Complex & Intelligent Systems
https://doi.org/10.1007/s40747-021-00295-z

ORIGINAL ARTICLE

Emotion classification from speech signal based on empirical mode decomposition and non-linear features

Speech emotion recognition

Palani Thanaraj Krishnan[1] · Alex Noel Joseph Raj[2] · Vijayarajan Rajangam[3]

Received: 11 October 2020 / Accepted: 2 February 2021

Abstract
Emotion recognition system from speech signal is a widely researched topic in the design of the Human–Computer Interface (HCI) models, since it provides insights into the mental states of human beings. Often, it is required to identify the emotional condition of the humans as cognitive feedback in the HCI. In this paper, an attempt to recognize seven emotional states from speech signals, known as sad, angry, disgust, happy, surprise, pleasant, and neutral sentiment, is investigated. The proposed method employs a non-linear signal quantifying method based on randomness measure, known as the entropy feature, for the detection of emotions. Initially, the speech signals are decomposed into Intrinsic Mode Function (IMF), where the IMF signals are divided into dominant frequency bands such as the high frequency, mid-frequency , and base frequency. The entropy measures are computed directly from the high-frequency band in the IMF domain. However, for the mid- and base-band frequencies, the IMFs are averaged and their entropy measures are computed. A feature vector is formed from the computed entropy measures incorporating the randomness feature for all the emotional signals. Then, the feature vector is used to train a few state-of-the-art classifiers, such as Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Gradient Boosting Machine. A tenfold cross-validation, performed on a publicly available Toronto Emotional Speech dataset, illustrates that the LDA classifier presents a peak balanced accuracy of 93.3%, F1 score of 87.9%, and an area under the curve value of 0.995 in the recognition of emotions from speech signals of native English speakers.

Keywords Speech signal · Emotion perception · Entropy measures · Linear discriminant analysis · Empirical mode decomposition

Introduction

Speech signals have a huge impact on the current modes of communication, such as emails and text messages. Even in written messages, emotion representations like emojis are inserted to reveal our emotional states. Speech communication is more prominent and effective when the text communication fails to reveal the emotional states.

The speech signal is of research interest over the decades for various applications such as emotion perception, HCI, bio metrics, and so on [1]. Also, emotion analysis from the auditory signal of humans has become more prominent research due to (a) the availability of fast computing systems, (b) the effectiveness of various signal processing algorithms, and (c) the acoustic differences in speech signals that are natu-

✉ Alex Noel Joseph Raj
jalexnoel@stu.edu.cn

Palani Thanaraj Krishnan
palanithanaraj.k@gmail.com

Vijayarajan Rajangam
viraj2k@gmail.com

[1] Department of Electronics and Instrumentation Engineering, St. Joseph's College of Engineering, Chennai, India

[2] Department of Electronic Engineering, Shantou University, Shantou, China

[3] Division of Healthcare Advancement, Innovation and Research, Vellore Institute of Technology, Chennai, India

Published online: 25 February 2021

rally embedded in various emotional situations. An in-depth analysis of speech signals in different domains is helpful in recognizing the emotions from the auditory signals of the people who are unable to communicate through proper speech signals. Furthermore, the speech signal analysis is also used to study the heart rate of the speaker [2]. The broader research perspective of Speech Emotion Classification (SEC) finds its applications in crime investigation, psychiatric diagnosis, human–computer interaction, fatigue detection, auxiliary disease diagnosis, bio metrics, and many more.

The basic emotions are categorized into sadness, fear, happiness, disgust, surprise, and anger [3]. The combination of basic emotions leads to other emotions such as love, affection, amusement, contempt, excitement, embarrassment, and so on. Over the decades, various studies have been conducted in the field of SEC where the general pipeline includes feature extraction, dimensionality reduction, and emotion classification. The broad literature for emotion analysis suggests two preferable features, known as statistical and temporal features. [4,5].

Speech Emotion Recognition (SER) system can be structured by analyzing well-crafted features that effectively expose each emotion in the speech signals [6]. The varying length and continuous nature of speech signals require local and global features for emotion recognition. The local features represent temporal dynamics, Whereas the global features expose the statistical aspects like standard deviation, mean, and minimum and maximum values. The features of SER system are categorized into prosodic features, spectral features, voice quality features, and Teager energy operator-based features. Prosodic features, such as rhythm and intonation, are the features based on human's perception. These features are based on energy, duration, and fundamental frequency. Spectral features are extracted in frequency domain using transforms and have received wide attention due to their ability of representing vocal card characteristics [5]. Short-term power spectrum is presented by Mel-frequency cepstral coefficients, whereas vocal tract characteristics are presented by linear prediction coefficients. Logarithmic filtering of auditory system is characterized by log-frequency power coefficients using Fourier transform [7]. Voice quality measurements, such as jitter, harmonics-to-noise ratio, and shimmer, exploit the relation between vocal tract characteristics and emotion content. Teager features detect stresses happening to the vocal tract muscles in the form of energy operator [8]. A few spectral and temporal feature-based SER systems are discussed below.

Fatemeh Daneshfar et al. proposed a hybrid SER system comprising of feature extraction, dimensionality reduction, and classification stages. In the feature extraction stage, three features, such as perceptual minimum variance distortion less response, perceptual linear prediction coefficient, and Mel-frequency cepstral coefficient, are extracted from each frame of the speech signal [9]. A high-dimensional feature vector is structured from the first- and second-order derivatives of the above-said feature vector. The dimension reduction of the feature vector is carried out by quantum behaved particle swarm optimization. The reduced feature vector is classified by a Gaussian elliptical basis function neural network classifier. Palo et al. proposed an SER system in wavelet domain based on Mel-frequency coefficients [10]. Both static and dynamic elements of the coefficients are combined for an SER system. The above-said feature coefficients are reduced in dimension using Principal Component Analysis (PCA) and linear discriminant analysis [11]. Jing et al. suggested an SER system using prominence features and traditional acoustic features [12]. The combined feature vector is reduced in dimension using PCA and non-parametric discriminant analysis. The features are classified using four types of supervised learning classifiers. Wavelet-based features, extracted from the speech signals, are used for SEC in [13]. In [14], spectral features with Naïve Bayes(NB) classifier is employed.

A set of methods on speech emotion classification is based on hidden Markov model [15], Gaussian Mixture Model (GMM) [16], Self-Organizing Map (SOM) [17], and neural network [18]. Singular Value Decomposition (SVD) classifier is used in [19], whereas, in [20], ensemble software regression model is proposed for emotion classification. A deep belief network based on high- and low-level features is also proposed for SEC [21]. Pao et al. proposed a method based on Support Vector Machine (SVM) and neural networks to classify five emotions such as anger, surprise, neutral, happiness, and sadness [22]. Xiao et al. suggested a classifier that uses several sub classifiers for the classification of seven types of emotions [23]. Lin and Wei presented a method that was experimented on gender-dependent and gender-independent experiments [24]. More recently, Xie et al. developed a frame-level emotion recognition system based on attention model in recurrent neural networks. They validated their system for English and non-English speech signals [25]. Demircan and Kahramanli proposed spectral features based on Mel Cepstral coefficients and linear prediction coefficients for speech emotion detection. Later, they used Fuzzy c-means for feature dimension reduction which was further given as input to machine learning classifiers. They used German speech emotion dataset for their work [26].

Our contributions are motivated by (a) the non-stationary nature of speech signals and classical signal processing methods such as Fourier and wavelet analysis use predefined basis functions failing to extract relevant information regarding emotions and (b) the above transformation techniques are block-based methods, wherein a group of samples surrounding the centre element are projected on to the respective basis function. Selection of an optimum window size is a additional requirement for improving the detection accuracy and elimination of artifacts for slow time-varying emotion like sadness. Therefore, there is a need to investigate the classification accuracy of human emotions through data-driven signal processing methods such as Empirical Mode Decomposition (EMD) and non-linear features. This paper investigates an SEC approach where emotions are recognized from speech signals by decomposing them into intrinsic mode functions. Later, five unique randomness measures are computed through entropy measures and state-of-the-art machine learning classifiers are trained on the entropy features. Finally, the performance of the model is validated using standard quantitative metrics on a publicly available emotion classification dataset.

The rest of the paper is organized as follows. “Materials and methods” elaborates on the proposed methodology, the extraction of IMFs through EMD, the computation of the randomness through entropy features, and the detailed analysis of the need for different entropy measures. Results and discussion are presented in “Results” and “Discussion” followed by the conclusion in “Conclusion”.

Materials and methods

Speech signal is a time-varying signal and requires proper selection of a signal processing method to extract the relevant features for emotion recognition. In this paper, the speech signals are analyzed in IMF domain using EMD. Unlike conventional signal processing methods using predefined basis function such as Fourier transform and Wavelet transform, EMD relies on the extraction of inherent patterns in the data for decomposing a signal into intrinsic signals [27]. Figure 1 shows the block diagram of the proposed speech recognition system. The speech signals of duration $\sim$ 2 s are initially decomposed into dominant, mid-, and base-band IMF frequencies. Here, windowing techniques are not involved, and hence, inherent features corresponding to the emotions are extracted with a higher confidence. Non-linear features based on entropy are extracted from the decomposed IMF signals. A feature vector is constructed from the entropy features and used to train a set of classifiers such as LDA, Naïve Bayes (NB), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting machine (GB). Finally the performance of the classifiers is evaluated through balanced accuracy, F1 score, recall, area under the curve, specificity, and precision.

