Computers, Materials & Continua
DOI:10.32604/cmc.2021.015070
Article

Tech Science Press

1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features

Mustaqeem and Soonil Kwon*

Interaction Technology Laboratory, Department of Software, Sejong University, Seoul, 05006, Korea
*Corresponding Author: Soonil Kwon. Email: skwon@sejong.edu
Received: 05 November 2020; Accepted: 12 January 2021

Abstract: Emotion recognition from speech data is an active and emerging area of research that plays an important role in numerous applications, such as robotics, virtual reality, behavior assessments, and emergency call centers. Recently, researchers have developed many techniques in this field in order to ensure an improvement in the accuracy by utilizing several deep learning approaches, but the recognition rate is still not convincing. Our main aim is to develop a new technique that increases the recognition rate with reasonable cost computations. In this paper, we suggested a new technique, which is a one-dimensional dilated convolutional neural network (1D-DCNN) for speech emotion recognition (SER) that utilizes the hierarchical features learning blocks (HFLBs) with a bi-directional gated recurrent unit (BiGRU). We designed a one-dimensional CNN network to enhance the speech signals, which uses a spectral analysis, and to extract the hidden patterns from the speech signals that are fed into a stacked one-dimensional dilated network that are called HFLBs. Each HFLB contains one dilated convolution layer (DCL), one batch normalization (BN), and one leaky_relu (Relu) layer in order to extract the emotional features using a hieratical correlation strategy. Furthermore, the learned emotional features are feed into a BiGRU in order to adjust the global weights and to recognize the temporal cues. The final state of the deep BiGRU is passed from a softmax classifier in order to produce the probabilities of the emotions. The proposed model was evaluated over three benchmarked datasets that included the IEMOCAP, EMO-DB, and RAVDESS, which achieved 72.75%, 91.14%, and 78.01% accuracy, respectively.

Keywords: Affective computing; one-dimensional dilated convolutional neural network; emotion recognition; gated recurrent unit; raw audio clips

1 Introduction

Speech signals are the most dominant source of human communication, and an efficient method of human-computer interaction (HCI) using 5G technology. Emotions express human behavior, which is recognized from various body expressions, such as speech patterns, facial expressions, gestures, and brain signals [1,2]. In the field of speech signal processing, speech emotion recognition (SER) is the most attractive area of research in this era. Speech signals

play an important role to recognize the emotional state and the human behavior during his\her speech. Many researchers have introduced various techniques for efficient SER systems in order to recognize individual speech patterns and to identify the state of the speaker in terms of emotions. Hence, a sufficient feature selection and extraction is an extremely challenging task in this area [3]. Artificial intelligence (AI) plays a crucial role with the development of the skills, and technologies, in the field of HCI, which includes robotics, and on-board systems, in order to detect human activity and recognize emotions. Similarly, call centers recognize the customer's expressions, health care centers recognize the emotional state of the patients, and virtual reality applications recognize the actions and the activities using sensors. In the field of SER, the emotions of the speaker depend on the paralinguistic features not on the lexical content or a speaker. Speech signals have 2 major types of signs that include paralinguistic signs, which have hidden messages about the emotions that are contained in the speech signals, and the linguistic signs, which are constantly referred to as a meaning of the speech signals or the context [4]. Recently, researchers introduced some techniques to improved the SER rate using high-level features by utilizing the deep learning approaches in order to extract the hidden cues [5]. In the last decade, the researchers have used many acoustic features, such as qualitative, spectral, and continuous features to investigate the best features of the speech signals [6].

In the current technological era, the researchers have utilized neural networks and deep learning tools in order to search for an efficient way to extract the deep features that ensure the emotional state of a speaker in the speech data [7,8]. Some researchers introduced hybrid techniques to evaluate the handcrafted features with CNN models in order to improve the recognition accuracy of the speech signals. The handcrafted features particularly ensure the accuracy, but this process is difficult due to the features engineering, the amount of time that is required, and this method is exclusive to manual selection, which is particularly depend on expert knowledge [9]. The deep learning approaches, which include the 2D-CNN models, special for the visual data, such as images and videos in computer vision [10], but the researchers adopted these models in speech processing and achieved better results than the classical models [7]. In addition, the researchers achieved good performances with the emotion recognition from the speech signals that utilize the deep learning approaches, such as the deep belief networks (DBNs), 2D-CNNs, 1D-CNNs, and the long short-term memory (LSTM) network [5,11–13]. The performance of the deep learning approaches is better than the traditional methods. Hence, Fiore et al. [14] developed an SER for on-board system to detect and analyze the emotional conditions of the driver, which involved taking the appropriate actions in order to ensure the passenger's safety. Badshah et al. [15] introduced an SER system for the smart health care centers in order to analyze the customer emotions using a fine-tuned Alex Net model with rectangular kernels. Kwon et al. [5] proposed a novel deep stride CNN network for speech emotion recognition that used the IEMOCAP [16] and the RAVDESS [17] datasets in order to improve the prediction accuracy and decrease the overall model complexity [8]. Kang et al. [18] developed an SER technique in order to analyze the emotion type and the intensity from the arousal features and the violence features using the content analysis in movies. Dias et al. [19] developed an SER method in order to recognize a speaker's privacy using a privacy-preserving-based hashing technique that used paralinguistic features.

The SER has recently become an emerging area of research in digital audio signal processing. Many researchers have developed a variety of techniques that utilize deep learning approaches, such as 2D-CNNs and 1D-CNNs models in order to increase the level of accuracy. Typically, the researchers utilized the pre-trained CNNs weights in order to extract the discriminative

high-level features, which were fed into the traditional RNNs afterwards for sequence learning [12,20]. The recognition performance was slightly increased when these pre-trained models were utilized, but the computational complexity was also increased with the use of huge pre-trained network weights. The current deep learning approaches, which include CNNs, DBNs, and CNN-LSTM architectures, have not shown enormous enhancements with respect to emotion recognition. In this study, we present a novel end-to-end one-dimensional CNN network with a BiGRU in order to identify the state of a speaker in term of the emotions. We assemble a DCNN-GRU network for the SER that utilizes one-dimensional dilated layers in stacked hierarchical features learning blocks (HFLBs) in order to extract the emotional features from the speech signals. The GRU network is connect to this network in order to learn and extract the long-term dependency from the time-varying speech signals. The design one-dimensional dilated CNN model accepts raw data, such as raw audio clips in order to remove noises and learn the high-level features while using the suggested model, which decrease of training time due to less parameters being used during the training. The proposed model is evaluated using three speech datasets, which included the IEMOCAP, EMO-DB, and RAVDESS, in order to ensure the effectiveness and the significance of the model. The key contributions of our work are illustrated below.

