Σ mathematics

MDPI

Article

CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network

Mustaqeem and Soonil Kwon *

Interaction Technology Laboratory, Department of Software, Sejong University, Seoul 05006, Korea; mustaqeemicp@gmail.com

* Correspondence: skwon@sejong.edu; Tel.: +82-2-3408-3847

Received: 18 October 2020; Accepted: 24 November 2020; Published: 30 November 2020

Abstract: Artificial intelligence, deep learning, and machine learning are dominant sources to use in order to make a system smarter. Nowadays, the smart speech emotion recognition (SER) system is a basic necessity and an emerging research area of digital audio signal processing. However, SER plays an important role with many applications that are related to human–computer interactions (HCI). The existing state-of-the-art SER system has a quite low prediction performance, which needs improvement in order to make it feasible for the real-time commercial applications. The key reason for the low accuracy and the poor prediction rate is the scarceness of the data and a model configuration, which is the most challenging task to build a robust machine learning technique. In this paper, we addressed the limitations of the existing SER systems and proposed a unique artificial intelligence (AI) based system structure for the SER that utilizes the hierarchical blocks of the convolutional long short-term memory (ConvLSTM) with sequence learning. We designed four blocks of ConvLSTM, which is called the local features learning block (LFLB), in order to extract the local emotional features in a hierarchical correlation. The ConvLSTM layers are adopted for input-to-state and state-to-state transition in order to extract the spatial cues by utilizing the convolution operations. We placed four LFLBs in order to extract the spatiotemporal cues in the hierarchical correlational form speech signals using the residual learning strategy. Furthermore, we utilized a novel sequence learning strategy in order to extract the global information and adaptively adjust the relevant global feature weights according to the correlation of the input features. Finally, we used the center loss function with the softmax loss in order to produce the probability of the classes. The center loss increases the final classification results and ensures an accurate prediction as well as shows a conspicuous role in the whole proposed SER scheme. We tested the proposed system over two standard, interactive emotional dyadic motion capture (IEMOCAP) and ryerson audio visual database of emotional speech and song (RAVDESS) speech corpora, and obtained a 75% and an 80% recognition rate, respectively.

Keywords: affective computing; artificial intelligence; deep learning; ConvLSTM; gated recurrent units (GRUs); speech emotion recognition; raw speech data

1. Introduction

Speech emotion recognition (SER) is an active area of research and a better way to communicate using among human–computer interaction (HCI). Speech signals play an important role in various real-time HCI applications, such as clinical studies, audio surveillance, lies detection, games, call centers, entertainment, and many more. However, the existing SER techniques still have some limitations, which include robust feature selection and advanced machine learning methods, for an efficient system. Thus, researchers are still working to find a significant solution in order to choose the right features and advance the artificial intelligence (AI) based classification techniques. Similarly, the background noise in a real-world voice could also be dramatically effective on the machine learning system [1,2]. Nevertheless, the development of a decent speech-based emotion recognition system can easily increase the user experience in different areas with the HCI, such as AI cyber security and mobile health (mHealth) [3]. The AI model has better potential than classical model to recognize the emotional state of the speaker from their signals during speech and shows a considerable impact on the SER in order to imitate these emotions [4]. Deep learning and the AI performed a significant improvement in the mobile health assistance field as well as increased their performance [5,6]. Nowadays, the researchers utilize the deep learning approaches in order to solve the recognition problems, such as voice recognition, emotion recognition, gesture recognition, face recognition, and image recognition [7–9]. The main advantage of the utilization of deep learning approaches is the automatic selection of the features. The suggested model adjusts the weights in the convolution operation according to the input data, and it finds the important and the task-specific attributes [10,11].

Recently, the researchers have introduced a various number of deep neural networks (DNNs) techniques to model the emotions recognition in the speech data. These models are fundamentally different in nature. For example, one group designed a DNN model that detects significant cues from raw audio samples [2], and another group utilized a particular representation of an audio recording to provide input for the model [12]. In order to make a robust and a significant model, the researchers utilized different types of feature combinations using diverse network strategies. Nowadays, deep learning and artificial intelligence methods are dominant toward extracting hidden cues and to recognize lines, curves, dots, shapes, and colors, because they utilize various types of convolution operations [13]. Thus, the researchers have utilized the modest end-to-end models for the emotion recognition, which include convolution neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), and deep belief networks (DBNs) [14,15]. These models extract high-level salient features from the speech signals that achieved a better recognition rate compared to the low-level features [8,13]. The researchers have used a deep learning model for an efficient SER, but the level of accuracy and the recognition rate are still quite low due to the data scariness and the worst model configuration. The current CNN approaches lack the ability to increase the accuracy of the emotion recognition in the SER domain. Furthermore, the researchers have utilized the RNN and the LSTM in order to learn the long-term dependencies and recognize the emotions. Nevertheless, these techniques have not revealed any significant changes in accuracy, but they increased the cost computation and training time of the whole model [16]. The SER domain still has many issues that need to be solved with an efficient and significant framework that recognizes the spatial emotion cues as well as the sequential cues.

In contrast, we addressed the low-accuracy issues and proposed a novel framework for the emotion recognition that utilized speech signals. We designed a system which recognized the local hidden emotional cues in the raw speech signal by utilizing the hierarchical ConvLSTM blocks [17]. We adopted two types of sequential learning in this framework. The first type is the ConvLSTM, which learns the local features using input-to-state, and the second is state-to-state transition using a convolution operation, which learns the global features using input sequences by a stacked BiGRU network. Furthermore, we utilized the center loss function for the final classification, which ensured the prediction performance and improved the recognition rate. For the evaluations of the proposed system, we utilized two benchmark databases that included the interactive emotional dyadic motion capture (IEMOCAP) dataset [18] and ryerson audio visual database of emotional speech and song (RAVDESS) [19] dataset, which obtained a 75% recognition rate and an 80% recognition rate, respectively. We compared the proposed SER system with other baselines in the experimental section, which clearly designates the robustness and the significance of the proposed system. The summarized contributions of the proposed framework are demonstrated below.

- We proposed a novel one-dimensional (1D) architecture for SER using the ConvLSTM layers, which find the spatial and the semantic correlation between the speech segments. We adopted an input-to-state and a state-to-state transition strategy that utilized a hierarchical connection that learns the spatial local cues by utilizing the convolution operation. Our model extracts the spatiotemporal hierarchical emotional cues by proposing several ConvLSTM blocks that are called the local features learning blocks (LFLBs). According to the best of our knowledge, this is the up-to-date invasion of AI and deep learning in SER domain.

- The sequence learning is very important to adjust the relevant global features' weights in order to find the correlation and the long-term contextual dependencies in the input features, which ensure the improvement of the feature extraction and the recognition performance. That's why we adaptively adjust the GRUs network with the ConvLSTM in order to compute the weights of the global features, since the global feature weights are re-adjusting from the local and the global features. According our knowledge, this is the first time that the GRUs network is seamlessly integrated with the ConvLSTM mechanism for the SER using raw speech signals.
- We introduced a center loss function, which learns and finds the center of the feature vector of each class and defines the distance among the features and their consistent class center. The usage of the center loss with softmax loss improves the recognition result with a high performance. We proved that the center loss is easily optimized and increased the performance of deep learning model. To the best of our knowledge, this is the first time that the center loss is used with the ConvLSTM in the SER domain.
- We tested our system on two standard corpora, which include the IEMOCAP [18] and the RAVDESS [19] emotional speech corpus, in order to evaluate the effectiveness and the significance of the model. We secured a 75% recognition rate and an 80% recognition rate for the IEMOCAP and RAVDESS corpora. The experimental result indicated the robustness and the simplicity of the proposed system and showed the applicability for actual application.

The rest of the article is divided as follows. Section 2 highlights the associated SER literature, and the proposed framework for emotion recognition is shown in Section 3. Detailed documentation of the practical results, experimentations, discussion, and comparison with baseline SER methods are presented in Section 4. The conclusion for this paper and the direction for future work is presented in Section 5.

