithms: Support Vector Machine and Decision Tree. The significant contributions of this work are outlined as follows.

- This work proposes a hybrid feature extraction method composed of three features representing emotion in multiple orientations. These features explore more and adequate information about feeling.
- To select only informative features, this work employed a correlation-based dependency assessment of one component on another such that their effect on emotion recognition is analyzed.
- To lessen the computational complexity, this work proposes finding the linear relation between inter-class speech signals and representing the emotion in low-dimensional space.
- This work executes different speech datasets to explore the robustness, and the performance is analyzed through several metrics.

Section II of the article explores the technicalities of the existing literature. The other sections of the paper are organized as follows. In Section III, the proposed methodology is thoroughly covered. Section IV explores the details of experimental validation on different datasets, and section V provides the concluding remarks.

## 2. LITERATURE SURVEY

A speech emotion recognition (SER) system with three stages—feature extraction, dimensionality reduction, and feature classification—was proposed by Daneshfar et al. [11]. Pitch prosodic feature, Mel-frequency Cepstral Coefficient (MFCC), and perceptual minimal variance distortion less response (PMVDR) coefficient is used to generate a feature vector from the speech signal and its glottal-waveform for each frame (F0). In addition, the feature vector is given first and second-order derivatives to create a high-dimensional hybrid feature vector. In the second stage, the authors developed a novel quantum-behaved particle swarm optimization (QPSO) method for dimensionality reduction that makes new particles using truncated Laplace distribution (TLD). The final step uses a neural network classifier based on the Gaussian elliptical basis function (GEBF) to identify the type of speech emotion. Abdel-Hamid [12] attempted to identify emotion using database-evaluated spectral, prosodic, and Wavelet features. The intensity, pitch, formants, and MFCC are employed in addition to the long-term average spectrum (LTAS) set parameters. Three stages make up Turker et al. [13] proposed work: multi-level feature generation using Tunable Q wavelet transform (TQWT), twine shuffle pattern (twine-shufpat) for feature generation, and iterative

neighborhood component analysis (INCA) selection and classification of discriminative features.

To extract new Cepstral features, two new triangular filter banks (TFBs): TFB-B (TFB-bark) and TFB-E (TFB-equivalent rectangular bandwidth (ERB)) are proposed by Nagarajan et al. [14]. Existing filter banks like the Mel-filter Bank (TFB-M) and the Human-factor Filter Bank (TFB-HF) are used along with the proposed two filter banks to extract four various kinds of TFBCC (TFB Cepstral Coefficients) features. Ozer [15] proposed a spectrogram-based SER system that includes log-power rate map features as an additive feature. They used the threshold function to concentrate on highly spectral energy regions and focused on a low-frequency region with the help of a rate map. Along with the above features, the variance between the parts and effects of user-dependent features is reduced by smoothing the spectral peak. Albanie et al. [16] proposed SER based on a speech-embedding technique without accessing any form of labeled audio. They proposed to use a simple hypothesis, i.e., the emotional content of speech matches the speaker's facial expression.

In three stages—feature extraction, feature selection, and classification—Er [17] innovative's hybrid SER architecture was proposed based on acoustic and deep features. Acoustic features, including Zero-crossing Rate, MFCC, and Root Mean Square Energy (RMS), are taken out of voice recordings in the first step. Spectrogram images are used as the input to pre-trained deep network architectures like VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201, from which deep features are retrieved. Additionally, a hybrid feature vector is produced by fusing deep features with acoustic characteristics, and the Relief algorithm is then used to select the most valuable features from this vector. Support vector machines are employed as a classifier in the final stage. Hamsa et al. [18] proposed a speaker-independent & text-independent SER system for real applications where the speech is noisy and talking conditions like stressful. They modeled the work with the combination of the pre-processing stage as pitch-correlogram-based noise reduction, the feature representation method as sparse-dense decomposition, and Random Forest as a classifier.

Christy et al. [19] proposed an SER system after applying fast Fourier transformation to acoustic speech signal frames, modulation spectral (M.S.) & MFCC to extract relevant features. Using methods such as linear regression, decision trees, random forests, support vector machines (SVM), and convolution neural networks, classification and prediction tasks are carried out after the pertinent properties are chosen from the audio stream (CNN).

The SER framework for evaluating similarities in clusters created by Sajjad and Kwon [20] includes key sequence segment selection based on radial-based function networks (RBFN). The STFT approach converts selected sequences into spectrogram pictures, which a CNN model subsequently processes to extract features. After normalizing the CNN features, temporal information is learned by moving the components to the deep bi-directional long-term memory (BiLSTM). Hamsa et al. [21] proposed a noise and interference-resilient SER system. The speaker's emotion is examined by considering the energy, spectral, and time parameters. The proposed work uses a wavelet packet transform (WPT) based cochlear filter bank.

Further Random Forest classifier is used for the classification of emotions. Shahin et al. [22] proposed a hybrid

classifier by cascading the Gaussian mixture model and deep neural network (GMM-DNN) for speaker-independent and text-independent SER. Sun [23] proposed a novel emotion recognition system that relies on speaker gender information rather than acoustic speech parameters. To improve accuracy, the gender information of the speaker is added. The Gender information block and residual convolution neural network (R-CNN) are combined, and the raw speech is passed through the above two blocks concurrently. Analyzing emotion is done based on R-CNN after getting the relevant information from the natural speech data.

To depict emotions, Kadiri and Alku [24] used excitation parameters such as the strength of excitation, the instantaneous fundamental frequency, and the energy of excitation. The Kullback-Leibler (K.L.) distance is measured to determine how closely emotional and neutral speech feature distributions resemble one another. The system bases its decision on the K.L. distance between the test speech and an utterance made by the same speaker in a neutral state. An autonomous training approach that makes use of deep learning models and human expertise was proposed by Zhong et al. [25]. Two distinct models—one an attention-based convolution long short-term memory neural network and the other a fully connected neural network—were each provided the MFCC and open SMILE features by the authors. Additionally, they created a feedback approach for each model to differentiate automatically taught features (ALFs) from empirically learned features (ELFs). A dynamic fusion framework using statistical features based on spectrograms and empirical characteristics based on auditory input was proposed by Guo et al. [26]. The classifier Kernel Extreme Learning Machine (KELM) distinguishes between emotions.