Emotion dataset

To present a realistic comparison, the proposed emotion recognition system from speech signals was trained and tested on publicly available dataset provided by University of Toronto, known as, Toronto Emotional Speech Set (TESS) [28]. The dataset consists of speech signals recorded from two native English participants of age 26 and 64 respectively speaking about 200 target words which completes the phrase ”Say the word—–”. These phrases are captured with seven different emotions of the speakers, namely anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The duration of the records vary between 2 and 3 s and is sampled at 22 KHz. Figure 2 illustrates the speech signals for different emotions from the TESS dataset. For analysis, 200 recordings of each emotion class were taken for the development of the speech recognition system. It should be noted that the original recordings are of high quality (recorded in a noise less environment) and, therefore, do not require additional pre-processing steps.

Empirical mode decomposition

In this section, the EMD of a signal is analyzed. Suppose $x(t)$ be a time-series speech signal that delivers the IMF signals $c(t)$ and the residue function $r(t)$ when decomposed by the EMD method. Equation (1) illustrates the decomposition process: [27]:

$$x(t) = \sum_{i=1}^{d} c(t) + r(t), \tag{1}$$

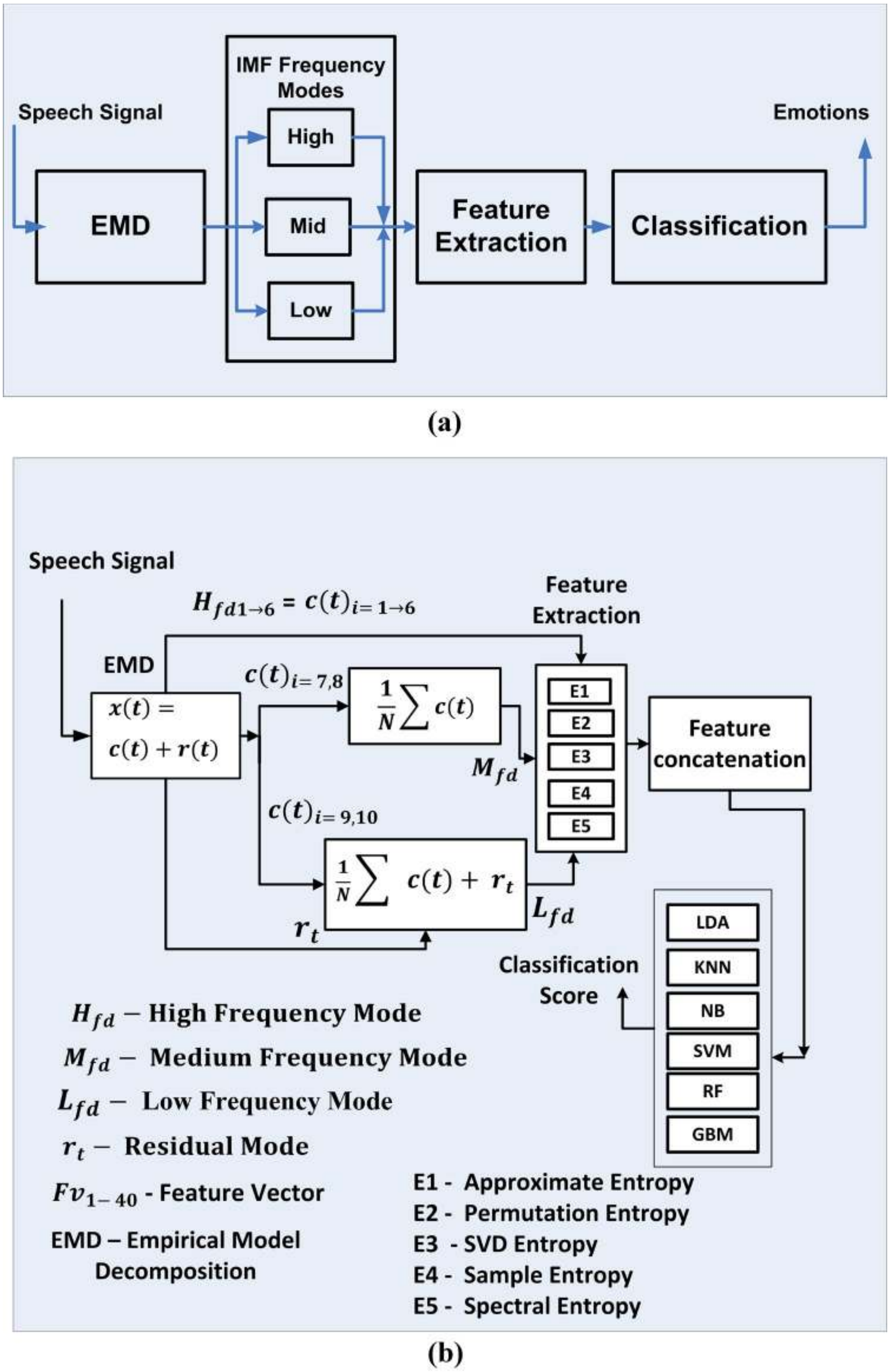
where ‘d’ is the number of IMFs generated for the input signal $x(t)$.

For the experiments, d is preset to a value of 10 components. Preliminary analysis illustrates that setting a lower value to ‘d’ leads to less number of decomposed IMFs resulting in the loss of information. On contrary, a large value for ‘d’ leads to higher levels of decomposition but at a considerable computational cost. Hence, an optimal value of 10 was chosen based on the ad hoc analysis at different levels of decomposition. Figure 3 shows the decomposed speech signal using EMD. The decomposed signal captures different oscillatory features of the speech signal in both temporal and frequency domain.

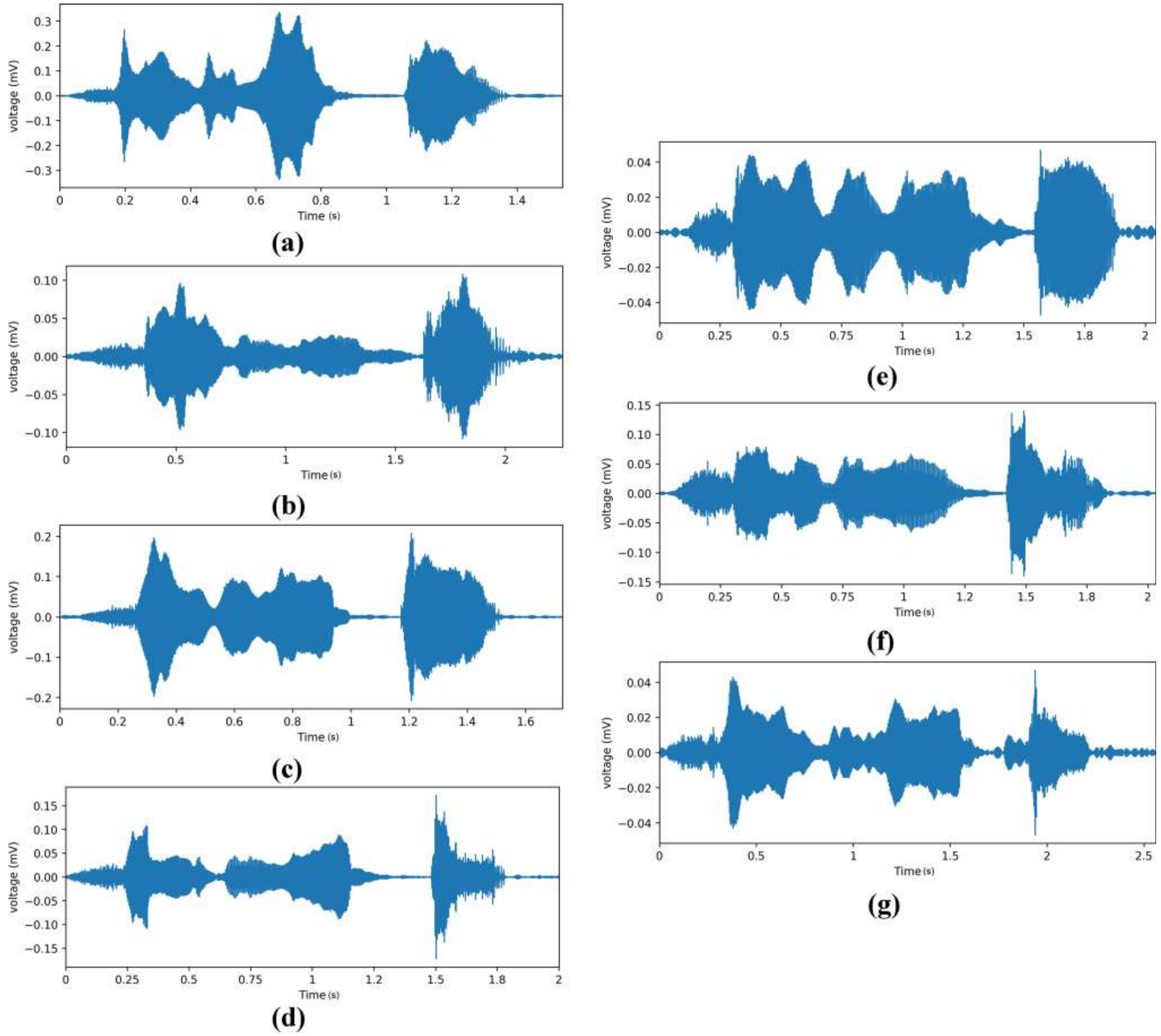
Principal frequency modes

EMD decomposes a time-series signal into IMFs that are localized in time and frequency domains. Since different emotions are captured in distinct frequency components of the IMF signal, the information content in each IMF signal is not uniform and varies depending on the input speech signal. Speech signals pertaining to happy and pleasant surprise are positive emotions. Meanwhile, negative emotions such as angry, fear, disgust, and sad are captured in different

