electronics

Article

Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network

Kishor Bhangale and Mohanaprasad Kothandaraman *

School of Electronics Engineering (SENSE), Vellore Institute of Technology, Chennai 600127, India
* Correspondence: kmohanaprasad@vit.ac.in

Abstract: Speech emotion recognition (SER) plays a vital role in human–machine interaction. A large number of SER schemes have been anticipated over the last decade. However, the performance of the SER systems is challenging due to the high complexity of the systems, poor feature distinctiveness, and noise. This paper presents the acoustic feature set based on Mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), wavelet packet transform (WPT), zero crossing rate (ZCR), spectrum centroid, spectral roll-off, spectral kurtosis, root mean square (RMS), pitch, jitter, and shimmer to improve the feature distinctiveness. Further, a lightweight compact one-dimensional deep convolutional neural network (1-D DCNN) is used to minimize the computational complexity and to represent the long-term dependencies of the speech emotion signal. The overall effectiveness of the proposed SER systems' performance is evaluated on the Berlin Database of Emotional Speech (EMODB) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets. The proposed system gives an overall accuracy of 93.31% and 94.18% for the EMODB and RAVDESS datasets, respectively. The proposed MFCC and 1-D DCNN provide greater accuracy and outpace the traditional SER techniques.

Keywords: affective computing; convolutional neural network; deep learning; MFCC; speech emotion recognition

Citation: Bhangale, K.; Kothandaraman, M. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics* **2023**, *12*, 839. https://doi.org/10.3390/electronics12040839

Academic Editor: Valeri Mladenov

Received: 4 January 2023
Revised: 27 January 2023
Accepted: 30 January 2023
Published: 7 February 2023

1. Introduction

Speech emotion recognition (SER) deals with the recognition of emotional content in the speech signal irrespective of its semantic content. Humans can naturally perform SER as a part of speech communication; the ability to perform automatic SER using computational strategies is still an enduring topic of research. SER systems are extensively utilized in various applications to understand the emotional status of humans such as call center operators, car drivers, customer care centers, pilots, narcotics analysis, online learning platforms, and many other human–machine interaction system users [1,2].

The generalized SER system encompasses two major phases: training and testing. Machine learning or deep learning techniques were used to learn the classifier based on hand-crafted characteristics of speech emotion signals during the training phase. During the testing step, the real-time samples are compared to the trained model to see if it can distinguish the specific emotion. Data preparation, feature extraction, feature selection, and classification are all important steps in the SER process. To improve raw voice signals, data preparation includes signal normalization, noise reduction, and artifact removal. Using various feature extraction strategies, the feature extraction step aids in capturing the key aspects of a certain emotion. The importance of feature selection in collecting crucial characteristics to reduce the SER system's complexity cannot be overstated. Lastly, different machine learning or deep learning classifiers are employed for SER [3,4].

Speech emotion signal is a continuous time-domain signal that contains emotion as well as information. Speech features can be local or global features depending upon the feature extraction approach. Local features are known as segmental features or short-term

features that represent temporal variations of the signal. Global attributes are also known as long-term features or supra-segmental features representing the signal's overall statistics. SER systems can analyze the local and global speech signal features in four categories: prosodic, spectral, voice-quality, and Teager energy operator (TEO)-based features [5].

Prosodic features are dependent on human hearing perception, such as rhythm and intonation. The most extensively utilized prosodic features are pitch, fundamental frequency, duration, and energy. Prosodic features are more indicative of happiness and anger and less indicative of fear and sadness. Heterogeneous sound features do not affect prosodic features [6]. Combination of prosodic features with spectral features has revealed significant improvement in the SER [7]. Spectral features of speech emotion signals are obtained by converting time-domain signal to frequency-domain signal. Spectral features represent the characteristics of the vocal tract. The Mel frequency cepstral coefficients (MFCC) scale is a well-accepted method for feature mining of speech signal. It gives the short-term power spectrum of speech signal and describes the phonemes in terms of the shape of the vocal tract. The Mel frequency scale correlates the perceived frequency with actual frequency. MFCC performs poor in cases of additive noise and background noise [8–10]. Linear predictive cepstral coefficients (LPCC) can be used to approximate the human vocal cord. LPCC gives a poor performance in emotion recognition compared with MFCC. Linear predictive coding (LPC) can be used for the encoding of low bit rate signal with higher efficiency [11,12]. Voice quality features are generally used to capture the physical characteristics of the vocal tract. Voice quality features comprise shimmer, jitter, harmonics to noise ratio (H.N.R.), etc. Jitter and shimmer represent the variability of frequency and amplitude of speech signal, respectively. Jitter is the quality of frequency unsteadiness whereas shimmer is the performance metric of amplitude variability [13]. TEO-based features are normally utilized for anger and stress emotion recognition. As per Teager, speech is shaped by a non-linear vortex-airflow association in the person's vocal system. Stressful conditions influence the muscle pressure of the speaker and lead to variation in airflow during speech production. TEO features are basically used for stress emotion detection [14,15].

The most commonly used classifiers for SER are support vector machine (SVM), random forest (RF), Gaussian mixture model (GMM), K-nearest neighbor (KNN), hidden arkov model (HMM), decision tree, dynamic time warping, etc. Traditional machine learning-based approaches have shown inferior performance because of their dependency on hand-crafted features, poor feature representation, inability to deal with complex and large data, etc. [16–18].

Deep learning-based approaches emerge as better solutions for SER because of their superior feature representation capability, ability to handle complex features, ability to learn unlabeled data, and ability to handle larger datasets. Distinct deep learning algorithms such as convolutional neural network (CNN), deep neural network (DNN), long short-term memory (LSTM), etc., are successfully presented for automatic SER [19–21].

Recently, various one- and two-dimensional convolutional neural network-based systems have been presented for the SER. Kwon presented 1-D dilated CNN to represent salient features and long-term dependencies of the speech emotion signal. It resulted in 73% and 90% accuracy on Interactive Emotional Dyadic Motion Capture (IEMOCAP) and EMODB databases, respectively [22]. Further, 1-D dilated CNN is used along with hierarchical feature learning blocks with the help of a bidirectional gated recurrent unit (BiGRU) to improve the signal quality in the spectral domain [23]. Zhao et al. [24] investigated 1-D and 2-D CNN along with LSTM for SER and observed that 2-D CNN–LSTM provides significantly better results compared with 1-D CNN–LSTM on the EMODB dataset. They suggested that the "black box" nature of the 2-D CNN poorly uncovers the details of speech emotion signal. Most of the deep learning-based SER systems use speech spectrogram or Mel frequency spectrogram as input, which escalates the computational difficulty because of a large number of features [25]. The Mel frequency logarithmic spectrogram (MFLS) has shown a better spectral representation of the emotion signal compared with the tradi-