Summary: In summary, the entire earlier developed SER methods are categorized into two categories based on the feature extraction methods employed by them; they are handcrafted features and deep learning features. The former category employed different methods like prosodic, MFCC, Wavelet, Fourier transform, etc. for feature extraction, while the latter category applied CNNs, RNNs, etc. for feature extraction followed by classification. From the past survey we observed that the deep learning-based methods are application specific and can't explore the full pledged discrimination between the emotions. Even though the complexity of handcrafted methods is more, they can provide sufficient discrimination between emotion and helps in the improvisation of accuracy. Moreover, speech signals re 1-D signals while Deep learning is more optimal for 2-D signals like images and videos, etc.

### 3. PROPOSED APPROACH

#### 3.1 Overview

As depicted in Figure 1, the proposed method comprises three phases: pre-processing, feature extraction, and classification. During pre-processing, the input speech signal is subjected to segmentation to divide into short-time segments. During the segmentation process, the size of the segment is kept in mind so that the overlapping is done so that no information loss is incurred. After the segmentation, each segment is subjected to feature extraction. At feature extraction, we employed three sets of feature extraction methods to represent the emotional information in each

segment. The three different features are namely Prosodic, Spectral, and Wavelet features. Once the segment is represented with three sets of features, they are subjected to feature selection followed by dimensionality reduction. We employed correlation analysis in the feature selection process, and for dimensionality reduction, we employed the Fischer criterion. Finally, the features with reduced dimensions are concatenated and represented with the final feature vector and then processed through an ensemble classifier to get the emotion label. At Ensemble, we employed two classifiers: a support vector machine and a decision tree. The SVM separates emotion from neutral. Next, the decision tree separates each emotion at one branch and, finally, the last emotion.

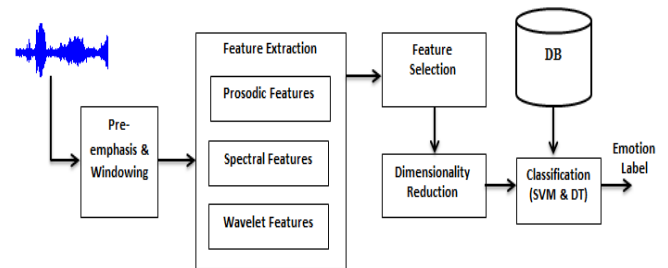


Figure 1. Overall block diagram of the proposed method

#### 3.2 Feature extraction

Before feature extraction starts, pre-emphasis is utilized to get the high-frequency components' effective resolution. Since the spectrum of the voice samples contains more energy at low frequencies than at high ones, pre-emphasis must be applied to obtain effective emotional characteristics. The point at high frequencies needs to be increased to balance the range of speech sounds. For pre-emphasis, we used a coefficient of 0.97, and the filter is defined in Eq. (1).

$$H(z) = 1 - 0.97Z^{-1} \quad (1)$$

The structure of speech signals changes concerning time and changes in the vocal tract. Therefore, speech is considered a non-stationary signal with a broad frequency range. However, for a brief time segment [27], it is believed that the signal properties are stationary. The choice of frame size is crucial because, as frame length increases, signal characteristics may change even within a frame. The overall frame size is maintained at 20 ms, and the overlapping is calculated as 50% of the frame size. Every frame has a window applied to it to make it smoother, and in this case, we employed a hamming window to keep the frames from having these discontinuities. We extract three-voice signal frames for each window frame. We extract three sets of features—prosodic characteristics, spectral features, and wavelet features—for each window frame of the speech signal. The following subsections investigate the specifics of computed features.

##### 3.2.1 Prosodic features

Acoustic features that are computed directly from discrete speech signals include prosodic features. Fundamental frequency ( $f_0$ ) or pitch and intensity are examples of prosodic features. It is believed that the vibration of a speaker's vocal cords produces the pitch known as  $f_0$ . Pitch ranges differ across individuals and between emotions. Adult females

typically have a higher range of pitch than adult males. When it comes to emotions, anger emotion is more intense than other emotions. Pitch was measured using the autocorrelation method in this study, with males' pitch ranges being 75–300 Hz and females' ranges being 100–500 Hz. Next, the volume of the speech is referred to as the intensity (energy). Due to their correlation with the speaker's emotion, pitch and intensity are employed frequently in speech emotion recognition [28]. Along with pitch and intensity, several statistical parameters that indicate fluctuations in pitch and intensity, such as mean, standard deviation, maximum, minimum, and range, are also computed.

### 3.2.2 Spectral features

Three spectral features—formants, MFCC, and LTAS—are incorporated in this paper. Formants are a representation of the vocal tract's resonance frequency at the moment when high energy is at its height. Formants are frequently employed in speech emotion recognition because they vary with emotion [29]. The relevance of low-frequency components relative to high-frequency components is next explored by measuring the MFCC from a nonlinear Me-scale. They are frequently employed in speaker and speech recognition systems because, by being sensitive to sound fluctuations at lower frequencies, they resemble the human auditory systems that correct the impact of pitch and the logarithmic signal power density of voiced sections of speech signals [30, 31]. In addition, the computational complexity of LTAS is lower than that of MFCC. The features of a segment are the last three formants, the average of the 12 MFCCs, the mean of the LTAS, and the range, maximum, and minimum.

### 3.2.3 Wavelet features

The wavelet transform is a multi-resolution analysis technique for studying acoustic data [32, 33]. Wavelet evaluation Create two sub bands from the input speech signal: an approximations sub band and a detailed sub band. After the voice signal has passed through a low pass filter and a high pass filter, respectively, these sub-bands are obtained. The voice signal is localized via the Wavelet Transform in both the temporal and frequency domains. Additionally, the sub-band wavelet transform supports the subsequence coefficients in various scales. In this study, Debauches 4 is used to perform the decomposition up to four levels (db4). Then, the entropy and energy are determined further for the approximation and precise sub-bands from the four scales [34].