Fig. 1 Proposed method for speech signal recognition system based on EMD and non-linear features: **a** simplified representation. **b** Detailed illustration of different processing steps involved in proposed SER system



frequency scales [29]. Hence, predefined selection of any IMF component or frequency scale will lead to a loss of information. The IMF signals are decomposed into three frequency groups, namely, the High-Frequency (HF), the Mid-Frequency (MF), and the Low-Frequency (LF) modes based on the frequency content, as shown in Fig. 3, to handle the loss of information. The categories are represented as follows: (a) the lower order IMFs starting from IMF-1 to IMF-6 represent the high-frequency modes $H_{\mathrm{fd}1\text{-}6}$, (b) IMF-7 and IMF-8 correspond to mid-frequency modes $M_{\mathrm{fd}7\text{-}8}$, and (c) the higher order IMFs, namely, IMF-9 and IMF-10, correspond to low-frequency modes $L_{\mathrm{fd}9\text{-}10}$. The last component r_{t}



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 2 Speech Signals for different emotions: **a** angry, **b** disgust, **c** fear, **d** happy, **e** neutral, **f** pleasant surprise, and **g** sad

is the residue mode which corresponds to the baseline activity of the signal. To enhance the discrimination ability of the speech signals, especially for negative emotions such as sad, fear, and disgust, with reduced number of trainable features, an averaging scheme is proposed for mid-frequency, low-frequency, and residual modes based on the power spectral density distribution.

Equations (2), (3), and (4) mathematically represent the different modes of IMFs: Here $H_{\mathrm{fd1\text{-}6}}$, M_{fd}, and L_{fd} denote the high-frequency mode, mid-frequency, and low-frequency modes respectively:

$$H_{\mathrm{fd1\text{-}6}} = c_i(t), i = 1, 2, 3, 4, 5, 6 \tag{2}$$

$$M_{\mathrm{fd}} = \frac{1}{N}\sum_{i=7}^{8} c_i(t) \tag{3}$$

$$L_{\mathrm{fd}} = \frac{1}{N}\sum_{i=9}^{10} c_i(t) + r(t). \tag{4}$$

The proposed frequency-based categorization technique can be validated by observing the Power Spectral Density (PSD) plot of IMFs, as shown in Fig. 4, of speech emotions like happy, angry, and sad. From the figure, it could be discerned that the IMF modes from IMF-1 to IMF-6 show unique power spectral density patterns compared to other higher order IMFs. Therefore, IMF-1–IMF-6 are considered separately and directly used for feature extraction. Mean-

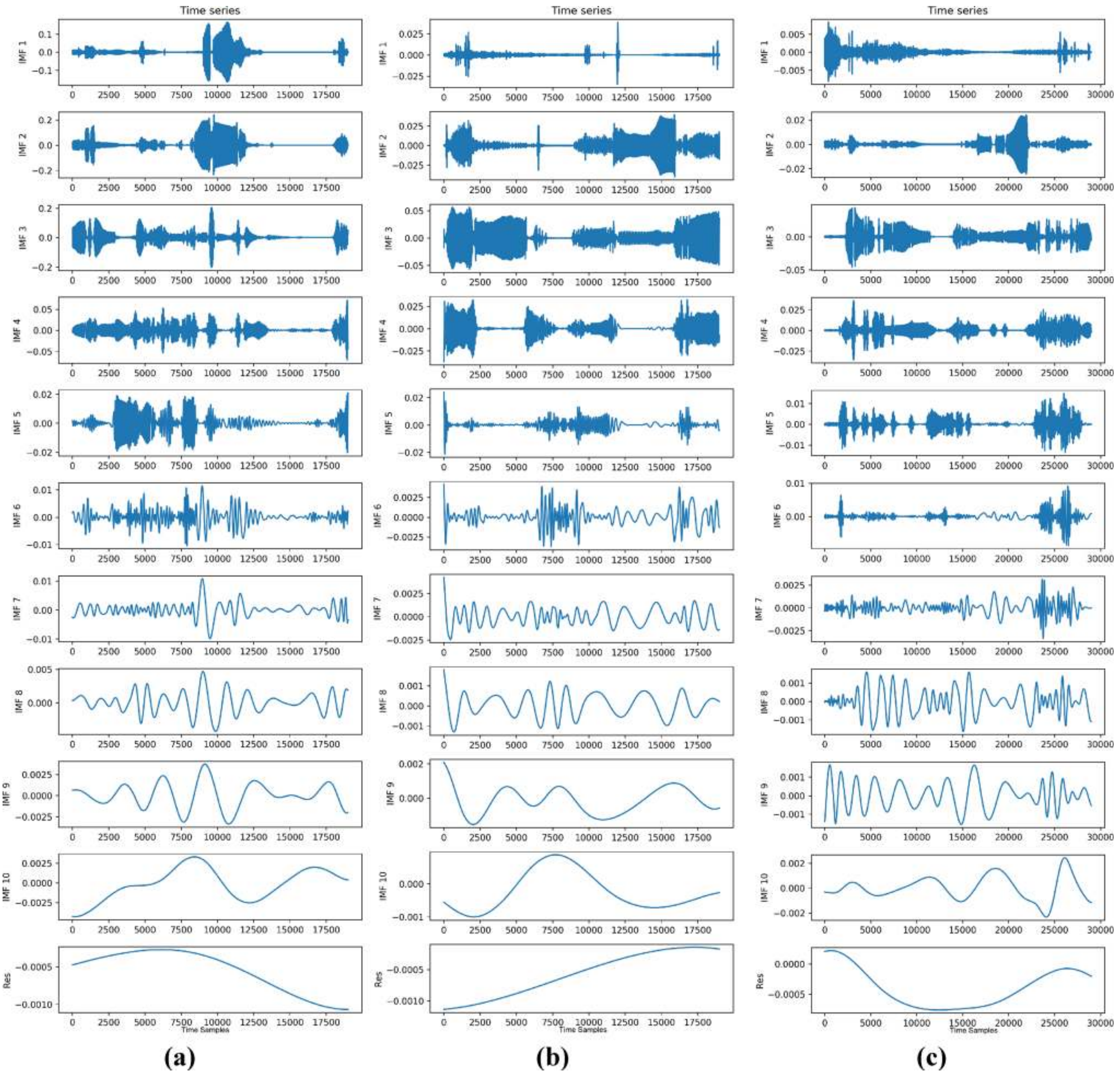


Fig. 3 Decomposed IMFs of speech signals: **a** angry, **b** happy, and **c** sad. Here, *x* axis represents the sample number and the *y* axis denotes the IMF amplitude. We could observe that different emotions show distinct IMF signals

while, mid-frequency, low-frequency modes, and the residue functions show similar PSD patterns. Therefore, these modes are averaged as explained in Eqs. (3) and (4), respectively. Hence, eight unique IMFs from the input speech signals for feature extraction process are selected based on the PSD distribution.