2. Related Works

In the literature, the majority of the researchers used CNNs, RNNs, DNNs, and LSTM architecture to develop an efficient system for emotion recognition that utilized speech data [20] or some researchers used a combination of them [2,21]. The combined architecture of the CNN and the LSTM extracted the hidden patterns as well as recognized the long-term contextual dependencies in the speech segments and learns the temporal cues [22]. The selection and the identification of the robust and the significant features is a challenging task for emotion recognition in the speech data that can be utilized for efficient model training. In this advanced era, researchers have used various approaches in order to improve the recognition accuracy and to solve the problems of the existing SER methods. In Reference [23], the authors utilized the CNN approach for the SER using the raw speech data. The CNN model processes the input signals to detect the noises, and it selects the specific regions of the audio sample for the emotion recognition, which achieved better results at that time [6]. In Reference [20], the authors utilized the predefined features, which included the eGeMaps [24] with a self-attention module, using global windowing techniques for the features extraction in order to recognize the emotional state of the speaker. Similarly, Reference [25] used a combination of the predefined features sets, which included the eGeMaps [24] and the ComParE [26], for the SER task. The authors enhanced the level of accuracy with these predefined features sets as well as increased the cost computations of the overall model. The authors utilized the Berlin emotion dataset (EMO-DB) [27] corpus in Reference [28] for the emotion recognition that utilized the time disturbed CNN layer and obtained highly favorable results over the state-of-the-art model of that time. The authors converted the speech signals into spectrograms by applying a short-term Fourier transform algorithm and fed it as an input into the time-distributed CNN model. A similar approach has been used in Reference [29] that utilizes speech data and recognizes emotions through spectrograms by applying the CNN model. The authors utilized a kernel shape filter in the convolution operation of the pre-trained CNN Alex-Net [30] model, which secures a satisfactory recognition rate on the EMO-DB [27] dataset. Nowadays, internet of things (IoT) has become popular in different fields such as privacy-preserving [31], stock traders using deep learning [32], and Q-learning approach for market forecasting [33].

However, the authors in Reference [22] developed an advanced two-dimensional (2D) CNN-LSTM model for an emotion recognition system that recognized the spatial and the temporal information in the speech data using log-Mel spectrograms. In Reference [34], the authors proposed an SER model and enhanced the recognition rate by utilizing a data augmentation approach that is based on a generative adversarial network (GAN) [35]. The authors produced synthetic spectrograms in order by applying the proposed technique using the IEMOCAP [18] emotional speech corpus. In the recent era, another proficient method was proposed in Reference [36], which extracted the Mel frequency cepstral coefficients (MFCCs) for the model training and reduced the features using a fuzzy c-mean clustering algorithm [37,38]. The authors developed several classification techniques, such as support vector machine (SVM), an artificial neural network (ANN), and k-nearest neighbors (KNN) in order to enhance the recognition accuracy of the existing SER methods. Similarly, the authors used text files and audio files of the IEMOCAP [18] dataset in order to develop a multi-model approach for the SER and in [39] developed a hybrid approach, deep learning, and machine learning combined in order to classify the emotional state of the speaker [6]. The authors used a semi-CNN for the feature learning, the SVM was utilized for the classification, and they performed speaker-dependent experiments and speaker-independent experiments to recognize emotions in speech signals. Furthermore, in Reference [39], the authors used the traditional machine learning method using the modulation spectral features (MSFs) and the prosodic features and classified them by multi-class linear discriminant analysis (LDA) classifier [40]. With the combination of these features, which the authors achieved an 85.8% accuracy with the testing set of the EMO-DB [27] corpus. Moreover, the researchers used the SVM classifier with different type features [41] that included the extracted timber and the MFCCs features [14] and the extracted Fourier parameter and the MFCCs features, and they classified them by the SVM classifier that achieved 83.93% and 73.3% accuracy respectively, on the EMO-DB [27] dataset.

With the development of the technologies, the researchers also applied the new techniques for the feature selection and the extraction. For example, Shegokar and Sircar [42] used the continuous wavelet transform (CWT) method for the RAVDESS [19] dataset in order to select the features and then classify them accordingly by the SVM classifier. In Reference [43], the authors proposed a multi-task learning model for the emotion recognition using the RAVDESS [19] speech dataset, and they achieved a 57.14% recognition rate for four classes, which included anger, neutral emotions, sadness, and happiness. Similarly, Zeng et al. [44] used the same learning strategy and dataset in order to enhance the recognition rate by utilizing deep learning and the multi-task residual network, and in Reference [45], they fine-tuned a pre-trained VGG-16 [46] model for the same task in order to classify the speech spectrograms. In the above literature, the researchers have developed many techniques in order to improve the prediction performance of the SER, but the recognition rates are still low. In this study, we developed an efficient and a significant SER system using hierarchical ConvLSTM blocks in order to extract the hidden emotional cues and then feed into a sequential model for the extraction of the temporal features. Finally, we produced the probabilities of the classes by using the center loss and the softmax loss, which ensure and increase the final classification results. A detailed description of the proposed framework is demonstrated in the upcoming part.

3. The Proposed SER Framework

In this section, we explain the detailed explanation of the proposed speech emotion recognition system and its main components, such as the ConvLSTM, sequential learning, and the center loss function. In contrast, the emotion recognition problem is treated as a multi-class classification based on the spatiotemporal and the sequential learning models. We designed a hierarchical ConvLSTM model with the GRUs and the center loss function in order to recognize the local and the global features in the speech signals using the raw speech segments. Our main framework is illustrated in Figure 1. The proposed model architecture consists of three modules that include the local features learning blocks (LFLBs) that use the ConvLSTM, the global features learning block (GFLB) that uses the GRUs, and the multi-class classification that uses the center and the softmax losses with a pre-processing step. In this step, we normalized the speech single by utilizing the root mean square (RMS) function, which is describe in Equation (1):

$$\mathrm{R} = \sqrt{\frac{1}{\mathrm{n}}\left[(\mathrm{fs}_1)^2 + (\mathrm{fs}_1)^2 + \ldots (\mathrm{fs}_\mathrm{n})^2\right]} \tag{1}$$

where "R" represents the desired results of RMS and "f" is a scaling factor of signal "s" to perform a linear gain change in the amplitude of signals. This is a common method for audio signal processing to scale the data according to the decidable corpus. Further, we designed four LFLBs, and each block has one ConvLSTM layer, one BN layer, and a pooling layer in order to extract the hidden emotional cues in the hierarchical manners. We connected the LFLBs in the hierarchical mode with each other in order to find the input-to-state and the state-to-state correlation among the speech segments. The ConvLSTM layer is utilized for the hidden step-by-step prediction in order to optimize the sequence and to keep the sequential information in the internal state in order to find the spatiotemporal correlation between the speech segments. Furthermore, the global weights are adjusted in order to extracted information by utilizing the GRUs network to find the long-term contextual dependencies in the speech segments and recognize them accordingly. Finally, we computed the probabilities for the final classification by utilizing the center and the softmax losses. A detailed explanation of the related components, which include the ConvLSTM, the GRUs, and the center loss, is illustrated in the subsequent sections.

