Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19674-y

A novel two-way feature extraction technique using multiple acoustic and wavelets packets for deep learning based speech emotion recognition

Kishor B. Bhangale[1] · Mohanaprasad Kothandaraman[2]

Received: 1 August 2023 / Revised: 25 March 2024 / Accepted: 10 June 2024

Abstract

Affective computing is crucial in various Human–Computer Interaction (HCI) and multimedia systems for comprehensive emotional assessment and response. The existing Speech Emotion Recognition (SER) provides limited performance due to inadequate frequency and time domain representation, poor correlation in global and local features, and contextual dependencies of components. The traditional SER techniques often result in poor accuracy due to spectral leakage, low-frequency resolution problems, and poor depiction of emotional speech's pitch, intonation, and voice timbre. This paper presents a novel two-way feature extraction (TWFR) based SER system using 2D-CNN and 1D-CNN to improve the distinctiveness of emotional speech. The first set of features, a 2-D representation of the wavelet packet decomposition (WPD) coefficients, is given to a 2-D Deep Convolution Neural Network (DCNN). The second set of features comprises various time-domain, spectral, and voice-quality features given to 1D-DCNN. The features from the last layer of 2D-DCNN and 1D-DCNN are concatenated and provided to a fully connected layer, followed by a softmax classifier for SER. The results of the TWFR-based SER scheme are assessed on EMODB and RAVDESS datasets based on recall, precision, accuracy, and F1-score. The proposed TWFR-based SER shows an overall accuracy of 98.48% for EMODB and 98.71% for RAVDESS datasets. The proposed TWFR-based SER helps improve the speech's pitch, intonation, and voice timbre in the spectral and time domain for SER and outpaces the current state of the arts.

Keywords Affective computing · Deep convolution neural network · Deep learning · Speech emotion recognition · Speech recognition · Speech analysis · Wavelet transform

✉ Mohanaprasad Kothandaraman
kmohanaprasad@vit.ac.in

[1] Vellore Institute of Technology, Chennai 600127, India

[2] Vellore Institute of Technology, Chennai, India

Published online: 17 June 2024

1 Introduction

Automatic SER is crucial in HCI, as it recognizes the emotion from the speech signal regardless of its semantic content. Speech emotion is an element of natural voice communication that humans can realize instinctively [1, 2]. Widespread research is being done on the capacity of programmable devices to identify this feeling using distinct behavioral and physiological modalities, including facial expression, muscle signals, Electroencephalography (EEG), body resistance, Electrocardiography (ECG), speech, etc. [3]. Speech is crucial for identifying emotions since it can be rapidly and cheaply collected. The SER is the process of the mapping of low-level speech information to high-level output class labels or emotion magnitudes in terms of arousal and valence or class labels or scalar values of emotion magnitudes, such as arousal and valence, is known as speech emotion recognition [4, 5]. SER is widely used in call centers, multimedia data analysis on social media, mobile phones, affective robots, HCI systems, clinical investigations, interactive games, banking, customer care centers, audio surveillance, audio conferencing, web-based e-learning, entertainment, etc. [6, 7].

All SER systems must be generalized because it is challenging to discern human speech emotions. Language and paralinguistic information are included in the human voice signal. The linguistics material illustrates the meaning and context of speech. Paralinguistic data, often independent of the speaker, age, language, dialect, gender, accent, and linguistic content, refers to implicit information such as mood, stress, etc. Age and accents are two important parameters affecting emotional voice. Increased age shows altered resonance, decreasing pitch, and lower vocal control. The change in accent leads to variation in intonation, rhythm and voice distinctiveness [8, 9]. Various emotions, including boredom, contempt, fear, melancholy, pleasure, excitement, surprise, and neutrality, are expressed via speech [10, 11]. Frequently, paralinguistic information is unconnected to the language, speaker, or linguistic content. Different emotions have a substantial influence on the numerous voice properties. The prosodic features comprises of intensity, voice quality, pitch, speaking rate, and voice variation [12]. The standard deviation and mean depicts the impact of emotions on the long term characteristics of speech [13]. With traditional machine learning (ML) techniques, a classifier is trained to provide the desired results by learning features from the raw speech signals. Different continuous features, prosodic features, spectral features, qualitative features, transform-based features, and hybrid features are all included in the feature extraction. Principal Component Analysis (PCA), Mel-Frequency Cepstrum Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), Linear Predictor Cepstral Coefficients (LPCC), Perceptual Linear Prediction coefficients (PLP), and other hand-crafted feature extraction techniques have all been presented in the past for the SER. In the classification step, the characteristics gleaned from the unprocessed voice signals are learned, and the particular emotion is predicted. The K-Nearest Neighbour classifier (KNN), Ensemble Classifier, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Classification Tree (CT), Dynamic Time Warping (DTW), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), etc., are commonly used classification algorithms that have been utilized in recent years for SER. Selecting features is a common issue with this approach, and the classifier's success heavily depends on these manually created features. It is challenging to foresee which characteristics would lead to higher performance. Traditional handmade elements cannot differentiate and correlate [14, 15].

The generalized SER includes the training and testing phase. The ML or deep learning (DL) algorithms are used during training to train the classifier based on specially created speech features. The trained model converts the unknown real-time samples to specific emotion labels in the testing phase. Data preprocessing, feature extraction, feature normalization, feature selection, and classification are essential steps in each stage of SER. Data preprocessing includes standardization, noise removal, and artifact removal to improve unprocessed voice signals. Using a variety of feature extraction approaches, the feature extraction is crucial for acquiring the key characteristics of the particular emotion. Feature selection is essential to acquire crucial characteristics and lessen SER system complexity. Finally, several classifiers based on DL or ML were used for SER. A DL is a potential field for SER that converts low-quality speech features into high-quality hierarchical abstract-level features. Numerous benefits are offered, including the capacity to handle unlabeled data, deal with complicated speech features and structure, needless parameter adjustment, and process more extensive datasets [16, 17].

Most SER systems suffer difficulties due to non-uniform and heavily skewed databases. Neutral speech samples are more often used while recording and interpreting SER data than other emotional speech samples, creating significantly unbalanced data. The data augmentation strategy, which uses artificial intelligence-based methods to create synthetic samples, is the most popular method for addressing the issue of data imbalance. The hand-crafted features' decreased dimensionality compared to spectrograms or raw waveforms makes them more suitable for data augmentation modeling. However, producing enhanced samples of a hand-crafted feature set limits its future application to compatible SER models. Future investigation may train models directly on the raw speech signal waveforms [18], on derived features [19], or on multi-channel audio [20] thanks to raw signals or spectrograms. The Generative Adversarial Network (GAN), invented by Goodfellow et al. in 2014 [21], is one of the most well-known methods to amplify and reinforce speech samples. Due to its capacity to provide a variety of solutions for a given sample, improved learning of the likelihood distribution of challenging real-time issues, and capacity to learn from noisy and unlabeled data, GAN is growing in popularity [22]. Different GAN-based methods have been proposed for enhancing speech data utilizing raw speech signal characteristics or spectrograms [23–25]. However, the produced samples often fail to maintain the real-time samples' marginal distribution.

Wavelet transform-based techniques have revealed a better spectral and time-domain depiction of the speech emotion and help to provide better resolution at lower frequencies. Wang et al. [26] presented WPC for the speaker-independent emotion recognition feature representation. It is observed that the Sequential Floating Forward Search (SFFS) based feature selection of WPC decomposed up to five levels (db2 filter) provides efficient feature selection and results in 79.2% and 79.5% accuracy for radial basis SVM (RSVM) and linear SVM (LSVM) respectively for EMODB dataset. Meng et al. [27] suggested that adding WPC features helps boost the spectral and time-domain properties of the speech signal, and recurrent neural network (RNN) assists in enhancing the contextual emotional dependencies of speech. It resulted in 82.26% and 66.90% accuracy for SER for the EMODB. Badshah et al. [28] presented an SER system based on 2-D DCNN that used a spectrogram. They used three layers of CNN and three fully connected layers, resulting in 84.3% accuracy of the EMODB. The sequential DCNN architecture provides less generalization capability and shows less results unseen data. Zhao et al. [29] proposed a combination of 1-D CNN and 2-D CNN with LSTM to enhance emotional speech's long-term dependencies and time-domain representation capability. The 2-D DCNN shows better spatial and spectral characteristics and substantially improved overall accuracy compared with the 1-D DCNN. It resulted in an overall accuracy of 95.33% for speaker-specific SER and 95.89% for speaker-independent SER for EMODB. Aftab et al. [30] investigated Full CNN (FCNN) to acquire the high-order features of the emotional voice. It encompasses two parts for speech representation. The first part uses MFC spectrograms to characterize the time–frequency feature depiction of speech. The second part describes the high-level emotion-related features by learning local features. It has given an overall accuracy of 94.21% and 79.89% for EMODB and IEMOCAP, respectively. Agrawal et al. [31] proposed a TWFR using PCA and MFCC for SER. It encompasses two DL frameworks, DNN and VGG16. The first stage consists of spectral feature selection using PCA and feature distinctiveness improvement using DNN. The second phase learns the emotion-specific attributes using MFCC and VGG16. It resulted in an overall SER accuracy of 81.94% for the RAVDESS dataset. It needs higher training parameters (138 M for VGG16 and 782 K for DNN) that limit its implementation flexibility on real-time devices with limited computational capacity. The MFCC is subjected to spectral leakage problems and low-frequency resolution, providing low results for lower arousal emotions.

Mustaqeem and Kwon [32] investigated 1-D dilated CNN, which can extract emotion-related features from raw speech. It has given SER accuracy of 90% for EMODB. The 1-D dilated CNN uses the bidirectional gated recurrent unit (BiGRU) to boost the time–frequency domain characteristics of voice. Farooq et al. [33] utilized a Mel log spectrogram (MLS) and 2-D DCNN for SER. The MLS-2-D DCNN shows a superior spatial and spectral depiction of emotional voice. It offers an overall accuracy of 90.5% and 73.5% for EMODB and RAVDESS, respectively. Further, Mustaqeem et al. [34] explored the combination of CNN and radial basis function network (RBFN) for SER. The CNN uses a short-time Fourier transform spectrogram (STFT) of speech to describe spectral-time domain properties. The Bidirectional LSTM is utilized to enhance SER precision. It delivers an overall SER accuracy of 85.57% for EMODB and 77.02% for RAVDESS. However, it results in higher trainable parameters (> 3 M). Chen et al. [35] suggested SER based on attention-based convolution RNN (ACRNN) and 3-D Mel spectrograms to enhance the feature representation of emotion-related content in speech. This resulted in SER accuracy of 82.82% for EMODB, but it needs a higher training time of 6811 s. Further, Meng et al. [36] explored dilated CNN-BLSTM with an attention layer (ADRNN) to increase the long-term dependency of speech. The 3-D Mel spectrogram and ADRNN offer an SER rate of 88.98% for EMODB.