- We investigated and studied the current literature of speech emotion recognition (SER), and we analyzed the performance of the classical learning approaches vs the deep learning approaches. As a result, we were inspired from the performance and the recent successes of the 2D-CNN models, so we planned a one-dimensional dilated convolutional neural network DCNN for an SER that can learn both the spatial features and the sequential features from the raw audio files by leveraging the DCNN with a deep BiGRU. Our proposed system accomplished automatically modeling the temporal dependencies using the learned features.
- The refining of the input data always plays a crucial role with an accurate prediction, which ensures improvement with the final accuracy. The existing methods for the SER lack the prior step of refining of the input data, which effectively assists boosting the final accuracy. In this study, we proposed a new preprocessing scheme in order to enhance and remove noise from the speech signals, which utilize the FFT and a spectral analysis, so our preprocessing step plays an important role with the SER system, which successfully dominated the state-of-the-art systems.
- We intensely studied the speech signals, the linguistic information, and the paralinguistic information for the SER, and proposed a method to extract the hierarchal emotional correlation. The existing literature for the SER lacks a focus on the hierarchal correlations, which easily recognize the emotional state and boost the final accuracy. In this paper, we proposed four stacked one-dimensional DCNN hierarchical features learning blocks (HFLBs) in order to raise the learned features and easily recognize the emotional signs. Each block consists of one dilated convolutional layer (DCN), one batch normalization (BN) layer, and one leaky_relu (Relu) layer.
- Our system is suitable to monitor the real-time processing in order to directly accept the raw audio speech files without reshaping them through high computations devices, which proved experimentally that our system can learn a lot of emotional features from the raw audio files. To the best of our knowledge, the proposed system is new, and this is the first contribution in this domain.
- We tested the robustness and the significance of our proposed system over three benchmarked datasets, which included the IEMOCAP, EMO-DB, and RAVDESS, and they

achieved 72.75%, 91.14%, and 78.01% accuracy, respectively. We also compared them with the baseline SER method. Our system outperformed from the other systems.

The rest of the paper is divided as follows. Section 2 highlights the literature about the SER using the low-level descriptors and the high-level descriptors. Section 3 represents the proposed system methods and the materials. The experimental results, the comparisons with the baseline, and the discussion is presented in Section 4. Section 5 concludes the paper and offers a future direction for the proposed system.

2 Literature Review

Speech emotion recognition (SER) is an active research area of digital signal processing that has been actively occurring throughout the last decade. The research frequently presents innovative techniques in this era in order to increase the performance of the SER and reduce the complexity of the overall system. Usually, an SER system has two core parts, which have challenges that need to be solved for an efficient emotion recognition system that include (i) the selection of the robust, discriminative, and salient features of the speech signals [21] and (ii) the classification methods in order to accurately classify them accordingly. Hence, the researchers currently use a dominant source, deep learning approaches for the robust features extraction and selection [22], which is quickly growing in this field. The state-of-the-art SER [23] developed methods for the SER in order to improve the performance of the existing systems, which utilize the CNN architectures that extract features from the speech spectrograms. Similarly, [24] used the end-to-end deep learning approach for the features extraction and classification [5] in order to recognize the emotional state of the distinct speakers.

In this technological era, the deep learning approaches have become popular in all fields and specifically in the field of SER that utilize the 2D-CNN models to extract the deep hidden cues from the spectrograms of the speech signals [24]. Actually, a spectrogram is a 2D plotting of the speech signal frequency with respect to time, and it is a more suitable representation for the 2D-CNN models [25]. Similarly, some researchers used the transfer learning techniques for the SER that utilizes spectrograms to train the pre-trained Alex Net [26] and VGG [27] models in order to identify the state of the speakers in term of the emotions [28]. Furthermore, the researchers used the 2D-CNNs to extract the special information from the spectrograms and the LSTM, or the RNNs were utilized to extract the hidden sequential and temporal information from the speech signals. Currently, the CNNs have increased the research interest of the SER. In this regard, [29] developed a new end-to-end method for an SER that utilizes a deep neural network (DNN) with the LSTM, which directly accepts the raw audio data and extracts the salient discriminative features rather than obtaining the handcrafted features. Most researchers used the joint CNNs with the LSTM and the RNNs for the SER [30] to capture the special cues and the temporal cues from the speech data in order to recognize the emotional information. The authors developed a technique in [31] for the SER that used a fixed variable-length, which utilizes the CNN-RNNs, where the CNN is used to extract the salient features from the spectrograms, and the RNNs controlled the length of the speech segment. Similarly, [32] used a hybrid approach for the SER, which utilized a CNN pre-trained Alex-Net model for the features extraction, and a traditional Support Vector Machine (SVM) was used for the classification. In [33], the authors suggested a deep learning model for a spontaneous SER that utilized the RECOLA usual emotions database for the model's evaluation.

The SER has many CNN methods it can use to take various types of inputs, which include spectrograms, log Mel spectrograms, speech signals, and Mel frequency cepstral coefficient

MFCCs [34] in order to recognized the emotional state of the speakers [35]. In this field, some researchers combined the traditional approaches with the advancements to utilize the pre-trained CNNs systems in order to extract the salient information from the audio data. Furthermore, they used the traditional machines to classify the emotions from the learned features [36] by using huge network parameters that boosted the complexity of the overall system. In [8], the authors developed techniques and introduced a new deep learning model for the SER that used the RAVDESS [17] and the IEMOCAP [16] datasets, which use less parameters to recognized the different emotions with high accuracy and less computational complexity. In this article, we proposed a new strategy for the SER that uses a one-dimensional DCNN model with four stacked hierarchical blocks. Each block consisted of one dilated convolutional layer with a rectified linear unit (relu) and one batch normalization layer (BN) with a proper dropout setting in order to reduce the model overfitting. Furthermore, we used a BiGRU to adjust the global weights and to extract the most salient high-level sequential information form the learned features. The final learned sequential information was passed from the last fully connected (FC) layer with a softmax activation in order to produce the probabilities of the different emotions. A detail description of the proposed technique is explained in the upcoming sections.

3 Proposed SER System

In this section, we properly demonstrate the proposed speech emotion recognition (SER) framework and its main components as well as provide a detailed description. In the field of SER, the selection of the appropriate features is a challenging problem for researchers to clearly define and distinguish among the emotional features and the non-emotional features. The features extraction methods are classified by the low-level (handcrafted) and the high-level (CNN) techniques. The handcrafted features extraction methods are designed by the feature engineering strategies, which are explained in more detailed in [37]. The high-level features are extracted using various deep learning approaches, such as the DBN, DNN, and CNN's in order to learn features from the input data by adjusting the weights automatically according to the input data. The prediction performance of the learned features is better than the handcrafted features [38]. In the field of SER, most researchers used the 2D-CNNs models for the feature extraction and the emotion classification [35], which requires more attention to the data preparation. The speech signal lays in one dimension, and the 2D-CNNs require two-dimensional input because of this. As a result, in order to preprocess the data, converting the speech signals into spectrograms is compulsory [39]. With the transformation of the 1D speech signals into the 2D speech spectrograms, some useful information may be lost. Similarly, the original speech signals have rich emotions cues instead of spectrograms. Due to this information, we proposed a novel 1D-DCNN-BiGRU system for the SER in order to directly accept the raw speech signals without transforming them, which is explained in the upcoming section.

3.1 One Dimensional DCNN-GRU System

The proposed one-dimensional CNN-GRU architecture is constructed in order to endure the raw speech signals in their original form. The system consists of three main chunks that includes a preprocessing step in order to enhance the speech signals that use the FFT and a spectral analysis. First, our model utilizes the 1D-DCNN network, which consists of four one-dimensional convolution layers with a rectified linear unit (Relu) and two batch normalization (BN) layers in order to learn the local features and prepare the initial tensor for a stacked network. We designed the network in order to place the first couple of convolution layers, which have 32 filters that are size 7 with stride setting one. Subsequently, we used the BN layer to normalize or to re-scale

the input in order to improve the speed and the performance. Similarly, the third and the fourth convolution layers were placed consecutively after the first BN layer. These convolutions had 64 filters that are size 5 with a stride setting two. The dropout and the L1 kernel regularization technique were used in the network to reduce the system overfitting [40]. Second, the stacked network had four (HFLBs), and each block consisted of one dilated convolution layer, one BN layer, and one leaky_relu layer. The stacked network extracted the salient features from the input tensor using a hierarchical correlation. Finally, a fully connected scheme that consisted of a BiGRU network and a fully connected (FC) layer with a softmax classifier was used to produce the probabilities of the classes. The learned spatiotemporal features were then directly passed from the FC layer, which can be stated as:

$$z^L = b^L + z^{L-1} \cdot w^L \tag{1}$$

Essentially, softmax was used as a classifier in this model, which calculated the prediction based on the maximum probability. We utilized the softmax classifier and generalized it for the multi-class classification in order to have more than two values with a label y, which can be expressed as:

$$x_i = \sum_j h_j w_{ji} \tag{2}$$

$$\text{softmax}(z)_i = p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_i}} \tag{3}$$

In Eqs. (2) and (3), z_i represents the input to the softmax, h_j shows the activation of the layer and the weight among the penultimate, and the softmax layer is illustrated by w_{ij}, which is connected to each other. Hence, the predicted emotion or label y would be:

$$y = \text{argmax}(p)_i \tag{4}$$

The overall structure of the suggested SER system is demonstrated in Fig. 1, and a detailed description of the sub-components is explained in the subsequent sections.