tional MFCC algorithm. The MFLS features along with 2-D DCNN have given 96.07% and 95.68% accuracy for the speaker-independent and speaker-dependent SER on the EMODB dataset [26]. Zhao et al. proposed a combination of 1-D and 2-D CNN to capture high-level emotional features along with Bayesian optimization to fasten the learning process. The merging of 1-D and 2-D CNN has given the accuracy of 91.78% and 92.72% accuracy for speaker-independent and speaker-dependent SER on the EMODB dataset [27]. Bilal [28] presented chroma, spectral, root mean square, and MFCC features for the SER along with spectrogram features to learn the emotional features from the speech. The set of these acoustic features along with ResNet provided SER accuracy of 79.41% and 90.21% for the RAVDESS and EMODB datasets, respectively. Chen at al. [29] investigated attention-based convolution RNN (ACRNN) that accepts 3-D Mel-spectrograms as input for emotion feature representation. It provided high-level feature representation and concentrated on the emotion specific content in speech. It resulted in SER accuracy of 82.82% for the EMODB dataset. Meng et al. [30] presented an SER system that used 3-D Mel-log spectrograms as an input to dilated CNN with BLSTM with attention mechanism to improve the long-term dependency. It provided SER accuracy of 88.08% for the EMODB dataset. Misbah et al., in [31], presented SER based on Mel-Log spectrogram and DCNN. It provides 81.30%, 83.80%, 83.80%, and 82.10% accuracy on the RAVDESS, IEMOCAP, SAVEE, and EMODB datasets, respectively. The classifiers used for classifications of emotions such as SVM, KNN, and random forest shows less generalization capability. Sonawane et al. [32] explored MFCC–CNN for the real time SER for datasets obtained from social media sites. It shows that the MFCC–CNN shows better performance than the traditional MFCC-based SER techniques. Sajjad et al. [33] investigated SER based on CNN feature extractor that accepts the short time Fourier transform spectrogram (STFT) as input and the radial basis function network (RBFN) for similarity computation. It used bidirectional LSTM (BiLSTM) to improve the precision of SER. It resulted in an accuracy of 72.25%, 85.57%, and 77.02% for the IEMOCAP, EMODB, and RAVDESS datasets, respectively. It is observed that the feature representation capability of the STFT is limited due to vast changes in spectral domain. Kwon et al. [34] proposed deep stride CNN (DSCNN) that extracts discriminative and important features from speech spectrogram to improve the SER precision and complexity of the network. It has shown better representation of the local and global features of speech signal. It has shown SER accuracy of 84.00% and 80.00% accuracy for the IEMOCAP and RAVDESS datasets, respectively. Vryzas et al. [35] suggested that CNN–SVM for continuous time frames of the speech signal. It is observed that the raw speech fails to provide better local and global representation of emotional content in time and frequency domain. Ngoc-Huynh et al. [36] presented SER scheme using a multi-level multi-head fusion (MLMHF) attention mechanism, and RNN. It uses MFCC features as input and has shown better resolution in time but fails to provide the generalization capability. Orhan et al. [37] proposed a 3-D CNN + LSTM using MFCC coefficients for SER that resulted in 96.18%, 87.50%, and 93.32% accuracy for the RAVDESS, SAVEE, and RML datasets, respectively. Liu et al. [38] suggested data augmentation to minimize the gross loss in SER recognition along with the CNN–LSTM model for feature representation. It used the log Mel spectrogram as the input to the CNN–LSTM model for SER. Various SER techniques used the common feature set and very little work is presented on the multiple acoustic features-based SER that provides the representation capability in the time and frequency domains. The distinctiveness of the SER depends upon the quality of speech features. Various SER techniques have used MFCC spectrogram or MFCC coefficients as the input to deep learning frameworks. The MFCC is capable of providing high-level features and neglects lower-order characteristics of the signals. The raw speech features have poor feature variability and fail to capture the precise arousal and valence level of the speech signal for SER [39,40]. Various machine and deep learning algorithms provide limited performance for the real time SER because of variability in the recording environment and language changes. Thus, there is need to improve the distinctiveness of the speech features in a way that combines the various local and global characteristics of the emotion signal. The huge trainable parameters of the

traditional deep learning architectures lead to higher training and testing time and make it less flexible for implementation on the standalone devices/processors [41–43].

This paper presents SER using MFCC and a one-dimensional convolutional neural network to minimize the computational complexity of SER. The chief contributions of the proposed article can be highlighted as:

- To present collaborative low-order and high-order features using various acoustic features such as MFCC, LPCC, WPT, ZCR, RMS, spectrum centroid, spectral roll-off, spectral kurtosis, formants, pitch, jitter, and shimmer for the improvement of speech signal's feature distinctiveness;
- To develop a lightweight 1-D deep convolutional neural network for complexity reduction of deep learning frameworks for SER.

The overall system's effectiveness is assessed using accuracy, recall, precision, and F1-score on the EMODB and RAVDESS datasets.

The remaining paper is structured as follows: Section 2 delivers a detailed description of various acoustics features considered for the implementation; Section 3 describes the proposed SER methodology based on 1-D DCNN; Section 4 depicts the dataset information, investigation results on the individual and cross-corpus datasets, and findings from the results; finally, Section 5 gives a concise conclusion and future scope.

2. Acoustic Features

Acoustics features of the speech signal represent the physical properties of the speech signal in terms of frequency, amplitude, and loudness. The proposed acoustic feature set consists of distinct spectral features, time-domain features, and voice quality features to characterize the speech emotion. Extracted acoustics features are Mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), wavelet packet transform (WPT), zero crossing rate (ZCR), spectrum centroid, spectral roll-off, root mean square (RMS), spectral kurtosis (SK), jitter, shimmer, pitch frequency, formants, mean and standard deviation of the formants. Before computing various features, the speech signal is passed through a moving average filter to minimize the noise and disturbances in the speech signal.

2.1. MFCC

MFCC provides the spectral information of the speech and characterizes the human hearing perception. Figure 1 shows the process flow of computation of MFCC coefficients [4,7].

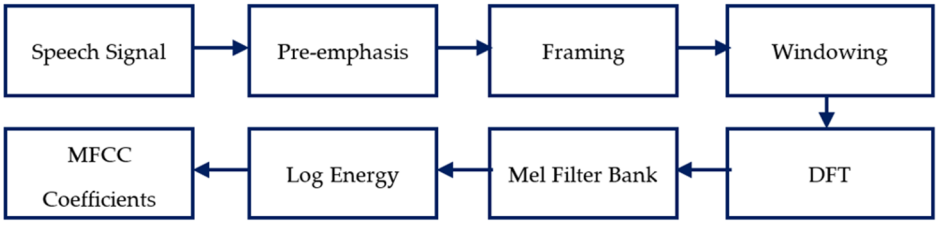

Figure 1. Process flow of MFCC.