Zhao et al. [37] combine 2-D CNN and 1-D CNN to boost the distinctiveness of emotion-specific features. The merged DCNN utilizes Bayesian optimization to optimize the learning process, which provided an accuracy of 91.78% for EMODB. It has shown that proposed DL frameworks need huge trainable parameters (> 10 M), increasing their computational volume. Bilal [38] utilized different speech features such as root mean square (RMS), chroma, spectral, MFCC features, and spectrogram representation for SER. These features are provided to ResNet, which offers an accuracy of 90.21% for EMODB and 79.41% for RAVDESS.

The wavelet transform has shown a superior spectral representation of the signals. It acquires the local information over a short period and spectral band to characterize the impact of emotion on speech [39–42]. The EMODB [43] and RAVDESS [44] datasets are widely used for the SER because of their distinctiveness, public availability, easy access, and availability for male/ female voice samples. The neural networks have shown the capability for high efficiency and high precision and can be effectively utilized for signal processing applications. The problem of low convergence rate and poor robustness in parallel neural networks has been solved using varying parameter DL frameworks [45–47]. The

DL-based Stacked Autoencoder (SAE) has shown a better global and local representation of emotional speech. The SAEs show the capability to distinguish the different emotions in the speech. However, the SAEs are computationally extensive, leading to over-fitting for higher dimensional data [48]. The real-time implementation of the SER systems is challenging because of the high dimensional feature vector that increases the computational intricacy of the SER systems. Therefore, selecting minimal and optimal speech features for emotion depiction is essential to lessen the computational volume and enhance the SER accuracy [49, 50].

The extensive survey of current SER systems shows that SER systems based on MFCC spectrogram suffer from reduced variance, frequency resolution issues, and spectrum leakage issues that lead to a poor SER identification rate. The intricate DL architecture raises the SER system's computational complexity. Due to the unequal distribution of emotion class samples in the training dataset, the performance of SER systems is constrained.

Thus, the proposed article presents a TWFR-based SER scheme based on a deep learning framework that improves feature representation, frequency resolution problems, poor feature variance, and spectral leakage problems. The proposed DL framework consists of the parallel combination of 2-D DCNN that improves the spatial representation of the spectral features and 1-D DCNN that acquires the emotion-specific patterns in Multiple Acoustic Features (MAFs). The chief offerings of the suggested article can be emphasized as follows:

- To improve the spectral and time-domain representation and enhance the low-frequency resolution of emotional speech signals using wavelet packet decomposition features and 2D-deep convolutional neural network
- To improve pitch, intonation, and voice timbre in the spectral and time-domain domain using Multiple Acoustic Features encompassing spectral, time-domain, and voice quality features along with 1D-DCNN.
- To improve the hierarchical feature representation and feature distinctiveness using a parallel combination of 2-D DCNN and 1-D DCNN.

The results of the proposed WPD-DCNN are evaluated for the different wavelet packet families such as Daubechies (dbN), Symlets (symN), Coiflets (coifN), and Fejer-Korovkin (fkN). It used three packets per family with different vanishing moments. The WPD uses 12 wavelet packets such as db1, db2, db3, coif1, coif2, coif3, sym4, sym5, sym6, fk4, fk6, and fk8. The overall system's effectiveness is evaluated using accuracy, F1-score, recall, precision, and selectivity on the public Emo-DB and RAVDESS database.

The remaining article is structured as follows: Section 2 details the proposed WPD-DCNN-based SER system. Section 3 discusses the database and experimental results. Lastly, Section 4 presents the conclusion and provides the scope for future improvement in the method.

2 Proposed methodology

The flow diagram of the proposed WPD-DCNN-based SER is illustrated in Fig. 1. It consists of a two-way feature extraction of the emotion signal. The first approach consists of 2-D WPD coefficients given to 2-D DCNN to capture the spectral and time-domain characteristics of the

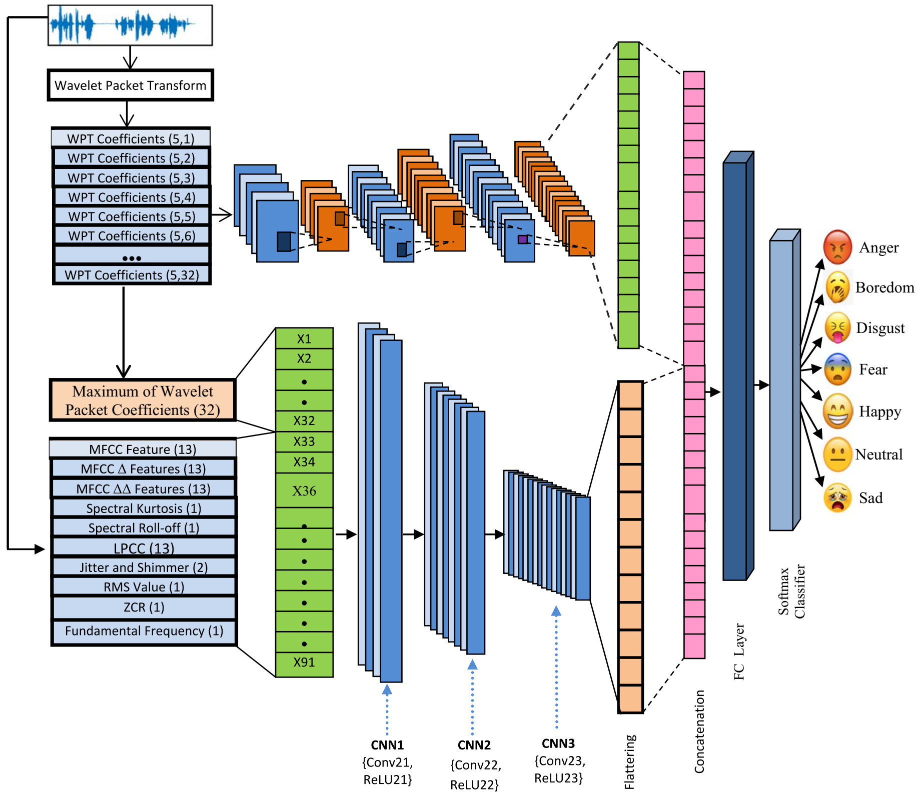


Fig. 1 Process of proposed WPT-DCNN-based SER

signal. The second approach consists of acoustic features such as spectral, time domain, voice-quality features, and a maximum of WPD packets.

These features are given to 1-D DCNN to improve the feature distinctiveness, learn emotion-specific patterns, minimize intra-class variance, and improve inter-class variance of the emotion features. Later, the flattened output of the 2D-DCNN and1D-DCNN is concatenated and given to the fully connected layer to connect each neuron with every other neuron and improve the features' local and global representation. The Softmax classifier is further used to classify emotion for two publics: EMODB and RAVDESS.

2.1 WPD + 2-D DCNN

The WPD gives a more precise frequency resolution than DWT for the speech signal. Unlike DWT, the WPD decomposes the low but also high-frequency sub-bands of the signal. The WPD maintains the smoothness, orthogonality, and localization properties of the signal and its parent wavelets [39–41]. The WPD decomposes the wavelet packet function $\Psi_j^i(n)$ up to L levels using $db3$ wavelet filters at various scales using Eq. 1 and 2.

$$\Psi_j^{2i}(n) = \sum_k h(k)\Psi_{j-1}^i(n - 2^{j-1}k) \tag{1}$$

$$\Psi_j^{2i+1}(n) = \sum_k g(k)\Psi_{j-1}^i(n - 2^{j-1}k) \tag{2}$$

Here, $g(k)$ denotes high pass quadrature mirror filter (QMF), whereas $h(k)$ stands for low-pass QMF, as described by Eq. 3 and 4, respectively. The QMF provides efficient orthogonal wavelet decomposition and better local features of the speech signal using a two-channel filter bank structure [42].

$$h(k) = \langle \Psi_j^{2i}(u), \Psi_{j-1}^i(u - 2^{j-1}k) \rangle \tag{3}$$

$$g(k) = \langle \Psi_j^{2i+1}(u), \Psi_{j-1}^i(u - 2^{j-1}k) \rangle \tag{4}$$

The emotional speech signal $x(n)$ is decomposed to level j using Eq. 5. The $X_j^i(k)$ represents k^{th} wavelet packet coefficient for i^{th} packet at level j.

$$x(n) = \sum_{i.k} X_j^i(k)\Psi_j^i(n - 2^j k) \tag{5}$$

where, $X_j^i(k)$ is kth wavelet packet coefficient at i$^{\text{th}}$ packet at j level that signifies the strength of the localized wavelet $\Psi_j^i\left(n - 2^j k\right)$ as given in Eq. 6.

$$X_j^i(k) = \langle x(n), \Psi_j^i(n - 2^j k) \rangle \tag{6}$$

The group of wavelet packets for the L level can be given in Eq. 7. The filter decimation recursion provides reduced time resolution and increased frequency resolution.

$$X_L(k) = \begin{bmatrix} X_L^0(k) \\ X_L^1(k) \\ . \\ . \\ X_L^{2^{L-1}}(k) \end{bmatrix} \tag{7}$$

In the WPD, approximation $g(n)$ and detailed information $h(n)$ signals are further decomposed at the next level to get a better local frequency resolution. The WPD considers the Shannon entropy function for the binary decomposition of the signal. Figure 2 illustrates the three-level WPD decomposition of the speech signal. The WPD generates 2^L sets of coefficients for the L levels.

Further, the coefficient sets are arranged in a two-dimensional matrix to form the WPD features. Each row represents the wavelet coefficients set, and each column denotes the coefficient value. The WPD feature set for fifth-level decomposition is described by Eq. 8. The fifth-level WPD decomposition creates the 2-D feature vector of $32 \times k$ where k represents the number of coefficients in every packet. The original signal has 64000 sample points in each speech signal. Therefore, the fifth level decomposition of the emotion speech signal consists of 2000 coefficients in every packet. Thus, the two-dimensional WPD matrix consists of dimensions of 32×2000, further provided to the 2D-DCNN to improve the connectivity and correlation between different packet coefficients. In Eq. 8, X_L^k Represents the k$^{\text{th}}$ packet of L$^{\text{th}}$ level WPD of speech such that $j = 1,2,3, ...2^L$ and $i = 1,2,3, \dots .2000$.