3.2 Speech Signals Enhancement

Speech signals need to be refined using an efficient technique that enhance the training performance and ensure the final prediction of the system. In our proposed system, we contributed a speech signal enhancement module for the SER, which is summarized in Fig. 2. We designed a module to enhance the high pass speech signals and the low pass speech signals using efficient algorithms. In the low pass speech signals, we estimated the spectrum by utilizing the fast Fourier transformation (FFT) in order to find the low-frequency band of the voice. We performed the spectral subtraction by utilizing the algorithm that is described in [41] in order to denoised and enhance the low pass speech signal, which is expressed as:

$$rdt[n] = \text{IFFT}\left\{\sqrt{S_l}[m].e^{j\theta_{l[m]}}\right\} \tag{5}$$

The inverse fast Fourier transformation is illustrated by the IFFT, Θ_l [m] represents the low pass noise speech signal phase of the FFT, and

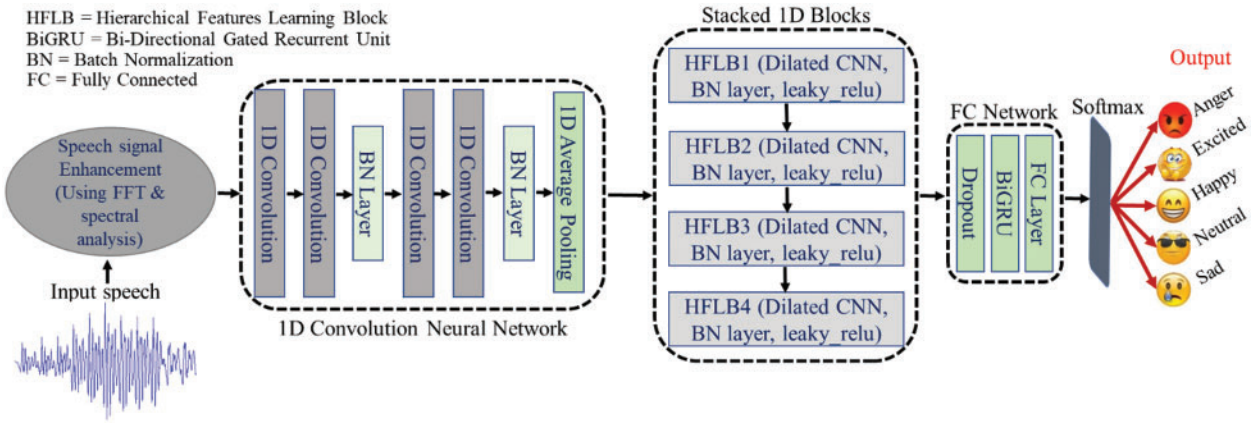

Figure 1: A detailed overview of the proposed one-dimensional architecture of the SER. In the 1D-CNN networks, we extracted the local features from the speech signals, and then passed from the stacked blocks. The blocks have connections with the last fully connected network where we adjusted the global weight and extracted the temporal or the sequential cues that utilized the gated recurrent unit (GRU) and the fully connected (FC) layers with softmax to calculate the probability of each class or emotion

$$S_l[m] = \begin{cases} R_l[m] - \alpha N_{l[m]}, & \text{if } R_{l[m]} > (\alpha + \beta) N_{l[m]} \\ \beta N_{l[m]}, & \text{otherwise.} \end{cases} \tag{6}$$

In the above equations, m and n indicated the frequency and the time indices. The squared magnitude of the FFT is represented by $R_{l[m]}$, and $N_{l[m]}$ shows the spectral estimate of a low pass speech signal. α indicates the positive subtraction factor, and β is the positive spectral floor parameters, which is described in [41] for the low pass speech signals. Similarly, in the high pass speech signal enhancement, we replaced $R_{l[m]}$ in Eq. 6 with:

$$R_h[m] = \sum_{k=1}^{k} r_h^k[m] R_h^k[m] \tag{7}$$

In additional, we used the following equation for the enhancement of the high pass speech signals.

$$R[m] = \sum_{k=1}^{k} r^k[m] R^k[m] \tag{8}$$

In the equations, $R_h^k[m]$ and $R^k[m]$ illustrated the squared magnitude of the FFT of the high pass speech signals with the k-th discrete spherical sequences, such as $r_h^k[m]$ and $r^k[m]$ in order to adoptively calculate the frequency-dependent weights. Similarly, phase Θ_l [m] is replaced by $r_h^l[m]$, and the noise spectral estimated by $N_l[m]$ is replaced by $R_h[m]$ in order to enhanced the high pass speech signals. A detailed overview of the proposed speech signal enhancement module is shown in Fig. 2 below.

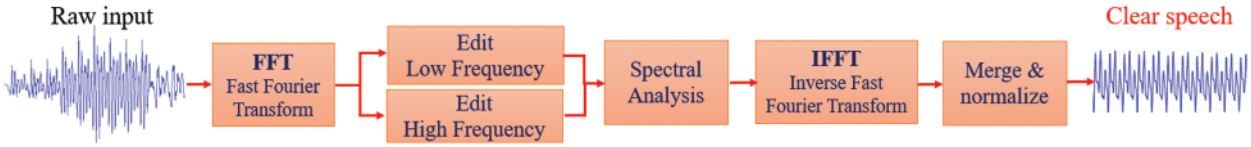

Figure 2: A detailed overview of the proposed speech signals enhancement technique that utilizes the fast Fourier transformational (FFT) and a spectral analysis (SA) that is based on the low pass and the high pass signals in order to remove noises and make it clear

3.3 Hierarchical Features Learning Blocks (HFLBs)

The proposed stacked one-dimensional network consists of four HFLBs, which were specially designed to extract the emotional information and the hierarchical correlation, and they boost the learned local features. Hence, all the blocks were stacked in a hierarchy, and each HFLB had one dilated convolution layer, one BN layer, and one leaky rectified linear unit. In the dilated convolution layer, we used a 1D filter of size 3 with a stride setting one to extract the high-level salient cues from the audio signals with a leaky_relu activation function. The batch normalization (BN) layer was used in all the HFLBs after the dilated convolution layer to normalize and re-scale the features map in order to increase the recognition performance and speed up the training process. In the pooling scheme, we proposed the average pooling strategy of filter size 4 with the same stride setting in the 1D-CNN architecture. We selected the average value in order to down-sample the input tensor by removing the redundancy and the distortion with this scheme. The dilated convolution layer played an important role in the HFLBs to extract the most salient cues and the emotional cues from the learned features, which produced a features map by computing the dot (.) product among the input value and the filters, which can be represented as:

$$z(n) = x(n) * w(n) = \sum_{m=-L}^{L} x(m) .w(n-m) \tag{9}$$

The proposed 1D convolution layer yields a signal x(n) as the input and produces the result z(n) by utilizing the convolution filters w(n) with respect to the L size. In the suggested model, the filter w(n) is randomly initialized in the dilated convolution layer in our experimentations.

