Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.037373
Article

TC-Net: A Modest & Lightweight Emotion Recognition System Using Temporal Convolution Network

Muhammad Ishaq[1], Mustaqeem Khan[1,2] and Soonil Kwon[1,*]

[1]Sejong University Software Convergence Department, Seoul, 05006, Korea
[2]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, 3838-111188, United Arab Emirates
*Corresponding Author: Soonil Kwon. Email: skwon@sejong.edu
Received: 01 November 2022; Accepted: 09 February 2023

Abstract: Speech signals play an essential role in communication and provide an efficient way to exchange information between humans and machines. Speech Emotion Recognition (SER) is one of the critical sources for human evaluation, which is applicable in many real-world applications such as healthcare, call centers, robotics, safety, and virtual reality. This work developed a novel TCN-based emotion recognition system using speech signals through a spatial-temporal convolution network to recognize the speaker's emotional state. The authors designed a Temporal Convolutional Network (TCN) core block to recognize long-term dependencies in speech signals and then feed these temporal cues to a dense network to fuse the spatial features and recognize global information for final classification. The proposed network extracts valid sequential cues automatically from speech signals, which performed better than state-of-the-art (SOTA) and traditional machine learning algorithms. Results of the proposed method show a high recognition rate compared with SOTA methods. The final unweighted accuracy of 80.84%, and 92.31%, for interactive emotional dyadic motion captures (IEMOCAP) and berlin emotional dataset (EMO-DB), indicate the robustness and efficiency of the designed model.

Keywords: Affective computing; deep learning; emotion recognition; speech signal; temporal convolutional network

1 Introduction

In digital audio processing, speech emotion recognition (SER) is an emerging research area and the most natural form of communication between humans with computer interaction. SER means understanding an individual's emotional state by detecting discriminative features in a voice. Nowadays, several scientists have utilized different deep learning techniques to improve the SER rate by quality features and reduce the overall model complexity through reliable and lightweight Convolution Neural Networks (CNN's) strategies. Similarly, researchers use a variety of machine learning and deep learning techniques for speech processing to recognize and understand individual speech. In contrast, many studies published in SER use pre-trained and fine-tuned CNNs models to

learn high-level cues. Inspired by these published works and therefore proposed a novel Temporal Convolution Network (TCN) architecture conducted extensive experiments using IEMOCAP and EMO-DB corpora to achieve high recognition accuracy.

Recently, researchers have explained several techniques which have increased the performance of SER in real-time for real-world problems. In this regard, researchers have a one-dimensional Dilated Convolution Neural Network (DCNN) with multi-learning strategies, such as a Residual Block with a Skip Connection (RBSC) and Sequence Learning (Seq_L) [1]. Similarly, the two-stream deep convolutional neural networks have achieved success in SER, where the authors used spectrum and spatial CNN architecture [2]. Some researchers have worked with other deep learning networks and technologies to improve the accuracy, efficiency, and performance of speech processing by 2D-CNN, Long-Short Term Memory (LSTM), and Recurrent Neural Networks (RNN) [3]. However, some researchers proposed an SER system that increased the prediction performance and improved the recognition rate by using Bi-GRU (Bidirectional Gated Recurrent Unit) with LSTM [4]. Hence, the researchers used many deep learning techniques such as designing a lightweight CNN architecture to adopt better deep frequency features from the speech spectrograms [5]. Similarly, hybrid techniques such as CNN-LSTM, and deep CNN models [6,7], Used attention mechanisms, which have increased the recognition accuracy in many fields by selectively focusing on salient points [8]. The contrast, SER is a challenging field of research, and researchers have proposed different deep networks for better results and feature selection. Therefore, researchers have frequently struggled to achieve high performance, accuracy, and results in speech emotion through 1D Dilated CNN, Deep Neural Networks (DNNs), and LSTM.

Continuously researchers work in different fields to address various limitations and develop networks to find a solution and clarify the efficiency and accuracy of publicly available data. In this regard, SER is a growing and challenging area of research where researchers proposed a lot of techniques, networks, and architecture to improve recognition performance and accuracy [9] by finding an effective feature. Usually, CNN is used as a Parallel Convolutional Neural Network (PCNN) to take out the better time and frequency dimensions of signals by SEnet and Self-attention Modules. Nowadays, some researchers work on deep learning approaches such as 2D-CNN, 1D-CNN, Convolutional LSTM (CLSTM), Deep-Net, Bidirectional (Bi-LSTM), Multi-Attention modules, and Two-Stream deep CNN [10]. All the upper mentioned deep learning modules improved the performance but there are some limitations due to the modules' configuration, time complexity, and pre-processing to prepare the input data. In this paper, authors develop a novel framework, namely, TCN: speech emotion recognition using temporal convolution network, for healthcare centers and cover limitations such as accuracy and cost computation. As a result of the created framework's efficient performance on speech signal identification tasks, the data dimension used in the inference technique is continuously reduced to increase computational efficiency as well as reference response, reference time, and frequency time analysis. It emphasizes the learning of time connections for each signal as well as the extraction of geographical information. The main contributions of this model are as follows:

- Propose a simple TCN: speech emotion recognition system using a temporal convolution network for a speech signal with the importance to recognize the temporal dependencies of each amplitude and enhance the extraction of spatial information.
- The authors introduce a new core block for TCN that takes advantage of temporal dependency extraction for long sequences and integrates it with impenetrable layers to handle the spatial and temporal information in speech signals.

- The authors utilize the inference method to reduce the data dimension, which increases the computational efficiency and reference response. Due to its intrinsic allocation, rationally summing data along with distinct levels could significantly improve the system's functioning and make it more instructive.
- Conducted extensive experimentation by two benchmarks, IEMOCAP and EMO-DB corpora, and secured a high recognition rate, 80.84%, and 92.31%, respectively. The obtained outcomes of the suggested SER techniques show a better generalization due to being evaluated over two different databases. Detailed explanations are mentioned in Section 4.

The remaining paper is divided into the following sections: The literature review of the SER presented in Section 2, the detail of the proposed framework and its main components explained in Section 3, and the experimental setup and the practical results of the proposed SER system illustrated in Section 4. The discussion and a comparative analysis are presented in Section 5, and finally, present conclusions with an outlook of future directions in Section 6.

2 Related Work

Speaker recognition and speech recognition are challenging areas of research in digital audio signal processing where researchers have proposed many techniques to cover and find an optimal solution for identification to recognize speaker and emotion. In the field of emotion recognition, many researchers have established certain well-organized techniques such as [10] introduced a CNN-based method, which can reduce the complexity and improve the performance of the existing SER systems. Similarly, the authors in [11] developed more techniques for SER called; multi-resolution texture image information (MRTII), acoustic activity detection (AAD strategy of BS-Entropy-based to improve and cover the amplitude of the signal for emotion using three public corpora and achieved higher accuracy than baseline models. Furthermore, [12] used stacked autoencoder and Recurrent Neural Network (RNN) methods to improve the recognition rate of the IEMOCAP corpus using spectrogram base representation for emotion. However, [13] investigated sequential learning and designed a new RNN and Bi-LSTM to improve the weighted accuracy. The researchers worked and proposed a deep analysis technique to model different emotions through Mel-frequency cepstral coefficients (MFCCs), 2D-CNN, 3D-CNN, LSTMS, and CNN-LSTM [14]. Furthermore, in [15] the authors developed a Conv-LSTM-RNN network for SER using IEMOCAP and further improved the performance and accuracy of the baseline models. Zhang et al. [16], proposed a deep learning-based approach with dilated convolutional neural layers and targeted only the performance of the model. Similarly [17] developed a deep method based on concatenated CNNs and RNNs and used an artificial neural network by applying an emotional speech database to recognize their emotional state.

Nowadays, scientists are working on an advanced SER system to address the real-time solution for real-world issues in this regard established a Conv-LSTM-based method for emotion recognition. The authors improved the recognition rate and tested the system on raw audio speech through IEMOCAP and Ryerson audio-visual database of emotional speech and song (RAVDESS) corpora. The work in [17] provided a comprehensive review of feature sets, classification algorithms of features, and accurate usage parameters for the SER. The work in [18] studied different models developed a deep learning approach called LSTM-GAN and tested it with EMO-DB and Democratic Electoral Systems (DES) datasets and secured a reasonable recognition rate. The work in [16] published another review article about emotion recognition performance and accuracy improvement by different systems such as belief networks, neural networks, RNNs, and LSTM using speech signals. With multiple features and classification techniques for emotions, Akçay et al. [19] developed a deep algorithm for speech signals.

The work in [20] used Hidden Markov Models (HMM), Gaussian Mixtures Models (GMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and K-nearest neighbor. Similarly, [21] proposed a method that used pre-trained weights to extract the features from the spectrogram and trained a separate classifier for recognition [22]. Due to the usage of high-level approaches and features, the researchers achieved superior performance and high accuracy.