Finally, after three sets of features are extracted, they are merged and formulated into a single feature vector. Then the final feature vector is passed for feature selection, followed by dimensionality reduction.

## 3.3 Feature selection

Feature selection executes to determine the importance of features during the feature selection. After extracting features from the input speech signal, the process for feature selection and only informative features are chosen, and the remaining features are discarded [35, 36]. At this phase, the significance of features is computed. In this work, we apply correlation analysis to the feature selection process. Further, we use the Fisher criterion, a linear discriminate analysis method for feature dimensionality reduction. At feature selection, initially, the features are selected based on Euclidean distance analysis, partial correlation analysis, and vicariate correlation analysis.

Then, the obtained resultant features are subjected to the fisher criterion to get the final features with reduced dimensions.

### 3.3.1 Correlation analysis

At this phase, Euclidean distance analysis is accomplished, and the entire features are divided into several groups. Then each group is subjected to partial correlation analysis to find the correlation between features in each group. Then the resultant characteristics are assessed for Spearman rank correlation (SRC) analysis to determine the final feature set. After the commencement of correlation analysis, the consequent features are more evident regarding emotion.

### 3.3.2 Euclidean distance analysis

Numerous characteristics can be used to describe a person's emotional state, but it is challenging to employ all of the attributes at once to conduct emotion recognition. Therefore, it is necessary to ascertain which characteristics are essential and substantially influence emotions and control [37]. All of the features are first analyzed to determine how they relate to other features, and then, based on the outcome, they are divided into several categories. Here, grouping the characteristics by distance analysis is used. In a feature set of  $n$  dimensions, the Euclidean distance illustrates the actual distance between two locations. The Euclidean distance between two points  $(x_1, y_1)$  and  $(x_2, y_2)$  is expressed in Eq. (2).

$$E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

where,  $E$  stands for the Euclidean distance, features with lower  $E$  values are grouped into the same cluster. In  $n$ -dimensional space,  $X=x_1, x_2, \dots, x_n$  and  $Y=y_1, y_2, \dots, y_n$  are the points. The closest-proximity features are selected as comparable features.

### 3.3.3 Partial correlation analysis

It might be challenging to determine how features affect the emotional state because numerous emotional aspects generally have the same traits as an emotional state. It is vital to remove or control the characteristics that adversely affect the other features before studying the relationship between features and their sentiments. Such analysis can be regarded as partial correlation analysis [38] or net correlation analysis. This kind of analysis determines the effect of one feature on the other based on the linear relation between them. Consider the group of independent variables as  $X=\{x_1, x_2, \dots, x_n\}$  the partial correlation is computed using Eq. (3).

$$R = (\rho_{ij})_{n \times n} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{bmatrix} \quad (3)$$

For the above Matrix, the inverse is calculated using Eq. (4).

$$R^{-1} = (\lambda_{ij})_{n \times n} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nn} \end{bmatrix} \quad (4)$$

Finally, the partial correlation between the two variables is calculated using Eq. (5).

$$Y_{ij} = \frac{-\lambda_{ij}}{\sqrt{\lambda_{ii}}\sqrt{\lambda_{jj}}} \quad (5)$$

So the partial correlation Coefficient defines the dependency of two independent variables. It indirectly shows how much they depend on and the necessity of selection or removal.

### 3.3.4 Nonlinear correlation analysis

For the computation of the Correlation Coefficient between two variables, several measures are available Pearson Product Moment Linear Correlation Coefficient (PLCC) [39], Kendall Coefficient of Concordance (KCC) [40, 41], and a Spearman Rank Correlation (SRC) [42]. But the features extracted from every frame in each interval are fully ranked and discrete variables, and to find their rank, we employed the SRC method. SRC Coefficient is one type of index that explores the statistical correlation between two variables under a monotonic function. For the variables that are strictly monotonic to each other, the SRC Coefficient is either +1 or -1 and called a variable and complete spearman correlation. Consider two variables  $X=\{x_1, x_2, \dots, x_n\}$ , and  $Y=\{y_1, y_2, \dots, y_n\}$ . SRC coefficient between them is calculated using Eq. (6).

$$\rho_s = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (6)$$

where,  $X_i$  is the  $i^{th}$  feature in the X variable, and  $Y_i$  is the  $i^{th}$  feature in the Y variable.  $\bar{X}$  and  $\bar{Y}$  are the Mean of X and Y, respectively. The SRC coefficient is generally measured as a non-parametric correlation coefficient. SRC Coefficient measures the accurate distribution of X and Y samples without knowledge about their joint probability distribution. SRC Coefficient exists between X and Y up to their monotonous relation. The SRC differs from PLCC, which is purely based on Statistics of linearity.

### 3.3.5 Fisher criterion

During the accomplishment of statistical methods for the applications related to pattern recognition, several problems exist due to the feature set's dimensionality. The ways that work in a low dimensional space have less computational complexity and can show an optimal performance. The features obtained at the feature selection are more significant and are transformed into a lower-dimensional space with less information loss. The primary issue of dimensionality reduction is information loss. So to get an optimal feature set with low dimensional space, we apply Fisher Criterion, which determines the linear relation-based dimensionality reduction. PCA is one more popular method for dimensionality reduction, but it cannot extract discriminative information from high-dimensional emotional features. In this work, we apply PCA and Fisher Criterion, and from the results, we found the optimal performance of the Fisher Criterion. Mathematically, the Fisher Criterion is calculated using Eq. (7) [43-45].

$$\lambda_F = \frac{\sigma_B}{\sigma_W} \quad (7)$$

where,  $\lambda_F$  is Fisher's rate of features,  $\sigma_B$  is the variance between different classes and  $\sigma_W$  is the variance within the class.  $\sigma_B$  is defined in Eq. (8).

$$\sigma_B = \sum_{c=1}^N (E_c - \bar{E})(E_c - \bar{E})^T \quad (8)$$

where,  $\bar{E}$  is the mean of the entire data set and is defined in Eq. (9).

$$\bar{E} = \frac{1}{M} \sum_{i=1}^M x_i \quad (9)$$

And  $E_c$  is the sample mean for  $i^{th}$  Emotion class  $E_i$ , defined in Eq. (10).

$$E_c = \frac{1}{N_p} \sum_{x \in E_c} x_i \quad (10)$$

The term M in Eq. (9) is the total number of emotions, and the term in Eq. (10) is the total number of samples in the emotional speech signal. Similarly, it is mathematically defined in Eq. (11).

$$\sigma_W = \sum_{c=1}^N \sum_{i=1}^{N_p} (x_i - E_c)(x_i - E_c)^T \quad (11)$$

Then they obtained distribution matrix  $\sigma_W$  is subjected to dimensionality reduction to remove the unnecessary feature while preserving the important information.