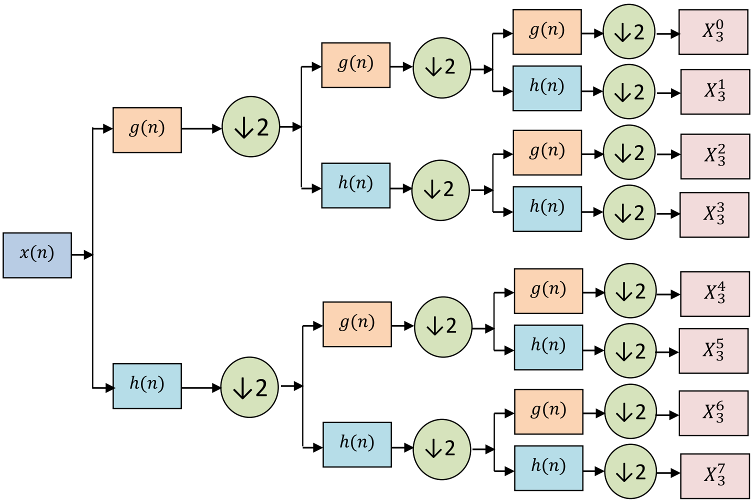


Fig. 2 Wavelet packet decomposition tree

$$f = \begin{bmatrix} X_5^0 \\ X_5^1 \\ X_5^2 \\ X_5^3 \\ X_5^4 \\ X_5^5 \\ \vdots \\ X_5^{32} \end{bmatrix} = \begin{bmatrix} X_5^0(0) & X_5^0(1) & X_5^0(2) & \cdots & X_5^0(k) \\ X_5^1(0) & X_5^1(1) & X_5^1(2) & \cdots & X_5^1(k) \\ X_5^2(0) & X_5^2(1) & X_5^2(2) & \cdots & X_5^2(k) \\ X_3^3(0) & X_5^3(1) & X_5^3(2) & \cdots & X_5^3(k) \\ X_5^4(0) & X_5^4(1) & X_5^4(2) & \cdots & X_5^4(k) \\ X_5^5(0) & X_5^5(1) & X_5^5(2) & \cdots & X_5^5(k) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ X_5^{32}(0) & X_5^{32}(1) & X_5^{32}(2) & \cdots & X_5^{32}(k) \end{bmatrix} \tag{8}$$

The proposed lightweight 2-D DCNN consists of three layers of CNN that encompass the convolution layer (*Conv*), rectified linear unit layer (*ReLU*), and maximum pooling layer (*MaxPool*). The *Conv* layer provides the localized features and connectivity in the different wavelet packets to capture the effect of emotion in the spectral domain. The *Conv* operation for WPD feature representation *(X)* and convolution filter *(F)* with $w \times w$ size is given by Eq. 9. The *Conv* layer provides the correlation between different spectral components decomposed using WPD at local level.

$$Conv(x, y) = \sum_{i=1}^{w} \sum_{j=1}^{w} X(x-i, y-j).F(i,j) \tag{9}$$

The ReLU layer improves the non-linearity by replacing the negative values with zero as shown in Eq. 10. The ReLU layer fastens the training and assits to avoid gradient vanishing problem.

$$ReLU(x, y) = max(Conv(x, y), 0) \tag{10}$$

Here, x and y represents the position of the neuron in WPD 2-D representation. Further, the MaxPool layer selects the prominent features and helps to minimize the feature dimensions.. Equation 11 provides the extraction of the maximum value from the ReLU layer.

$$\mathrm{MaxPool(x,y)} = \max_{\substack{x=1\,:\,row-wm,\\ y=1\,:\,col-wm}} \{\mathrm{ReLU(x+wm-1,y+wm-1)}\} \tag{11}$$

where *wm* is the pooling window, and *MaxPool* is the output of the MaxPool layer. The proposed 2-D DCNN includes 64, 128, and 256 convolution filters with a stride of one pixel in the first, second, and third CNN layers, respectively.

2.2 Multiple acoustic features + 1-D DCNN

Different spectral, voice quality and time-domain features form the MAF set. The MAFs enhance the feature distinctiveness, minimize intra-class variance, and improve inter-class variance of the emotion features. The fundamental frequency depicts the rise and fall in emotional expression. The time-domain features include fundamental frequency and Zero Crossing Rate (ZCR). The fundamental frequency is a higher pitch for joy, excitement, and boredom, whereas there is a lower pitch for calm, sadness, and anger. The ZCR provides abrupt changes and noise measures for the voice. The ZCR has a higher value for high-arousal emotions and a low value for low-arousal emotions such as calm or serious. The voice quality feature encompasses emotional speech's jitter, shimmer, and root mean square value (RMS).

The jitter and shimmer depict the emotional voice's disparity and stability in time and amplitude. The RMS provides the overall intensity or energy of the speech. Anger, excitement, and stress have higher RMS values because of high voice intensity. The low arousal emotions, such as calmness and sadness, have lower RMS values. The spectral features include Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Spectral Kurtosis (SK), Spectral Rolloff (SR), and WPD features. MFCC provides the spectral attributes of timbre, pitch, and formants of emotional voice. MFCC Δ gives MFCC variations to depict prosodic and phoneme variation transitions over time. The MFCC$\Delta\Delta$ provides the MFCC acceleration to describe the minor time-domain variation in speech due to emotion [51, 52].

LPCC provides an emotion-specific compact representation of speech to characterize intonation and prosody. The SK offers detailed information regarding energy distribution over the distinct frequency bands. High SK indicates that energy is accumulated at significantly fewer frequency components, whereas lower SK describes uniform energy distribution. The SR denotes the frequency value below which 85% of the power of the speech spectrum is accumulated.

Further, a maximum of fifth-level WPD packets are added to MAFs to enhance the special and time-domain resolution at a lower frequency [39, 53, 54]. The emotions with higher arousal values have high SR. The components extracted are summarized in Table 1.

The one-dimensional features that characterize the speech signal's time domain, spectral domain, and voice quality features are further provided by 1-D DCNN. The 1-D DCNN consists of three CNN layers encompassing *Conv* and *ReLU* layers. After three convolution layers, the flattening layer output is concatenated with the flattening layer output of the first approach. The combined two-way features are later given to the FC and softmax layers for emotion recognition. The softmax is simple and provides probabilistic interpretations of the output classes. It is applicable for multiclass classification and compatible with different learning optimization algorithms. The representations of the probability function, which computes the probabilities associated with each class in the network, may be found in Eq. 12 and 13, respectively. The Softmax

Table 1 Details of MAFs

Type of the Features	Features	Number of Features
Time-domain Features	Fundamental Frequency	1
	ZCR	1
Voice Quality Features	Jitter	1
	Shimmer	1
	RMS value	1
Spectral Features	MFCC	13
	MFCC Δ	13
	MFCCΔΔ	13
	LPCC	13
	Spectral Kurtosis	1
	Spectral Rolloff	1
	WPD features	32
Total Features		91

classifier uses this set of probabilities to depict output emotion as given in Eq. 14. The class label with the highest probability provides output emotion label.

$$z_i = \sum_j h_j w_{ji} \tag{12}$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \tag{13}$$

$$\hat{y} = arg \overset{max}{i} p_i \tag{14}$$

Here, $\hat{y}$ represents the output class, z_i represents dense layer output, h_j denotes the hidden layer inputs, w_{ji} represents the final dense layer's weights and p_i denotes the likelihood of the output class.

3 Experimental results and discussions

The proposed SER scheme is implemented on the NVidia GPU system with 64 GB RAM and a 512 tensor core. The effectiveness of the suggested approach is estimated based on accuracy, recall, precision, and F1-score. The proposed algorithm is trained using ADAM optimizer for 200 epoch, initial learning rate of 0.001, batch size of 64 and crossentropy loss function. The configuration of the two parallel arms of the proposed TWFR-based SER model are described in Table 2.

Table 2 Configurations of 2-D DCNN and 1-D DCNN of TWFR-based SER scheme

DCNN Framework	Layer	Filter Size	Stride	Padding	Activations Map	Trainable Parameters	Total Trainable Parameters
WPD-2-D DCNN	WPD Representation	-	-	-	91×2000×1	-	2,161,671 ~ 2.16 M
	Conv11	3×3×64	[1 1]	[1 1]	32×2000×64	640	
	ReLU11	-	[1 1]	-	32×2000×64	-	
	MaxPool11	-	[2 2]	-	16×1000×64	-	
	Conv12	3×3×128	[1 1]	[1 1]	16×1000×128	73,856	
	ReLU12	-	[1 1]	-	16×1000×128	-	
	MaxPool12	-	[2 2]	-	8×500×128	-	
	Conv13	3×3×256	[1 1]	[1 1]	8×500×256	295,168	
	ReLU13	-	[1 1]	-	8×500×256	-	
	MaxPool12	-	[2 2]	-	4×250×256	-	
	FC Layer1	-	-	-	256,000×7	1,792,007	
	Output Layer	-	-	-	1×7	-	
MAF-1-D DCNN	MAF Features	-	-	-	91×1×1	-	858,880 ~ 858 K
	Conv21	3×3×64	[1 1]	[1 1]	91×3×64	640	
	ReLU21	-	[1 1]	-	91×3×64	-	
	Conv22	3×3×128	[1 1]	[1 1]	91×3×128	73,856	
	ReLU22	-	[1 1]	-	91×3×128	-	
	Conv23	3×3×256	[1 1]	[1 1]	91×3×256	295,168	
	ReLU23	-	[1 1]	-	91×3×256	-	
	FCLayer2	-	-	-	69,888×7	489,216	
	Output Layer	-	-	-	1×7	-	

3.1 Dataset

The outcomes of TWFR-based SER are evaluated on the two open-source SER datasets, EMODB and RAVDES. EMODB is a dataset recorded for ten professional actors in the German language. It consists of 535 samples of 10 male and 10 female actors recorded at 48 kHz, further sampled to 16 kHz. It encompasses seven emotions: anger, boredom, fear, anxiety, sadness, happiness, and disgust [43].

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset consists of 1440 samples recorded in English. It includes audio recordings of eight emotions from 12 male and 12 female actors at the 48 kHz sampling rate. It contains eight emotions: calm, sad, angry, happy, fear, neutral, disgust, and surprise [44].

The dataset is split in the ratio of 70:15:15 for training, testing, and validation purposes. The experiments are carried out for the tenfold cross-validation. The suggested system provides a training accuracy of 99.50% and 100% for the EMODB and RAVDESS. The proposed model resulted in 99.50% and 99.80% validation accuracy for the EMODB and RAVDESS datasets. The sampling frequency is kept at 16 kHz, and the signal duration is maintained at 4 s by cropping the longer speech or appending the shorter speech samples to maintain uniformity in the dataset. A detailed description of the dataset is given in Table 3.

3.2 Experimental results and discussions for the EMODB dataset

Figure 3 and 4 provide the original speech signal and its first-level decomposition using WPD (db2). At every level, 2^L Packets are generated. The length of packet coefficients is down-sampled to half of its original size at every level.