Feature extraction

Non-linear features based on randomness measure and chaos theory have been widely used in signal classification problems. They have been reported with good classification performance for many biomedical applications that involves ECG and EEG signals. Though the speech signals are inherently different from biological signals, the oscillations within the signals define each emotions. Here, it is attempted to quantify the randomness measure by computing entropy functions. Thus, five entropy measures such as Approximate entropy (ApEnt), Sample entropy (SamEnt), singular value decomposition entropy (SVDEnt), Permutation entropy (PermEnt), and Spectral entropy (SpecEnt) are used for extracting randomness features from the speech signals. For each IMF,

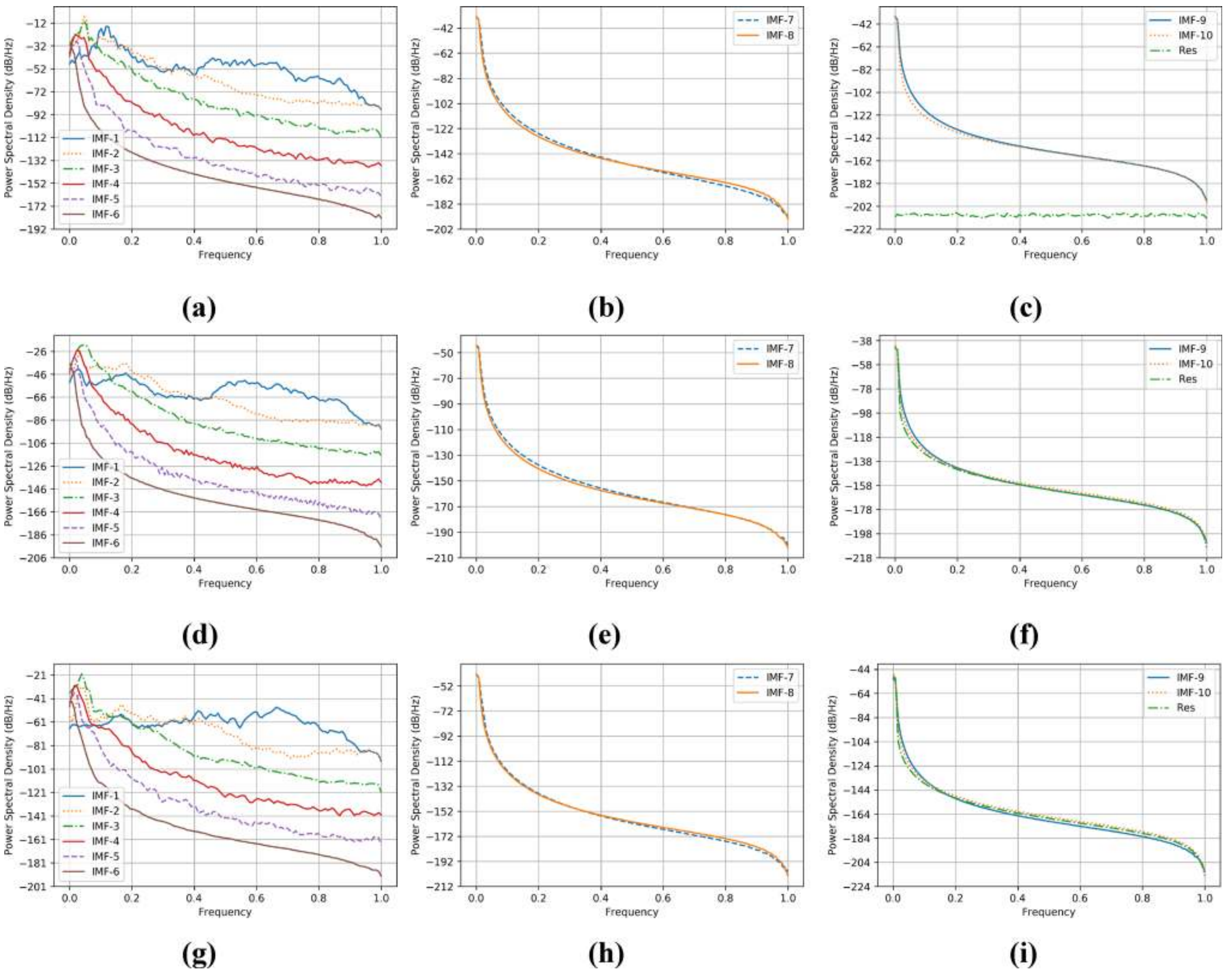


Fig. 4 Power spectral density of some of the speech signals with emotions: **a**–**c** angry, **d**–**f**, and happy **g**–**i** sad for the three frequency groups, namely, HF (IMF-1–IMF-6), MF (IMF-7–IMF-8), LF (IMF-9, IMF-10), and residue modes

a total of 5 different entropy measures are computed resulting to 40 different trainable features:

$$H_{\mathrm{fd1\text{-}6}} \in \mathbb{R}^{1\times 30},\ M_{\mathrm{fd}} \in \mathbb{R}^{1\times 5},\ L_{\mathrm{fd}} \in \mathbb{R}^{1\times 5}. \tag{5}$$

The proceeding sections present (a) the different entropy measures, (b) a detailed investigation on the need for different entropy measures, and (c) a brief analysis of the different types of classifiers used in this study.

Approximate entropy

Approximate entropy is a complexity measure widely used in the regularity analysis of time-series signal. It quantifies the amount of randomness based on signal fluctuations. A lower value of ApEnt suggests that the time-series signal is regular and a higher value demonstrates the randomness. The parameters 'm' and 'r' define the delay parameter and similarity value, respectively. ApEnt be computed through Eq. (6) [30]:

$$\mathrm{ApEnt}_{x(t)} = \phi_r^m - \phi_r^{m+1}. \tag{6}$$

Here, $x(t)$ is the speech signal and ϕ_r is the correlation integral function for the phase space vectors (embedded signal). For experiments, the chosen values are $m = 3$ and $r = 0.2\mathrm{std}(x(t))$ based on the work of [31]. Here, 'std' refers to the standard deviation of the input signal.

Sample entropy

Sample entropy is a modified version of approximate entropy where the limitations such as self-similar pattern bias are overcome [32]. Here, the similarity measure is computed based on various embedded time-series samples and avoids computing the self-similarity measure between the samples.

It reduces the bias which is inherent in approximate entropy. The representation of sample entropy is provided in Eq. 7 [32]:

$$\text{SamEnt}_{x(t)} = -\log \frac{z(m+1, r)}{z(m, r)}. \tag{7}$$

Here, 'z' is the measure of similarity of a embedded time-series for 'm' and '$m+1$' and r is the tolerance level. The values of 'm' and 'r' are identical to the values used for approximate entropy.

SVD entropy

Singular value decomposition represents the dimensionality of the data. It decomposes the high-dimensional data into orthogonal matrices based on the singular values 'σ'. The time-series signal is converted to a embedded matrix based on different time delayed template vector taken from the speech signal. The embedded matrix is decomposed into various orthogonal matrices. However, the SVD entropy measure is computed only on the diagonal matrix which contain the singular values. Equation (8) represents the SVD computation [33]:

$$\text{SVDEnt}_{x(t)} = -\sum_{i=1}^{L} \sigma_i \log_2(\sigma_i). \tag{8}$$

Here, 'L' represents the number of singular values of the embedded matrix and 'σ_i' denote the singular values.

Spectral entropy

Spectral entropy measures the randomness by employing Fourier transform to the time-series signal. For the computation of this entropy measure, the power spectral density, $S(f)$, of the speech signal is obtained. The spectral entropy is calculated using the formulation of Shannon entropy measure as given below [34]:

$$\text{SpecEnt}_{x(t)} = -\sum_{f=0}^{f_n} S(f) \log_2[S(f)]. \tag{9}$$

Here, 'f_n' is the sampling frequency of the signal.

Permutation entropy

Permutation entropy computes the randomness of the time-series signal based on ordinal patterns of the signal. It is a non-parametric approach and provides a robust estimation of irregularity information of the signal. The approach involves creating a embedded time delayed matrix based on τ and D denotes the size of the embedded matrix. Usually, τ and D are set to 1 and 3, respectively. The different ordinal patterns are tabulated and verified with the column vector of the embedded matrix. The number of occurrence of the ordinal patterns in the matrix is counted and the probability of occurrence, 'ψ_i', of each pattern is tabulated. The permutation entropy is computed as given below [35] :

$$\text{PermEnt}_{x(t)} = -\sum_{i} p(\psi_i) \log_2 p(\psi_i). \tag{10}$$

The following section briefly explains the contribution of each entropy toward the calculation of randomness attribute which plays a significant role in the analysis of the speech emotions. Table 1 provides a summary of each entropy measure and its implications in speech analysis.