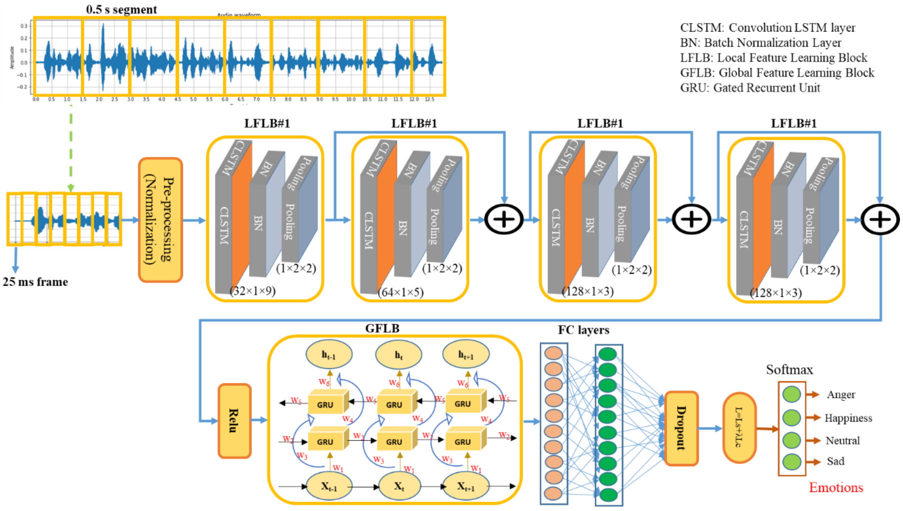


Figure 1. An overall overview of the proposed architecture, which consists of mainly three parts that include the local feature learning blocks (LFLBs) that uses the convolutional long short term memory (ConvLSTM) for the hidden emotional spatiotemporal features extraction, and the global feature learning blocks (GFLBs) that uses the stacked gated recurrent units (GRUs) network to readjust the global weights and find the long-term contextual dependencies in the speech segments. Finally, a fully connected (FC) network recognizes the emotions and classifies them accordingly with the help of the center and the softmax losses.

3.1. ConvLSTM in the SER Model

The emotion recognition in the speech segments is a sequential task that has a strong correlation among the consecutive speech segments. The proposed SER model recognizes the emotions in a raw speech signal by processing the one-by-one segments and catching the correlation in the process of the sequential prediction. Even though the LSTM [47] is known as a better network to use to perform the sequential modeling tasks, the general LSTM network ignores the spatial cues in the input data during the processing. Thus, it models the sequential cues into a one-dimensional vector that leads to the loss of the spatial cues. Therefore, it is not supportive to improve the performance of the recognition in the speech data. For the emotion recognition in the speech segments, the spatial cues will be very useful to improve the model performance. The CNN network maintains and finds the only spatial structure in the speech data, while it ignores the sequential cues during the initial processing. To maintain both the spatial and the temporal cues, such as the spatiotemporal information in the consecutive speech segments, the convolution LSTM (ConvLSTM) [48] is utilized in this framework instead of a standard CNN network. The convolution operation is used to perform the input-to-state and the state-to-state transition in the ConvLSTM. The ConvLSTM captures the spatiotemporal cues during the convolution operation to better mine the correlation among the speech segments. The ConvLSTM computes the weight by using the following Equations (2)–(7).

$$i_t = \sigma(w_{ix} * x_t + w_{ih} * h_{t-1} + w_{ic}(c)c_{t-1} + b_i) \tag{2}$$

$$f_t = \sigma\left(w_{fx} * x_t + w_{fh} * h_{t-1} + w_{fc}(c)c_{t-1} + b_f\right) \tag{3}$$

$$o_t = \sigma(w_{ox} * x_t + w_{oh} * h_{t-1} + w_{oc}(c)c_t + b_0) \tag{4}$$

$$g_t = tanh\left(w_{gx} * x_t + w_{gh} * h_{t-1} + b_g\right) \tag{5}$$

$$c_t = f_t(c)c_{t-1} + i_t(c)g_t \tag{6}$$

$$h_t = o_t(c)\tanh(c_t) \tag{7}$$

In the above equations, we utilized different symbols in order to represent various functionality during the convolution process, for example, the "σ" represents the sigmoid function, * signifies the convolution operation, the element-wise operation is indicated by (c), and tanh represents the hyperbolic tangent function. Similarly, the different gates, which include the input gate, the forget gate, the output gate, and the input modulation gate, are represented by $\mathrm{i_t}$, $\mathrm{f_t}$, $\mathrm{o_t}$, and $\mathrm{g_t}$ respectively, by applying the tth step of the ConvLSTM, which is represented by the subscript "t" in the equations. Furthermore, the input data is illustrated by $\mathrm{x_t}$, and the cell state and the hidden state are represented by $\mathrm{c_t}$ and $\mathrm{h_t}$, respectively. From the input tensor, the first dimension utilizes the temporal information, and the second-dimension size and the third dimension utilize the spatial information during the convolution operation in the ConvLSTM. The core concept of the ConvLSTM and the standard LSTM network is the same as the output of the previous layer, which is utilized for the input of the next layer. The main difference lies in their strategy. The ConvLSTM added the convolution operation during the state-to-state transition that acquires the spatiotemporal feature. Hence, the ConvLSTM can focus on the key elements in the speech segments that can easily recognize the consecutive sequences of the speech signals and ensure the prediction performance of the SER system. We found a better recognition performance using the ConvLSTM layer, which is shown in Figure 2.

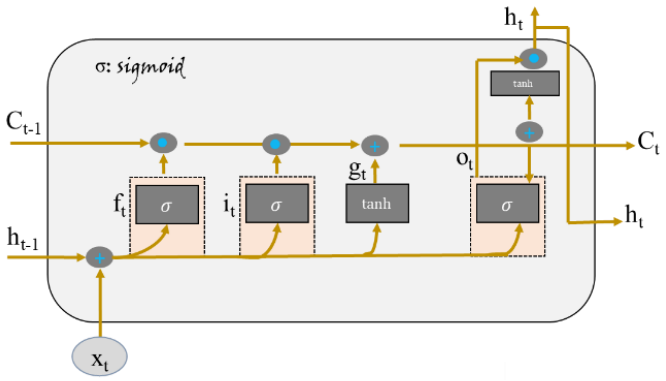


Figure 2. The internal structure of the ConvLSTM architecture.

3.2. The Global Feature Learning Block (GFLB)

In the proposed emotion recognition framework, we adjusted the gated recurrent units (GRUs) in order to learn the global cues in the learned features and to recognize the long-term contextual dependencies. We used a stacked GRUs network as a global feature learning block, which is a modified and a simplified module for sequence learning, and it is widely utilized for time series data [49]. The GRUs network is a special version of the LSTM [47], which is quickly spread for sequential learning in partial sequential data. The GRUs module consists of two gates, which include update gates and reset gates, such as the LSTM. The update gate of a GRU acts like the forget and input gate of the LSTM, and the reset gate works the same as an LSTM reset gate in order to retune the units. The internal mechanism of the GRUs module is completely changed from the LSTM module, and it does not utilize a distinct cell of memory for the information modification inside units, such as the LSTM. The creation h_t^j at interval "t" shown between the candidate $\hat{h}_t^j$ and the earlier creation h_{t-1}^j at the direct interpolation, can be computed using Equation (8).

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j\hat{h}_t^j \tag{8}$$

The position of the units apprising and the stimulation in the update gate is denoted by z_t^j, which illustrates their level that is represented by Equation (9).

$$z_t^j = \sigma(W_x x_t + U_z h_{t-1})^j \tag{9}$$

The activation of $\hat{h}_t^j$ is performed by an update gate in order to calculate the activation of the candidates utilizing Equation (10) in the GRUs.

$$\hat{h}_t^j = \tan h(Wx_t + U(r_t{*}h_{t-1}))^j \tag{10}$$

In Equation (10), the reset gate is represented by r_t^j, and the element-wise multiplication is represented by (*). When the gate is off, the value must be equal to zero, such as ($z_t^j = 0$), then the reset gate will agree with the unit in order to forget the previous information. This is the mechanism that is used to inform the unit to search the head sign of an input sequence. We can compute the reset gate status by applying Equation (11).

$$r_t^j = \sigma\left(W_r x_t + r^{h_{t-1}}\right)^j \tag{11}$$

The short-term dependencies are activated by the reset gate, and the previous state of a sequence is controlled by the update gate that is represented by "r" and "z", respectively. The update gate in the GRUs network is also responsible for controlling the long-term contextual information. In contrast, we utilized a stacked bidirectional gated recurrent unit (BiGRU) network in order to re-adjust the global weights, which is shown in Figure 3.