The first level WPD packets (1,0) and packet (1,0) have 32,000 coefficients in each packet, as shown in Fig. 4. Figure 5 shows that the second level WPD packets (2,0), (2,1), (2,2), and (2,3) consist of 16,000 coefficients in every packet. In contrast, third-level WPD packets (3,0), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), and (3,7) include 8000 coefficients in each packet as illustrated in Fig. 6.

The efficiency of the suggested WPD-DCNN is evaluated for the different wavelet packet families such as coif1, coif2, coif3, sym4, sym5, sym6, db1, db2, db3, fk4, fk6 and fk8. The proposed WPD-DCNN provides highest overall accuracy of 98.45% for db2, whereas it results in an accuracy of 95.15% for db1, 97.14% for db3, 92.78% for sym4, 93.11% for sym5, 93.60% for sym6, 92.20% for coif1, 92.95% for coif2, 92.55% for coif3,

Table 3 Description of EMODB and RAVDESS dataset

Description	EMODB	RAVDESS
Number of Samples	535	1440
Type of Dataset	Acted	Acted
Number of Subjects	20 (10 Male and 10 Female)	24 (12 Male and 12 Female)
Sampling Rate	48 kHz	48 kHz
Duration	4 s	4 s
Total Emotion	7	8
Emotions	Anger, boredom, happiness, anxiety, fear, sadness, and disgust	Calm, sad, anger, happy, fear, disgust, and surprise
Availability	Public	Public

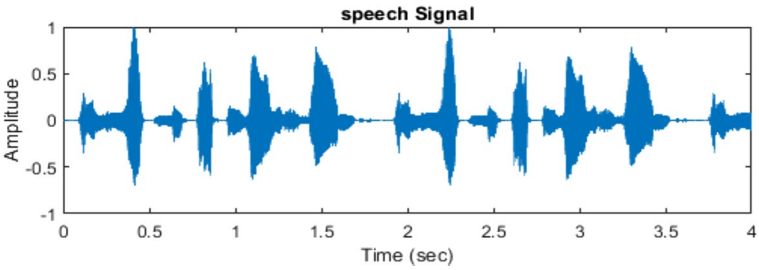
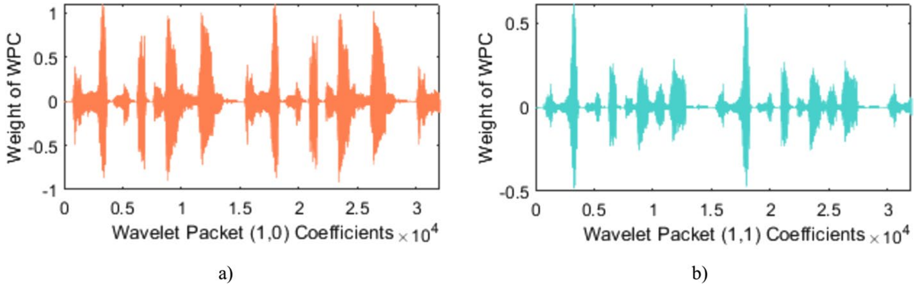


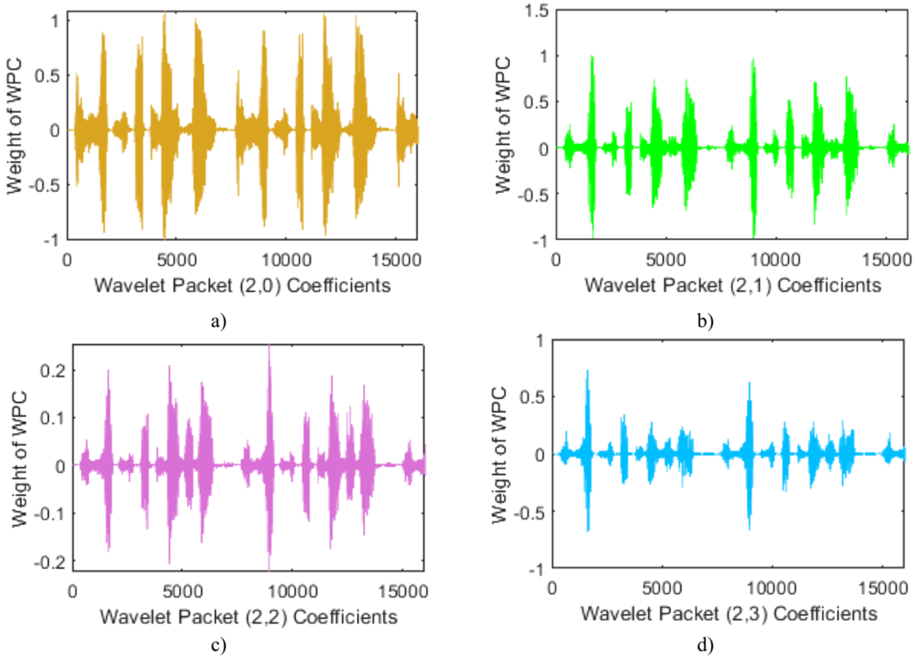
Fig. 3 Original Speech Signal (EMODB: Happy-09a01Fa.wav)



a)

b)

Fig. 4 WPD packet coefficients for 1.st level decomposition a) Wavelet packet (1,0) b) Wavelet packet (1,1)



a)

b)

c)

d)

Fig. 5 WPD packet coefficients for 2.nd level decomposition a) Wavelet packet (2,0) b) Wavelet packet (2,1) c) Wavelet packet (2,2) d) Wavelet packet (2,3)

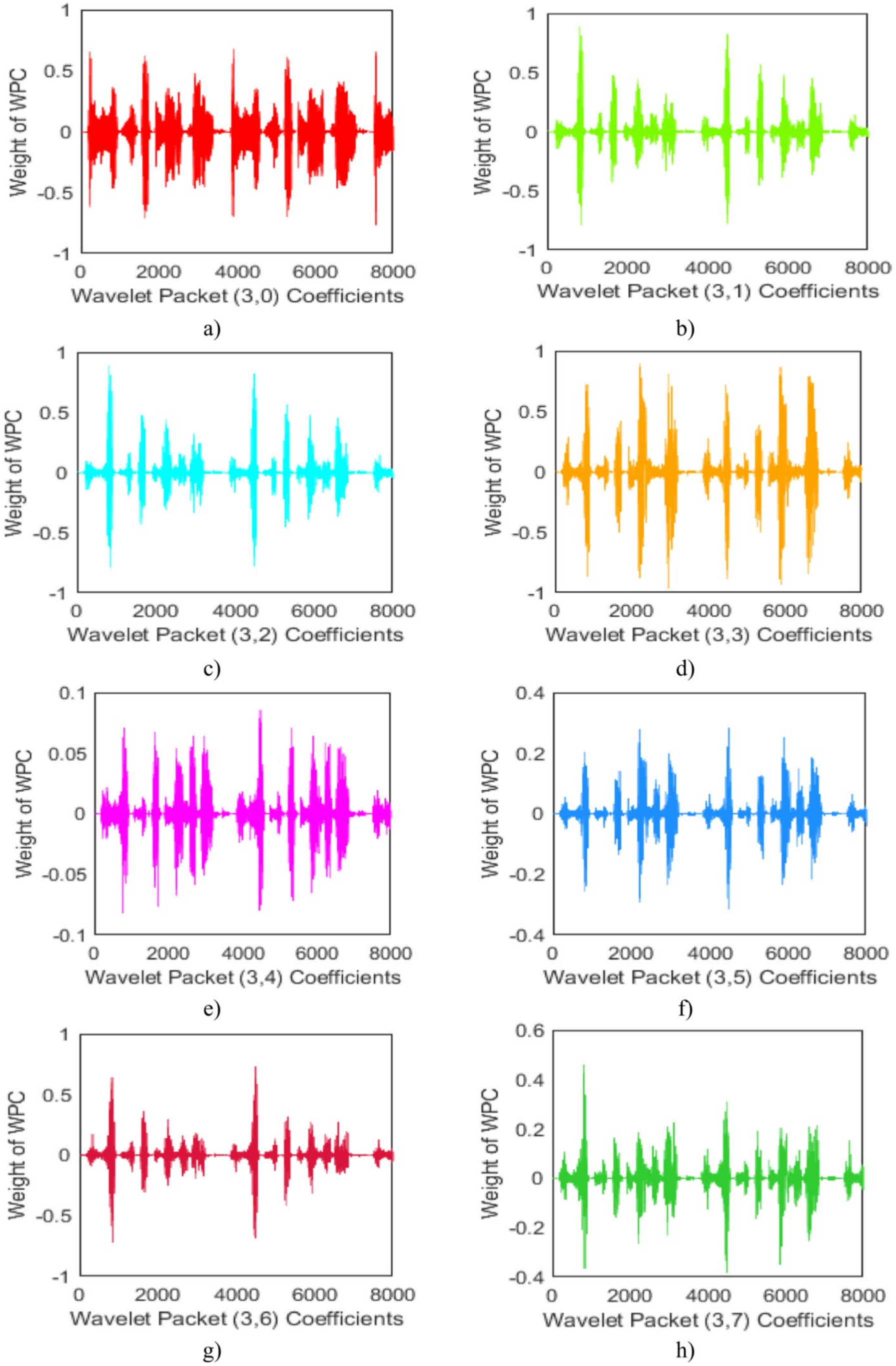
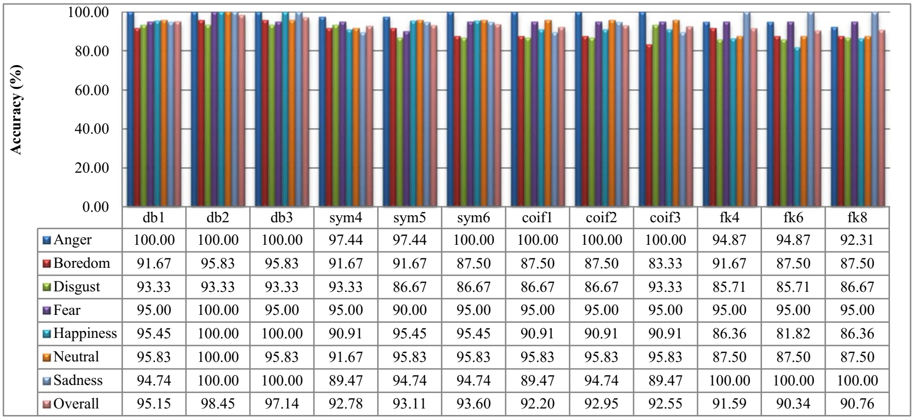
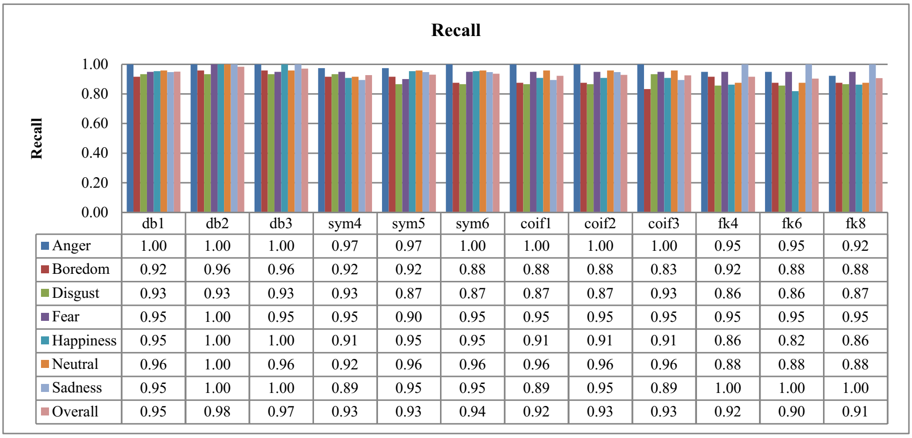