In contrast, researchers developed and utilized many methods for emotion recognition with diverse types of input to extract salient features from speech signals that boost the recognition accuracy and reduced the computations. Accordingly, Zhang et al. [23] used a multi-task residual network as a learning strategy and enhanced the recognition rate of emotion recognition. Hence, studied the literature on SER, investigated types of methods and their learning strategies, and proposed a novel framework, called TCN: a modest and lightweight emotion recognition system using a temporal convolution network. The proposed methods extract high-level discriminative cues through TCN and constantly reduce the data depth in the inference technique, such as frequency-time analysis. The suggested structure has succeeded in obtaining better results on IEMOCAP and EMO-DB corpora compared with recent SER competitive methods. A detailed description of the proposed methodology will be explained in the subsequent sections.

3 Methodology

High-level characteristics are extracted using several high-level learning methods, such as Deep Belief Network (DBN), Deep Neural Network (DNN), and CNN. The features from input data automatically adjust the weight according to the input data, which provides better performance than the handcraft features. The study proposed a deep-learning model for speech signals. The proposed model consists of 3 parts, input, TCN, and a Fully Connected Network (FCN) that has identified sentiments in speech signals. The study introduces a novel core block of TCN in the framework, which has many advantages, such as temporal dependencies extraction and sequence learning in speech signals. The extracted features feed to FCN for global cues and then pass from the SoftMax to produce the final probabilities of each class. The inference procedures reduce data dimension, which is computationally efficient and produce a reference response. In Building the right network, it is necessary to reflect local connections along with time dependencies in speech signals. The network introduces the basic block of the TCN to efficiently recognize the temporal cues in speech sequences and identify them accordingly. The proposed framework is shown in Fig. 1. A detailed explanation of other components of the framework is described in the upcoming section.

3.1 Core Block

Nowadays, a deep learning layer, such as the Convolutions layers and recurrent layer has been recommended to deal with vibrant things of raw speech for mining meaningful cues. A temporal convolution network is utilized as a cues mining component to deal with the speech for SER classification. This work fine-tuned the basic block in this framework for two purposes, temporal density, and downsampling. First, the temporal density in the core unit can handle the data sequentially to extract the high-level features for emotion classification. Second, add a pooling strategy to downsample the input pad for further processing. As part of the design, a core block has three TCN layers and one pooling layer, where each temporal convolution layer is made up of three-time convolutions (1×3 each) and linear activation. Therefore, to improve the performance of speech signals each block shares similar hyper-parameters for the temporal convolutions.

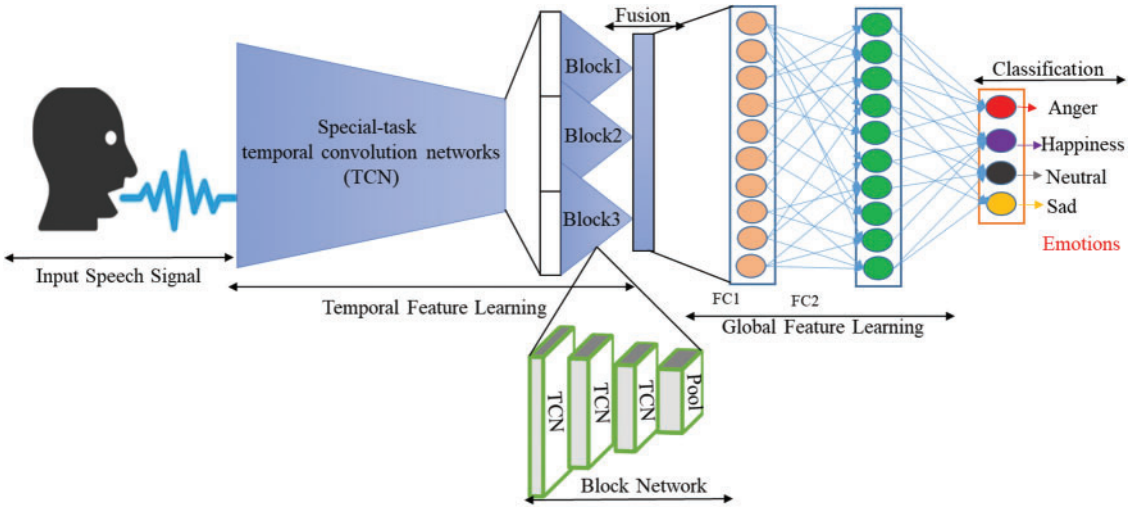

Figure 1: The overall architecture of the proposed speech emotion recognition using a temporal convolution network

The core block can be classified according to its feature level as arises. Each temporal convolution layer in the main block conducts subjective features combined with non-linear activation. After that, the extracted features are combined respectively, and repeatedly the next temporal convolution layers contribute to and extract high-level information in a global view. Additionally, the pooling layer followed to downsample the input map and reduce the overfitting of the model during training.