## 3.4 Classification

We employed two machine learning algorithms for classification purposes, such as the SVM and decision tree. The simple classification through SVM and decision tree is shown in Figure 2.

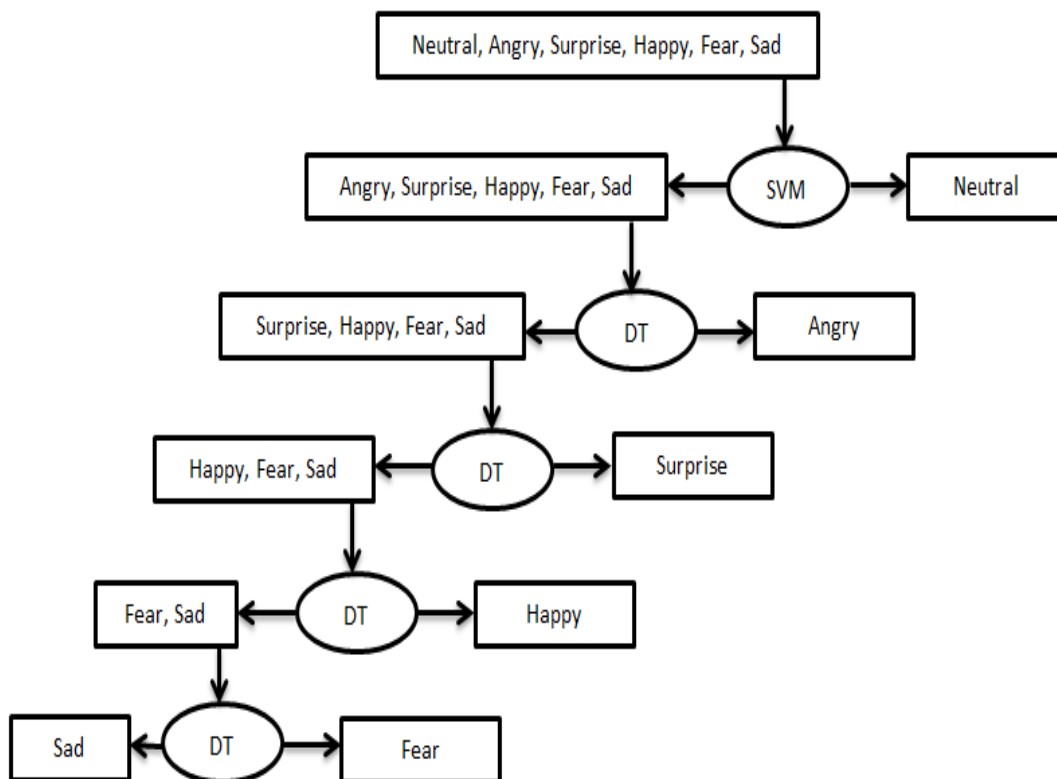
The SVM is a binary and nonlinear classifier, and it classifies the signal into two classes' either neutral or emotional. Since the features of neutral much deviate from emotional features, we employed a nonlinear algorithm for classification. Next, we used a decision tree algorithm for further classification. In the testing phase, if the input speech signal label is determined as emotional, it is processed for a binary tree for further emotional determination.

## 4. RESULTS AND DISCUSSION

### 4.1 Speech emotion databases

For the experimental validation of the proposed method, we referred to four standard databases, namely RAVDESS database [46], SAVEE Database [47], EMOVO Emotional speech corpus [48], and the Urdu database [49]. The details of these databases are shown in Table 1.

RAVDESS database consists of eight emotions: surprise, disgust, fear, anger, sadness, happiness, calm, and neutral. It is created with the help of 24 professional actors, among which 12 are male actors, and 12 are female actors. This database is created by asking actors to act for emotion and talk. This is an audio-visual database created based on audio-only and video-only modules. All the feelings are acquired under two modes such as strong and regular. The speech file has been recorded with a sampling rate of 48 kHz and encoded through 16 bits. The audio database consists of a total of 1440 speech files, with a total of 60 utterances that were uttered in North American English. The total number of emotions in this database is eight; they are neutral, calm, sad, happy, fearful, disgusted, surprised, and angry.



**Figure 2.** Classification through ensemble learning

**Table 1.** Different speech emotion databases and their attributes

Feature/Database	RAVDESS	SAVEE	EMOVO	Urdu
<b>Total number of emotions</b>	8	7	8	4
<b>Emotions</b>	neutral, calm, sad, happy, fearful, disgusted, surprised, anger	surprise, sadness, neutral, happiness, fear, disgust and anger	surprise, sadness, joy, neutral, happiness, fear, disgust and anger	happy, angry, neutral, and sadness
<b>Speakers</b>	24 (12 M and 12 F)	4 (M)	6 (3M and 3F)	38 (27M and 11F)
<b>Number of samples</b>	1440	480	588	400
<b>Language</b>	North American English	British English	Italian English	Urdu
<b>Sampling rate</b>	48 kHz	44.1 kHz	48 kHz	44.1 kHz
<b>Advantage</b>	Free of available and all basic emotions are acquired	Free of available and all basic emotions are acquired	Free of available and all basic emotions are acquired	Free of available and emotions are pretty natural
<b>Disadvantage</b>	Subjects are only adult	Subjects are only adult	Subjects are only adult	All basic emotions are not covered

SAVEE database is also an acted database acquired based on the Utterances spoken in British English. This audio-visual database was developed with the help of four male actors in the Visual media lab. The total number of emotions present in this database is seven; they are namely: surprise, sadness, neutral, happiness, fear, disgust, and anger. This database consists of a total of 480 speech audio files, and they are acquired at the sampling rate of 441.1 kHz and encoded with 16 bits. The utterances of this database are validated through ten subjects to validate the quality of the emotion corpus.