Fig. 6 WPD packet coefficients for 3.rd level decomposition a) Wavelet packet (3,0) b) Wavelet packet (3,1) c) Wavelet packet (3,2) d) Wavelet packet (3,3) e) Wavelet packet (3,4) f) Wavelet packet (3,5) g) Wavelet packet (3,6) h)Wavelet packet (3,7)



	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Anger	100.00	100.00	100.00	97.44	97.44	100.00	100.00	100.00	100.00	94.87	94.87	92.31
Boredom	91.67	95.83	95.83	91.67	91.67	87.50	87.50	87.50	83.33	91.67	87.50	87.50
Disgust	93.33	93.33	93.33	93.33	86.67	86.67	86.67	86.67	93.33	85.71	85.71	86.67
Fear	95.00	100.00	95.00	95.00	90.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
Happiness	95.45	100.00	100.00	90.91	95.45	95.45	90.91	90.91	90.91	86.36	81.82	86.36
Neutral	95.83	100.00	95.83	91.67	95.83	95.83	95.83	95.83	95.83	87.50	87.50	87.50
Sadness	94.74	100.00	100.00	89.47	94.74	94.74	89.47	94.74	89.47	100.00	100.00	100.00
Overall	95.15	98.45	97.14	92.78	93.11	93.60	92.20	92.95	92.55	91.59	90.34	90.76

Fig. 7 Accuracy for WPD-DCNN based SER for different filters for EMODB dataset (L = 3)



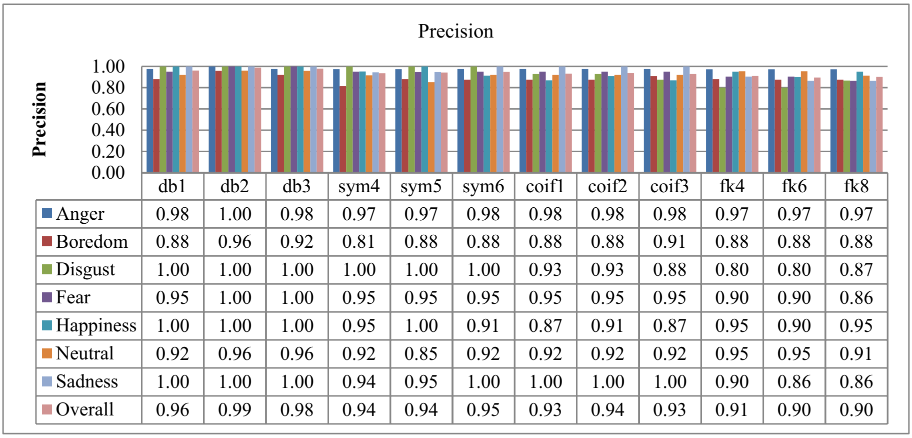
	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Anger	1.00	1.00	1.00	0.97	0.97	1.00	1.00	1.00	1.00	0.95	0.95	0.92
Boredom	0.92	0.96	0.96	0.92	0.92	0.88	0.88	0.88	0.83	0.92	0.88	0.88
Disgust	0.93	0.93	0.93	0.93	0.87	0.87	0.87	0.87	0.93	0.86	0.86	0.87
Fear	0.95	1.00	0.95	0.95	0.90	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Happiness	0.95	1.00	1.00	0.91	0.95	0.95	0.91	0.91	0.91	0.86	0.82	0.86
Neutral	0.96	1.00	0.96	0.92	0.96	0.96	0.96	0.96	0.96	0.88	0.88	0.88
Sadness	0.95	1.00	1.00	0.89	0.95	0.95	0.89	0.95	0.89	1.00	1.00	1.00
Overall	0.95	0.98	0.97	0.93	0.93	0.94	0.92	0.93	0.93	0.92	0.90	0.91

Fig. 8 Recall for WPD-DCNN based SER for different filters for EMODB dataset (L = 3)

91.59% for fk4, 90.34% for fk6 and 90.76% for fk8. Combining WPD-based features and MAFs improves the feature distinctiveness of the emotional features and provides better results for the proposed method than the traditional state of arts. The *db*2 filter provides superior speech representation capability compared with other wavelet filters and results in the highest accuracy of 98.45% for the EMODB dataset. The WPD (db2) and 2-D DCNN provide 100% accuracy for the anger, fear, happiness, neutral, and sadness emotions of EMODB (Fig. 7).

Recall rate provides the quantitative analysis of the SER as given in Fig. 8. The proposed WPD-DCNN provides overall recall of 0.95 for db1, 0.98 for db2, 0.97 for db3, 0.93 sym4, 0.93 for sym5, 0.94 for sym6, 0.92 for coif1, 0.93 for coif3,0.92 for fk4, 0.90 for fk6 and 0.91 for fk8.

The precision indicates the qualitative measures of the proposed WPD-DCNN-based SER, as shown in Fig. 9. It provides a higher accuracy of 0.99 for the db2 wavelet packet. In contrast, it gives a lower precision of 0.90 for fk6 and fk8 packets. It is observed that



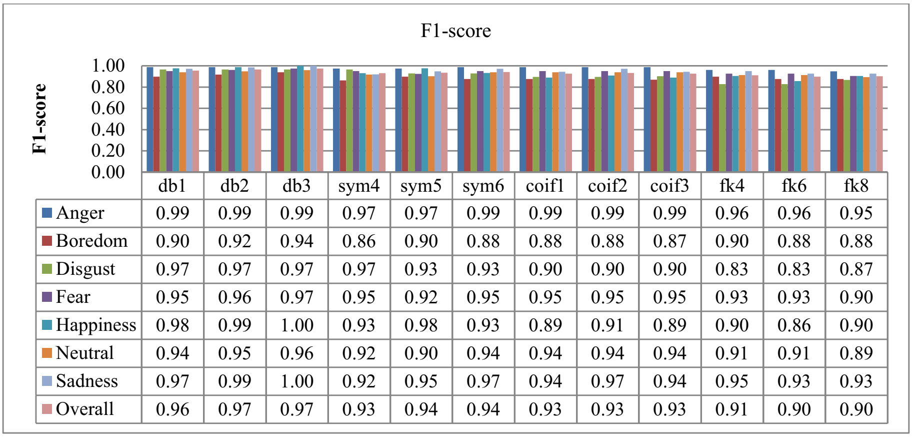
	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
■ Anger	0.98	1.00	0.98	0.97	0.97	0.98	0.98	0.98	0.98	0.97	0.97	0.97
■ Boredom	0.88	0.96	0.92	0.81	0.88	0.88	0.88	0.88	0.91	0.88	0.88	0.88
■ Disgust	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.93	0.88	0.80	0.80	0.87
■ Fear	0.95	1.00	1.00	0.95	0.95	0.95	0.95	0.95	0.95	0.90	0.90	0.86
■ Happiness	1.00	1.00	1.00	0.95	1.00	0.91	0.87	0.91	0.87	0.95	0.90	0.95
■ Neutral	0.92	0.96	0.96	0.92	0.85	0.92	0.92	0.92	0.92	0.95	0.95	0.91
■ Sadness	1.00	1.00	1.00	0.94	0.95	1.00	1.00	1.00	1.00	0.90	0.86	0.86
■ Overall	0.96	0.99	0.98	0.94	0.94	0.95	0.93	0.94	0.93	0.91	0.90	0.90

Fig. 9 Precision for WPD-DCNN based SER for different filters for EMODB dataset (L = 3)

sadness, anger, and disgust have higher precision, whereas boredom and neutral emotion have lower overall precision.

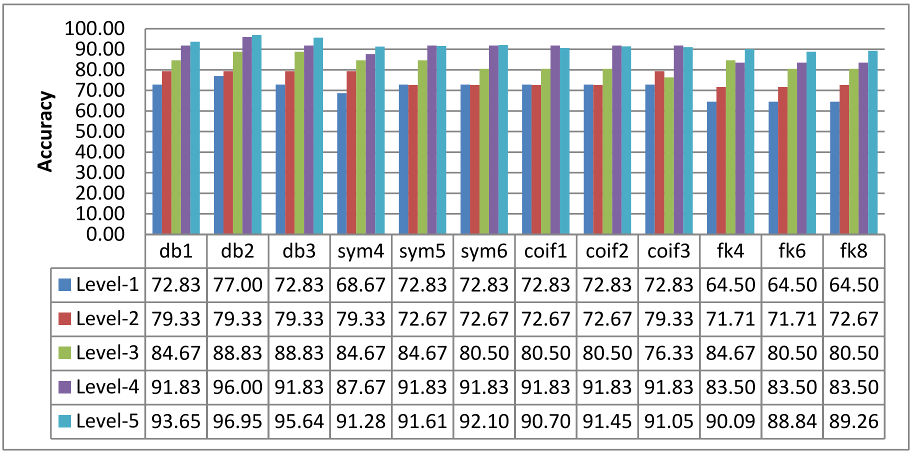
Figure 10 illustrates that TWFR-based SER with db2 (0.97) and db3 (0.97) wavelet packets provides a good balance between qualitative and quantitative results for the SER on the EMODB dataset compared with db1 (0.96), sym4 (0.93), sym5(0.94), sym6 (0.94), coif1 (0.93), coif2 (0.93), coif3 (0.93), fk4 (0.91), fk6 (0.90) and fk8 (0.90).