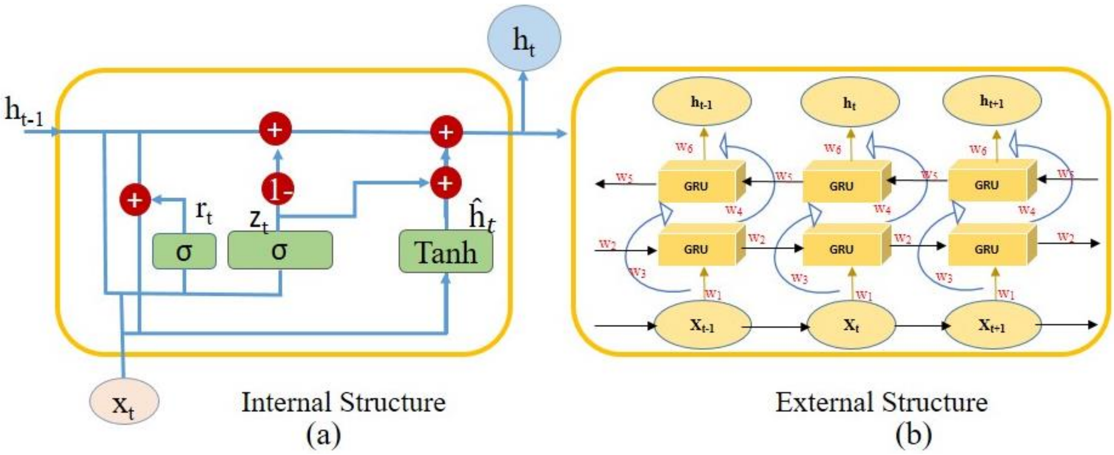


Figure 3. The internal and the external mechanism of the GRU, which shows (**a**) the internal structure with the gates and (**b**) the external flow of the information using different layers.

3.3. Center Loss Function

For a better model performance, we calculated the center loss of the deep features in conjunction with the softmax loss for the final classification of the speech emotions. Both losses collectively produce the final probabilities of the classes that ensure a high prediction performance. The model prediction performance with the softmax loss is a bit lower than the fused losses due to their large distances within the classes. In this framework, we utilize the mean "λ" setting for an update loss function that calculates and adjusts the inter/intra-class distance. The suggested function calculates the loss between the deep features and their consistent class center [50]. We calculated the minimum distance within the classes using the center loss function and the maximum distance among the classes by the softmax loss function, which is shown in Equations (12) and (13).

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{w_{yi}^{T}.x_i + b_{yi}}}{\sum_{j=1}^{n} e^{w_{yi}^{T}.x_i + b_{yi}}} \tag{12}$$

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{yi}\|_2^2 \tag{13}$$

We represent the number of classes and the minimum batch size by "n" and "m" to the classifier, and c_{yi} denotes the center of the class y_i in the ith sample. We used the "λ" sign for the center loss to calculate the minimum distance needed to avoid misclassifications in real-time scenarios. The center loss with softmax loss function achieved a good result for the emotion classification, which is shown in Equation (14).

$$L = L_S + \lambda L_c \tag{14}$$

where "L" represents the final fused loss function. The significance and the effectiveness of the center loss function utilization have been proven by experiments. The "λ" is used as a hyper parameter.

3.4. Our Model Configuration

We used a scikit-learn python library for machine learning with additional connected supportive libraries to implement the proposed framework and to ensure their prediction performance. We utilized the raw audio speech signals for the model training. We converted the raw audio file into different segments with respect to time, and we inputted a one-by-one speech segment into the proposed SER model in order to extract the high-level discriminative features and to recognize them accordingly. For the training and validation purpose of the proposed AI model, we split the entire dataset into an 80:20% ratio, and 80% of the data was utilized for the training and 20% was utilized for the testing or the validating. For the model generalization, we use a k-fold cross-validation technique [51] to investigate the robustness and significance of the system. In the k-fold cross-validation technique, we utilized an automatic technique to split the database in five-fold and to consequently train the model. We utilized an eight gigabytes GTX 1070 GeForce NVIDIA GPU machine for the proposed model preparation and for the analysis process during the 5-fold cross-validation. We trained our model for 100 epochs and selected the decay rate as one after 10 epochs with a 0.0001 learning rate. We tried different batches for the model during the configuration and finally achieved better recognition results on 64, so we selected this one for the entire model training procedure. For the model optimization, we utilized an Adam optimizer, which has secured the best accuracy with a 0.346 training loss and a 0.57 testing loss for the speech emotion recognition. The architecture of the proposed framework took the raw speech signal as an input and recognized them accordingly without any preprocessing or converting to other forms. This behavior of the model indicates the applicability and the capacity for real-time emotion recognition applications.

4. Experimental Evaluation and Discussion

In this section, we experimentally proved our proposed recognition approach for speech emotions using the speech signals and experimentally assessed them utilizing various matrices of accuracy, such as weighted and un-weighted accuracy, precision, recall, and F1_score. We evaluated the proposed SER approach over different benchmarks, which included the IEMOCAP [18] and the RAVDESS [19] speech corpora. We deeply assessed the prediction performance of the proposed approach with different accuracies, matrices that included class-wise balance accuracies, overall accuracy, and a confusion matrix among the actual emotions and the predicted emotions. The performance of the proposed emotion recognition system is compared with the state-of-the-art systems. The model performance, the training, the validation accuracies, and losses graph of both datasets are presented in Figure 4. The detailed investigational consequences with the discussion and the comparisons with baseline state-of-the-art methods of each dataset are explained in separate sections.

In this study, we tried different types of AI and deep learning architecture with a sequence learning module and without a sequential module in order to select an efficient approach, which can easily recognize the depth of emotion in the speech signals. After extensive experimentations, we proposed this framework for the SER, which ensures a high-level performance with a better prediction rate using the different classes. For the convenience of the readers, we conducted an ablation study in order to select the best model and investigate further with a different number of emotions. A detailed description and results for the different AI and the deep learning architecture and their results in the form of un-weighted accuracy are illustrated in Table 1, which utilize the IEMOCAP [18] and the RAVDESS [19] emotional speech corpora.

Table 1 presents the recognition results of the different deep learning architectures utilizing the ConvLSTM layer using speech signals. In the ablation study, we proposed the best model or architecture for a selected task in order to process for further investigation. In this study, we can see the different results in order to evaluate our data, and we selected the best model, and we further processed them. First, we utilized only the ConvLSTM with a fully connected layer and recognized the emotions using softmax. The results of this model were not convincing, so we then applied a residual learning strategy with a skip connection and the result slightly increased. In skip connection, we concatenate the early layer or block output with later layer or block by utilizing an addition of essential connections. In this strategy, for essential connection, we made a straight up connection between early and later layers or blocks. Furthermore, we adjusted the global weight with the ConvLSTM for the long-term contextual dependences in order to efficiently identify the sentiments in long dialogue sequences.

We applied a bi-directional LSTM and a bi-Directional GRUs network in order to re-adjust the global weights with the learned features. The recognition rates were increased with this sequential learning model. We selected the GRUs model for sequential leaning, because it is quite good for partial data and the results of the GRUs were better and more convenient for this task, which is illustrated in Table 1.