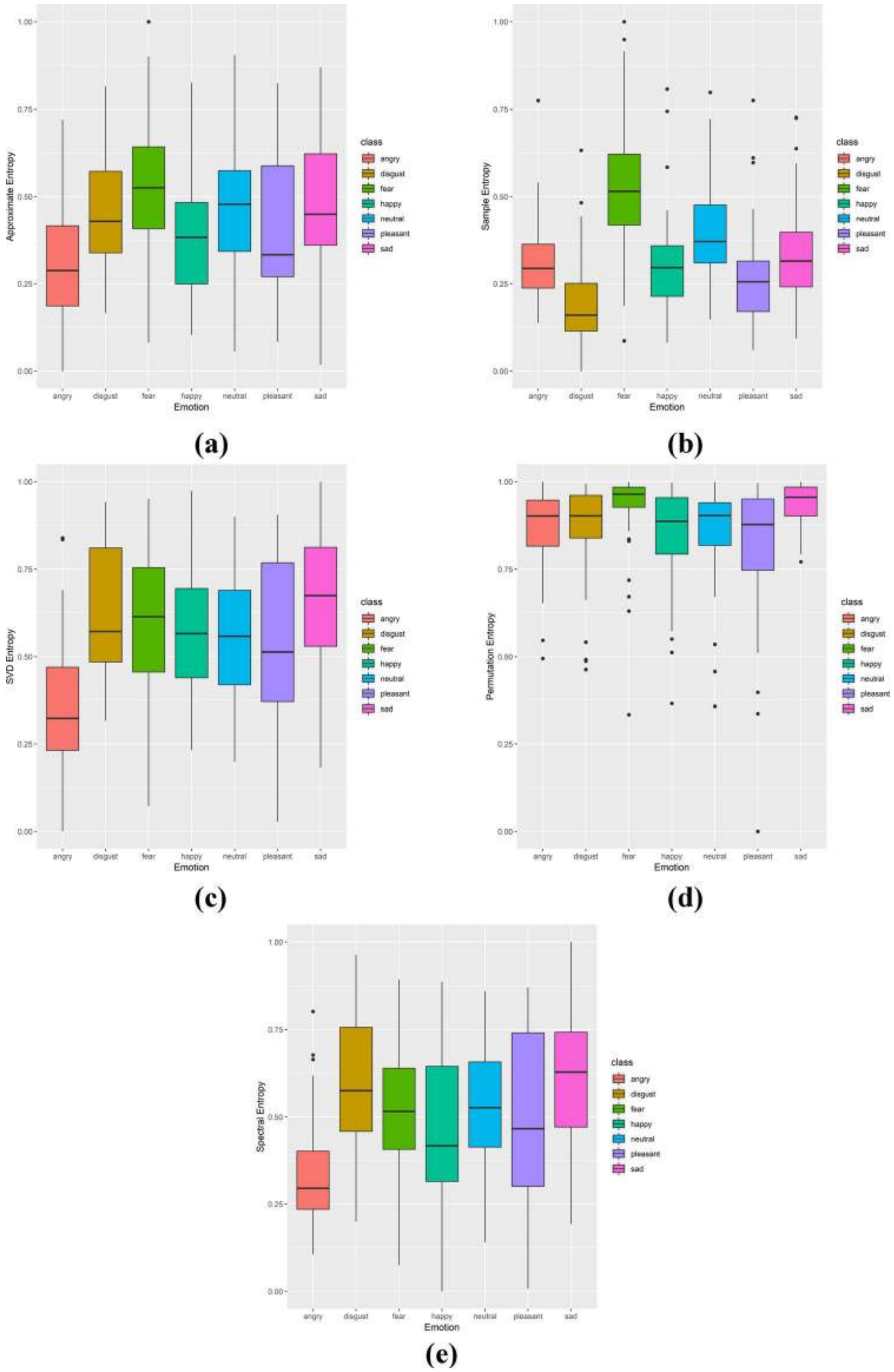
Table 1 Entropy measure and its attributes in speech analysis

Randomness measure	Implication
Approximate entropy	Computes irregularity in the speech signal, however considers self-similar patterns in the input signal
Sample entropy	Modification of Approximate entropy with no bias of self-similar patterns
SVD entropy	Computes the randomness measure based on decomposition of high-dimensional data using singular values
Spectral entropy	computes the randomness in the power spectral density function of the speech signal
Permutation entropy	Uses ordinal patterns in the speech signal to detect the emotions

Entropy and their relation with IMFs

As explained in the previous section, the entropy features are extracted from the principal mode IMFs through decomposition of the original speech signal. Figure 5 illustrates the box-plot of different emotions captured by non-linear entropy features. From the figure, It is observed that the median values of all the entropy features differ for different class of emotions, and therefore, the computed entropy features could be readily used as discriminators for classification of emotions. We analyzed the emotion classification accuracy of different entropy features by varying the number of extracted IMFs. Figure 6 illustrates the variation of classification accuracy of each of the entropy measure for different IMF lengths. It could be observed that no single entropy measure provides good classification accuracy for all the speech signal emotions and the classification accuracy

Fig. 5 Distribution of randomness value based on entropy of different speech emotion signals: **a** approximate entropy, **b** sample entropy, **c** SVD entropy, **d** permutation entropy, and **e** spectral entropy



(a)

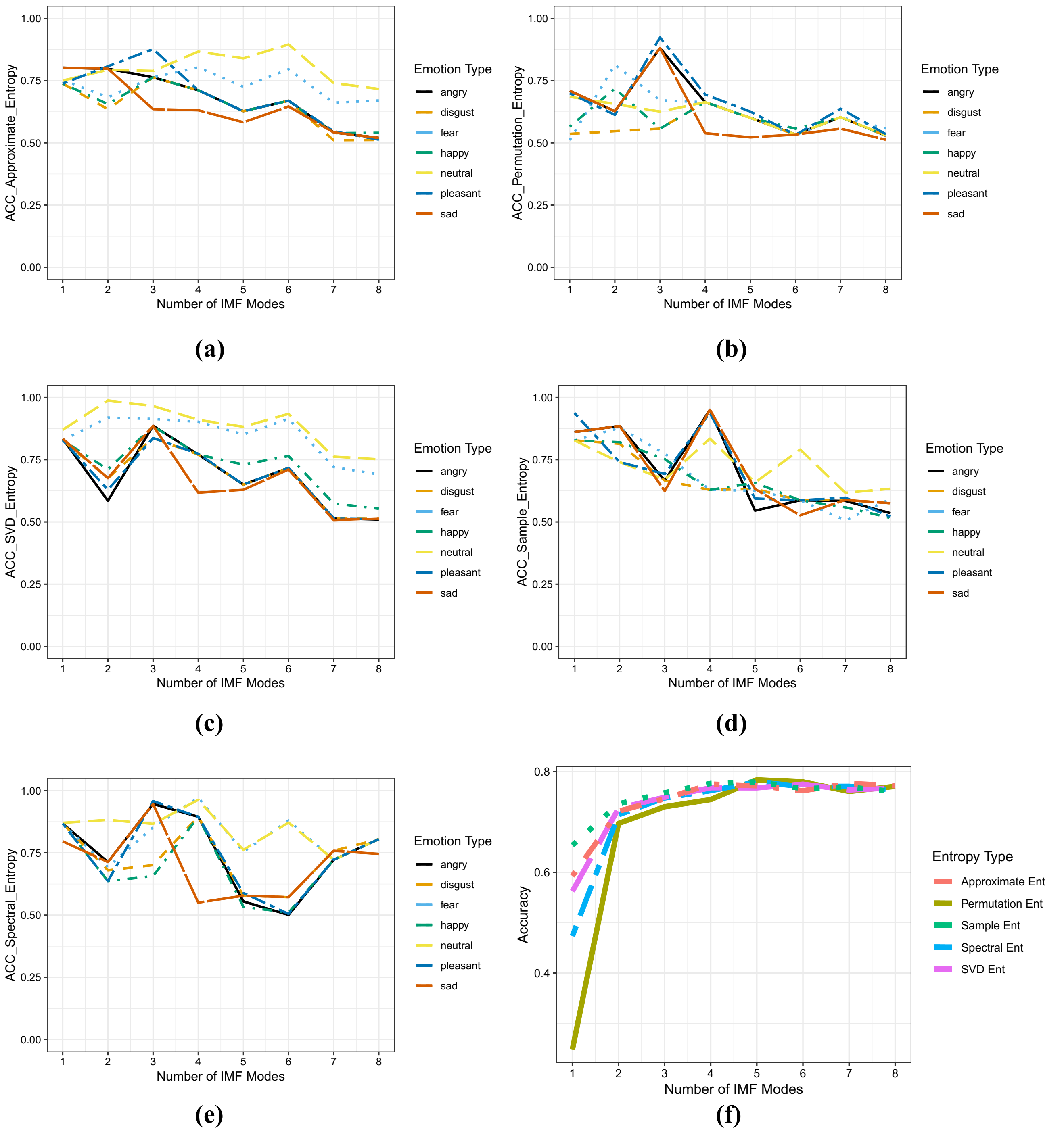
(b)

(c)

(d)

(e)

depends on the choice of number IMFs. For example, permutation entropy provides good accuracy for emotions such as pleasant surprise, angry, and sad speech signals decomposed upto IMF-3, however, for the same decomposition, its accuracy is less for fear, disgust, neural, and happy. Similarly, when the speech signal is decomposed upto IMF-4, sample entropy provides good discrimination of sad, angry, neutral, and pleasant surprise and less accuracy for other emotions. Likewise, entropies such as approximate, SVD, and spectral provide higher discrimination abilities from IMF-3 to IMF-6, respectively. Therefore, the experimental analysis suggests that entropy feature extracted from the EMD of speech sig-



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Fig. 6 Performance of different entropy measures of speech emotion signals for different IMF modes: **a** approximate entropy, **b** permutation entropy, **c** SVD entropy, **d** sample entropy, **e** spectral entropy, and **f** average accuracy of each entropy measure for all emotions

nals present complimentary information at different IMFs. Thus, it is prudent to include all the IMFs to improve the classification accuracy.