The results of the TWFR-based SER are validated for the different levels of decomposition for other wavelet packets for the EMODB dataset, as given in Fig. 11. It is observed that increasing the decomposing level increases the lower frequency resolution and helps to acquire the local characteristics of the emotion signal. It provides superior accuracy for the db2 (96.9%) packet over db1 (93.6%), db3 (95.6%), sym4 (91.2%), sym5 (91.6%), sym6 (92.1%), coif1 (90.7%), coif2 (91.4%), coif3 (91%), fk4 (90%), fk6 (88.8%), and fk8 (89.2%). The WPD decomposition level 1 to 5 provides 2, 4, 8, 16, and 32 decomposed packets for the original signal.



	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
■ Anger	0.99	0.99	0.99	0.97	0.97	0.99	0.99	0.99	0.99	0.96	0.96	0.95
■ Boredom	0.90	0.92	0.94	0.86	0.90	0.88	0.88	0.88	0.87	0.90	0.88	0.88
■ Disgust	0.97	0.97	0.97	0.97	0.93	0.93	0.90	0.90	0.90	0.83	0.83	0.87
■ Fear	0.95	0.96	0.97	0.95	0.92	0.95	0.95	0.95	0.95	0.93	0.93	0.90
■ Happiness	0.98	0.99	1.00	0.93	0.98	0.93	0.89	0.91	0.89	0.90	0.86	0.90
■ Neutral	0.94	0.95	0.96	0.92	0.90	0.94	0.94	0.94	0.94	0.91	0.91	0.89
■ Sadness	0.97	0.99	1.00	0.92	0.95	0.97	0.94	0.97	0.94	0.95	0.93	0.93
■ Overall	0.96	0.97	0.97	0.93	0.94	0.94	0.93	0.93	0.93	0.91	0.90	0.90

Fig. 10 Precision for WPD-DCNN based SER for different filters for EMODB dataset (L = 3)

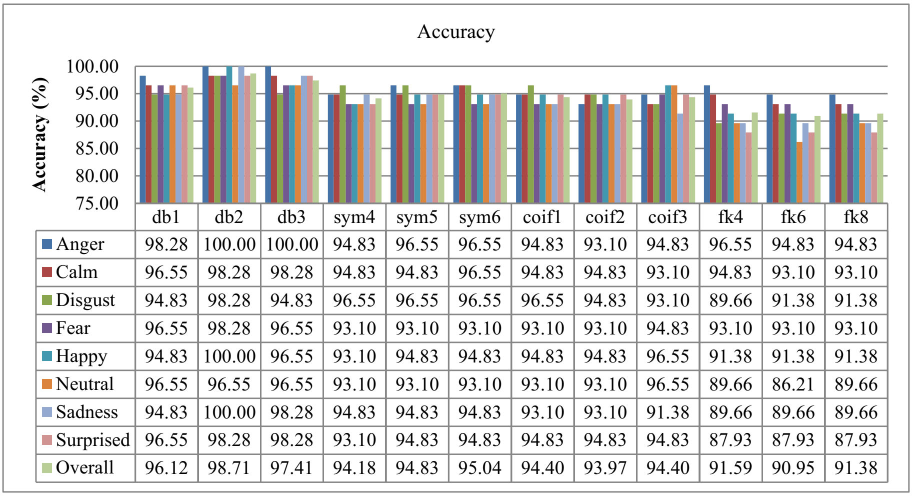


	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Level-1	72.83	77.00	72.83	68.67	72.83	72.83	72.83	72.83	72.83	64.50	64.50	64.50
Level-2	79.33	79.33	79.33	79.33	72.67	72.67	72.67	72.67	79.33	71.71	71.71	72.67
Level-3	84.67	88.83	88.83	84.67	84.67	80.50	80.50	80.50	76.33	84.67	80.50	80.50
Level-4	91.83	96.00	91.83	87.67	91.83	91.83	91.83	91.83	91.83	83.50	83.50	83.50
Level-5	93.65	96.95	95.64	91.28	91.61	92.10	90.70	91.45	91.05	90.09	88.84	89.26

Fig. 11 Overall accuracy for WPD-DCNN-based SER for different levels of decomposition for the EMODB dataset

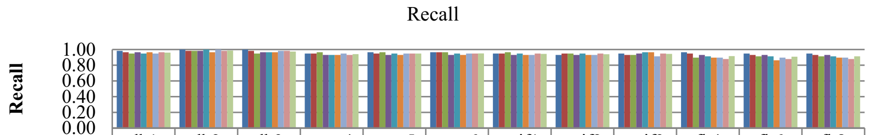
3.3 Experimental results and discussions for RAVDESS dataset

The performance of the proposed WPD-DCNN is evaluated on the RAVDESS dataset for four types of wavelet packets. Figures 12, 13, 14 and 15 show the accuracy, recall, precision, and F1-score of the proposed WPD-DCNN for SER for the RAVDESS dataset, respectively. The proposed scheme provides an overall accuracy of 98.71%, recall of 0.99, precision of 0.99, and F1-Score of 0.99 for the RAVDESS dataset for *db*2 filter. It gives 100% accuracy in describing anger, sadness, and happiness. However, it substantially increases the accuracy of low-arousal emotions such as disgust (98.28%)



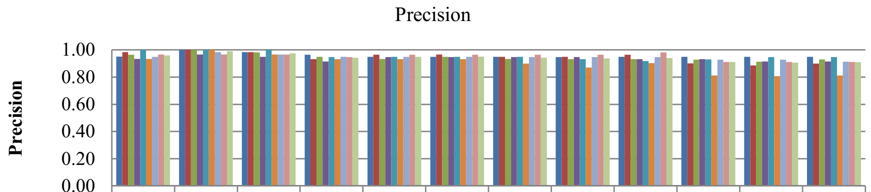
	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Anger	98.28	100.00	100.00	94.83	96.55	96.55	94.83	93.10	94.83	96.55	94.83	94.83
Calm	96.55	98.28	98.28	94.83	94.83	96.55	94.83	94.83	93.10	94.83	93.10	93.10
Disgust	94.83	98.28	94.83	96.55	96.55	96.55	96.55	94.83	93.10	89.66	91.38	91.38
Fear	96.55	98.28	96.55	93.10	93.10	93.10	93.10	93.10	94.83	93.10	93.10	93.10
Happy	94.83	100.00	96.55	93.10	94.83	94.83	94.83	94.83	96.55	91.38	91.38	91.38
Neutral	96.55	96.55	96.55	93.10	93.10	93.10	93.10	93.10	96.55	89.66	86.21	89.66
Sadness	94.83	100.00	98.28	94.83	94.83	94.83	93.10	93.10	91.38	89.66	89.66	89.66
Surprised	96.55	98.28	98.28	93.10	94.83	94.83	94.83	94.83	94.83	87.93	87.93	87.93
Overall	96.12	98.71	97.41	94.18	94.83	95.04	94.40	93.97	94.40	91.59	90.95	91.38

Fig. 12 Accuracy for WPD-DCNN based SER for different filters for RAVDESS dataset (L = 3)



	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
■ Anger	0.98	1.00	1.00	0.95	0.97	0.97	0.95	0.93	0.95	0.97	0.95	0.95
■ Calm	0.97	0.98	0.98	0.95	0.95	0.97	0.95	0.95	0.93	0.95	0.93	0.93
■ Disgust	0.95	0.98	0.95	0.97	0.97	0.97	0.97	0.95	0.93	0.90	0.91	0.91
■ Fear	0.97	0.98	0.97	0.93	0.93	0.93	0.93	0.93	0.95	0.93	0.93	0.93
■ Happy	0.95	1.00	0.97	0.93	0.95	0.95	0.95	0.95	0.97	0.91	0.91	0.91
■ Neutral	0.97	0.97	0.97	0.93	0.93	0.93	0.93	0.93	0.97	0.90	0.86	0.90
■ Sadness	0.95	1.00	0.98	0.95	0.95	0.95	0.93	0.93	0.91	0.90	0.90	0.90
■ Surprised	0.97	0.98	0.98	0.93	0.95	0.95	0.95	0.95	0.95	0.88	0.88	0.88
■ Overall	0.96	0.99	0.97	0.94	0.95	0.95	0.94	0.94	0.94	0.92	0.91	0.91

Fig. 13 Recall for WPD-DCNN based SER for different filters for RAVDESS dataset (L = 3)

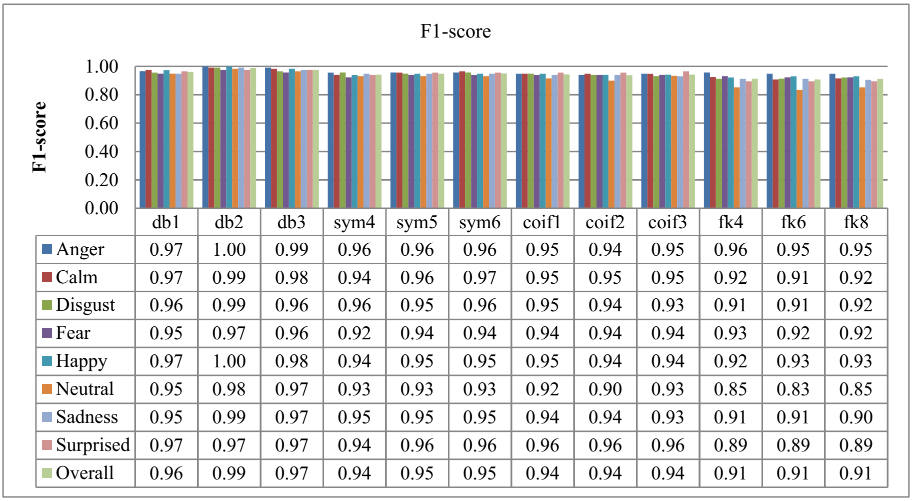


	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
■ Anger	0.95	1.00	0.98	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
■ Calm	0.98	1.00	0.98	0.93	0.96	0.97	0.95	0.95	0.96	0.90	0.89	0.90
■ Disgust	0.96	1.00	0.98	0.95	0.93	0.95	0.93	0.93	0.93	0.93	0.91	0.93
■ Fear	0.93	0.97	0.95	0.92	0.95	0.95	0.95	0.95	0.93	0.93	0.92	0.92
■ Happy	1.00	1.00	1.00	0.95	0.95	0.95	0.95	0.93	0.92	0.93	0.95	0.95
■ Neutral	0.93	1.00	0.97	0.93	0.93	0.93	0.90	0.87	0.90	0.81	0.81	0.81
■ Sadness	0.95	0.98	0.97	0.95	0.95	0.95	0.95	0.95	0.95	0.93	0.93	0.91
■ Surprised	0.97	0.97	0.97	0.95	0.96	0.96	0.96	0.96	0.98	0.91	0.91	0.91
■ Overall	0.96	0.99	0.97	0.94	0.95	0.95	0.94	0.94	0.94	0.91	0.91	0.91

Fig. 14 Precision for WPD-DCNN based SER for different filters for RAVDESS dataset (L = 3)

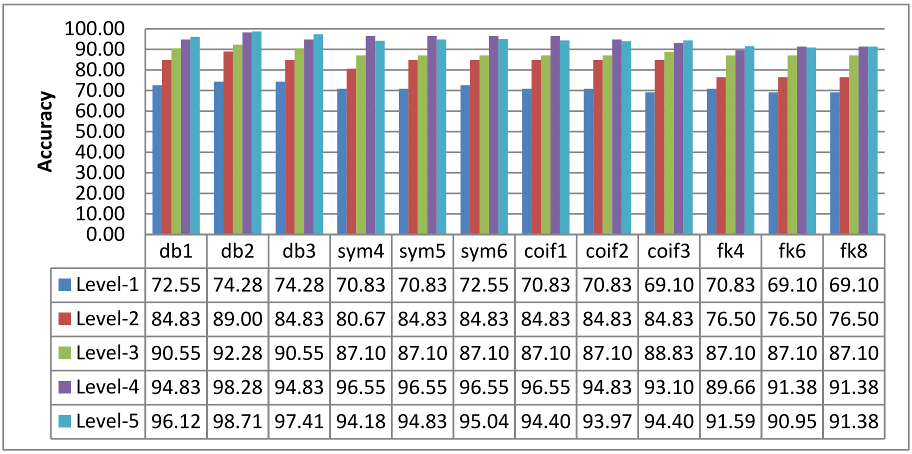
and calm (98.28%). It results in superior accuracy and better balance in recall and precision because of the even dataset size of RAVDESS compared with EMODB.