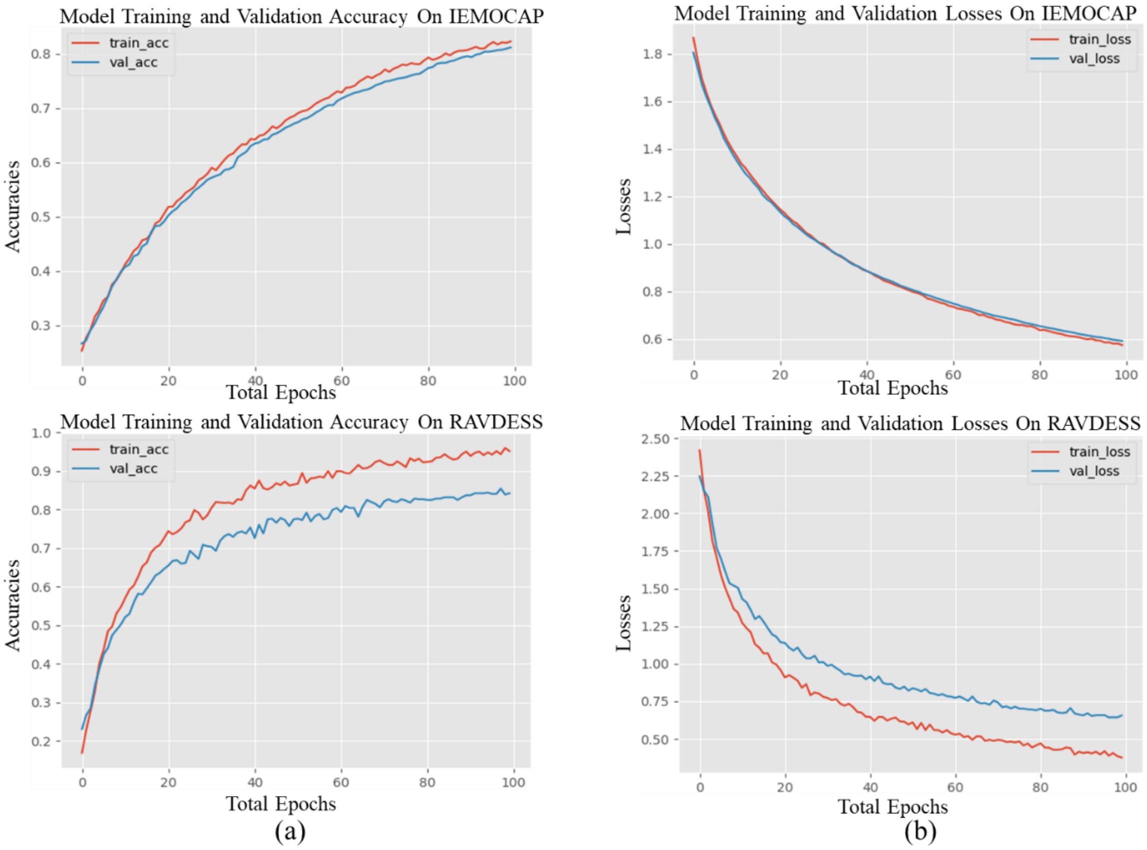


Figure 4. The graphical representation of the proposed model training and the validation performance using interactive emotional dyadic motion capture (IEMOCAP) [18] and ryerson audio visual database of emotional speech and song (RAVDESS) [19] emotional speech datasets. (**a**) indicates the model accuracies, and (**b**) shows the losses for the training and for the validation using the IEMOCAP and RAVDESS speech datasets.

Table 1. The ablation study and a detailed overview of the different deep learning architecture and their recognition results using the IEMOCAP and the RAVDESS emotional speech datasets are presented in this table.

Input	Deep Learning Architectures	IEMOCAP	RAVDESS
Audio clip or raw waveform	ConvLSTM + FCN + Softmax	65.18%	69.21%
	ConvLSTM + Skip connection + FCN + Softmax	66.33%	69.67%
	ConvLSTM + BiLSTM + Softmax	65.44%	69.37%
	ConvLSTM + Skip connection + BiLSTM + Softmax	67.01%	73.29%
	ConvLSTM + BiGRU + Softmax	69.26%	76.67%
	ConvLSTM + Skip connection + BiGRU + Softmax	73.86%	78.66%
	ConvLSTM + Skip connection +BiGRU + Center loss	75.03%	80.05%

4.1. *The Results of the IEMOCAP (Interactive Emotional Dyadic Capture)*

IEMOCAP [18] is an emotional and acted corpus of audio speech data, which has eight different emotions that are collected in five sessions with ten (10) professional actors. Each session used two actors, which included one male actor and one female, in order to record the diverse expressions, such as anger, happiness, sadness, neutral emotions, and surprise. These expert actors recorded 12-hour audiovisual data with a sixteen kHz sampling rate. All the emotions in this dataset are pre-scripted, and the actors just read with different expressions and categorized them in the scripted and the improvised versions. We evaluated our model using four emotions, which are frequently used in the literature, and compared our model with an existing state-of-the-art SER model. The detailed overview of selected four emotions is illustrated in Table 2.

Table 2. The detailed overview of all the utilized emotions of the IEMOCAP [18] dataset with the total numbers of utterances and the percentage of participation of each class.

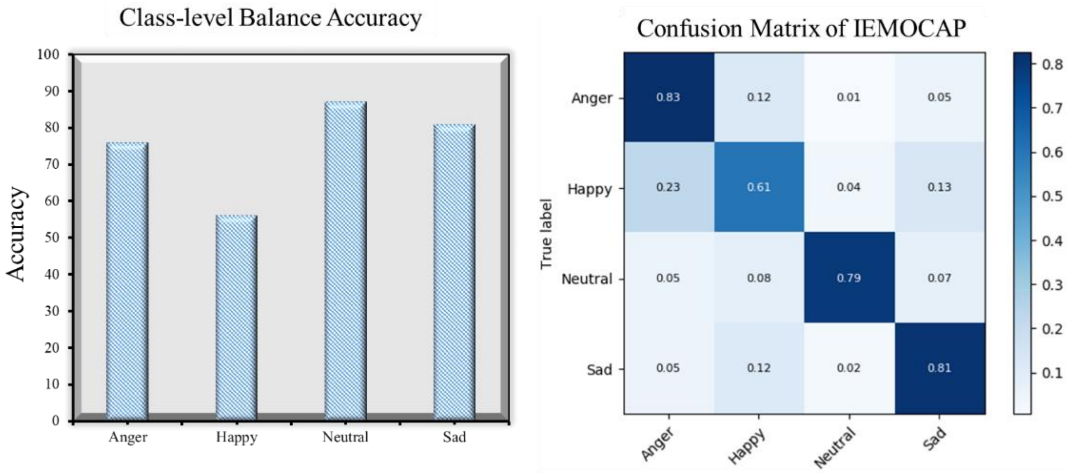
Emotions/Classes	Whole Utterances	Participation
Anger	1103	19.94%
Sadness	1084	19.60%
Happy	1636	29.58%
Neutral	1708	30.88%

We conducted extensive experiments on the IEMOCAP corpus utilizing the five-fold cross-validation technique in order to evaluate the proposed system, and tested the prediction performance on the unseen data. We are using 20% of the data for the model testing or validation and 80% of the data for the model training in each fold. For the model prediction performance, we used different evaluation matrices, which included the weighted, un-weighted, F1_score, class-wise balance accuracy, and the confusion matrix, in order to show the strength of the proposed system for the selected task. The class-wise results represent the balance accuracy in each class of recognition—how the model accurately recognizes each class. The weighted and un-weighted accuracy shows the actual and the predicted ratio among the dataset and the classes. Similarly, the F1_score shows the balance among the precision and the recall or the harmonic mean between these in each class. Finally, we constructed the confusion matrix in order to show the confusion among the actual emotions and the predicted emotions for further investigation. In the confusion matrix, the diagonal values show the actual prediction performance, and the confusion among other classes are shown in the corresponding rows. The classification report, the confusion matrix, and the class-wise balance accuracy of the IEMOCAP dataset are illustrated in Table 3 and Figure 5.