EMOVO database is an acted database created with the help of six professional actors by making them utter 14 Italian sentences in a total of seven emotional states. This database is created in the Fondazione Ugo Bordoni laboratory. The total number of emotions present in this database is seven; they are namely: surprise, sadness, joy, neutral, happiness, fear, disgust, and anger. The total number of audio files present in this database is 580, and all of them are acquired at the 48 kHz sampling rate and encoded with 16 bits.

Urdu database in one more speech database captured from the Urdu T.V. talk shows is regarded as a natural database as it is acquired naturally. This dataset considered a total of 38 speakers, with 11 being females and 27 being males. The total number of emotions in this database is seven: happy, anger, neutral, and sadness. This database consists of 100 speech files for every emotion, and they are acquired at the 44.1 kHz sampling rate with 16-bit encoding.

#### 4.2 Results and discussion

Under the results section, we explore the details of results derived after the simulation of the proposed methodology on different types of databases. At every validation, we represent the results in a confusion matrix in which the diagonal elements denote the correct recognized samples. We describe the results in a separate table for each database, shown below. Table 2 shows the confusion matrix of the proposed method on the RAVDESS database. Next, Tables 3, 4, and 5 show the

confusion matrices of the proposed method on SAVEE, EMOVO, and Urdu databases, respectively. In each table, the diagonal elements represent the True Positives (T.P.s), which means the correctly recognized emotions. The last value in each row denotes the total number of samples used for testing. Next, the previous value in each Column represents the total number of signals resulting in testing as corresponding emotion. In each row, expect the T.P.; the sum of the remaining values denotes the False Negatives (F.N.s). Similarly, in each column, expect the T.P.; the sum of the remaining values denotes the False Positives (F.P.s). Based on these values, the performance metric is measured. Here we used four performance metrics for performance assessment: Recall, Precision, F1-Score, and FNR.

As shown in the Table 2 (confusion matrix formed based on the results obtained after the simulation of the proposed method on the RAVDESS database), more T.P.s are observed for calm emotion as the proposed system recognized 155 samples out of 184 samples. Next, the least T.P.s are determined to have Sad emotions; only 115 are correctly identified out of 175 samples processed for testing. From the confusion matrix shown in the Table 3 belongs to the results of the proposed method after its accomplishment on the SAVEE database; the more number of T.P.s are observed for Neutral as the proposed system recognized a total of 115 samples out of 120 samples. Next, the least T.P.s are determined at Surprise emotion; only 40 are correctly identified out of 58 samples processed for testing. Next, from

Table 4, the maximum T.P.s are observed for Sad emotion since the total number of T.P.s are 75 and the total samples processed for testing are 85.

Further, the Joy emotion is recognized less in number as its T.P.s are only 60, but the samples processed for testing are 80. For the simulation of the Urdu database, we have used an equal number of samples for testing each emotion. From the results in the confusion matrix Table 5, the happy emotion is recognized perfectly as its T.P.s are 100 out of the test samples 100. The least number of T.P.s are observed at sad emotion, as it resembles the neutral emotion.

Further, to check the performance effectiveness of the proposed method, we have simulated the emotions in each database with individual features such as Prosodic features, Spectral features, and Wavelet features. Next, the same emotions are processed through hybrid features by merging them as a single feature vector, followed by feature selection through correlation analysis and fisher criterion. The recall rate observed in this simulation is demonstrated in Figure 3. From the results, the maximum recall is monitored by the proposed hybrid features as it reflects the attributes of all features. Next is the Figure 4 demonstrates the precision, and Figure 5 demonstrates the FNR at different features. From the database point of view, we can see that poor performance is observed in the RAVDESS database, and the ultimate version is observed in the Urdu database. The remaining two databases, SAVEE and EMOVO, maintained a better performance.

**Table 2.** Confusion matrix of the proposed method on the RAVDESS database

	Neutral	Calm	Sad	Happy	Fear	Disgust	Surprise	Anger	Total
Neutral	62	6	17	3	0	3	3	0	94
Calm	15	155	10	0	0	4	0	0	184
Sad	11	10	115	10	10	10	9	0	175
Happy	0	0	8	114	20	5	15	10	172
Fear	0	0	15	18	118	5	8	7	171
Disgust	3	0	15	5	3	130	6	10	172
Surprise	2	0	9	15	13	9	122	3	173
Anger	0	0	0	11	5	10	11	140	177
Total	93	171	189	176	169	176	174	170	1318

**Table 3.** Confusion matrix of the proposed method on the SAVEE database

	Neutral	Sad	Happy	Fear	Disgust	Surprise	Anger	Total
Neutral	115	3	0	0	2	0	0	120
Sad	6	52	0	0	0	0	0	58
Happy	0	0	45	8	0	2	3	58
Fear	2	0	3	45	1	9	4	64
Disgust	4	2	0	2	45	0	3	56
Surprise	2	0	5	10	0	40	1	58
Anger	0	1	5	3	0	0	52	61
Total	129	58	58	68	48	51	63	475

**Table 4.** Confusion matrix of the proposed method on the EMOVO database

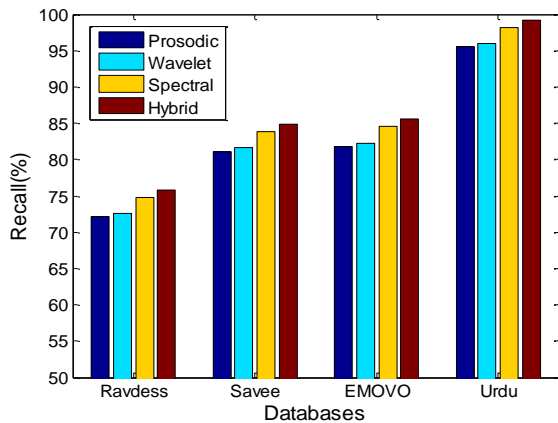
	Neutral	Sad	Joy	Fear	Disgust	Surprise	Anger	Total
Neutral	70	2	4	0	5	0	0	81
Sad	3	75	0	5	0	2	0	85
Joy	2	2	60	2	3	3	8	80
Fear	0	3	0	61	6	9	0	79
Disgust	3	0	3	3	65	3	0	77
Surprise	0	1	5	9	4	60	0	79
Anger	1	1	5	1	2	2	71	83
Total	79	84	77	81	85	79	79	564