Subsequently, Fig.6f illustrates the average emotion classification accuracy for different entropies considered for IMFs 1–8 (only 8 features). the entropies present complementary information at different decomposition levels, employing them individually for emotion classification presents a peak accuracy of only $\sim$ 79% for the 7 different emotions considered in this study. Hence, the entropy features computed from different IMFs are combined as a feature vector (considering 40 features) and presented to the classifier for improved signal classification which is explained in the proceeding section.

Feature classification

State-of-the-Art (SoA) classifiers such as LDA, NB, K-NN, SVM, RF, and GB are used for classification of emotions from speech signals. The following section briefly describes the classifiers.

Linear discriminant analysis

Linear discriminant analysis is a well-known machine learning algorithm for classification and prediction task. The method is simple, and therefore, prediction process is trivial compared to some of the other classification algorithms. LDA is a dimensionality reduction technique with focus on projecting higher dimension space to lower dimension. The following steps are involved in LDA: initially, the separability between the different classes is calculated by finding a distance between mean and the elements of each class which is referred as the intra class variance [36]. Finally, a lower dimension space is created using the Fisher's criterion by reducing the intraclass variance and increasing the distance between the interclasses.

Naïve Bayes

It is a classification method which is based on Bayes theorem. This classifier works on the assumption that the presence of a particular feature of a class is not related to the presence of other features and, therefore, represents a probabilistic machine learning model. It is easier to build and very useful for very large dataset. Using Bayes theorem, probability of an happening event 'A' can be found out, given another event 'B' that has already occurred. Since the presence of one particular feature does not effect others, it is referred as Naïve [14] .

K-nearest neighbor

K-NN is a supervised machine learning algorithm which is widely used for classification as well as prediction problems. It is considered as a lazy learning technique, since there is no specialized training data. Generally, the entire data are used for the training purpose. It is also a non-parametric method, because there are no assumptions involved where the similarity between features is used for prediction of a new data point. The value of 'K' is the number of neighbors selected initially, which can be any integer, based on number of classes in the dataset [37]

The distance between the training and test data is calculated using the distance measures such as Euclidean, Hamming, etc. The computed distance is sorted in the ascending order and the queried data are presented with the class label having the least distance.

Support vector machine

Support vector machine is a machine learning model used for classification and regression challenges [38]. Each data point is plotted as a point in a n-dimensional space with feature value expressed as the coordinate value. Hyperplane- based decision boundaries are found separating the two classes. While finding the hyperplane, many possibilities are considered and the plane that has maximum margin separating the two classes is selected. The separation plane classifies the future test point with utmost confidence [39].

Random forest

Random forest is a supervised machine learning technique where multiple decision trees are built and combined together to present a stable prediction. It can be used for both classification and regression challenges. Generally, the more the number of decision trees, better the model's accuracy [37].

Gradient boosting machine

It is one of the powerful models designed for predictions. The technique involves three parts. (a) Differentiable loss function; (b) a decision tree to boost the weak learners; (c) a additive model along with the decision trees for selection of the best decision tree model.

The nodes in each decision tree take a different subset of features for selecting the best split. In this technique, all the trees are unique and they are able to capture different signal from the data points. Also each new tree is based on the errors of the previous tree and all these operations are executed in a sequential order [40].

Results

From the preliminary analysis discussed in the methodology section, it is understood that detection of emotions from speech signals requires to use the complementary information provided by all the entropy measures. Hence, the entropy features are combined to form a composite feature vector. Each emotion class consists of 200 speech signals that are used for forming the feature space of dimension 1400 × 40. The feature matrix is used to train a collection of state-of-the-art machine learning classifiers to detect the seven emotions from the speech signal. A tenfold cross-validation technique is used to obtain the performance measures of the classifiers in emotion classification. To evaluate the classifiers, the performance metrics such as balanced accuracy, F1 score, recall, AUC , specificity, and precision are used [41]. Figure 7 illustrates the box-plot of performance measures for different classifiers used in the work. Table 2 provides the classifica-

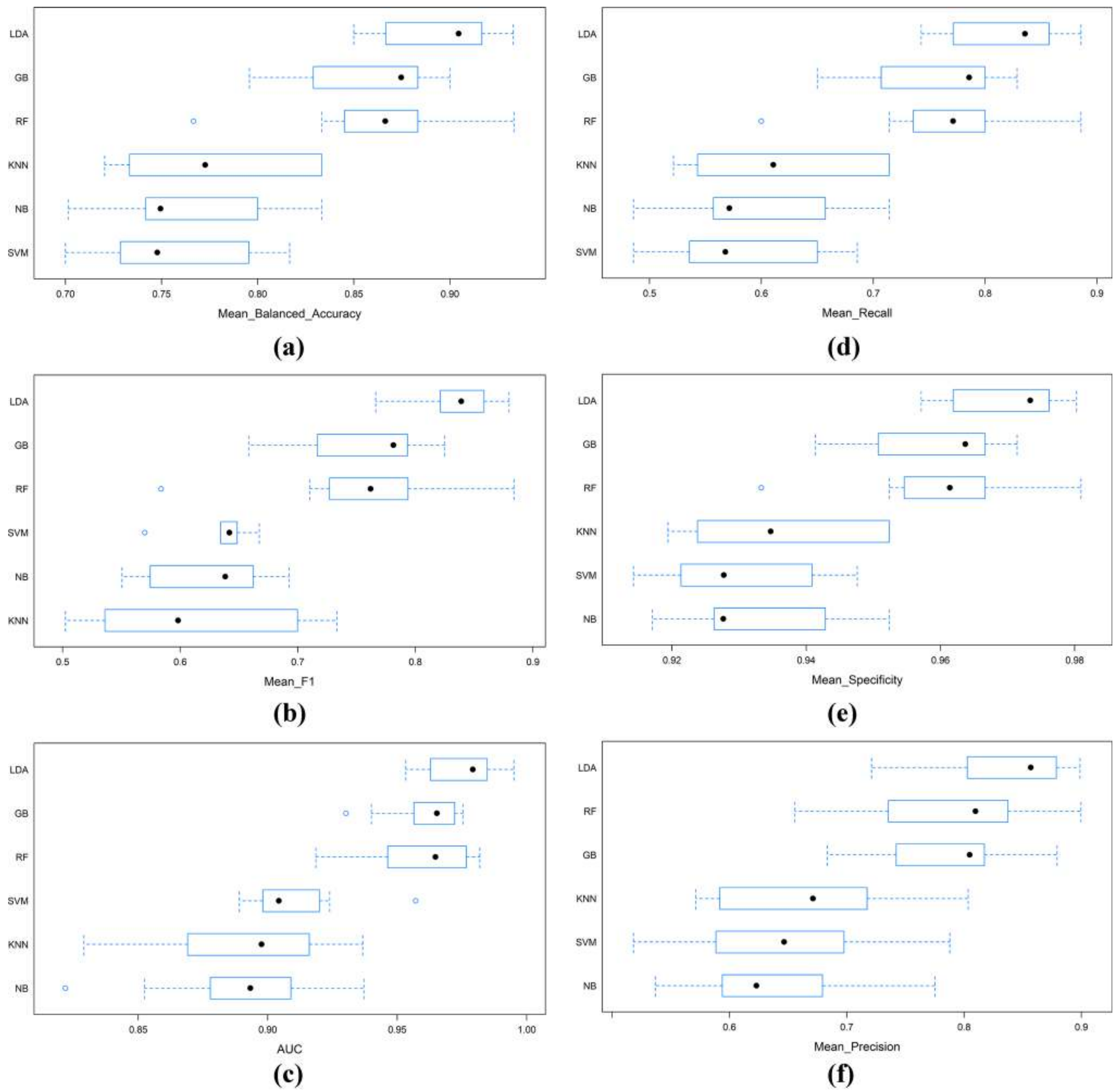


Fig. 7 Performance metrics of the state-of-the-art classifiers for speech emotion classification based on entropy measure: **a** balanced accuracy, **b** F1 score, **c** AUC, **d** sensitivity, **e** specificity, and **f** precision

tion performance metrics of the SoA models with respect to balanced accuracy. Table 3 compares the area under the curve value of the Receiver-Operating Characteristic (ROC) curve, and finally, Table 4 tabulates the F1 score for all the classifiers considered for emotion classification. Table 5 shows the specificity of the proposed system.