During the MFCC coefficient extraction process, pre-emphasis normalizes the raw signal speech signal. The pre-emphasis minimizes the noise and disturbances present in the raw emotional speech ($x(n)$). Further, the filtered signal is alienated into 40 ms frames with a frame shift of 50% (i.e., 20 ms). For 4 s speech signals, a total of 199 frames is generated considering 40 ms frame width and 50% overlapping. Further, a single hamming window with $\alpha = 0.46$ and N number of samples per frame length (N) of 30 ms gathers the closest frequency components together, which is given by Equation (1).

$$H(n) = (1 - \alpha) - \alpha \times \cos\left(\frac{2\pi n}{(N-1)}\right),\ 0 \leq n \leq N-1 \quad (1)$$

In the next step, Discrete Fourier Transform (DFT) is employed to convert time-domain emotion speech signal into the frequency-domain equivalent ($X(k)$), as given in Equation (2). Equation (3) provides the power spectrum of the DFT which exemplifies the vocal tract characteristics. Then, the signal is passed through M(24) number of Mel Frequency triangular filter banks ($\nabla_m(k)$) to provide the speech-hearing perceptual information as given in Equation (4). Equations (5) and (6) provides the conversion of linear to Mel frequency and vice versa.

$$X(k) = \sum_{n=0}^{N-1} x(n) \times H(n) \times e^{-j2\pi nk/N},\ 0 \leq n, k \leq N-1 \tag{2}$$

$$X_k = \frac{1}{N}|X(k)|^2 \tag{3}$$

$$ET_m = \sum_{k=0}^{k=1} \nabla_m(k) \times X_k;\ m = 1, 2, \ldots M \tag{4}$$

$$Mel = 2595 \log\left(1 + \frac{f}{700}\right) \tag{5}$$

$$f = 70\left(10^{\frac{Mel}{2595}} - 1\right) \tag{6}$$

Afterward, discrete cosine transform (DCT) of log-filter bank energy signal provides L number of cepstral coefficients as given by Equation (7).

$$MFCC_i = \sum_{m=1}^{M} \log_{10}(ET_m) \times \cos j\left((m + 0.5)\frac{\pi}{m}\right) \quad \text{for } j = 1, 2, \ldots L \tag{7}$$

The MFCC provides a total of 39 features that encompass one feature as the energy of speech signal, 12 MFCC coefficients, and 26 first- and second-order derivatives of the MFCC features. The derivative features are essential for characterizing the transition in the emotional speech [26,27].

2.2. RMS

RMS (x_{rms}) provides the loudness of the emotion signal that is computed by considering the root mean squares of the amplitudes of the emotion speech samples (x_i) [43,44]. Equation (8) provides the estimation of RMS of the emotion speech signal with N samples.

$$x_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{8}$$

2.3. ZCR

ZCR provides the transition of signal over the zero line that indicates noisiness measure in the speech signal. Equation (9) provides computation of ZCR in the time domain [44]. The sign function provides 1 value for positive sample amplitude and 0 for negative sample amplitude over a time frame (t).

$$ZCR_t = \frac{1}{2}\left(\sum_{n=1}^{N} (\text{sign}(x[n]) - \text{sign}(x[n-1]\right) \tag{9}$$

2.4. Spectrum Centroid

The spectrum centroid represents the center of gravity of the scale invariant Fourier transform (SIFT) spectrum. It provides the spectral shape characteristic of the speech signal. The higher value of spectrum centroid indicates the accumulation of higher frequency

values [45]. Equation (10) gives the spectrum centroid (SC) for nb frequency bins, SIFT magnitude ($M_t(nb)$) over time frame t.

$$SC_t = \frac{\sum_{n=1}^{N} M_t(nb) \times nb}{\sum_{n=1}^{N} nb} \tag{10}$$

2.5. Spectral Roll-off

Spectral roll-off frequency ($F_{rolloff}$) is a measure of spectral shape that provides the frequency below which 85% of SIFT magnitude is concentrated. Equation (11) provides the computation of spectral roll-off [44].

$$\sum_{n=1}^{F_{rolloff}} M_t(nb) = 0.85 \times \sum_{n=1}^{N} M_t(nb) \tag{11}$$

2.6. LPCC

The LPCC is the spectral feature derived from the linear predictive analysis to represent the emotion-specific phonetic representation of the speech signal. The LPCC is good at providing human vocal tract characteristics that help to uniquely characterize the emotional content in the speech [46–48]. In linear predictive analysis, the n^{th} samples can be estimated from the knowledge of previous p samples as given in Equation (12).

$$x(n) = a_1x(n-1) + a_2x(n-2) + a_3x(n-3) + \ldots \ldots + a_px(n-p), \tag{12}$$

where a_1, a_2, $\ldots..a_p$ are the constants over the speech frame. These linear predictor coefficients predict the speech sample. Equation (13) is used to analyze the error between predicted $\hat{x}(n)$ and actual sample x(n).

$$x(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{13}$$

To obtain the unique predictive coefficients, the sum of the squared difference of error (e_n) between predicted $\hat{x}(n)$ and actual sample x(n) is computed using Equation (14). Here, m represents the number of samples in the frame.

$$e_n = \sum_{m} \left[x(m) - \sum_{k=1}^{p} a_k x(m-k) \right]^2 \tag{14}$$

The LP coefficients are computed by solving Equation (15). The LPCC coefficients are computed using Equations (15)–(18).

$$\frac{dE_n}{da_k} = 0 \text{ for } k = 1, 2, 3, \ldots\ldots..p \tag{15}$$

$$C_0 = \log_e(p) \tag{16}$$

$$LPCC_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}; \text{ for } 1 < m < p \tag{17}$$

$$LPCC_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k}; \text{ for } m > p \tag{18}$$

The proposed approach considers total of 13 LPCC coefficients as features [46,47].

2.7. Spectral Kurtosis

The spectral kurtosis (SK) provides the series of transients along with their locations in the spectral domain. It characterizes the non-Gaussianity or flatness of the speech spectrum

around its centroid that shows the effects of variations in arousal and valence in emotion on the speech spectrum [5,7,9]. Equation (19) is used to estimate the spectral kurtosis of the speech signal.

$$SK = \frac{\sum_{k=b1}^{b2}(f_k - \mu_1)^4 s_k}{(\mu_2)^4 \sum_{k=b1}^{b2} s_k} \tag{19}$$

Here, μ_1 and μ_2 represents the spectral centroid and spectral spread, respectively, s_k is spectral value over k bins, and b_1 and b_2 are the lower and upper bound of the bins where spectral skewness of speech is estimated.

2.8. Jitter and Shimmer

Jitter and shimmer provide the changes over frequency and amplitude of the emotion signal, respectively, caused due to irregular vocal fold vibrations. Jitter and shimmer depict the breathiness, roughness, and hoarseness in the emotional sound. Equation (20) provides the average absolute value of jitter [13].

$$\text{Jitter} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \tag{20}$$

where T_i stands for the time period in sec and N represents number of periods. Equation (21) represents average value of shimmer.

$$\text{Shimmer} = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N} \sum_{i=1}^{N} A_i}, \tag{21}$$

where A_i is peak to peak amplitude of emotional speech and N depicts number of periods.

2.9. Pitch Frequency

Pitch (f_0) is significant to exemplify the voiced part of speech. The pitch of the speech is estimated by computing the difference between the peaks derived from autocorrelation of the speech signal [5,7].