IEMOCAP [18] is a challenging dataset in the field of speech emotion recognition, which has different emotions, and most emotions overlap, so the recognition rate is therefore very low according to the literature. Our system shows a significant improvement with the recognition accuracy of each class as well as with the overall model prediction performance. The system recognized anger, sadness, and neutral emotions with more than a 75% rate and happiness with a 61% rate, which is better than the state-of-the-art methods. Figure 5a shows the class-wise balance recognition accuracy of the proposed system over the IEMOCAP speech corpus. The x-axis direction indicates the whole classes, and the y-axis shows each class' recognition results. Similarly, Figure 5b shows the confusion matrix of the IEMOCAP corpus, which illustrates the confusion between the actual emotions and the predicted emotions. The x-axis direction shows the actual label, and the y-axis direction shows the predicted labels of the system. The recall values represent the actual recognition rate in the confusion matrix, which are shown diagonally. The comparative analysis of the proposed SER system is represents in Table 4.

Table 3. The overall performance over the IEMOCAP corpus during prediction or testing of the system is presented, that shows the classification report including the F1_score accuracy matrixes that show the strength of the model, and the overall prediction accuracy is illustrated using percentages.

Emotion	Precision	Recall	F1 Score
Anger	0.70	0.83	0.76
Happiness	0.51	0.61	0.56
Neutral	0.96	0.79	0.87
Sadness	0.81	0.81	0.81
Weighted	0.77	0.78	0.77
Un-weighted	0.76	0.74	0.75
Accuracy		**78%**	



(a) Classes (b) Predicted label

Figure 5. The class-wise balance accuracy of the proposed system is shown in (**a**) the confusion matrix among the actual emotions and the (**b**) predicted emotions during the model testing illustrated using the IEMOCAP speech corpora.

Table 4. A comparative analysis of the proposed speech emotion recognition method with the baseline, state-of-the-art speech emotion recognition (SER) method over the IEMOCAP emotional speech corpora.

Dataset	Reference	Year	Un-Weighted Accuracy
IEMOCAP	[22]	2019	52.14%
//	[52]	2017	64.78%
//	[53]	2019	57.10%
//	[54]	2015	40.02%
//	[55]	2014	51.24%
//	[56]	2019	69.32%
//	[57]	2019	66.50%
//	[58]	2018	63.98%
//	[59]	2019	61.60%
//	[50]	2018	64.74%
//	[60]	2020	64.03%
//	[2]	2020	71.25%
//	[61]	2020	73.09%
Proposed	[Ref#]	2020	**75.00%**

The confusion rate among the other classes is shown in the corresponding rows of each emotion in the confusion matrix. Utilizing this dataset, we compared our proposed system with the different CNN and classical models, and it showed results that outperformed, which are shown in Table 4.

Table 4 represents the comparative analysis of the proposed system over the baseline method [61]. In the recent literature, 71.25% reported high accuracy using the 2D-CNN approach, and the current result of the 1D-CNN approach is 52.14% according to Reference [22]. Our proposed system increased the level of accuracy by 3.75% from the recent deep learning approach and more than 20% from the 1D baseline model [22], because the existing model used low-level features, and our system hierarchically learned high-level discriminative spatiotemporal features. The high recognition rates in 2019 were 66% and 67% using the IEMOCAP emotional speech dataset, and in 2020, it was 71.25%, which is still lower and needs to improve for real-time applications. In contrast, we focused on accuracy and developed a novel 1D deep learning model that uses ConvLSTM and the GRUs to acquire high discriminative features from the raw speech signals and recognize them consequently. Due to this significant improvement, we claim that our system is a recent success of deep learning and the most suitable for real-world problem monitoring.

4.2. Results of the RAVDESS (Ryerson Audiovisual Database of Emotional Speech and Song)

RAVDESS [19] is an acted British linguistic emotional speech database, which is broadly utilized in the emotion recognition systems in speech and songs. The RAVDESS dataset consists of 24 professional actors, which includes 12 male actors and 12 female actors, to record the pre-scripted text using different emotions. The suggested dataset has eight (8) different types of speech emotions, which includes anger, fear, sadness, and happiness. In this dataset, the number of audio files is similar in all the classes except for neutral, which has less audio files than the other classes. All the emotions were recorded using a 48 kHz sampling rate, and the average length of an audio is 3.5 seconds. In contrast, we used this dataset for our proposed model testing and training in order to estimate the prediction concert. A detailed description of the suggested dataset is illustrated in Table 5.

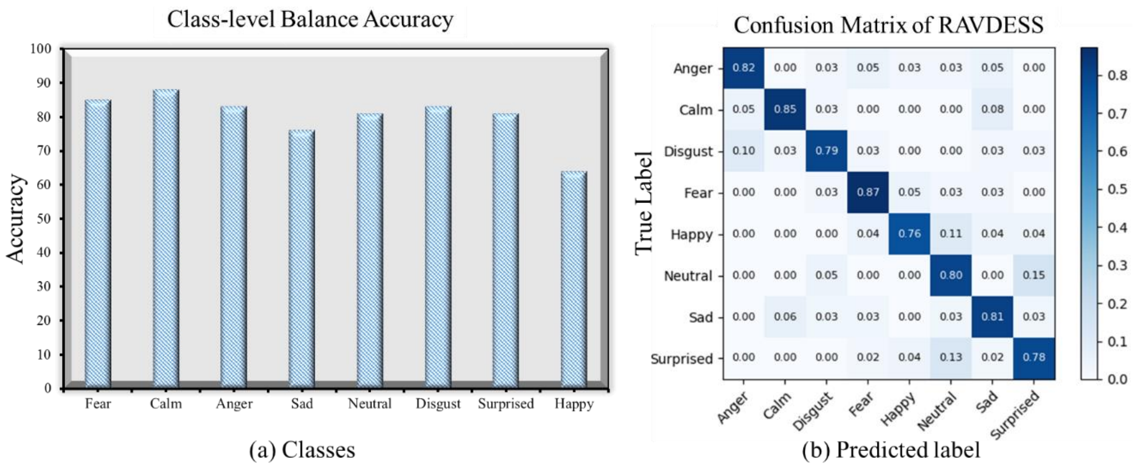
Nowadays, the RAVDESS [19] dataset is widely used for the emotional speech and song recognition systems in order to check the significance and the effectiveness. In contrast, we utilized this dataset and evaluated our proposed system using a 5-fold cross-validation technique. The prediction performance was evaluated through different accuracy matrices, such as a classification report, class-wise accuracy, and a confusion matrix. The classification report shows each category precision, recall, and F1_score, and the class-wise balance accuracy represents the actual prediction accuracy of each class using the suggested dataset. Similarly, we constructed the confusion matrix of the RAVDESS dataset in order to deeply analyze the performance of the actual and the predicted emotions for further investigation. For the model comparison, we used the weighted and un-weighted precision matrix that is frequently utilized in the related works for the comparative analysis. A detailed description and a visual evaluation, which includes the classification report, class-wise accuracy, and the confusion matrix of the proposed SER system, are illustrated in Table 6 and Figure 6 using the RAVDESS emotional speech dataset.

Table 5. A detailed overview of all the utilized emotions of the RAVDESS [19] dataset with the total numbers of utterances and the percentage of the participation of each class.

Emotions/Classes	Whole Utterances	Participation
Anger	192	13.33%
Fear	192	13.33%
Happy	192	13.33%
Sad	192	13.33%
Neutral	96	6.667%
Surprise	192	13.33%
Calm	192	13.33%
Disgust	192	13.33%

Table 6. The overall performance over RAVDESS corpus during prediction or testing of the system is presented, that shows the classification report including the F1_score accuracy matrices that show the strength of the model, and the overall prediction accuracy is illustrated using percentages.