Here, the Mean Balanced Accuracy (MBA) metric is used to analyze the goodness of the binary classifier and the MBA for SVM classifier is 0.74, while LDA gives the best accuracy of 0.899. In regards to F1 score, K-NN gives the lower score of 0.6 and LDA gives the best score of 0.84. When considering AUC, which provides an overall measure of classifiers performance across all possible classification thresholds, Naïve Bayes classifier records the lowest value of 0.89, while LDA scores the highest value of 0.98. Considering recall metric which is a measure of sensitivity, SVM provides the lowest performance score of 0.58, while LDA delivers the best score of 0.85. The model ability to predict the true negatives is recorded by the specificity metric and LDA gives best score of 0.97. For precision, a measure of rel-

Table 2 Balanced accuracy of some of the state-of-the-art (SoA) classifiers trained on entropy features for emotion recognition

Model	Min	Mean	Max
LDA	0.8500000	0.8991203	0.9330049
K-NN	0.7204258	0.7766033	0.8333333
SVM	0.7000000	0.7566092	0.8166667
RF	0.7666667	0.8619875	0.9333333
NB	0.7013957	0.7662931	0.8333333
GB	0.7956897	0.8586084	0.9000000

Table 3 AUC of ROC plot of some of the SoA classifiers trained on entropy features for emotion recognition

Model	Min	Mean	Max
LDA	0.9533333	0.9760250	0.9950739
K-NN	0.8290476	0.8908201	0.9366667
SVM	0.8890394	0.9099204	0.9571429
RF	0.9185714	0.9590877	0.9819048
NB	0.8219048	0.8912003	0.9371429
GB	0.9302135	0.9604885	0.9753695

Table 4 F1 score of some of the SoA classifiers trained on entropy features for emotion recognition

Model	Min	Mean	Max
LDA	0.7662338	0.8327253	0.8797314
K-NN	0.5025308	0.6105503	0.7332112
SVM	0.5700371	0.6324655	0.6673469
RF	0.5838384	0.7542363	0.8838384
NB	0.5506081	0.6236622	0.6925170
GB	0.6582973	0.7561204	0.8249433

Table 5 Specificity of some of the SoA classifiers trained on entropy features for emotion recognition

Model	Min	Mean	Max
LDA	0.9571429	0.9710978	0.9802956
K-NN	0.9194229	0.9360638	0.9523810
SVM	0.9142857	0.9303612	0.9476190
RF	0.9333333	0.9604035	0.9809524
NB	0.9170772	0.9333005	0.9523810
GB	0.9413793	0.9593596	0.9714286

evancy, LDA scores the best with a high value of 0.89. Based on the performance scores of all the classifiers, it could be stated that the proposed model delivers the best prediction accuracy using LDA classifier.

Discussion

The performance of an SER system highly depends on the type of emotion signal databases. There are three types of speech signal databases on which classification algorithms are experimented. Simulated databases or acted databases consist of speech signals generated by professional actors for various emotions. Induced databases, otherwise known as elicited databases, are collected from fresh induced emotions by artificial situations [42]. In this case, emotion signals are recorded without informing the speakers, and the emotions are induced by playing audios, videos, digital games, and images. The third type of database is known as natural database in which the emotions are recorded naturally like talk shows, natural conversations, and so on. Among the databases, the first one is prominent for emotion analysis, whereas natural databases are often influenced by surrounding noises and artifacts [43].

In this section, the related works in speech emotion recognition are compared based on TESS and other publicly available speech emotion datasets for English and other spoken languages, as shown in Table 6. Verma et al. studied the impact of age in recognition of emotions and trained a set of support vector machine classifier based on Mel-frequency coefficients [44]. They showed an overall accuracy of 96%. However, their method considered only five emotions and required to categorize the dataset based on age before applying the classifier. Sundarprasad used Mel-Frequency Cepstrum Coefficients and PCA for dimensionality reduction followed by SVM classifier to classify the emotions. They have used the TESS dataset and reported an accuracy of 90% [45]. Gao et al. used a combination of features based on Mel frequency and time domain features, and reported a highest overall classification accuracy of 81%. The lower classification accuracy of neutral emotions reduced the overall accuracy of their method [46]. Xie et al. showed that frame-level features based on attention model of long-short time memory (LSTM) could be more discriminant in emotion detection [25]. They reported an unweighted average recall (UAR) measure of 89.6% based on LSTM for a English emotion database (Enterface). Demircan et al. proposed speech recognition system based on Mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) and used fuzzy c-means for dimensionality reduction. They reported a peak classification accuracy of 92.86% for seven emotions for German language dataset, EmoDB using SVM classifier [26].

Venkataramanan et al. used convolution neural network based on log-Mel spectrogram and reported an accuracy of 70% based on TESS dataset [47]. Praseetha et al. proposed a model based on MFCC features, Deep Neural Network (DNN), and Gated Recurrent Unit (GRU) [48]. The two models were tested on TESS dataset for five emotions: anger, happiness, fear, sadness, and neutral. The DNN model with MFCC delivered a classification accuracy of 89.96%; on the other hand, GRU model gave an accuracy of 95.82%. Though they have reported with high accuracy for the TESS dataset, they have considered only five emotions. Moreover, DNN and GRU are computationally intensive methods that could not be easily implemented for real-time emotion recognition.

Kerkeni et al. proposed an automated SER system based on combination of features obtained in EMD domain. They have used modulation spectral features and modulation frequency features based on the IMF signal and combined them with cepstral features [49]. Their methodology was initially evaluated on the Spanish emotional database using RNN classifier and reported an accuracy of 91.16%. For the Berlin database, they have used SVM classifier and reported an accuracy of 86.22%. However, their method is not tested for the English language.

In comparison, the proposed approach used non-linear features such as the entropy measure and selection of principal IMF modes for feature extraction. The proposed system obtained a highest balanced accuracy of 93.3% with a mean of 89.9%, F1 score of maximum 87.9%, and a mean of 83.3% using LDA classifier. The method obtained a peak AUC value of 0.995 and a mean value of 0.976 for LDA classifier in recognizing the seven speech emotions. Other related works which reported high accuracy involved only five emotions.Other methods which considered all seven emotions from TESS dataset reported a lower accuracy since neutral emotions were considered. The proposed method based on entropy features from principal modes performs well in recognizing human emotions from speech signals. In this work, the impact of age and gender in recognition of emotions from speech is not considered, though it is known to influence the emotions in the speech signal. More investigations are required in advancing the proposed method to account for the age and gender bias in the speech emotion recognition. Moreover, it is also planned to use deep learning methods for time-series and sequence analysis such as recurrent neural networks for speech emotion recognition as an extension of this work.

Conclusion

In this paper, the application of non-linear features such as the entropy measure in recognition of human emotions from speech signals is demonstrated. This work investigated the entropy feature extraction using EMD for recognizing seven emotions of native English speakers. As the positive and negative emotions are captured in different frequency scales of the speech signal, the IMFs are categorized into various frequency groups and selected principal modes for entropy feature extraction. The reported metrics are a peak balanced

Table 6 Comparison of some of the notable works in the speech emotion recognition (English)

Author(s)	Method	Dataset	# of Emotions (types)	Classification accuracy
Verma et al. [44]	Mel-frequency cepstral coefficients (MFCC) and SVM classifier	TESS	5 (angry, sad, happy, neutral, fear)	96%
Sundarprasad [45]	Mel-frequency cepstrum features, PCA and SVM	TESS	7 (angry, sad, happy, surprise, disgust, neutral, fear)	90%
Gao et al. [46]	MFCC and time domain features	TESS	4 (angry, sad, happy, neutral)	81%
Xie et al. [25]	Frame level features and LSTM (attention-based)	eNTERFACE	6 (angry, sad, happy, surprise, disgust, fear)	89.6%
Venkataramanan et al. [47]	Convolution neural network based on log-Mel spectrogram	TESS	7 (angry, sad, happy, surprise, disgust, neutral, fear)	70%
Praseetha et al. [48]	MFCC features, Deep Neural Network (DNN) and Gated Recurrent Unit (GRU)	TESS	5 (anger, happy, fear, sad, neutral)	89.96% (DNN) 95.82% (GRU)
Proposed method	Entropy features from principal IMF modes and SoA machine learning classifier	TESS	7 (angry, sad, happy, surprise, disgust, neutral, fear)	93.3%

accuracy of 93.3%, a peak F1 score of 87.9%, and a peak AUC value of 0.995 using LDA classifier. This proves that the proposed method of dividing the frequency components in the speech signal into three frequency groups such as the high-frequency, mid-frequency, and low-frequency modes could recognize different emotions existing in different frequency scales of a speech signal.