2.10. Formants

Formants indicate the peak frequencies in the speech spectrum that has higher energy. It characterizes the resonance phenomenon of the vocal tract, which is very helpful in characterizing the effect of emotion on the resonance phenomenon. The formants are derived from the MFCC spectrogram and 3 formants—f_1, f_2, and f_3—are considered for evaluation. Further, the mean and standard deviation of the formants are computed using three formants to provide the variations in the formants [5,7]. Equations (22)–(24) provide formants (fm), mean of formants (fm_u), and standard deviation of formants (fm_σ), respectively.

$$fm = \{f_1, f_2, f_3\} \tag{22}$$

$$fm_u = \frac{f_1 + f_2 + f_3}{3} \tag{23}$$

$$fm_\sigma = \sqrt{\frac{\sum_{i=1}^{3} (f_i - fm_u)^2}{3}} \tag{24}$$

2.11. Wavelet Packet Decomposition Features

WPT permits complex information such as speech, images, music, emotion, and patterns to be decomposed into basic forms at diverse positions and scales and consequently reconstructed with high precision. WPT helps us to analyze the variations over the speech due to different emotions. The Daubechies (db2) wavelet at various scales is used to

decompose the wavelet packet basis function $\Psi_j^i(n)$ using WPD for L-levels, as shown in Equations (25) and (26) [49,50].

$$\Psi_j^{2i}(n) = \sum_k h(k)\Psi_{j-1}^i\left(n - 2^{j-1}k\right) \tag{25}$$

$$\Psi_j^{2i+1}(n) = \sum_k g(k)\Psi_{j-1}^i\left(n - 2^{j-1}k\right) \tag{26}$$

where $g(k)$ and $h(k)$ denotes, respectively, the high and low pass quadrature mirror filters shown in Equations (27) and (28).

$$h(k) = \langle \Psi_j^{2i}(u), \Psi_{j-1}^i\left(u - 2^{j-1}k\right)\rangle \tag{27}$$

$$g(k) = \langle \Psi_j^{2i+1}(u), \Psi_{j-1}^i\left(u - 2^{j-1}k\right)\rangle \tag{28}$$

The emotion speech is separated into segments at level j using Equation (29).

$$x(n) = \sum_{i.k} X_j^i(k)\Psi_j^i\left(n - 2^j k\right) \tag{29}$$

where $X_j^i(k)$ is kth WPT at ith packet at j level. Equation (30) indicates the energy of the local wavelet.

$$X_j^i(k) = \langle x(n), \Psi_j^i\left(n - 2^j k\right)\rangle \tag{30}$$

The wavelet coefficient $X_j^i(k)$ describes the localized WPT weights denoted by $\Psi_j^i(n - 2^j k)$ as given in Equation (31).

$$X_j^i(k) = \langle x(n), \Psi_j^i\left(n - 2^j k\right)\rangle \tag{31}$$

Equation (32) gives distinct WPT set for L level.

$$X_L(k) = \begin{bmatrix} X_L^0(k) \\ X_L^1(k) \\ . \\ . \\ X_L^{2^{L-1}}(k) \end{bmatrix} \tag{32}$$

The speech signal is decomposed up to three levels using db2 filter. Seven statistical features are extracted for the last decomposed level: mean, median, standard deviation, variance, skewness, kurtosis, and energy of every wavelet packet. The different WPT features provide the spectral changes in the speech signal due to changes in prosody and intonation of emotion speech signal. The three-level decomposition of speech signal results in total 56 WPT features.

Thus, the final feature vector consists of total 715 features. which is the concatenation of 39 MFCC features, 1 RMS feature, 199 ZCR features, 199 spectrum centroid features, 13 LPCC features, 56 WPT features, 1 spectral roll-off feature, 199 spectral kurtosis features, 1 jitter, 1 shimmer, pitch frequency, three formants, mean and standard deviation of formants. The feature representation (Feat), which is given as input to DCNN for SER, is given by Equation (33).

$$\text{Feat} = \{\text{MFCC}_{1-39},\ x_{rms},\ \text{ZCR}_{1-199},\ \text{SC}_{1-199},\ \text{LPCC}_{1-13},\ \text{WPT}_{1-56},\ F_{rolloff},\ \text{SK}_{1-199},\ \text{Jitter},\ \text{Shimmer},\ f_0,\ \text{fm}_{1-3},\ \text{fm}_u,\ \text{fm}_\sigma\ \} \tag{33}$$

3. Proposed Methodology

Figure 2 illustrates the process of the proposed SER system. The proposed 1-D Deep convolutional neural network (1-D DCNN) comprises three convolution layers (Conv),

three rectified linear unit (ReLU) layers, two fully connected layers (FC), and a softmax classifier layer.

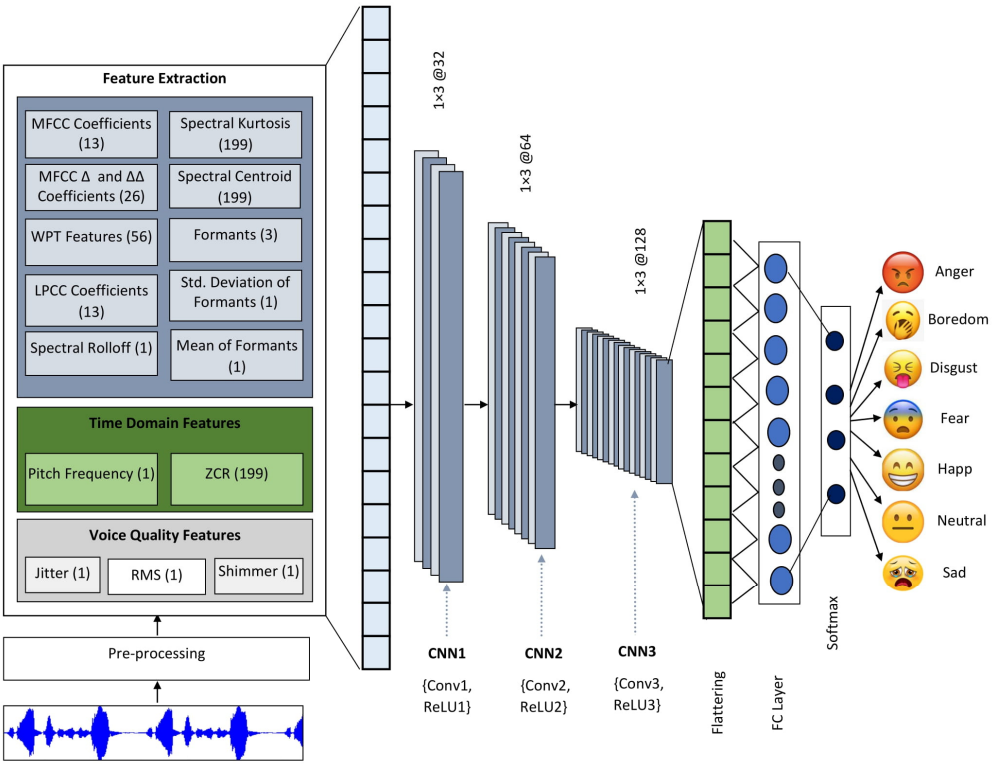

Figure 2. Illustration of proposed SER system.

The proposed compact 1-D DCNN accepts the 715 acoustic features (Feat) as input. The first layer of CNN consists of two layers {Conv1(Filters: 32, Filter Size: 1×3, Stride: 1, Padding: Yes) → ReLU1 } that provide the output of 715 × 1 × 32. The convolution filter size is selected as 1 × 3, which helps to provide the local changes in the feature set and combines the correlation between different features. The filter is stride with one pixel and the original feature vector is zero padded to maintain the original dimension after convolution operation. The second CNN layer includes {Conv2 (Filters: 64, Filter Size: 1×3, Stride: 1, Padding: Yes) → ReLU2} that produces the feature map of 715 × 1 × 64; the third CNN layer encompasses {Conv3 (Filters : 128, Filter Size : 1 × 3, Stride : 1, Padding : Yes) → ReLU3} that results in feature maps of 715 × 1 × 128. The increasing number of filters in each layer assists in improving the connectivity in the local and global features of the emotion. In each convolution layer, the one-dimensional input is convolved with the convolution filter. It provides the high-level characteristics of the speech emotion signal [51,52].