**Table 5.** Confusion matrix of the proposed method on the Urdu database

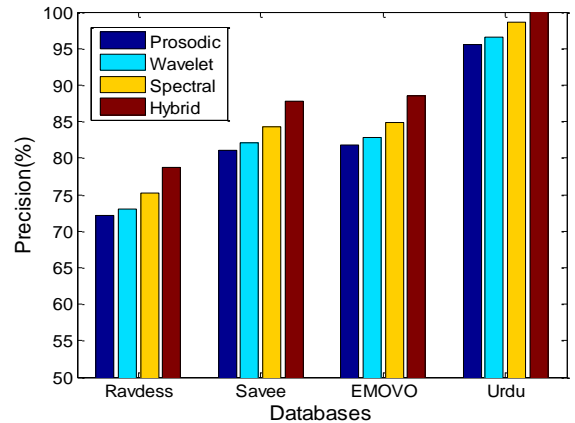
	Neutral	Sad	Happy	Anger	Total
Neutral	98	2	0	0	100
Sad	5	95	0	0	100
Happy	0	0	100	0	100
Anger	1	1	1	97	100
Total	104	98	101	97	400

**Table 6.** Comparison between proposed and conventional methods

Reference	Features	Classifier	Database	Recognition Rate (%)
Daneshfar et al. [11]	1. PMVDR, PLPC and MFCC 2. QPSO for feature selection	GEBF neural Network	EMODB, SAVEE and IEMOCAP	SAVEE-59.38, EMOVB-76.81, IEMOCAP – 65.71
Abdel-Hamid [12]	1. Wavelet features, MFCC and Prosodic features	SVM with linear kernel and k-NN	EYASE	90.70
Turker et al. [13]	1. TQWT features 2. INCA for features selection	3 <sup>rd</sup> -degree polynomial kernel SVM	RAVDESS Speech, Berlin, SAVEE, and EMOVO).	RAVDESS-87.43, Berlin –90.09, SAVEE – 84.79, and EMOVO – 79.08
Er [17]	RMS, MFCC and ZCR	VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201	RAVDESS, Berlin, and IEMOCAP	RAVDESS – 79.41, Berlin –90.21 IEMOCAP – 85.37
Christy et al. [19]	MFCC	Random forest, Decision trees, SVM CNN	RAVDESS	RF –72.35 DT –62.33 SVM – 68.96 CNN –78.20
Sajjad and Kwon [20]	STFT and CNN features	Bi-LSTM	IEMOCAP, EMO-DB, and RAVDESS	IEMOCAP – 72.25, EMO-DB – 85.57, RAVDESS – 77.02
Kadiri and Alku [24]	Excitation features	KL-Distance	EMO-DB IIIT-H	EMDDB – 77.33 IIITH – 83.02
Proposed	1. Prosodic, Spectral, and Wavelet features 2. Correlation analysis for feature selection 3. Fisher criterion for dimensionality reduction	SVM and Decision Tree	RAVDESS, SAVEE, EMOVO and Urdu	RAVDESS –79.66 SAVEE – 88.99 EMOVO – 87.68 Urdu – 95.78



**Figure 3.** Recall of proposed method for the simulation of different databases with different features

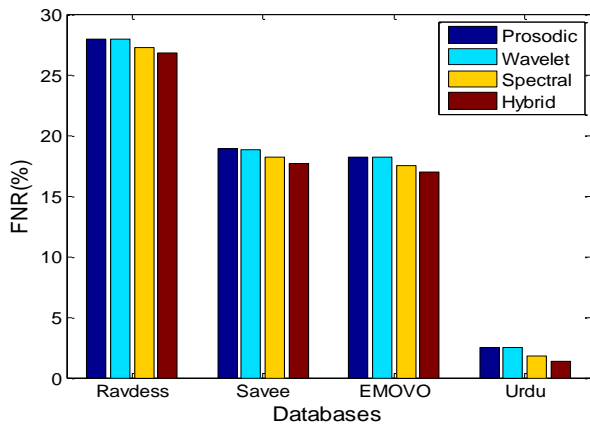


**Figure 4.** Precision of the proposed method on simulation of different databases with different features

From Figure 3, the maximum recall is observed in the Urdu database at the hybrid feature combination, which is approximately 98.5523%. In contrast, the least recall rate is kept in the RAVDESS database, about 77.3365%. The SAVEE and EMOVO databases have gained an approximate recall of 81.3316% and 81.5678%, respectively, at the Hybrid feature set simulation. Next, From Figure 4, the maximum precision is observed at the Urdu database's hybrid feature

combination, which is approximately 99.5856%. In comparison, the least accuracy is observed in the RAVDESS database, about 79.6325%. The SAVEE and EMOVO databases have gained an approximate recall of 87.4281% and 88.6637%, respectively, at the Hybrid feature set simulation. Finally, the least F.R. is observed for Happy emotion in Urdu; its average FNR is 2.5215%, while the larger FNR is kept in the RAVDESS database and approximated as 27.3345%.





**Figure 5.** FNR of the proposed method on simulation of different databases with different features

### 4.3 Comparison

The comparison is shown in the Table 6 reveals the effectiveness of the proposed method for emotion recognition from speech signals. Even though most methods use MFCC features for emotion representation through the human audiology system, the feature section is not done much effectively. In the feature selection process, the discrimination provision is necessary between emotional and neutral states and different emotions. The accomplishment of feature selection can provide such discrimination through correlation analysis.