Emotions	**Precision**	**Recall**	**F1_score**
Anger	0.84	0.82	0.83
Calm	0.92	0.85	0.88
Disgust	0.86	0.79	0.83
Fearful	0.83	0.87	0.85
Happy	0.88	0.76	0.81
Neutral	0.53	0.80	0.64
Sad	0.72	0.81	0.76
Surprise	0.84	0.78	0.81
Weighted	0.83	0.82	0.81
Un-weighted	0.80	0.81	0.80
Accuracy		**82%**	



	Anger	Calm	Disgust	Fear	Happy	Neutral	Sad	Surprised
Anger	0.82	0.00	0.03	0.05	0.03	0.03	0.05	0.00
Calm	0.05	0.85	0.03	0.00	0.00	0.00	0.08	0.00
Disgust	0.10	0.03	0.79	0.03	0.00	0.00	0.03	0.03
Fear	0.00	0.00	0.03	0.87	0.05	0.03	0.03	0.00
Happy	0.00	0.00	0.00	0.04	0.76	0.11	0.04	0.04
Neutral	0.00	0.00	0.05	0.00	0.00	0.80	0.00	0.15
Sad	0.00	0.06	0.03	0.03	0.00	0.03	0.81	0.03
Surprised	0.00	0.00	0.00	0.02	0.04	0.13	0.02	0.78

(a) Classes

(b) Predicted label

Figure 6. The class-wise balance accuracy of the proposed system is shown in (**a**) a confusion matrix among the actual emotions and the (**b**) predicted emotions during the model testing illustrated using the RAVDESS [19] speech corpora.

RAVDESS [19] is a new speech dataset that is used the most in order to recognize emotions in speech and songs. The level of accuracy of the speech emotion recognition system on this data is still low, and the researchers have developed various techniques in order to increase the accuracy for commercial use. In contrast, we targeted the accuracy of the existing system and proposed a novel deep learning architecture that utilizes ConvLSTM and a bi-directional GRUs network. Our system extracts the deep local and the global features from the raw audio signals through a hierarchal manner that can easily improve the accuracy and ensure the prediction performance. With the use of this architecture, we increased the level of accuracy, which shows the significance of the proposed SER system. A detailed classification report of the system using the RAVDESS corpus is shown in Table 6, which indicated the precision, the recall, the weighted, and the un-weighted accuracy of each class. Furthermore, the class-wise balance accuracy of the system is presented in Figure 6a, which shows each class recognition rate. The x-axis shows the class label, and the y-axis shows each class recognition rate. Similarly, Figure 6b illustrates the confusion matrix, which shows the confusion between the actual emotions and the predicted emotions, for example, the system predicts how many emotions are correctly and incorrectly predicted in each class. The actual predicted emotions are represented diagonally in the confusion matrix, and the incorrectly predicted emotions are shown in the corresponding rows of each class. The overall performance of the proposed system is better than the state-of-the-art methods. We compared our system with different baseline state-of-the-art SER systems that are based on the traditional and the deep learning approaches in order to display the effectiveness and the strength of the proposed method over these baseline system. The comparative analysis of the proposed SER model on the RAVDESS dataset is illustrated in Table 7.

Table 7. A comparative analysis of the proposed speech emotion recognition method with the baseline, state-of-the-art SER method over the RAVDESS emotional speech corpora.

Dataset	Reference	Year	Un-Weighted Accuracy
RAVDESS	[44]	2019	64.48%
//	[62]	2019	69.40%
//	[63]	2019	75.79%
//	[64]	2019	67.14%
//	[60]	2020	71.61%
//	[4]	2020	79.01%
//	[2]	2020	77.01%
Proposed	[Ref#]	2020	**80.00%**

Table 7 represents the recent statistics of the existing speech emotion recognition model and the proposed SER model using the RAVDESS emotional speech dataset as well as their comparison. The researchers used different deep learning and classical approaches to create an efficient SER system, which accurately predicts different emotions. In the last year, the highest recognition rate that was reported in 2019 for RAVDESS corpus was 75.79%, and in the most recent year, 2020, the highest recorded rate was 79.01%, which still needs further improvement. Due to this limitation, we proposed a unique deep learning methodology in this study and increased the level of accuracy up to 80% using a one-dimensional CNN model. Our system recognized emotions from raw speech signals and classified them accordingly, with high accuracy. Our system is the most suitable for industrial applications, because we used a 1D network to recognize the emotional state of an individual's speech patterns while speaking, which does not need to convert the speech signals into other forms, such as speech spectrograms and log Mel spectrograms. Due to this pretty structure, our proposed system outperformed the results of IEMOCAP and RAVDESS speech corpora, which proved its strength to monitor real-time emotions for real-world problems. According to the best of our knowledge, this is a recent success of deep learning that utilizes a one-dimensional CNN strategy with the ConvLSTM layer in the speech recognition domain.

5. Conclusions and Future Direction

In this study, we developed an end-to-end artificial intelligence (AI) based emotion recognition model using one-dimensional CNN strategy by utilizing ConvLSTM in order to extract the spatiotemporal cues of the speech signals in order to improve the prediction performance of the speech emotion recognition (SER) system. In this approach, we used four ConvLSTM blocks, which are called the local features learning blocks (LFLBs), in order to learn the spatial semantic correlation of the emotions in a hierarchical manner by utilizing a residual learning strategy. We extracted the most salient discriminative emotional features using these blocks. After that, we fed these extracted features into a stacked GRUs network in order to re-adjust the global weight with these learned features. We further processed the extracted refined spatiotemporal features using a fully connected network and fused softmax in order to produce the probabilities of the classes. We conducted extensive experiments over IEMOCAP [18] and RAVDESS [19] speech corpora in order to check and evaluate the effectiveness and the significance of the proposed model over the state-of-the-art models. Our proposed system showed outperformed results and obtained a 75% recognition accuracy and an 80% recognition accuracy for IEMOCAP and REVDESS, respectively which is clearly indicated the robustness of the system.

In the future, we plan to design a multi-model speech emotion recognition system that utilizes the ConvLSTM layer with novel architecture or structures. In this paper, we designed and introduced the ConvLSTM network for speech signals, which can be easily applied to other speech processing-related problems, such as the automatic speaker recognition system and energy forecasting [65]. The proposed framework can be easily adopted in other speech classification tasks in order to find a mutual understanding, and it will be further extended for more efficient SER systems.

Author Contributions: Conceptualization, M. and S.K.; methodology, M. and S.K.; software, M.; validation, M. and S.K.; formal analysis, M.; investigation, M.; data curation, M.; writing—original draft preparation, M.; visualization, M.; resources, S.K.; writing—review and editing, S.K.; supervision, S.K.; project administration, S.K.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: 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 year (2020) Faculty Research Fund of Sejong University.

Conflicts of Interest: The authors declare that there are no conflict of interest.