The leaky rectified linear unit (leaky_relu) activation function σ is utilized in order to remove and replace the negative value with zero, which can be represented as:

$$\sigma(x_i) = \begin{cases} x_i, & \text{if } x_i, \geq 0 \\ \sigma(e^x - 1) & \text{if } x_i, < 0 \end{cases} \tag{10}$$

After the leaky_relu function, we used the BN layer to normalize the input features map of the previous layer for each batch. The transformation is applied in the BN layer by utilizing the mean and the variance of the convolved features, which is represented as:

$$z_i^L = \sigma(BN(b_i^L + \sum_j z_j^{L-1} .w_{ij}^L)) \tag{11}$$

In the above equation, z_i^L and z_j^{L-1} denotes the output features of the i-th input feature at the L-th layer, and the j-th input features is represented at the $(L-1)$th layer. The convolution filters

are represented by w_{ij}^{L}, between the i-th and the j-th features. In the 1D-CNN architecture, we passed the normalized features from the average pooling layer in order to reduce the dimensionality of the features map, which is a non-linear down-sampling technique that is represented as:

$$z_k^L = \text{avg } z_p^L \parallel \forall \rho \in \Omega_k \tag{12}$$

In the above Equation, Ω_k shows the pooling area with index k, and z_p^L represents the input features map of the Lth average pooling layer with index k and P. The final output results of the mechanism are represented by z_k^L, which is shown in the equation. The pooling scheme is a core operation of the 1D-CNN, which is represented in Eq. (12).

3.4 Bi-Directional Gated Recurrent Units (BiGRU)

The gated recurrent units (GRUs) are a special and more simplified network for the time series data in order to recognize the sequential information or the temporal information [42]. The GRU is a simplified version of the long short-term memory (LSTM), and very popular for sequential learning. The GRU is the combination of two gates, which include the update gate and the reset gate. The internal mechanism of the gates is different than LSTM. For example, the GRU update gate is the combination of the forget gate and the input gate of the LSTM, and the reset gate remained the same. The model is becoming gradually popular due to its simplicity over the standard LSTM network. The GRU network modifies the information inside the units, which is similar to the LSTM network, but it doesn't consume the distinct memory cells. The activation of the GRUs h_t^j at time t represent the linear interpolation among candidate $\hat{h}_t^j$ and the previous activation h_{t-1}^j, which can be represented as:

$$h_t^j = \left(1 - z_t^j\right) h_{t-1}^j + z_t^j \hat{h}_t^j \tag{13}$$

z_t^j represent the update gate, and it decides how much the units update and its activation, which is represented by:

$$z_t^j = \sigma \left(W_x x_t + U_z h_{t-1}\right)^j \tag{14}$$

Similarly, the update gate in the GRUs also computes the activation, and $\hat{h}_t^j$ utilizes the following Equation:

$$\hat{h}_t^j = \tanh \left(W x_t + U(r_t .* h_{t-1})\right)^j \tag{15}$$

In Eq. (15), r_t^j represents the reset gate, and an element-wise multiplication is denoted by (.*). The reset gate agrees the unit to forget in the previous cues when $(z_t^j == 0)$, which means the gate is off. This is the mechanism that informs the unit in order to search the head sign of an input sequence. We can calculate the reset gate by using the following equation, which is easily expressed as:

$$r_t^j = \sigma \left(W_r x_t + r^{h_{t-1}}\right)^j \tag{16}$$

where z represents the update gate, which controls the previous state, and the reset gate r is used to activate the short-term dependencies in units. The long-term dependencies are activated by updating gate z in units. In this paper, we utilized the BiGRU network for the SER in order to recognize the temporal information that used the learned features. The structure of the BiGRU network is illustrated in Fig. 3.

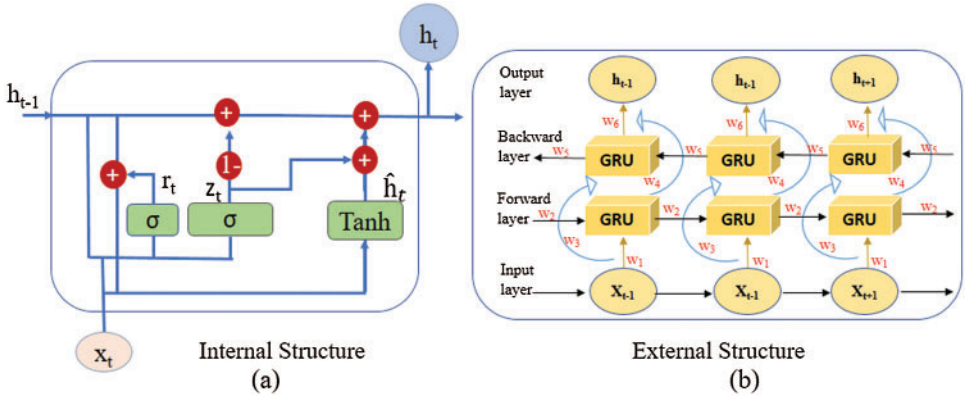

Figure 3: The overall internal and external architecture of the gated recurrent unit (GRUs). The internal structure is illustrated in (a), and the external structure is illustrated in (b)

3.5 Model Optimization and Computational Setup

In this section, we explained the detailed implementation of the suggested framework that was implemented in a python environment, which utilized a scikit-learn library and other libraries for machine learning. We tuned the model with different hyperparameters in order to make it sufficient for the speech emotion recognition (SER). We evaluated our model with different batch sizes, learning rates, optimizers, number of epochs, and regularization factors, such as L1 and L2 with different values. In the dataset, we divided the data into an 80:20 ratio, the 80% of the data was used for the model training, and the remaining 20% of the data was used for the testing. In these experiments, we performed the emotion prediction directly from the raw audio data or the speech rather than conducting any pre-processing. We used a single GeForce GTX 1070 NVIDIA GPU with an 8 GB memory for the model training and the evaluation. We trained our model using an early stopping method in order to save the best model and set the learning rate to 0.0001 with one decay after 10 epochs. We set the batch size to 32 for all the datasets, and 32 was selected for the number of hidden units for the GRUs with an Adam optimizer, which was used in the overall process for the training, and it achieved the best precision, which produced only a 0.3351 training loss and a 0.5642 validation loss.

4 Experimental Evaluation and Results

The proposed one-dimensional DCNN-GRU architecture was experimentally estimated using three benchmark speech emotions datasets, which included the IEMOCAP [16], EMO-DB [43], and RAVDESS [17]. All of these datasets are acted, and the actors expressed and read from scripts using different emotions. We documented the detailed experimental evaluations and the results in order to prove the effectiveness and the robustness of the system in the SER domain.

A detailed explanation of the datasets, the model evaluations, and the performances are included in the subsequent sections.

4.1 Datasets

The IEMOCAP dataset, which is an interactive emotional dyadic motion capture [16], is a challenging and well-known emotional speech dataset. The dataset consists of audios, videos, different facial motions, and text transcriptions, which were recorded by 10 different actors. The dataset has 12 hours of total audio-visual data and five different sessions in the dataset, and each session consists of two actors, which include one male and one female. The dataset was annotated by three field experts in order to assign the labels individually, and we selected the files that at least two experts agreed upon them. In contrast, we selected four emotions categories that included anger, happy, sad, and neutral, with 1103, 1084, 1636, and 1708 number of utterances, respectively, for the comparative analysis, which is frequently used in the literature.