The convolution output z(n) of features Feat(n) and filter w(n) having size l is given in Equation (34). Equation (35) represents convolution feature map where z_i^l describes the i^{th} feature map of the lth layer, z_j^{l-1} stands for jth feature of the $(l-1)^{th}$ layer, w_{ij}^l describes the filter kernel of l^{th} layer linked to jth feature, b_i^l stands for bias, and σ depicts ReLU activation function. The ReLU layer is simple and has a faster activation function that overcomes the problem of vanishing gradient as given in Equation (36).

$$z(n) = \text{Feat}(n) \times w(n) = \sum_{m=0}^{i-1} \text{Feat}(m) \times w(n-m) \tag{34}$$

$$z_i^l = \sigma\left(b_i^l + \sum_j z_j^{l-1} \times w_{ij}^l\right) \tag{35}$$

$$\sigma(z) = \max(0, z) \tag{36}$$

Following the 3 CNN layers, 2 fully connected layers are used having 20 and 7 hidden layers, respectively. In the FC layer, the linear transformation is applied to the input feature vector using the weight matrix. The non-linear activation function is applied for non-linear transformation as given in Equation (37).

$$y_{jk}(x) = f(\sum_{i=1}^{n_H} W_{jk} x_i + wj_0) \tag{37}$$

where x_i represent value from flattening feature vector, w_0 stands for bias term, w represent weight matrix, f stands for non-linear activation function, y is non-linear transformation output, and n_H provides no hidden layers.

Finally, the softmax classifier provides the probability of the output where maximum probability of class label provides output class label, as given in Equations (38)–(40) [15].

$$z_i = \sum_j h_j w_{ji} \tag{38}$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \tag{39}$$

$$\hat{y} = \arg\max_i p_i \tag{40}$$

Here, h_j is weight of penultimate layer and w_{ji} represent the weights of softmax and penultimate layer, z_i is input of softmax layer, p_i is probability of class label, and $\hat{y}$ is predicted class label.

Table 1 provides the particulars of the feature maps of the distinct layers of the proposed 1-D DCNN. The raw speech signal consists of 64,000 samples in the input vector; whereas the multiple acoustic feature vector consists of 715 values. The length of the input feature vector affects the total number of trainable parameters and the computation time. The proposed network is trained using stochastic gradient descent with momentum (SGDM). The system is trained for a batch size of 64 to cope with the memory limit. The training process considers 200 epochs, cross entropy loss function, an initial learning rate of 0.001, and momentum of 0.9. The training accuracy and training loss of the proposed 1-D DCNN are illustrated in Figures 3 and 4, respectively.

Table 1. Details of different layers of proposed 1-D DCNN.

Network Layers	Raw Speech + 1-D DCNN		MFCC + 1-D DCNN	
	Size	**Stride**	**Size**	**Stride**
Input Layer	64,000 × 1	-	715 × 1	-
Conv1	64,000 × 1 × 32	1	715 × 1 × 32	1
ReLU-1	64,000 × 1 × 32	1	715 × 1 × 32	1
Conv2	64,000 × 1 × 64	1	715 × 1 × 64	1
ReLU-2	64,000 × 1 × 64	1	715 × 1 × 64	1
Conv3	64,000 × 1 × 128	1	715 × 1 × 128	1
ReLU-3	64,000 × 1 × 128	1	715 × 1 × 128	1
FC	20 × 1	-	20 × 1	-
FC	7 × 1	-	7 × 1	-
Output	7 × 1	-	7 × 1	-

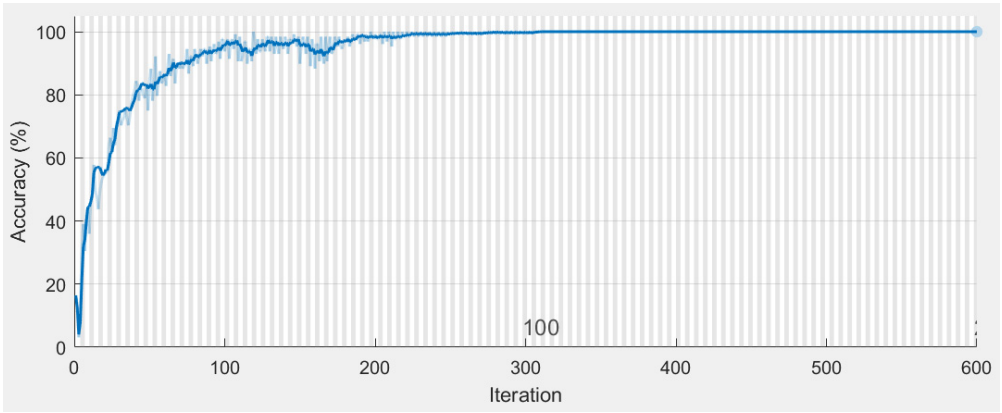

Figure 3. Training accuracy of proposed 1-D DCNN.

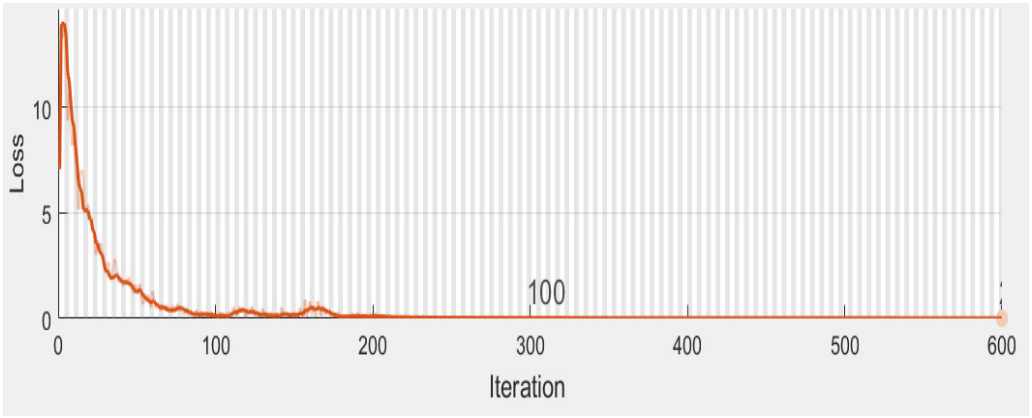

Figure 4. Training loss of proposed 1-D DCNN.

4. System Implementation and Results

The proposed SER scheme is instigated using MATLAB R2021b (Mathworks, Bengaluru, India) on the NVIDIA Volta GPU (NVIDIA, Bengaluru, India) with a tensor core (512 cores). The MATLAB deep learning toolbox is used for the construction of the deep learning algorithm.