The outcomes of the proposed WPD-DCNN are validated for the different levels of decomposition for other wavelet packets for the RAVDESS dataset, as given in Fig. 16. It provides superior accuracy for the db2 (98.71%) over db1 (96.1%), db3 (97.4%), sym4 (94.1%), sym5 (94.8%), sym6 (95%), coif1 (94.4%), coif2 (93.9%), coif3 (94.4%), fk4 (91.5%), fk6 (90.9%), and fk8 (91.3%). The WPD decomposition level 1 to 5 provides 2, 4, 8, 16, and 32 decomposed packets for the original signal.



	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Anger	0.97	1.00	0.99	0.96	0.96	0.96	0.95	0.94	0.95	0.96	0.95	0.95
Calm	0.97	0.99	0.98	0.94	0.96	0.97	0.95	0.95	0.95	0.92	0.91	0.92
Disgust	0.96	0.99	0.96	0.96	0.95	0.96	0.95	0.94	0.93	0.91	0.91	0.92
Fear	0.95	0.97	0.96	0.92	0.94	0.94	0.94	0.94	0.94	0.93	0.92	0.92
Happy	0.97	1.00	0.98	0.94	0.95	0.95	0.95	0.94	0.94	0.92	0.93	0.93
Neutral	0.95	0.98	0.97	0.93	0.93	0.93	0.92	0.90	0.93	0.85	0.83	0.85
Sadness	0.95	0.99	0.97	0.95	0.95	0.95	0.94	0.94	0.93	0.91	0.91	0.90
Surprised	0.97	0.97	0.97	0.94	0.96	0.96	0.96	0.96	0.96	0.89	0.89	0.89
Overall	0.96	0.99	0.97	0.94	0.95	0.95	0.94	0.94	0.94	0.91	0.91	0.91

Fig. 15 F1-score for WPD-DCNN based SER for different filters for RAVDESS dataset (L = 3)



	db1	db2	db3	sym4	sym5	sym6	coif1	coif2	coif3	fk4	fk6	fk8
Level-1	72.55	74.28	74.28	70.83	70.83	72.55	70.83	70.83	69.10	70.83	69.10	69.10
Level-2	84.83	89.00	84.83	80.67	84.83	84.83	84.83	84.83	84.83	76.50	76.50	76.50
Level-3	90.55	92.28	90.55	87.10	87.10	87.10	87.10	87.10	88.83	87.10	87.10	87.10
Level-4	94.83	98.28	94.83	96.55	96.55	96.55	96.55	94.83	93.10	89.66	91.38	91.38
Level-5	96.12	98.71	97.41	94.18	94.83	95.04	94.40	93.97	94.40	91.59	90.95	91.38

Fig. 16 Overall accuracy for WPD-DCNN-based SER for different levels of decomposition for the RAVDESS dataset

3.4 Performance comparison with the previous state of arts

The results of the TWFR-based SER is compared with previous deep learning-based schemes utilized for SER for EMODB and RAVDESS datasets, as given in Table 4. The TWFR-based SER shows an improvement of 2.66%, 4.5%, 16.78%, 8.78%, and 7.26% over the 2D-CNN-LSTM [29], FCNN [30], DCNN-3 Layer [28], DCNN [33] and Merged DCNN [37] which has used Mel spectrogram based representation for the emotional speech representation for the EMODB dataset. It improves by 18.87% and 10.64% in SER accuracy over ACRNN [35] and ADRNN [36] using a 3-D Mel spectrogram representation of speech. The TWFR-based SER shows 9.38% superiority over the 1-D Dilated CNN

Table 4 Results comparison of proposed TWFR-based SER with previous techniques

Author and Year	Feature Representation	Deep Learning Model	% Accuracy		Total Trainable Parameters	Total Training Time (sec)
			EMODB	RAVDESS		
Badshah et al. (2017) [28]	Spectrogram	2-D DCNN (3-Layers)	84.3		-	-
Zhao et al. (2019) [29]	MLS	2D-CNN-LSTM	95.89		-	-
Aftab et al. (2022)[30]	MFCC Spectrogram	FCNN	94.21		-	-
Aggrawal et al. (2022) [31]	TWFR using spectral features and Mel Spectrogram	DNN- VGG16	-	81.94	782 K for DNN and 138 M for VGG16	-
Mustaqeem, and Kwon (2021) [32]	Raw Speech	1-D Dilated CNN	90	-	-	3150
Farooq et al. (2020) [33]	MLS	2-D DCNN	90.5	73.5	-	-
Mustaqeem et al. (2020) [34]	STFT	RBFN-BiLSTM	85.57	77.02	> 3 M	-
Chen et al. (2018) [35]	3-D Mel Spectrogram	ACRNN	82.82	-	-	6811
Meng et al. (2019) [36]	3-D Mel Spectrogram	ADRNN	88.98	-	-	7187
Zhao et al. (2018) [37]	MLS	Merged DCNN	91.78	-	> 10 M	-
Bilal et al. (2020) [38]	MFCC, Croma Features, RMS, Spectral Features	ResNet101	90.21	79.41	44.5 M	-
Bhangale and Mohanaprasad (2023) [39]	MAFs	DCNN	93.31	94.18	1.77 M	2132 (EMODB), 2180 (RAVDESS)
Proposed Method	WPD (db2)	2-D DCNN	94.00	95.83	2.16 M (EMODB), 2.41 (RAVDESS)	2520 (EMODB), 2634 (RAVDESS)
	MAFs	1-D DCNN	93.20	94.10	858 K (EMODB), 928 K (RAVDESS)	2132 (EMODB), 2290 (RAVDESS)
	WPD + MAFs	DCNN	98.45	98.71	3 M (EMODB), 3.33 (RAVDESS)	2845 (EMODB), 2930 (RAVDESS)

[32], which used raw speech signals. In recent years, many multiple feature extraction techniques have been utilized for the SER. The TWFR-based SER improved by 9.13% and 5.5% over the ResNet101 [38] and DCNN [39], which used MAFs for the EMODB dataset. EMODB dataset consists of limited samples for each class where the TWFR-based SER provides better performance by providing superior frequency resolution over a wide range of frequencies. When the consequences of the TWFR-based SER is evaluated on the RAVDESS dataset, it shows an improvement of 33.94%, 27.82%, 23.97%, 20.14%, and 4.53% over the DCNN [33], RBFN-BLSTM [34], ResNet101 [38], DNN-VGG16 [31] and DCNN [39] respectively.

The TWFR-based model needs total trainable parameters of 3 M for EMODB and 3.33 M for the RAVDESS, respectively. The trainable parameters are lower compared with traditional techniques such as ResNet101 (44.5 M), Merged DCNN (> 10 M), VGGNet (138 M), and RBFN-BiLSTM (> 3 M), which increase the deployment flexibility of the suggested SER scheme for real-time systems with restricted computational resources. The algorithm's effectiveness is also evaluated based on the total training time of the model, with the proposed model needing a total training time of 2845 s and 2930 s for EMODB and RAVDESS, respectively, considering WPD and MAFs. It requires a total training time of 2132 s and 2290 s for EMODB and RAVDESS when MAFs are considered for SER using 1-D DCNN. For WPD-based speech representation, the 2-D DCNN needs 2520 s and 2634 s for EMODB and RAVDESS, respectively. The combination of lightweight 2-D DCNN and 1-D DCNN provides significant improvement in total training of the model compared with 1-D Dilated CNN (3150 s), ACRNN (6811 s), and ADRNN (7187 s) for the EMODB dataset.

4 Conclusions and future scopes

This paper presents a two-way feature representation of the speech signal using wavelet packet coefficients and MAFs. WPD features help effectively capture emotions' vocal characteristics and reflect nonlinear vortex-flow interactions. The WPD provides superior time–frequency characteristics and better contextual dependencies using the 2-D DCNN algorithm. The 2-D DCNN helps to acquire the spatial information of spectral domain properties of emotional speech, whereas the 1-D DCNN learns the pattern of MAFs for different emotions. It provides robustness against spectral leakage problems, low-frequency resolution problems, and poor intonation, timbre, and emotional speech prosody representation. It substantially increases the accuracy of low-arousal emotions such as disgust, calm, and boredom. The proposed TWFR-based SER provides 98.45% and 98.71% accuracy for SER for EMODB and RAVDESS datasets. It provides recall of 0.98 and 0.99, precision of 0.99 and 0.99, and F1-Score of 0.97 and 0.99 for EMODB and RAVDESS, respectively. It shows superior SER accuracy performance compared with state-of-the-art techniques. In the future, the results of the proposed scheme can be improved by utilizing an efficient scheme for feature selection and hyper-parameter tuning of the DCNN architecture. The effectiveness of the SER scheme can be validated for the real-time and cross-corpus dataset in the future.

Author contribution **Kishor Bhangale:** Conceptualization, Methodology, Software, Validation, Visualization, Formal analysis, Writing – original draft. **Mohanaprasad Kothandaraman:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration.

Funding This research received no external funding.

Data availability The open access EMODB dataset can be found at http://emodb.bilderbar.info/showresults/index.php, and open access RAVDESS dataset can be found at https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

Declarations

Informed consent Not applicable.