The SRC coefficient between any two emotions determines their nonlinear relation, while the fisher criterion determines the linear relation between them. Hence the proposed method outperformed all the earlier forms. Even though Er [17] employed deep learning statistics for feature extraction, they showed a limited recognition rate. Deep learning algorithms constitute more complexity in the recognition system due to the complex computational processing of convolution operations. Wavelet features [12], Turker et al. [13] explore the localization of emotion in a speech by decomposing it into the low pass and high sub-bands. Still, it has significantly less significance in emotion recognition when compared to the prosodic features and spectral features like MFCC [11, 19]. Unlike all these methods, Kadiri and Alku [24] proposed applying excitation features and achieving optimal performance even on the self-created dataset IIIT-H.

## 5. CONCLUSIONS

In this study, we presented a new Speech emotion recognition system based on the fusion of multiple features, such as Prosodic, Spectral, and Wavelet features. These three features describe an emotion in multiple orientations so that the system can effectively discriminate between emotions. This study's major novelty is correlation-assisted feature selection and linear discrimination-assisted dimensionality reduction. These two methods helped to improve recognition accuracy with less computational time-space. SVM and D.T. are employed for the simultaneous categorization of emotions. Simulation on the standard databases like RAVDESS, SAVEE, EMOVO, and URDU proves the effectiveness as they achieved recognition rates of 79.66%, 88.99%, 87.68%, and 95.78%, respectively. Thus, the effectiveness of the proposed method is validated. In summary, the proposed method

combined three sets of features, while the earlier methods employed only a single set of features. Hence our method has gained an improved accuracy compared to the earlier methods. On average, the proposed methods gained an improvement of 4-5% from earlier methods over different datasets.

## REFERENCES

- [1] Swain, M., Routray, A., Kabisatpathy, P. (2018). Databases, features, and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, 21: 93-120. <https://doi.org/10.1007/s10772-018-9491-z>
- [2] Nardelli, M., Valenza, G., Greco, A., Lanata, A., Scilingo, E. (2015). Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, 6(4): 385-394. <https://doi.org/10.1109/TAFFC.2015.2432810>
- [3] Polap, D. (2018). Model of identity verification support system based on voice and image samples. *Journal of Universal Computer Science*, 24(4): 460-474. <https://doi.org/10.3217/jucs-024-04-0460>
- [4] Thagard, P. (2019). *Mind society: From brains to social sciences and professions*. Oxford University Press. United Kingdom. <https://doi.org/10.1093/oso/9780190678722.001.0001>
- [5] Mohammadi, Z., Frounchi, J., Amiri, M. (2017). Wavelet-based emotion recognition system using EEG signal. *Neural Computing and Applications*, 28: 1985-1990. <https://doi.org/10.1007/s00521-015-2149-8>
- [6] Tawari, A., Trivedi, M.M. (2010). Speech emotion analysis: Exploring the role of context. *IEEE Transactions on Multimedia*, 12(6): 502-509. <https://doi.org/10.1109/TMM.2010.2058095>
- [7] Khalil, A., Al-Khatib, W., El-Alfy, E.S., Cheded, L. (2018). Anger detection in Arabic speech dialogs. *Proceedings of the International Conference on Computing Sciences and Engineering*, pp. 1-6. <http://dx.doi.org/10.1109/ICCSE1.2018.8374203>
- [8] Meddeb, M., Karray, H., Alimi, A.M. (2017). Content-based Arabic speech similarity search and emotion detection. *Proceedings of the international conference on advanced intelligent systems and informatics*, pp. 530-539. [https://doi.org/10.1007/978-3-319-48308-5\\_51](https://doi.org/10.1007/978-3-319-48308-5_51)
- [9] Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S. (2016). Emotion recognition from audio signals using Support Vector Machine. *Proceedings of the IEEE recent advances in intelligent computational systems*, pp. 139-144. <https://doi.org/10.1109/RAICS.2015.7488403>
- [10] Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U. (2017). Speech-based human emotion recognition using MFCC. *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking*, pp. 2257-2260. <https://doi.org/10.1109/WiSPNET.2017.8300161>
- [11] Daneshfar, F., Kabudian, S.J., Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. *Applied Acoustics*, 166: 1-14. <https://doi.org/10.1016/j.apacoust.2020.107360>
- [12] Abdel-Hamid, L. (2020). Egyptian Arabic speech