4.1. Dataset

The experimentations are carried on an open-source EMODB speech emotion database that comprises 535 utterances of 7 emotions recorded from 10 professional actors in the German language [53]. Additionally, the RAVDESS emotional speech dataset is used for the performance evaluation of the proposed SER scheme. The RAVDESS consists of total 1440 samples recorded from 24 professional actors (12 male and 12 female). It encompasses eight emotions: calm, surprise, neutral, happy, angry, sad, fearful, and disgust [54]. The original EMODB database samples have variable lengths and are down-sampled at 16 kHz. Therefore, all samples are cropped/appended to make each sample of 4 s duration. The data is split in the ratio of 70:30 for training and testing, respectively, as shown in Tables 2 and 3.

Table 2. Details of EMODB database.

Samples	Speech Emotions (EMODB)							
	Anger	**Boredom**	**Disgust**	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Total**
Total Samples	126	79	46	70	73	78	63	535
Training Samples (70%)	87	55	31	50	51	54	44	372
Testing samples (30%)	39	24	15	20	22	24	19	163

Table 3. Details of RAVDESS database.

Samples	Speech Emotions (RAVDESS)								
	Anger	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sadness**	**Surprised**	**Total**
Total Samples	192	192	192	192	192	96	192	192	1440
Training Samples (70%)	134	134	134	134	134	67	134	134	1005
Testing samples (30%)	58	58	58	58	58	29	58	58	435

4.2. Results and Discussions

The effectiveness of the proposed 1-D DCNN with multiple acoustic features as input is compared with 1-D DCNN with raw speech as input. The outcomes of the proposed method are compared based on various performance metrics such as precision, recall, accuracy, and F1-score as given in Equations (41)–(44). The precision and recall provide the qualitative and quantitative performance of the proposed SER system. The F1-score provides the harmonic mean of the precision and recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{41}$$

$$\text{Recall} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{42}$$

$$\text{Accuracy}(\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{43}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{44}$$

The outcomes of the proposed 1-D DCNN-based SER for the various emotions with raw speech as input and MFCC coefficient and multiple acoustic features as input is given in Table 4.

Table 4. Results of proposed system on EMODB database.

Emotion	Raw Speech + 1-D DCNN				MFCC + 1-D DCNN				Multiple Acoustic Features + 1-D DCNN			
	Accuracy	**Recall**	**Precision**	**F1-Score**	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**Accuracy**	**Recall**	**Precision**	**F1-Score**
Anger	97.44	0.97	0.97	0.97	100.00	1.00	0.98	0.99	100.00	1.00	0.98	0.99
Boredom	87.50	0.88	0.78	0.82	87.50	0.88	0.84	0.86	87.50	0.88	0.84	0.86
Disgust	86.67	0.87	0.93	0.90	80.00	0.80	1.00	0.89	93.33	0.93	1.00	0.97
Fear	85.00	0.85	0.89	0.87	90.00	0.90	0.90	0.90	95.00	0.95	0.95	0.95
Happiness	86.36	0.86	0.95	0.90	90.91	0.91	1.00	0.95	90.91	0.91	1.00	0.95
Neutral	91.67	0.92	0.88	0.90	95.83	0.96	0.88	0.92	91.67	0.92	0.88	0.90
Sadness	89.47	0.89	0.89	0.89	94.74	0.95	0.90	0.92	94.74	0.95	0.95	0.95
Overall	89.16	0.89	0.90	0.89	91.28	0.91	0.93	0.92	93.31	0.93	0.94	0.94

The experimental results show that compared to the raw speech signal and MFCC coefficients, multiple acoustic features along with 1-D DCNN give a better representation of the emotion signal and provide a higher discriminant feature. The proposed compact 1-D DCNN multiple acoustic features provide 93.31% accuracy for SER on the EMODB dataset. It shows 4.61% and 2.21% improvement over the 1-D DCNN with raw features and MFCC coefficients (39 coefficients), respectively. It gives highest 100% accuracy for anger emotion and lowest 90.91% accuracy for the happiness emotion.

Table 5 illustrates the SER performance for the RAVDESS dataset. The effectiveness of the proposed 1-D DCNN-based SER system provides average accuracy of 90.52%, 92.03%, and 94.18% for 1-D DCNN with raw speech, MFCC, and multiple acoustic features, respectively, for the RAVDESS dataset. The 1-D DCNN along with multiple acoustic features gives the highest (98.28%) accuracy for the anger emotion, and lowest accuracy (91.38%) for the neutral emotion. It shows an overall improvement of 4.04% and 2.34%

over the SER-based 1-D DCNN using raw speech and MFCC coefficients, respectively, for the RAVDESS dataset. The proposed 1-D DCNN along with multiple acoustic features provides superior accuracy for the RAVDESS dataset (94.10%) over the EMODB dataset (93.31%). The higher number of samples and variability in the training samples of the RAVDESS dataset provide finer SER performance compared with the EMODB dataset.

Table 5. Results of proposed system on RAVDESS database.

Emotion	Raw Speech + 1-D DCNN				MFCC + 1-D DCNN				Multiple Acoustic Features + 1-D DCNN			
	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score
Anger	94.83	0.95	0.93	0.94	98.28	0.98	0.89	0.93	98.28	0.98	0.93	0.96
Calm	91.38	0.91	0.90	0.91	91.38	0.91	0.93	0.92	93.10	0.93	0.93	0.93
Disgust	89.66	0.90	0.93	0.91	93.10	0.93	0.96	0.95	91.38	0.91	0.96	0.94
Fear	89.66	0.90	0.90	0.90	93.10	0.93	0.93	0.93	93.10	0.93	0.93	0.93
Happy	91.38	0.91	0.95	0.93	91.38	0.91	0.91	0.91	94.83	0.95	1.00	0.97
Neutral	82.76	0.83	0.77	0.80	82.76	0.83	0.86	0.84	93.10	0.93	0.82	0.87
Sadness	91.38	0.91	0.93	0.92	94.83	0.95	0.95	0.95	91.38	0.91	0.96	0.94
Surprised	93.10	0.93	0.92	0.92	91.38	0.91	0.95	0.93	96.55	0.97	0.95	0.96
Overall	90.52	0.91	0.90	0.90	92.03	0.92	0.92	0.92	94.18	0.94	0.94	0.94

The performance of the offered scheme is compared for different learning strategies such as SGDM, Adaptive Moment Estimation (ADAM) and Root Mean Square Propagation (RMSPROP) optimization algorithms, as shown in Figure 5. The proposed 1-D DCNN provides noteworthy improvement in the SER accuracy for SGDM with mini batch training (batch size of 64 and initial learning rate of 0.001) over ADAM and RMSPROP algorithm. It provides SER accuracy of 94.18%, 92.00%, and 91.23% for SGDM, RMSPROP, and ADAM optimization algorithm, respectively, for the RAVDESS dataset, whereas it results in 93.31%, 91.23%, and 92.5% accuracy for SGDM, RMSPROP, and ADAM optimization algorithm, respectively, for the EMODB dataset.