The EMO-DB dataset, which is the Berlin emotion database [43], was recorded by ten (10) experienced actors, and it includes 535 utterances with different emotions. The dataset contains five (5) male and five (5) female actors who read pre-determined sentences in order to express the different emotions. The approximate time of the utterances in the EMO-DB is three to five seconds with a sixteen kHz sampling rate. The EMO-DB dataset is very popular, and it is frequently used for the SER in machine learning and deep learning approaches.

The RAVDESS dataset, which is the Ryerson audiovisual database of emotional speech and songs [17], is a British language dataset. It is a simulated dataset that is broadly used in recognition systems to identify the emotional state of the speaker during his/her speech and songs. The dataset is recorded by twenty-four experienced actors, which include twelve male actors and twelve female actors who use eight different emotions. Emotions, such as anger, being calm, sadness, happiness, neutral, disgust, fearful, and surprised contain 192, 192, 192, 192, 96, 192,192, and 192 number of audio files, respectively.

4.2 Experimental Evaluation

In this section, we practically evaluated our suggested system using three benchmark databases in order to test the model's robustness and effectiveness regarding the emotion recognition. All the datasets consist of a different number of speakers, so we split the data into an 80:20 ratio in each fold. 80% of the data was utilized for the model training, and the remaining 20% of the data was utilized for the model testing. We used a new strategy for the model training in this framework. First, we removed the un-important cues, such as noises from the raw audio files that utilize the FFT and a spectral analysis. Furthermore, we used a new 1D-CNN architecture to extract the hidden patterns from the speech segments and then feed them into a stacked dilated CNN network in order to extract the high-level discriminative emotional features with a hierarchical correlation. Similarly, we used the GRU network to extract the temporal cues that utilize the learned emotional features, which are explained in detail in Section 3.4. The suggested architecture of the SER uses various evaluation matrixes, such as precision, recall, F1_score, weighted accuracy, un-weighted accuracy, and confusion matrix for each dataset in order to checked the model prediction performance. In the weighted accuracy, we computed the model performance, which correctly predicted the labels divided by the whole labels in the corresponding class. Similarly, the un-weighted accuracy represents the model performance in order to compute the prediction among the total correct predicted labels divided by the whole labels in the dataset. The F1_score actually represents the weighted average of the precision and

the recall value, which always show the balance among the precision and the recall values. The confusion matrix illustrated the actual predicted values and the confusion with the other emotions in the corresponding class in a certain row and column. We evaluated all the datasets in order to generate the model prediction performance, the confusion matrix, and the class-wise accuracy. In addition, we conducted an ablation study to select the best architecture for the SER. A detailed evaluation and the model configuration is given in Tab. 1 for all the suggested datasets.

Table 1: An ablation study of our proposed model configuration for three standard speech emotion datasets using an audio clip or a waveform

Input	Architecture	IEMOCAP (%)	EMO-DB (%)	RAVDESS (%)
Audio clip or raw waveform	CNN + Softmax	64.20	83.83	68.56
	CNN + Stacked CNN + Softmax	65.18	83.90	69.21
	CNN + Stacked dilated CNN + Softmax	66.33	85.39	69.67
	CNN + BiLSTM + Softmax	65.44	86.83	69.37
	CNN + Stacked CNN + BiLSTM + Softmax	66.01	87.21	71.29
	CNN + Stacked dilated CNN + BiLSTM + Softmax	67.26	88.02	74.67
	CNN + BiGRU + Softmax	66.64	87.72	73.96
	CNN + Stacked CNN + BiGRU + Softmax	68.86	89.32	75.66
	CNN + Stacked dilated CNN + BiGRU + Softmax	72.75	91.14	78.01

The ablation study indicates the different architectures and the recognition results, which statistically represent the best model for the suggested task. We evaluated different CNN architectures in the ablation study, selected the best one, and started to further investigation. A detailed experimental evaluation of the IEMOCAP dataset is given below.

Tab. 2 represents the prediction performance of the system, which clearly shows the classification summary of each class in order to compute the precision, recall, and the F1 score of the IEMOCAP dataset. The overall prediction performance of the model is computed using the weighted accuracy and the un-weighted accuracy. Our model predicts the anger emotion and the sad emotion with high priority, and it predicts the happy emotion with low priority. The happy emotion has less linguistic information compared to the other emotions. Also, the neutral emotion is the most related to the other emotions. Due to this characteristic, our model is confused and missed recognizing these emotions. The confusion matrix shows the actual labels and the predicted labels of each emotion, and the class-wise prediction performance of the system for the IEMOCAP database is shown in Fig. 4.

The graph in Fig. 4a shows the class-wise recognition results, and (b) shows the confusion matrix of the IEMOCAP dataset. The model recognizes anger, sad, neutral, and happy emotions with 83%, 79%, 70%, and 59% accuracy, respectively. Our model is mostly confused with the happy emotion, which 12% of the anger class, 9% of the neutral class, and 10% of the sad emotion are recognized as happy. The model mixed the sad emotions and the happy emotions with each other and recognized 26% of happy as sad and 10% of sad as happy, so the overall performance of the model for the IEMOCAP dataset is better than the state-of-the-art SER models.

The overall average recall rate of the suggested model is 72.75%. Similarly, the classification performance of the proposed system for the EMO-DB dataset is illustrated in Tab. 3.

Table 2: Prediction performance of the proposed model in terms of precision, recall, F1_score, the weighted score, and the un-weighted score of the IEMOCAP dataset

Classes	Recall	Precision	F1_Score
Anger	0.83	0.85	0.84
Happy	0.59	0.54	0.57
Neutral	0.70	0.88	0.78
Sad	0.79	0.66	0.72
Weighted Acc	0.74	0.74	0.74
Un-weighted Acc	0.73	0.73	0.73

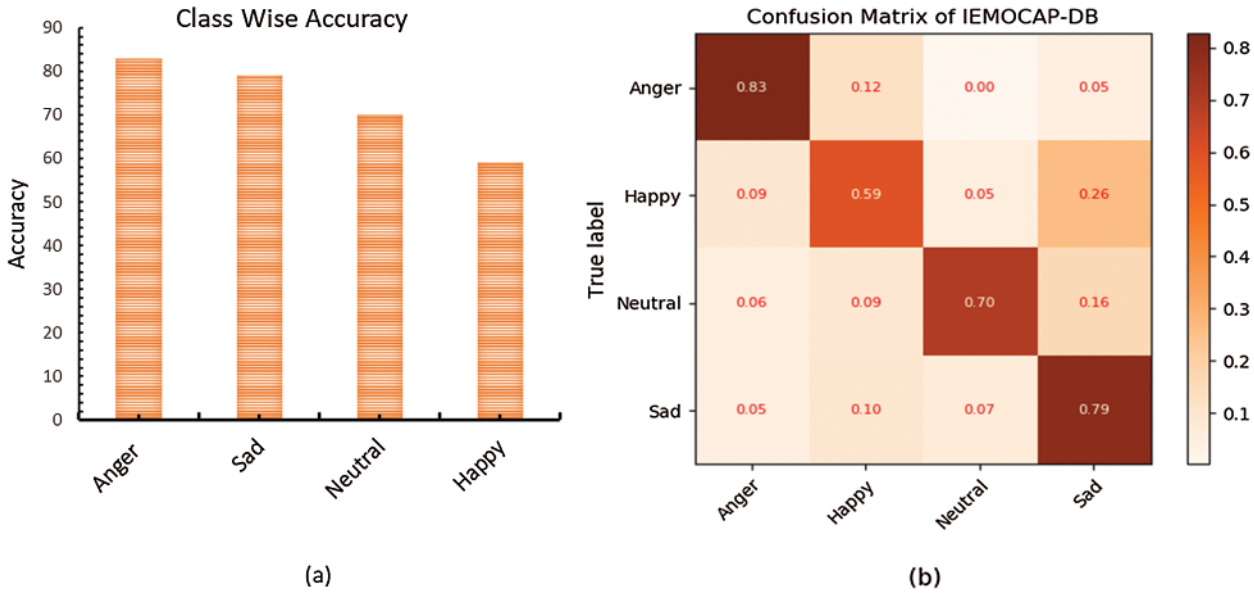

Figure 4: The class-wise accuracy of the proposed technique is illustrated in (a), and the confusion matrix between the actual labels and the predicted labels is shown in (b)