- emotion recognition using prosodic, spectral and wavelet features. *Speech Communication*, 122: 19-30. <https://doi.org/10.1016/j.specom.2020.04.005>
- [13] Turker, T., Sengul Dogan, U., Acharya, R. (2021). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowledge-Based Systems*, 211: 1-15. <https://doi.org/10.1016/j.knosys.2020.106547>
- [14] Nagarajan, S., Nettimi, S.S., Kumar, L.S., Nath, M.K., Kanhe, A. (2020). Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales. *Digital Signal Processing*, 104: 1-10. <http://dx.doi.org/10.1016/j.dsp.2020.102763>
- [15] Ozer, I. (2021). Pseudo-colored rate map representation for speech emotion recognition. *Biomedical Signal Processing and Control*, 66: 1-10. <https://doi.org/10.1016/j.bspc.2021.102502>
- [16] Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. *ACM International Conference on Multimedia*, pp. 292-301. <https://doi.org/10.1145/3240508.3240578>
- [17] Er, M. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, 8: 221640-221653. <https://doi.org/10.1109/ACCESS.2020.3043201>
- [18] Hamsa, S., Iraqi, Y., Shahin, I., Werghi, N. (2021). An enhanced emotion recognition algorithm using pitch correlogram, deep sparse matrix representation and random forest classifier. *IEEE Access*, 9: 87995- 88010. <https://doi.org/10.1109/ACCESS.2021.3086062>
- [19] Christy, A., Vaithyasubramanian, S., Jesudoss, A., Praveena, M.D. (2020). Multimodal speech emotion recognition and classification using convolutional neural network techniques. *International Journal of Speech Technology*, 23(2): 381-388. <https://doi.org/10.1007/s10772-020-09713-y>
- [20] Sajjad, M., Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, 8: 79861-79875. <https://doi.org/10.1109/ACCESS.2020.2990405>
- [21] Hamsa, S., Shahin, I., Iraqi, Y., Werghi, N., Emotion. (2020). Recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier. *IEEE Access*, 8: 96994-97006. <https://doi.org/10.1109/ACCESS.2020.2991811>
- [22] Shahin, I., Nassif, A.B., Hamsa, S. (2019). Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access*, 7: 26777-26787. <https://doi.org/10.1016/j.knosys.2022.108659>
- [23] Sun, T.W. (2020). End-to-end speech emotion recognition with gender information. *IEEE Access*, 8: 152423-152438. <https://doi.org/10.1109/ACCESS.2020.3017462>
- [24] Kadiri, S.R., Alku, P. (2020). Excitation features of speech for speaker-specific emotion detection. *IEEE Access*, 8: 60382-60391. <https://doi.org/10.1109/ACCESS.2020.2982954>
- [25] Zhong, S., Yu, B., Zhang, H. (2020). Exploration of an independent training framework for speech emotion recognition. *IEEE Access*, 8: 222533-222543. <https://doi.org/10.1109/ACCESS.2020.3043894>
- [26] Guo, L., Wang, L., Dang, J., Liu, Z., Guan, H. (2019). Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access*, 7: 75798-75809. <https://doi.org/10.1109/ACCESS.2019.2921390>
- [27] Rabiner, L.R., Schafer, R.W. (2004). *Digital processing of speech signals*. Pearson Educ(Singapore) Pte. Ltd., (Indian reprint). <https://doi.org/10.1121/1.384160>
- [28] Meftah, A., Alotaibi, Y., Selouani, S.A., (2014). Designing, building, and analyzing an Arabic speech emotional corpus. *Work. Free. Arab. Corpora Corpora Process. Tools Work. Program.*, 22: 1-4.
- [29] Koolagudi, S.G., Murthy, Y.V., Bhaskar, S.P. (2018). Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, 21(1): 167-183. <https://doi.org/10.1007/s10772-018-9495-8>
- [30] Bahmanbiglu, S.A., Mojiri, F., Abnavi, F. (2017). The Impact of language on voice: An LTAS study. *Journal of Voice*, 31(2): 249.e9-249.e12. <https://doi.org/10.1016/j.jvoice.2016.07.020>
- [31] Yüksel, M., Gündüz, B. (2018). Long term average speech spectra of Turkish. *Logop. Phoniatr. Vocology*, 43: 101-105. <https://doi.org/10.1080/14015439.2017.1377286>
- [32] Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, 90: 250-271. <https://doi.org/10.1016/j.eswa.2017.08.015>
- [33] Haridas, A.V., Marimuthu, R., Sivakumar, V.G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 22(1): 39-57. <https://doi.org/10.3233/KES-180374>
- [34] Coifman, R.R., Wickerhauser, M.V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2): 713-718. <https://doi.org/10.1109/18.119732>
- [35] Atmaja, B.T., Sasou, A., Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140: 11-28.
- [36] Hans, A.S., Rao, S. (2021). A CNN-LSTM based deep neural networks for facial emotion detection in videos. *International Journal of Advances in Signal and Image Sciences*, 7(1): 11-20. <https://doi.org/10.29284/ijasis.7.1.2021.11-20>
- [37] Aparicio, J., Pastor, J.T. (2014). On how to properly calculate the Euclidean distance-based measure in DEA. *Optimization*, 63(3): 421-432. <https://doi.org/10.1080/02331934.2012.655692>
- [38] Chen, L., Zheng, S.K. (2009). Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biology*, 10(1): 1-20. <https://doi.org/10.1186/gb-2009-10-1-r3>
- [39] Derrick, T.R., Bates, B.T., Dufek, J.S. (1994). Evaluation of time-series data sets using the Pearson product-moment correlation coefficient. *Medicine and Science in Sports and Exercise*, 26(7): 919-928. <https://doi.org/10.1249/00005768-199407000-00018>
- [40] Baumgartner, R., Somorjai, R., Summers, R., Richter, W. (1999). Assessment of cluster homogeneity in fMRI data using Kendall's coefficient of concordance. *Magnetic*

- Resonance Imaging, 17(10): 1525-1532. [https://doi.org/10.1016/S0730-725X\(99\)00101-0](https://doi.org/10.1016/S0730-725X(99)00101-0)
- [41] Sedgwick, P. (2014). Statistical question spearman rank correlation coefficient. *British Medical Journal*, 349(5): 27-37.
- [42] Zar, J.H. (2015). Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339): 578-580. <https://doi.org/10.2307/2284441>
- [43] Malina, W. (1981). On an extended fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5: 611-614. <https://doi.org/10.1109/TPAMI.1981.4767154>
- [44] Yang, J., Yang, J. (2003). Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36(2): 563-566. [https://doi.org/10.1016/S0031-3203\(02\)00048-1](https://doi.org/10.1016/S0031-3203(02)00048-1)
- [45] Zhang, S.Q., Lei, B.C., Chen, A.H. (2010). Spoken emotion recognition using local fisher discriminant analysis. *IEEE International Conference on Signal Processing*, pp. 538-550. <https://doi.org/10.1109/ICOSP.2010.5656091>
- [46] Zhang, S., Lei, B., Chen, A., Chen, C., Chen, Y. (2010). Spoken emotion recognition using local fisher discriminant analysis. *IEEE 10th International Conference on Signal Processing*, pp. 538-540. <https://doi.org/10.1109/ICOSP.2010.5656091>
- [47] Haq, S., Jackson, P.J.B. (2010). *Machine audition: Principles, algorithms and systems*. Hershey: IGI Global Press, pp. 398-423. <https://doi.org/10.4018/978-1-61520-919-4>
- [48] Costantini, G., Iadarola, I., Paoloni, A., Todisco, M. (2014). EMOVO corpus: An Italian emotional speech database. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3501-3504.
- [49] Siddique, L., Adnan, Q., Muhammad, U., Junaid, Q. (2018). Cross-lingual speech emotion recognition: Urdu vs. western languages. *International Conference on Frontiers of Information Technology*, pp. 88-93. <https://doi.org/10.1109/FIT.2018.00023>