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Maithri M, Raghavendra U, AnjanGudigar, Jyothi Samanth, Prabal DattaBarua, MurugappanMurugappan, YashasChakole, and U. Rajendra Acharya (2022) Automated Emotion Recognition: Current Trends and Future Perspectives. Computer Methods and Programs in Biomedicine*:* 106646. https://doi.org/10.1016/j.cmpb.2022.106646
2. Schuller BW (2018) Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun ACM 61(5):90–99. https://doi.org/10.1145/3129340
3. Dzedzickis A, Kaklauskas A, Bucinskas V (2020) Human emotion recognition: Review of sensors and methods. Sensors 20(3):592. https://doi.org/10.3390/s20030592
4. Swain M, Routray A, Kabisatpathy P (2018) Databases, features, and classifiers for speech emotion recognition: a review. Int J Speech Technol 21(1):93–120. https://doi.org/10.1007/s10772-018-9491-z
5. Gupta, Nehul, Vedangi Thakur, Vaishnavi Patil, Tamanna Vishnoi, and Kishor Bhangale (2023) Analysis of Affective Computing for Marathi Corpus using Deep Learning. In 4th International Conference for Emerging Technology (INCET) (1–8). https://doi.org/10.1109/INCET57972.2023.10170346
6. Bhangale, Kishor, and K. Mohanaprasad (2022) Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network. In Futuristic Communication and Network Technologies (241–250). Springer, Singapore. https://doi.org/10.1007/978-981-16-4625-6_24
7. Issa D (2020) FatihDemirci M, and Adnan Yazici (2020) Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control 59:101894. https://doi.org/10.1016/j.bspc.2020.101894
8. A. Bastanfard, D. Amirkhani and M. Hasani (2019) Increasing the Accuracy of Automatic Speaker Age Estimation by Using Multiple UBMs. In 5th Conference on Knowledge Based Engineering and Innovation (KBEI) (592–598). https://doi.org/10.1109/KBEI.2019.8735005.
9. R. Mahdavi, A. Bastanfard and D. Amirkhan (2020) Persian Accents Identification Using Modeling of Speech Articulatory Features. In 25th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, (1–9). https://doi.org/10.1109/CSICC49403.2020.9050139.
10. Sonawane, Anagha, Inamdar MU, and Kishor B. Bhangale (2017) Sound-based human emotion recognition using MFCC & multiple SVM. In International conference on information, communication, instrumentation and control (1–4). https://doi.org/10.1109/ICOMICON.2017.8279046
11. Anagnostopoulos CN, Iliou T, Giannoukos I (2012) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev 43(2):155–177. https://doi.org/10.1007/s10462-012-9368-5
12. Zhou Y, Sun Y, Zhang J, and Yan Y (2009) Speech emotion recognition using both spectral and prosodic features. In Information Engineering and Computer Science, 2009, IEEE (1–4). https://doi.org/10.1109/ICIECS.2009.5362730
13. Chattopadhyay S, Dey A, Singh PK, Ahmadian A, Sarkar R (2023) A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. Multimedia Tools and Applications 82(7):9693–9726. https://doi.org/10.1007/s11042-021-11839-3
14. Bhangale KB, Mohanaprasad K (2021) A review on speech processing using machine learning paradigm. Int J Speech Technol 24(2):367–388. https://doi.org/10.1007/s10772-021-09808-0

15. Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116(2020):56–76. https://doi.org/10.1016/j.specom.2019.12.001
16. Deng L, Yu D (2014) Deep learning: methods and applications. Found Trends Signal Process 7(3–4):197–387. https://doi.org/10.1561/2000000039
17. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003
18. Dinkel, Heinrich, Nanxin Chen, Yanmin Qian, and Kai Yu (2017) End-to-end spoofing detection with raw waveform CLDNNS. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (4860–4864). https://doi.org/10.1109/ICASSP.2017.7953080
19. Guo, Jinxi, Kenichi Kumatani, Ming Sun, Minhua Wu, Anirudh Raju, Nikko Ström, and Arindam Mandal (2018) Time-delayed bottleneck highway networks using a dft feature for keyword spotting. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (5489–5493). https://doi.org/10.1109/ICASSP.2018.8462166
20. Minhua, Wu, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, and BjörnHoffmeister (2019) Frequency domain multi-channel acoustic modeling for distant speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (6640–6644) https://doi.org/10.1109/ICASSP.2019.8682977
21. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, A. Courville, and Y. Bengio (2014) Generative adversarial nets. Advances in neural information processing systems :2672–2680
22. S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. EspyWilson (2017) Adversarial auto-encoders for speech based emotion recognition. In Proc. Interspeech (1243–1247). https://doi.org/10.48550/arXiv.1806.02146
23. Yi, Lu, and Man-Wai Mak (2019) Adversarial data augmentation network for speech emotion recognition. In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (529–534). https://doi.org/10.1109/APSIPAASC47483.2019.9023347
24. Bakır H, Çayır AN, Navruz TS (2023) A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification. Multimedia Tools and Applications, 1–28. https://doi.org/10.1007/s11042-023-16200-4
25. Su, Bo-Hao, and Chi-Chun Lee (2021) A Conditional Cycle Emotion Gan for Cross Corpus Speech Emotion Recognition. In IEEE Spoken Language Technology Workshop (SLT) (351–357). https://doi.org/10.1109/SLT48900.2021.9383512
26. Wang K, Guoxin Su, Liu Li, Wang S (2020) Wavelet packet analysis for speaker-independent emotion recognition. Neurocomputing 398:257–264. https://doi.org/10.1016/j.neucom.2020.02.085
27. Meng H, Yan T, Wei H, Ji X (2021) Speech emotion recognition using wavelet packet reconstruction with attention-based deep recurrent neutral networks. Bulletin of the Polish Academy of Sciences. Tech Sci 69(1):1–12. https://doi.org/10.24425/bpasts.2020.136300
28. Badshah, A, M., Jamil, A., Nasir, R., Sung, W (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In International conference on platform technology and service (PlatCon) (1–5).
29. Zhao J, Xia M, Lijiang C (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control 47:312–323. https://doi.org/10.1016/j.bspc.2018.08.035
30. Aftab, Arya, AlirezaMorsali, ShahrokhGhaemmaghami, and Benoit Champagne. (2022) Light-SERNet: A lightweight fully convolutional neural network for speech emotion recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (6912–6916). https://doi.org/10.1109/ICASSP43922.2022.9746679.
31. Aggarwal A, Srivastava A, Agarwal A, Chahal N, Singh D, Alnuaim AA, Alhadlaq A, Lee HN (2022) Two-way feature extraction for speech emotion recognition using deep learning. Sensors 22(6):2378. https://doi.org/10.3390/s22062378
32. Kwon S (2021) 1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features. CMC-Comput Mater Con 67(3):4039–4059
33. Farooq M, Hussain F, Baloch NK, Raja FR, Yu H, Zikria YB (2020) Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors 20(21):6008. https://doi.org/10.3390/s20216008
34. Mustaqeem SM, 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
35. Chen M, He X, Yang J, Zhang H (2018) 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. IEEE Signal Process Lett 25(10):1440–1444. https://doi.org/10.1109/LSP.2018.2860246

36. Meng H, Yan T, Yuan F, Wei H (2019) Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. IEEE Access 7:125868–125881. https://doi.org/10.1109/ACCESS.2019.2938007
37. Zhao J, Mao X, Chen L (2018) Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Proc 12(6):713–721. https://doi.org/10.1049/iet-spr.2017.0320
38. Mehmet B (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
39. Bhangale K, Kothandaraman M (2023) Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. Electronics 12(4):839. https://doi.org/10.3390/electronics12040839
40. Gokhale MY, Khanduja DK (2010) Time domain signal analysis using wavelet packet decomposition approach. Int'l J Commun, Net Syst Sci 3(3):321. https://doi.org/10.4236/ijcns.2010.33041
41. Cody MA (1994) The wavelet packet transform: Extending the wavelet transform." Dr. Dobb's J 19:44–46
42. Shi J, Liu X, Xiang W, Han Mo, Zhang Q (2020) Novel fractional wavelet packet transform: theory, implementation, and applications. IEEE Trans Signal Process 68:4041–4054. https://doi.org/10.1109/TSP.2020.3006742
43. Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss (2005) A database of German emotional speech. In Interspeech 5:1517–1520. http://emodb.bilderbar.info/showresults/index.php
44. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):e0196391
45. Zhang Z, Tingzhong Fu, Yan Z, Jin L, Xiao L, Sun Y, Zhuliang Yu, Li Y (2018) A varying-parameter convergent-differential neural network for solving joint-angular-drift problems of redundant robot manipulators. IEEE/ASME Trans Mechatron 23(2):679–689. https://doi.org/10.1109/TMECH.2018.2799724
46. Zhang Z, Yeyun Lu, Zheng L, Li S, Zhuliang Yu, Li Y (2018) A new varying-parameter convergent-differential neural-network for solving time-varying convex QP problem constrained by linear-equality. IEEE Trans Autom Control 63(12):4110–4125. https://doi.org/10.1109/TAC.2018.2810039
47. Zhang Z, Zheng L, Weng J, Mao Y, Wei Lu, Xiao L (2018) A new varying-parameter recurrent neural-network for online solution of time-varying Sylvester equation. IEEE Trans Cybern 48(11):3135–3148. https://doi.org/10.1109/TCYB.2017.2760883
48. Bastanfard A, Abbasian A (2023) Speech emotion recognition in Persian based on stacked autoencoder by comparing local and global features. Multi Tool Appl 82(23):36413–36430. https://doi.org/10.1007/s11042-023-15132-3
49. M. Savargiv and A. Bastanfard (2016) Real-time speech emotion recognition by minimum number of features. Artificial Intelligence and Robotics (IRANOPEN), Qazvin, Iran, 2016, (72–76). https://doi.org/10.1109/RIOS.2016.7529493.
50. Savargiv M, Bastanfard A (2014) Study on unit-selection and statistical parametric speech synthesis techniques. J Comput Robot 7(1):19–25
51. Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS (2023) Speech emotion recognition through hybrid features and convolutional neural network. Appl Sci 13(8):4750
52. Marik A, Chattopadhyay S, Singh PK (2022) A hybrid deep feature selection framework for emotion recognition from human speeches. Multi Tool Appl 82(8):11461–11487. https://doi.org/10.1007/s11042-022-14052-y
53. Bhangale KB, Kothandaraman M (2023) Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network). Appl Acous 212:109613. https://doi.org/10.1016/j.apacoust.2023.109613
54. Patnaik S (2022) Speech emotion recognition by using complex MFCC and deep sequential model. Multi Tool Applic 82(8):11897–11922. https://doi.org/10.1007/s11042-022-13725-y

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