References

1. Kim, J.-Y.; Cho, S.-B. Towards Repayment Prediction in Peer-to-Peer Social Lending Using Deep Learning. *Mathematics* **2019**, *7*, 1041. [CrossRef]
2. Sajjad, M.; Kwon, S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
3. Lin, Y.-C.; Wang, Y.-C.; Chen, T.-C.T.; Lin, H.-F. Evaluating the suitability of a smart technology application for fall detection using a fuzzy collaborative intelligence approach. *Mathematics* **2019**, *7*, 1097. [CrossRef]
4. Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2020**, *20*, 183.
5. Baioletti, M.; Di Bari, G.; Milani, A.; Poggioni, V. Differential Evolution for Neural Networks Optimization. *Mathematics* **2020**, *8*, 69. [CrossRef]
6. Anvarjon, T.; Kwon, S. Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. *Sensors* **2020**, *20*, 5212. [CrossRef]
7. Das Antar, A.; Ahmed, M.; Ahad, A.R. Challenges in Sensor-based Human Activity Recognition and a Comparative Analysis of Benchmark Datasets: A Review. In Proceedings of the 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Spokane, WA, USA, 30 May–2 June 2019; pp. 134–139.
8. Khalil, R.A.; 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. [CrossRef]
9. Pandey, S.K.; Shekhawat, H.S.; Prasanna, S.R.M. Deep Learning Techniques for Speech Emotion Recognition: A Review. In Proceedings of the 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019; pp. 1–6.
10. Ji, S.; Kim, J.; Im, H. A Comparative Study of Bitcoin Price Prediction Using Deep Learning. *Mathematics* **2019**, *7*, 898. [CrossRef]
11. Khan, N.; Ullah, A.; Haq, I.U.; Menon, V.G.; Baik, S.W. SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network. *J. Real Time Image Process.* **2020**, 1–15. [CrossRef]
12. Jara-Vera, V.; Sánchez-Ávila, C. Cryptobiometrics for the Generation of Cancellable Symmetric and Asymmetric Ciphers with Perfect Secrecy. *Mathematics* **2020**, *8*, 1536. [CrossRef]
13. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
14. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors* **2017**, *17*, 1694. [CrossRef] [PubMed]
15. Ullah, W.; Ullah, A.; Haq, I.U.; Muhammad, K.; Sajjad, M.; Baik, S.W. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimed. Tools Appl.* **2020**, 1–17. [CrossRef]
16. Zhang, J.; Jiang, X.; Chen, X.; Li, X.; Guo, N.; Cui, L. Wind Power Generation Prediction Based on LSTM. In Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence—ICMAI 2019, Chegndu, China, 12–15 April 2019; pp. 85–89.
17. Kurpukdee, N.; Koriyama, T.; Kobayashi, T.; Kasuriya, S.; Wutiwiwatchai, C.; Lamsrichan, P. Speech emotion recognition using convolutional long short-term memory neural network and support vector machines.

In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1744–1749.

18. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
19. 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]
20. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018. [CrossRef]
21. Liu, B.; Qin, H.; Gong, Y.; Ge, W.; Xia, M.; Shi, L. EERA-ASR: An Energy-Efficient Reconfigurable Architecture for Automatic Speech Recognition With Hybrid DNN and Approximate Computing. *IEEE Access* **2018**, *6*, 52227–52237. [CrossRef]
22. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323. [CrossRef]
23. Yu, Y.; Kim, Y.-J. Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database. *Electronics* **2020**, *9*, 713. [CrossRef]
24. Eeyben, F.; Scherer, K.R.; Schuller, B.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
25. Triantafyllopoulos, A.; Keren, G.; Wagner, J.; Steiner, I.; Schuller, B.W. Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019. [CrossRef]
26. Schuller, B.; Steidl, S.; Batliner, A.; Hirschberg, J.; Burgoon, J.K.; Baird, A.; Elkins, A.; Zhang, Y.; Coutinho, E.; Evanini, K. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; Volumes 1–5, pp. 2001–2005.
27. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
28. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and Recurrent Neural Networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 13–16 December 2016; pp. 1–4.
29. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Muhammad, K. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589. [CrossRef]
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
31. Osia, S.A.; Shamsabadi, A.S.; Sajadmanesh, S.; Taheri, A.; Katevas, K.; Rabiee, H.R.; Lane, N.D.; Haddadi, H. A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics. *IEEE Internet Things J.* **2020**, *7*, 4505–4518. [CrossRef]
32. Carta, S.M.; Corriga, A.; Ferreira, A.; Podda, A.S.; Recupero, D.R. A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning. *Appl. Intell.* **2020**, 1–17. [CrossRef]
33. Carta, S.M.; Ferreira, A.; Podda, A.S.; Recupero, D.R.; Sanna, A. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting. *Expert Syst. Appl.* **2020**, *164*, 113820. [CrossRef]
34. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019. [CrossRef]
35. Bao, F.; Neumann, M.; Vu, N.T. CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 35–37.

36. Fahad, M.; Yadav, J.; Pradhan, G.; Deepak, A. DNN-HMM based Speaker Adaptive Emotion Recognition using Proposed Epoch and MFCC Features. *arXiv* **2018**, arXiv:1806.00984.
37. Kourbatov, A.; Wolf, M. Predicting maximal gaps in sets of primes. *Mathematics* **2019**, *7*, 400. [CrossRef]
38. Demircan, S.; Kahramanli, H. Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. *Neural Comput. Appl.* **2018**, *29*, 59–66. [CrossRef]
39. Wu, X.; Liu, S.; Cao, Y.; Li, X.; Yu, J.; Dai, D.; Ma, X.; Hu, S.; Wu, Z.; Liu, X.; et al. Speech Emotion Recognition Using Capsule Networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6695–6699.
40. Jukic, S.; Saračević, M.; Subasi, A.; Kevric, J. Comparison of Ensemble Machine Learning Methods for Automated Classification of Focal and Non-Focal Epileptic EEG Signals. *Mathematics* **2020**, *8*, 1481. [CrossRef]
41. Ahmad, J.; Sajjad, M.; Rho, S.; Kwon, S.; Lee, M.Y.; Baik, S.W. Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. *Multimed. Tools Appl.* **2016**, *77*, 4883–4907. [CrossRef]
42. Shegokar, P.; Sircar, P. Continuous wavelet transform based speech emotion recognition. In Proceedings of the 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), Gold Coast, Australia, 19–21 December 2016; pp. 1–8.
43. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019. [CrossRef]
44. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. *Multimed. Tools Appl.* **2019**, *78*, 3705–3722. [CrossRef]
45. Popova, A.S.; Rassadin, A.G.; Ponomarenko, A.A. Emotion Recognition in Sound. In Proceedings of the International Conference on Neuroinformatics, Moskow, Russia, 2–6 October 2017; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2017; Volume 736, pp. 117–124.
46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
47. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
48. Zapata-Impata, B.S.; Gil, P.; Torres, F. Learning Spatio Temporal Tactile Features with a ConvLSTM for the Direction Of Slip Detection. *Sensors* **2019**, *19*, 523. [CrossRef] [PubMed]
49. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G.W. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2627–2633.
50. 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]
51. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2011**, *21*, 137–146. [CrossRef]
52. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
53. Guo, L.; Wang, L.; Dang, J.; Liu, Z.; Guan, H. Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine. *IEEE Access* **2019**, *7*, 75798–75809. [CrossRef]
54. Zheng, W.Q.; Yu, J.S.; Zou, Y.X. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 827–831.
55. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Fifteenth Annual Conference of The International Speech Communication Association, Singapore, 14–18 September 2014.
56. 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]
57. Zhao, Z.; Bao, Z.; Zhao, Y.; Zhang, Z.; Cummins, N.; Ren, Z.; Schuller, B. Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access* **2019**, *7*, 97515–97525. [CrossRef]

58. Luo, D.; Zou, Y.; Huang, D. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018.
59. Jiang, S.; Zhou, P.; Li, Z.; Li, M. Memento: An Emotion Driven Lifelogging System with Wearables. In Proceedings of the 2017 26th International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada, 31 July–3 August 2017; Volume 15, pp. 1–9. [CrossRef]
60. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [CrossRef]
61. Mustaqeem, S.K. MLT-DNet: Speech Emotion Recognition Using 1D Dilated CNN Based on Multi-Learning Trick Approach. *Expert Syst. Appl.* **2020**, 114177. [CrossRef]
62. Jalal, A.; Loweimi, E.; Moore, R.K.; Hain, T. Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 1701–1705. [CrossRef]
63. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl. Based Syst.* **2019**, *184*, 104886. [CrossRef]
64. Zamil, A.A.A.; Hasan, S.; Baki, S.M.J.; Adam, J.M.; Zaman, I. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames. In Proceedings of the 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 10–12 January 2019; pp. 281–285.
65. Khan, Z.A.; Hussain, T.; Ullah, A.; Rho, S.; Lee, M.; Baik, S.W. Towards Efficient Electricity Forecasting in Residential and Commercial Buildings: A Novel Hybrid CNN with a LSTM-AE based Framework. *Sensors* **2020**, *20*, 1399. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.