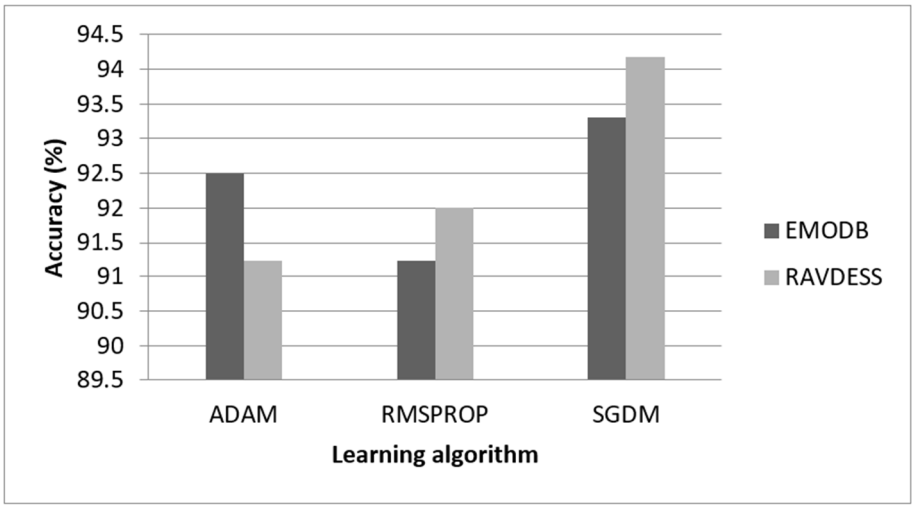

Figure 5. Performance of proposed 1-D DCNN with multiple acoustic features for different learning algorithms.

Additionally, the effectiveness of the proposed approach is evaluated on the real time Marathi speech emotion dataset, as given in Table 6. The Marathi dataset consist of 250 samples per emotion for anger, happiness, neutral, and sadness, which are recorded at 16 Hz sampling frequency for a 4 s duration. The proposed 1-D DCNN along with multiple acoustic features provides an overall accuracy of 89.94% for the four class SER. It provides superior performance compared to DCNN with raw speech (83.33%) and DCNN with MFCC coefficients (86.11%) for the real-time dataset without any pre-processing. The DCNN with multiple acoustic features provides the highest (95.56%) accuracy for the anger

emotion and the lowest (88.64%) accuracy for the happiness emotion. The results of the real-time dataset can be improved in the future by considering an effective speech-enhanced approach and more emotions.

Table 6. Results of proposed system on in-house Marathi speech emotion database.

Emotion	Raw Speech + 1-D DCNN				MFCC + 1-D DCNN				Multi-Feature + 1-D DCNN			
	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score
Anger	86.67	0.87	0.83	0.85	91.11	0.91	0.84	0.87	95.56	0.96	0.86	0.91
Happy	80.00	0.80	0.95	0.87	84.44	0.84	0.95	0.89	88.64	0.89	0.95	0.92
Neutral	84.44	0.84	0.75	0.79	86.67	0.87	0.81	0.84	88.89	0.89	0.93	0.91
Sadness	82.22	0.82	0.84	0.83	82.22	0.82	0.86	0.84	86.67	0.87	0.87	0.87
Overall	83.33	0.83	0.84	0.83	86.11	0.86	0.86	0.86	89.94	0.90	0.90	0.90

The proposed 1-D DCNN with multiple acoustic features results in 1.77 M trainable parameters; those are lower compared with DCNN with raw speech as input and the traditional state of the art. The lower trainable parameters help to minimize the training time of the network and increase the implementation flexibility of the proposed algorithm on the standalone devices. The computational complexity of the SER system is hugely dependent upon overall trainable parameters and training time. The proposed architecture needs 2980 s time for training the network. It is observed that the use of multiple distinctive acoustic feature sets and lightweight CNN helps to minimize the training time of the system compared with the existing state of the art. The proposed algorithm performance is compared with previously used 1-D CNN architectures for SER on the EMODB database, as given in Table 7. The multiple acoustic features help to improve the long-term dependencies of the speech emotion signal. It provides the better phonetic representation of the emotional signal by considering low-level and high-level features. The multiple acoustic features provide spectral properties, a loudness measure, a spectral shape measure, and an asymmetry measure, and roll-off frequency provides distinctive features for the SER and improves the SER performance. The voice quality features such as RMS, Jitter, and shimmer provide the intonation changes due to emotions on the speech signal.

Table 7. Comparison of the proposed SER with earlier approaches on the EMODB and RAVDESS datasets.

Methods	Features	Accuracy (%)		Total Trainable Parameters (Million)	Total Training Time (s)
		EMODB	RAVDESS		
1-D Dilated CNN [23]	Raw Speech	90.00	-	-	3150
1-D CNN + LSTM [24]	Raw Speech	86.73	-	-	-
DCNN [31]	Mel Log Spectrogram	85.57	77.02	-	-
RBFN-BiLSTM [33]	STFT	85.57	77.02	>3 M	-
ACRNN [29]	3-D Mel Spectrogram	82.82	-	-	6811
ADRNN [30]	3-D Mel Spectrogram	88.98	-	-	7187
Merged DCNN [27]	Log Mel Spectrogram	91.78	-	>10 M	-
ResNet101 [28]	MFCC, RMS, Croma Features, Spectral Features, Spectrogram	90.21	79.41	44.5 M	-
Proposed Method	Raw Speech	89.16	90.52	163 M	8650
Proposed Method	MFCC speech input-39 features	91.28	92.03	0.192 M	2650
Proposed Method	Multiple Acoustic Features (Spectral, Time domain, Voice quality)	93.31	94.18	1.77 M	2980

5. Conclusions and Future Scopes

This paper presents SER based on multiple acoustic features and 1-D deep convolutional neural network. The multiple acoustic features set includes the distinct low-order and high-order features of spectral, temporal, and voice quality feature domains to characterize the effect of emotion on the various spectral, time-domain, and voice quality features.

It assists in improving the quality of distinctiveness of the traditional low-order speech features. In the case of the EMODB database, our model for raw speech gives 89.16%, for MFCC features, 91.28% accuracy, and the proposed combination of multiple acoustic features gives an improved accuracy of 93.31%. While evaluating the RAVDESS dataset, the proposed system provides 80.52%, for MFCC features, 92.03% accuracy, and the proposed combination of multiple acoustic features gives an improved accuracy of 94.18%.

The proposed 1-D DCNN improves the feature discriminancy of the MFCC coefficients. The 1-D DCNN provides a simple, compact, and cost-effective solution for the hardware implementation of the SER system. The proposed SER scheme shows an improvement of 1.91–7.85% in SER accuracy over the traditional SER techniques for the EMODB dataset. The proposed algorithms need lower trainable parameters (1.77 M) and training time (2980 s), and have shown superior improvement over recent traditional SER techniques. In future, the class imbalance caused due to uneven training in the dataset can be minimized using effective data augmentation techniques. Further, outcomes of the system can be evaluated for the cross-corpus SER under stationary and non-stationary noisy conditions.