The classification summary shows that the overall prediction performance of the system over the EMO-DB *corpus* is good, but the model also shows a lower performance for the happy emotion due to less linguistic information. Furthermore, the model denotes a high performance for other emotions, such as anger, sad, fear, disgust, neutral, and boredom, which produced more than a 90% recognition rate. Similarly, again the model confused the happy emotion, which mixed it with the other emotions. In order to further investigate this, we generated the class-wise accuracy and the confusion matrix of the EMO-DB dataset to check the confusion ratio of each class with the other emotions. The confusion matrix and the class level accuracy are shown in Fig. 5.

Table 3: Prediction performance of the proposed model in terms of precision, recall, F1_score, weighted, and un-weighted score of the EMO-DB dataset

Classes	Recall	Precision	F1_Score
Anger	0.98	0.96	0.97
Boredom	0.93	0.99	0.96
Disgust	0.91	1.00	0.95
Fear	0.97	0.94	0.95
Happy	0.69	0.95	0.80
Neutral	0.94	0.81	0.87
Sad	0.96	0.61	0.75
Weighted Acc	0.90	0.92	0.92
Un-weighted Acc	0.89	0.91	0.90

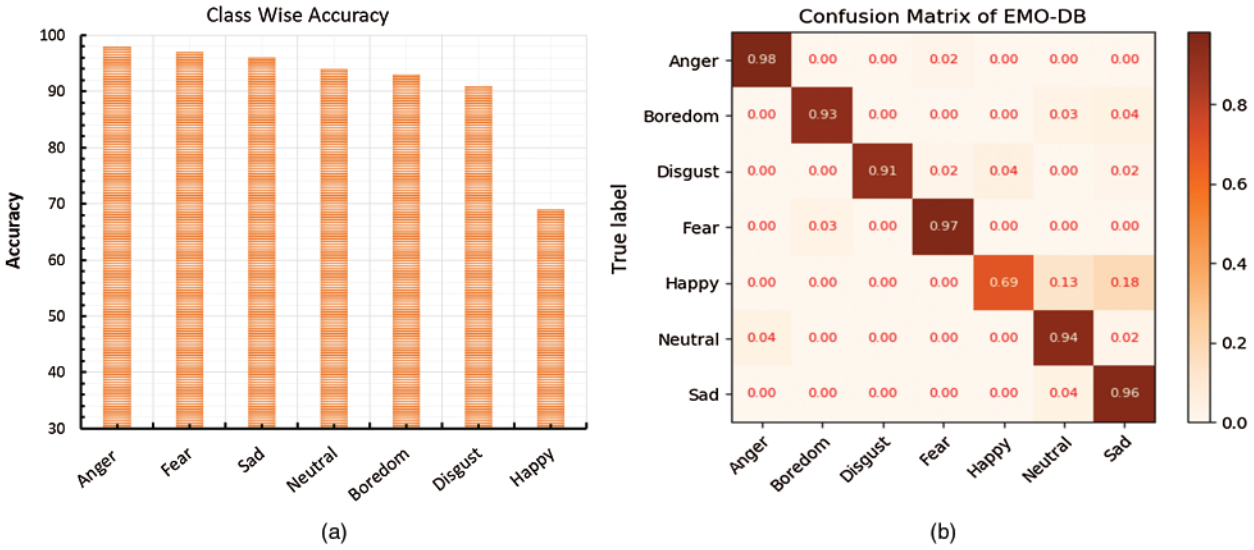

Figure 5: Class-wise accuracy of the proposed technique is illustrated in (a), and the confusion matrix between the actual labels and the predicted labels is shown in (b)

The class-wise accuracy is represented in Fig. 5a, which shows each class recognition ratio. The x-axes illustrate the directions toward the classes, and the y-axes show the recognition accuracy of the corresponding class. Fig. 5b illustrates the confusion matrix among the actual labels and the predicted labels of the Emo-DB dataset. The diagonal values of the confusion matrix represents the actual recall value of the proposed model for each emotion. The recognition rates for anger, fear, sadness, boredom, and disgust are more than 90% each, respectively, which is much greater than the recognition rate from the happy emotion. The proposed model increases the precision rate of the happy emotion more than the baseline model, which is relatively lower than the other emotions. The average recall rate of the system is 91.14% for the Emo-DB dataset. The prediction performance of the suggested technique for the RAVDESS dataset is presented in Tab. 5.

Tab. 4 represents the model prediction summary of the RAVDESS dataset. It was recently launched for the emotion recognition in the natural speeches and songs. The model's secure weighted accuracy and the un-weighted accuracy produced 80% and 78% scores for the overall classes. The individual recognition ratio of all the classes is much better than the state-of-the-art SER methods. Our model improves the overall performance, which includes the happy emotion. Similarly, the happy emotion recognition rate is relatively lower than the others, which is due to less linguistic information. The model mixed the happy emotion with the others, and it was confused during the prediction stage. We further investigated the model performance in order to find a class-wise accuracy and confusion matrix for the efficiency evaluation, which is shown in Fig. 6.

Table 4: Prediction performance of the proposed model in terms of precision, recall, F1_score, weighted score, and un-weighted score of the RAVDESS dataset

Classes	Recall	Precision	F1_Score
Anger	0.91	0.87	0.89
Calm	0.90	0.52	0.66
Disgust	0.80	0.98	0.88
Fearful	0.79	0.82	0.80
Happy	0.50	0.92	0.65
Neutral	0.67	0.63	0.65
Sad	0.80	0.94	0.86
Surprise	0.91	0.81	0.86
Weighted Acc	0.83	0.80	0.80
Un-weighted Acc	0.78	0.81	0.78