Acknowledgements This research was financially supported by the Scientific Research Grant of Shantou University, China, Grant no: NTF17016.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Huang W, Wu Q, Dey N, Ashour A, Fong SJ, González-Crespo R (2020) Adjectives grouping in a dimensionality affective clustering model for fuzzy perceptual evaluation. Int J Interact Multimedia Artif Intell 6(2):10. https://doi.org/10.9781/ijimai.2020.05.002
2. Anttonen J, Surakka V (2005) Emotions and heart rate while sitting on a chair. In: Proceedings of the SIGCHI conference on Human factors in computing systems—CHI '05, ACM Press, New York, New York, USA, p 491. https://doi.org/10.1145/1054972.1055040, http://portal.acm.org/citation.cfm?doid=1054972.1055040
3. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76. https://doi.org/10.1016/j.specom.2019.12.001
4. Sailunaz K, Dhaliwal M, Rokne J, Alhajj R (2018) Emotion detection from text and speech: a survey. Soc Netw Anal Min 8(1):28. https://doi.org/10.1007/s13278-018-0505-2
5. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117. https://doi.org/10.1007/s10772-011-9125-1
6. Yang N, Dey N, Sherratt RS, Shi F (2020) Recognize basic emotional statesin speech by machine learning techniques using mel-frequency cepstral coefficient features. J Intell Fuzzy Syst. https://doi.org/10.3233/jifs-179963
7. Nwe TL, Foo SW, De Silva LC (2003) Detection of stress and emotion in speech using traditional and FFT based log energy features. In: ICICS-PCM 2003—Proceedings of the 2003 joint conference of the 4th international conference on information, communications and signal processing and 4th Pacific-Rim conference on multimedia, institute of electrical and electronics engineers Inc., vol 3, pp 1619–1623. https://doi.org/10.1109/ICICS.2003.1292741
8. Teager HM, Teager SM (1990) Evidence for nonlinear sound production mechanisms in the vocal tract. In: Speech production and speech modelling. Springer Netherlands, pp 241–261. https://doi.org/10.1007/978-94-009-2037-8_10
9. Daneshfar F, Kabudian SJ, 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. Appl Acoust 166:107360. https://doi.org/10.1016/j.apacoust.2020.107360
10. Palo HK, Behera D, Rout BC (2020) Comparison of classifiers for speech emotion recognition (SER) with discriminative spectral features, pp 78–85. https://doi.org/10.1007/978-981-15-2774-6_10
11. Nazid Mohd H, Muthusamy H, Vijean V, Yaacob S (2018) Improved speaker-independent emotion recognition from speech using two-stage feature reduction—UUM Repository. J Inf Commun Technol 14:57–76. http://repo.uum.edu.my/24081/
12. Jing S, Mao X, Chen L (2018) Prominence features: effective emotional features for speech emotion recognition. Digit Signal Proc 72:216–231. https://doi.org/10.1016/j.dsp.2017.10.016
13. Roy T, Marwala T, Chakraverty S (2020) Speech emotion recognition using neural network and wavelet features, pp 427–438. https://doi.org/10.1007/978-981-15-0287-3_30
14. Khan A, Roy UK (2018) Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. Institute of Electrical and Electronics Engineers (IEEE), pp 1017–1021. https://doi.org/10.1109/wispnet.2017.8299916
15. Song P, Jin Y, Zhao L, Xin M (2014) Speech emotion recognition using transfer learning. IEICE Trans Inf Syst E97D(9):2530–2532. https://doi.org/10.1587/transinf.2014EDL8038
16. Partila P, Tovarek J, Voznak M (2016) Self-organizing map classifier for stressed speech recognition, p 98500A. https://doi.org/10.1117/12.2224253
17. Lanjewar RB, 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 Comput Sci 49:50–57. https://doi.org/10.1016/j.procs.2015.04.226
18. Patel P, Chaudhari AA, Pund MA, Deshmukh DH (2017) Speech emotion recognition system using gaussian mixture model and improvement proposed via boosted gmm. IRA Int J Technol Eng (ISSN 2455-4480) 7(2 (S)):56–64
19. Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W, Sturge-Apple M (2017) Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. Int J Speech Technol 20(1):27–41. https://doi.org/10.1007/s10772-016-9364-2
20. Sinith MS, Aswathi E, Deepa TM, Shameema CP, Rajan S (2016) Emotion recognition from audio signals using Support Vector Machine. In: 2015 IEEE recent advances in intelligent computational systems, RAICS 2015, Institute of Electrical and Electronics Engineers Inc., pp 139–144. https://doi.org/10.1109/RAICS.2015.7488403
21. Wen G, Li H, Huang J, Li D, Xun E (2017) Random deep belief networks for recognizing emotions from speech signals. Comput Intell Neurosci 2017:1–9. https://doi.org/10.1155/2017/1945630
22. Tsang-Long Pao YC, Jun-Heng Yeh PL (2006) Mandarin emotional speech recognition based on SVM and NN. In: 18th International conference on pattern recognition (ICPR'06), IEEE, pp 1096–1100. https://doi.org/10.1109/ICPR.2006.780
23. Xiao Z, Dellandrea E, Dou W, Chen L (2010) Multi-stage classification of emotional speech motivated by a dimensional emotion

model. Multimedia Tools Appl 46(1):119–145. https://doi.org/10.1007/s11042-009-0319-3

24. Lin YL, Wei G (2005) Speech emotion recognition based on HMM and SVM. In: 2005 International conference on machine learning and cybernetics, IEEE, vol 8, pp 4898–4901. https://doi.org/10.1109/ICMLC.2005.1527805
25. Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B (2019) Speech emotion classification using attention-based lstm. IEEE/ACM Trans Audio Speech Lang Proc 27(11):1675–1685. https://doi.org/10.1109/TASLP.2019.2925934
26. Demircan S, Kahramanli H (2018) Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech. Neural Comput Appl 29(8):59–66. https://doi.org/10.1007/s00521-016-2712-y
27. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond Ser A Math Phys Eng Sci 454(1971):903–995. https://doi.org/10.1098/rspa.1998.0193
28. Dupuis K, Kathleen Pichora-Fuller M (2010) Toronto emotional speech set (TESS) | TSpace Repository. https://doi.org/10.5683/SP2/E8H2MF
29. Hassouneh A, Mutawa AM, Murugappan M (2020) Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Inform Med Unlock 20:100372. https://doi.org/10.1016/j.imu.2020.100372
30. Pincus SM (1991) Approximate entropy as a measure of system complexity. Proc Nat Acad Sci 88(6):2297–2301. https://doi.org/10.1073/pnas.88.6.2297
31. Delgado-Bonal A, Marshak A (2019) Approximate entropy and sample entropy: a comprehensive tutorial. Entropy 21(6):541. https://doi.org/10.3390/e21060541
32. Richman JS, Lake DE, Moorman J (2004) Sample entropy. In: Methods in enzymology, pp 172–184. https://doi.org/10.1016/S0076-6879(04)84011-4
33. Gu R, Shao Y (2016) How long the singular value decomposed entropy predicts the stock market—evidence from the dow jones industrial average index. Phys A 453:150–161
34. Tian Y, Zhang H, Xu W, Zhang H, Yang L, Zheng S, Shi Y (2017) Spectral entropy can predict changes of working memory performance reduced by short-time training in the delayed-match-to-sample task. Front Hum Neurosci 11:437. https://doi.org/10.3389/fnhum.2017.00437
35. Yang Y, Zhou M, Niu Y, Li C, Cao R, Wang B, Yan P, Ma Y, Xiang J (2018) Epileptic seizure prediction based on permutation entropy. Front Comput Neurosci. https://doi.org/10.3389/fncom.2018.00055
36. Izenman AJ (2013) Linear discriminant analysis. Springer, New York, pp 237–280. https://doi.org/10.1007/978-0-387-78189-1_8
37. Pohjalainen J, Räsänen O, Kadioglu S (2015) Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. Comput Speech Lang 29(1):145–171. https://doi.org/10.1016/j.csl.2013.11.004
38. Bellamkonda S, Np G (2020) An enhanced facial expression recognition model using local feature fusion of gabor wavelets and local directionality patterns. Int J Ambient Comput Intell 11(1):48–70. https://doi.org/10.4018/ijaci.2020010103
39. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422. https://doi.org/10.1023/A:1012487302797
40. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat. https://doi.org/10.1214/aos/1013203451
41. Angadi S, Nandyal S (2020) Human identification system based on spatial and temporal features in the video surveillance system. Int J Ambient Comput Intell 11(3):1–21. https://doi.org/10.4018/ijaci.2020070101
42. Sapinski, Tomasz; Kaminska D, Pelikant A, Ozcinar C, Avots E, Anbarjafari G (2018) Multimodal database of emotional speech, video and gestures
43. Saratxaga I, Navas E, Hernáez I, Aholab I (2006) Designing and recording an emotional speech database for corpus based synthesis in Basque. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, http://www.lrec-conf.org/proceedings/lrec2006/pdf/19_pdf.pdf
44. Verma D, Mukhopadhyay D (2017) Age driven automatic speech emotion recognition system. In: Proceeding—IEEE international conference on computing, communication and automation, ICCCA 2016, Institute of Electrical and Electronics Engineers Inc., pp 1005–1010. https://doi.org/10.1109/CCAA.2016.7813862
45. Sundarprasad N (2018) Speech emotion detection using machine learning techniques. Master's thesis, San Jose State University, San Jose, CA, USA. https://scholarworks.sjsu.edu/etd_projects/628
46. Gao Y (2019) Speech-Based Emotion Recognition. Master's thesis, https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1_Gao_Ye_2019_MS.pdf
47. Venkataramanan K, Rajamohan HR (2019) Emotion recognition from speech. arXiv: 1912.10458
48. Praseetha V, Vadivel S (2018) Deep learning models for speech emotion recognition. J Comput Sci 14(11):1577–1587. https://doi.org/10.3844/jcssp.2018.1577.1587
49. Kerkeni L, Serrestou Y, Raoof K, Mbarki M, Mahjoub MA, Cleder C (2019) Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. Speech Commun 114:22–35. https://doi.org/10.1016/j.specom.2019.09.002

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.