Author Contributions: Conceptualization, K.B.; Methodology, K.B.; Software, K.B.; Validation, M.K.; Formal analysis, M.K.; Investigation, M.K.; Writing—original draft, K.B.; Writing—review & editing, M.K.; Visualization, M.K.; Supervision, M.K.; Project administration, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Marathi dataset can be made available on reasonable request to authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* **2021**, *10*, 1163. [CrossRef]
2. Berkehan, A.M.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76.
3. KishorBarasu, B.; Kothandaraman, M. Survey of Deep Learning Paradigms for Speech Processing. *Wirel. Pers. Commun.* **2022**, *125*, 1913–1949.
4. Shah, F.M.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]
5. Michalis, P.; Spyrou, E.; Giannakopoulos, T.; Siantikos, G.; Sgouropoulos, D.; Mylonas, P.; Makedon, F. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation* **2017**, *5*, 26.
6. Turgut, Ö. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326.
7. Abdel-Hamid, L.; Shaker, N.H.; Emara, I. Analysis of Linguistic and Prosodic Features of Bilingual Arabic–English Speakers for Speech Emotion Recognition. *IEEE Access* **2020**, *8*, 72957–72970. [CrossRef]
8. Ben, A.S.; Mary, L.; Babu, B.P. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. *Circuits Syst. Signal Process.* **2020**, *39*, 5681–5709.
9. Atreyee, K.; Roy, U.K. Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; pp. 1017–1021.
10. Likitha, M.S.; Gupta, S.R.R.; Hasitha, K.; Raju, A.U. Speech based human emotion recognition using MFCC. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; pp. 2257–2260.
11. Renjith, S.; Manju, K.G. Speech based emotion recognition in Tamil and Telugu using LPCC and hurst parameters—A comparitive study using KNN and ANN classifiers. In Proceedings of the 2017 International Conference on Circuit, Power and Computing Technologies (I.C.C.P.C.T.), Kollam, India, 20–21 April 2017; pp. 1–6.
12. Monica, F.S.; Zbancioc, M.D. Emotion recognition in Romanian language using LPC features. In Proceedings of the 2013 E-Health and Bioengineering Conference (E.H.B.), Iasi, Romania, 21–23 November 2013; pp. 1–4.
13. Roddy, C.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80.
14. Li, X.; Li, X. Speech Emotion Recognition Using Novel HHT-TEO Based Features. *J. Comput.* **2011**, *6*, 989–998. [CrossRef]

15. Drisya, P.S.; Rajan, R. Significance of TEO slope feature in speech emotion recognition. In Proceedings of the 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), Thiruvananthapuram, India, 20–22 July 2017; pp. 438–441.
16. Barasu, B.K.; Mohanaprasad, K. A review on speech processing using machine learning paradigm. *Int. J. Speech Technol.* **2021**, *24*, 367–388.
17. Majid, W.T.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access* **2021**, *9*, 47795–47814.
18. Sonawane, A.; Inamdar, M.U.; Bhangale, K.B. Sound based human emotion recognition using MFCC & multiple SVM. In Proceedings of the 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Indore, India, 17–19 August 2017; pp. 1–4.
19. Amin, K.R.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345.
20. Rashid, J.; WahTeh, Y.; Hanif, F.; Mujtaba, G. Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimed. Tools Appl.* **2021**, *80*, 23745–23812.
21. Anuja, T.; Dhull, S. Speech Emotion Recognition: A Review. *Adv. Commun. Comput. Technol.* **2021**, *4*, 815–827.
22. Soonil, K. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Syst. Appl.* **2021**, *167*, 114177.
23. Mustaqeem; Kwon, S. 1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features. *Cmc-Comput. Mater. Contin.* **2021**, *67*, 4039–4059. [CrossRef]
24. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
25. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
26. Kishor, B.; Mohanaprasad, K. Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. In *Futuristic Communication and Network Technologies*; Springer: Singapore, 2022; pp. 241–250.
27. Zhao, J.; Mao, X.; Chen, L. Learning deep features to recognise speech emotion using merged deep CNN. *IET Signal Process.* **2018**, *12*, 713–721. [CrossRef]
28. Bilal, E.M. A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access* **2020**, *8*, 221640–221653.
29. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
30. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech Emotion Recognition From 3D Log-Mel Spectrograms with Deep Learning Network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
31. Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, Y.B. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors* **2020**, *20*, 6008. [CrossRef] [PubMed]
32. Sonawane, S.; Kulkarni, N. Speech emotion recognition based on MFCC and convolutional neural network. *Int. J. Adv. Sci. Res. Eng. Trends* **2020**, *5*, 18–22.
33. Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
34. Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2019**, *20*, 183.
35. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous speech emotion recognition with convolutional neural networks. *J. Audio Eng. Soc.* **2020**, *68*, 14–24. [CrossRef]
36. Ho, N.-H.; Yang, H.-J.; Kim, S.-H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attentionbased recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686. [CrossRef]
37. Atila, O.; Şengür, A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoust.* **2021**, *182*, 108260. [CrossRef]
38. Liu, J.; Wang, H. A speech emotion recognition framework for better discrimination of confusions. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 4483–4487.
39. Gintautas, T.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A study of cross-linguistic speech emotion recognition based on 2D feature spaces. *Electronics* **2020**, *9*, 1725.
40. Huang, S.; Dang, H.; Jiang, R.; Hao, Y.; Xue, C.; Gu, W. Multi-Layer Hybrid Fuzzy Classification Based on SVM and Improved PSO for Speech Emotion Recognition. *Electronics* **2021**, *10*, 2891. [CrossRef]
41. Fazliddin, M.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047.
42. Bhangale, K.B.; Titare, P.; Pawar, R.; Bhavsar, S. Synthetic speech spoofing detection using MFCC and radial basis function SVM. *IOSR J. Eng. (IOSRJEN)* **2018**, *8*, 55–62.
43. Chaturvedi, I.; Noel, T.; Satapathy, R. Speech Emotion Recognition Using Audio Matching. *Electronics* **2022**, *11*, 3943. [CrossRef]
44. George, T.; Cook, P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 293–302.

45. Emery, S.; Wolfe, J.; Tarnopolsky, A. Spectral centroid and timbre in complex, multiple instrumental textures. In Proceedings of the International Conference on Music Perception and Cognition; North Western University: Evanston, IL, USA, 2004; pp. 112–116.
46. Harshita, G.; Gupta, D. LPC and LPCC method of feature extraction in Speech Recognition System. In Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 498–502.
47. Olla, E.; Elbasheer, E.; Nawari, M. A comparative study of MFCC and LPCC features for speech activity detection using deep belief network. In Proceedings of the 2018 International Conference on Computer, Control, Electrical, And Electronics Engineering (ICCCEEE), Khartoum, Sudan, 12–14 August 2018; pp. 1–5.
48. John, M. Linear prediction: A tutorial review. *Proc. IEEE* **1975**, *63*, 561–580.
49. Rupali, K.; Bhalke, D.G. Speech Emotion Recognition Based on Wavelet Packet Coefficients. In *ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering*; Springer Nature Singapore: Singapore, 2022; pp. 823–828.
50. Shibani, H.; Shahin, I.; Iraqi, Y.; Werghi, N. Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier. *IEEE Access* **2020**, *8*, 96994–97006.
51. Sumita, N.; Kulkarni, V. Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. *Int. J. Speech Technol.* **2021**, *24*, 809–822.
52. Chowdhury, S.M.M.A.R.; Nirjhor, S.M.; Uddin, J. Bangla speech recognition using 1D-CNN and LSTM with different dimension reduction techniques. In *International Conference for Emerging Technologies in Computing*; Springer: Cham, Switzerland, 2020; pp. 158–169.
53. Felix, B.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. *Interspeech* **2005**, *5*, 1517–1520.
54. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [CrossRef]