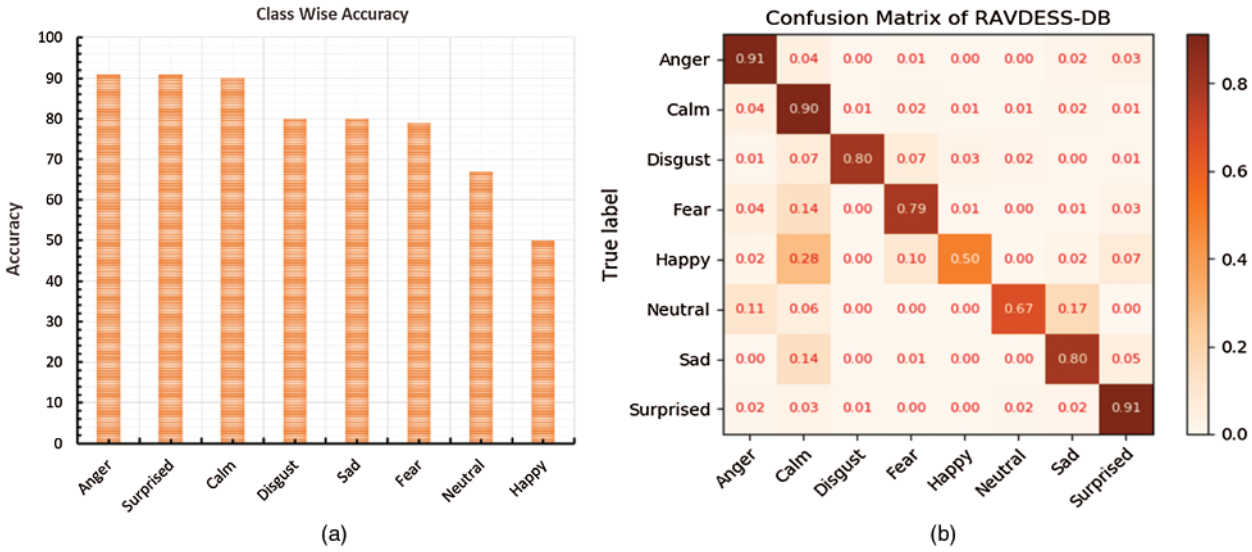

Figure 6: Class-wise accuracy of the proposed technique is illustrated in (a), and the confusion matrix between the actual and the predicted labels is shown in (b)

The class-wise accuracy represents the actual performance of the corresponding class in percentages, which is shown in Fig. 6a. The x-axes represent the number of emotions, and the y-axes represent the accuracy in percentages. Fig. 6b shows the RAVDESS dataset confusion matrix, which illustrates the model representations among the actual emotions and the predicted emotions. The model secured 91%, 90%, 80%, 79%, 50%, 67%, 80%, and 91% recognition scores for anger, calm, disgust, fear, happy, neutral, sad, and surprised, respectively. Similarly, the system recognizes the happy emotion with a low accuracy, but the recognition rate for happiness is better than the baseline methods. Hence, the overall prediction performance of the proposed system for the SER is better than the state-of-the-art methods.

4.3 Comparative Analysis

We compared the proposed model with the baseline SER methods in order to show the improvement, the robustness, and the effectiveness of the system. We applied the un-weighted evaluation matrix, which is mostly used in the literature to find the recognition rate of the system. We compared the un-weighted accuracy of the proposed system with the other baseline methods in this regard. The literature for the SER lacks the 1D-CNN architectures except for some limited articles, which still don't show good improvements with the performance. In contrast, we suggested a new one-dimensional DCNN-GRU system for the SER that uses HFLBs with a dilated convolution layer that efficiently recognized the emotional features. We evaluated the suggested system using three standard SER datasets and compared the results with the baseline techniques in order to show the model efficiency. Essentially, our model utilizes the new types of architectures, such as the dilated DCNN blocks, which are used to extract the emotional features and reduce the processing time for the model training. The comparison of the proposed system is illustrated in Tabs. 5 and 6 with the state-of-the-art systems in terms of the accuracy and the processing time.

Table 5: Comparative analysis of the proposed SER system with other baseline methods over three benchmark speech datasets. The output of our model outperformed the baseline methods.

IEMOCAP			EMO-DB			RAVDESS		
Year	Reference	Accuracy (%)	Year	Reference	Accuracy (%)	Year	Reference	Accuracy (%)
2019	[44]	52.14	2019	[45]	84.49	2019	[39]	64.48
2017	[20]	64.78	2019	[47]	88.99	2019	[52]	69.40
2019	[45]	57.10	2018	[49]	82.82	2019	[53]	75.79
2015	[46]	40.02	2019	[15]	80.79	2019	[54]	67.14
2014	[24]	51.24	2019	[51]	84.53	2020	[50]	71.61
2019	[47]	69.32	2020	[50]	86.10	2020	[7]	77.01
2019	[48]	66.50	2020	[7]	85.57			
2018	[38]	63.98						
2019	[21]	61.60						
2018	[49]	64.74						
2020	[50]	64.03						
2020	[7]	71.25						
Proposed		**72.75**			**91.14**			**78.01**

Table 6: A comparison of the processing time of the proposed 1D-DCNN system that is analyzed with other SER models. Our system has less processing time due to the simple architecture

Scheme	IEMOCAP-DB (s)	RAVDESS-DB (s)	EMO-DB (s)
ACRNN [49]	13487	–	6811
ADRNN [47]	13887	–	7187
CL-SER [7]	10452	6250	5396
Prop 1D-CNN	**8200**	**3970**	**3150**

We compared our model with different one-dimensional and two-dimensional CNN models, which are shown in the table above. In [44], the authors used 1D-CNN architecture, and they used local features learning blocks to extract the hand-crafted features from the speech signals, which pass to the sequential network to extract the temporal cues in order to recognize the emotions. The authors tried to improve the recognition rate but due to the simple and local features, they are not suitable for efficient SER models. The authors in [44] achieved an accuracy rate of 52% for the IEMOCAP dataset, which utilized the 1D-CNN model. The rest of the other studies that were compared [45,46] used 2D [48,50] and 3D-CNN [47,49] models for the SER [51], which had some significant changes by using a bagged support vector machine [52,53] and a capsule network [54]. These models secured up to 70% accuracy for the IEMOCAP, 85% accuracy for the EMO-DB, and they had approximately a 76% accuracy for the RAVDESS dataset. The prediction performance of the suggested system was reported as 72.75%, 91.14%, and 78.01% for the IEMOCAP, the EMO-DB, and the RAVDESS databases, respectively.

5 Discussion

We used a novel strategy for the SER in the proposed framework contribute a new 1D-DCNN model with hierarchical features learning blocks (HFLBs) in order to recognize the emotional features. Tab. 5 illustrates a comparative study of the model with the other state-of-the-art systems under the same conditions, such as datasets and an accuracy matrix. In Tab. 5, the researchers solved the SER problem using different techniques in this era, but they mostly used the 2D-CNN architectures to address the stated problem. Overall, the 2D-CNN strategies are built for visual data recognition and classification in the field of computer vision [55]. With the use of this strategy, we lost some paralinguistic cues with the speech signals, and we didn't achieve better accuracy for the emotion recognition. In order to address this limitation, we proposed a 1D-DCNN model that can accept the direct speech data in order to extract the features and recognize the paralinguistic information, such as emotions. Our model is able to predict emotions with a high accuracy rate compared to the other prior models that involve emotion recognition, which is shown in Tab. 5. Our system computes the probabilities for each segment to predict the class label for an emotion, and a class has a maximum average probability that is selected as a label. We utilized the equation given below in order to compute the class label in this paper.

$$L_u = \underset{k=1,\ldots k}{\text{argmax}} \frac{\sum_{t=1}^{T} p(y_t|x)}{T} \tag{17}$$

In the Equation, k represents the number of classes except for the silent class, and Lu is the predicted label of the corresponding class. We evaluated our proposed method using three different SER datasets in order to prove the robustness and the effectiveness of the system and also to

show the generalization of the system. Furthermore, our suggested model deals with additional items, which include providing real-time output to process the data sufficiently, because it is not contingent on the prospective situation. In addition, our system is capable of dealing with speech that has an arbitrary size without a reduction in the performance, and it has the ability to handle speech that contains more than one emotion class.

6 Conclusion and Future Direction

Speech emotion recognition (SER) has many challenges and limitations in the literature that need to be solved by using an efficient approach. We explored various deep learning approaches for the SER tasks, conducted different experimentations, and proposed a new architecture for the SER, which utilizes a one-dimensional plain DCNN strategy with stacked HFLBs and GRUs network. Our model investigates the emotional cues in the speech signals using the local and global hierarchical correlations that utilize the raw audio signals. We used the 1D-DCNN architecture with deep BiGRU networks in order to find sequential or temporal dependencies, so the proposed model is capable of learning local information and global contextual cues from the speech signals and recognize the state of the speakers. We tested our suggested system on three benchmark databases and achieved recognition accuracies of 72.75%, 91.14%, and 78.01% for the IEMOCAP, EMO-DB, and RAVDESS datasets, respectively, which proved the robustness and the significance of the system.

This work has many future directions, which include being used with an automatic speech recognition (ASR) system. We can easily integrate our work in this domain to employ a mutual understanding of the paralinguistic and linguistic elements of speech in order to develop a superior model for speech processing. Similarly, the proposed architecture can be further explored for the SER, which utilized deep belief networks, graphs, and spike networks, and it is also useful for the speaker recognition and identification in order to achieve better results with satisfactory computational costs.

Funding Statement: This work was supported by the National Research Foundation of Korea funded by the Korean Government through the Ministry of Science and ICT under Grant NRF-2020R1F1A1060659 and in part by the 2020 Faculty Research Fund of Sejong University.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. A. Naqvi, M. Arsalan, A. Rehman, A. U. Rehman, W. K. Loh *et al.,* "Deep learning-based drivers emotion classification system in time series data for remote applications," *Remote Sensing*, vol. 12, no. 3, pp. 587, 2020.

[2] S. Z. Bong, K. Wan, M. Murugappan, N. M. Ibrahim, Y. Rajamanickam *et al.,* "Implementation of wavelet packet transform and non linear analysis for emotion classification in stroke patient using brain signals," *Biomedical Signal Processing and Control*, vol. 36, no. 12, pp. 102–112, 2017.

[3] B. Wei, W. Hu, M. Yang and C. T. Chou, "From real to complex: Enhancing radio-based activity recognition using complex-valued CSI," *ACM Transactions on Sensor Networks*, vol. 15, no. 3, pp. 35, 2019.

[4] M. Swain, A. Routray and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[5] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, pp. 183, 2020.

[6] S. Demircan and H. Kahramanli, "Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech," *Neural Computing and Applications*, vol. 29, no. 8, pp. 59–66, 2018.
[7] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[8] Mustaqeem and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications*, pp. 114177, 2020.
[9] L. Chen, X. Mao and H. Yan, "Text-independent phoneme segmentation combining EGG and speech data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1029–1037, 2016.
[10] S. U. Khan and R. Baik, "MPPIF-Net: Identification of plasmodium falciparum parasite mitochondrial proteins using deep features with multilayer bi-directional lstm," *Processes*, vol. 8, pp. 725, 2020.
[11] S. Tripathi, A. Kumar, A. Ramesh, C. Singh and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions. Arxiv Preprint Arxiv:1906.05681, 2019.
[12] F. Karim, S. Majumdar and H. Darabi, "Insights into lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 7, pp. 67718–67725, 2019.
[13] H. Zhiyan and W. Jian, "Speech emotion recognition based on deep learning and kernel nonlinear PSVM," in *2019 Chinese Control And Decision Conf.*, Nanchang, China, pp. 1426–1430, 2019.
[14] U. Fiore, A. Florea and G. Pérez Lechuga, "An interdisciplinary review of smart vehicular traffic and its applications and challenges," *Journal of Sensor and Actuator Networks*, vol. 8, no. 1, pp. 13, 2019.
[15] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad *et al.,* "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, 2019.
[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower *et al.,* "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.
[17] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, pp. e0196391, 2018.
[18] S. Kang, D. Kim and Y. Kim, "A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show," *International Journal of Distributed Sensor Networks*, vol. 15, pp. 1550147719864886, 2019.
[19] M. Dias, A. Abad and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada, pp. 2057–2061, 2018.
[20] H. M. Fayek, M. Lech and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
[21] S. Jiang, Z. Li, P. Zhou and M. Li, "Memento: An emotion-driven lifelogging system with wearables," *ACM Transactions on Sensor Networks*, vol. 15, no. 1, pp. 8, 2019.
[22] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar *et al.,* "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.
[23] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey *et al.,* "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.
[24] K. Han, D. Yu and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Fifteenth Annual Conf. of the Int. Speech Communication Association*, vol. 1, pp. 1, 2014.
[25] P. Cao, W. Xia and Y. Li, "Heart ID: Human identification based on radar micro-Doppler signatures of the heart using deep learning," *Remote Sensing*, vol. 11, pp. 1220, 2019.
[26] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1097–1105, 2012.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Arxiv Preprint Arxiv: 1409. 1556, 2014.
[28] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93847–93857, 2019.
[29] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 4580–4584, 2015.
[30] Mustaqeem and S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical convlstm network," *Mathematics*, vol. 8, pp. 2133, 2020.
[31] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng *et al.,* "Emotion recognition from variable-length speech segments using deep learning on spectrograms," *Interspeech*, vol. 1, pp. 3683–3687, 2018.
[32] A. Zhang, W. Zhu and J. Li, "Spiking echo state convolutional neural network for robust time series classification," *IEEE Access*, vol. 7, pp. 4927–4935, 2018.
[33] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao and J. P. Xu, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, no. 10, pp. 271–280, 2018.
[34] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *International Journal for Advance Research in Engineering and Technology*, vol. 1, pp. 1–4, 2013.
[35] Q. Mao, M. Dong, Z. Huang and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.
[36] P. Liu, K. K. R. Choo, L. Wang and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification," *Soft Computing*, vol. 21, no. 23, pp. 7053–7065, 2017.
[37] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang *et al.,* "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, no. 5, pp. 27–35, 2018.
[38] D. Luo, Y. Zou and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," *Interspeech*, vol. 1, pp. 152–156, 2018.
[39] Y. Zeng, H. Mao, D. Peng and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, 2019.
[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
[41] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, no. 2, pp. 574–584, 2015.
[42] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Arxiv Preprint Arxiv: 1412.3555, 2014.
[43] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A database of german emotional speech," *Ninth European Conf. on Speech Communication and Technology*, vol. 1, pp. 1–10, 2005.
[44] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, no. 4, pp. 312–323, 2019.
[45] L. Guo, L. Wang, J. Dang, Z. Liu and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.
[46] W. Zheng, J. Yu and Y. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," *Int. Conf. on Affective Computing and Intelligent Interaction*, vol. 1, pp. 827–831, 2015.
[47] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[48] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins *et al.,* "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[49] M. Chen, X. He, J. Yang and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[50] D. Issa, M. F. Demirci and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, pp. 101894, 2020.

[51] P. Jiang, H. Fu, H. Tao, P. Lei and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.

[52] M. A. Jalal, E. Loweimi, R. K. Moore and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," *Proc. Interspeech*, vol. 1, pp. 1701–1705, 2019.

[53] A. Bhavan, P. Chauhan and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, no. 3, pp. 104886, 2019.

[54] A. A. A. Zamil, S. Hasan, S. M. J. Baki, J. M. Adam and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," *International Conf. on Robotics, Electrical and Signal Processing Techniques*, vol. 1, pp. 281–285, 2019.

[55] N. Khan, A. Ullah, I. U. Haq, V. G. Menon and S. W. Baik, "SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network," *Journal of Real-Time Image Processing*, vol. 1, pp. 1–15, 2020.