3.2 Proposed SER Techniques

Fig. 1 shows the proposed method for SER using the temporal network. The proposed model can learn effective information from the framework through a combination of TCN, cores units, and FCN layers. Let input matrix samples (for 1×8000 kHz) with balancing tags where E represents taped sentiments and C indicates the random sample that differs from 0 to 1 with a connected tag for extrapolation.

In the experimentations, proposed a network with a depth selected to accept 8000 samples of speech as an input to process from three TCN main units and two FC-L with SoftMax classification to generate the possibilities accordingly. The details of the suggested network structure are presented in Table 1 and are shown visually in Fig. 1.

A pooling layer goes only into the spatial dimension without overlapping. The first two set to max pooling with a filter size of 2, and the third is to average pooling with a filter size of 2 and stride setting of 4, respectively. There are 128 neurons in the first dense layer and 64 in the second dense layer. The flattened layer reshapes the feature vector for further processing, and the number of cells in the SoftMax layer will be equal to the number of classes in the input data. Additionally, regulation is embraced in the input speech processing process to normalize the features for better recognition. Furthermore, in the temporal convolution layer set with default stride setting and $32\times$ filter where x is modified to 1 and every unit is doubled accordingly.

Table 1: Detail of the TCN structure. The sign "()" comprises the filter range and the quantity of element padding along layer types

Block/layers	Output dimension	Parameters/Attributes
Input	1×8000	$1/2$ second sample per iteration
Block 1	1×8000 1×4000	(Filter, 32, 1×9, pad, 0, conv) $\times 3^{TCN}$ 1×2, max-pool, stride setting, 2
Block 2	1×4000 1×2000	(Filter, 64, 1×5, pad, 0, conv) $\times 3^{TCN}$ 1×2, max-pool, stride setting, 2
Block 3	1×2000 1×1000	(Filter, 128, 1×3, pad, 0, conv) $\times 3^{TCN}$ 1×5, avg-pool, stride setting, 4
Fusion	Block 3, the final output	Flatten layer
Fully connected network	128 (D1) 64 (D2) No of classes (SoftMax)	Dense layer Dense layer Classification layer

3.3 TCN Model Technical Details

A unit in the TCN is denoted by $xl, k, (n, m)$, where l is the layer, k is the provided maps, and (n, m) represents their position in the map. Similarly, $\sigma l, k, (n, m)$ represents the scalar effect among groups of input cells. Next, $xl,\ k, (n, m)$ can be obtained as:

$$xl,\ k,\ (n,\ m) = \int (\sigma l, k, (n, m) \tag{1}$$

where $\int$ is the activation function [24] utilized for the entire network accordingly. Every neuron of the feature pad is exemplified in worldly temporal convolution. Weight parameters for neurons in the anterior layer is shared across all layers to reduce the number of parameters for each subset of neurons following the correct position of each neuron. The weights of neurons are independently trained in their respective fields. Introducing the first block and other reasoning components sequentially is like the structure of the TCN. The information transmission process would look like this: let Lm be the layer, and let K be the key component as L1:

$$\sigma 1,\ \mathrm{k},\ (\mathrm{n},\ \mathrm{m}) = \omega 1,\ \mathrm{k}, 0 + \sum\nolimits_{j=1}^{A} lm + j - 1, \omega 1, k,\ j \tag{2}$$

where $\omega 1$, k, 0 is a threshold and $\omega 1$, k, I denote a set of weights with $1 \leq \mathrm{I} \leq \mathrm{A}$ (A = 3). With C + 1 weights for each pad, this layer extracts more useful temporal features from all electrodes using the spatiotemporal convolution kernel m employed in this framework as L2:

$$\sigma 2, k, (n, m) = \omega 2, k, 0 + \sum\nolimits_{J=1}^{P1} \sum\nolimits_{i=1}^{B} X1, j, (n + i - 1, m) \cdot \omega 2, k, j \tag{3}$$

Layer L2 further processes the temporal information extracted from layer L1 by cross-channel processing. P1 is the number of feature padding in L1, and B is the filter size such as L3:

$$\sigma 3, k, (n, m) = \omega 3, k, 0 + \sum\nolimits_{j=1}^{P2} \sum\nolimits_{i=1}^{C} X2, i, (n, +i - 1, m) \cdot \omega 3, k, j \tag{4}$$

This layer is used to extract high-level information about temporal dependencies based on the feature padding in layer L2, where P2 is the number of feature padding in layer L2, and C is the filter size of L3.

$$\sigma 4, \mathrm{k}, (\mathrm{n}, \mathrm{m}) = \max\left(x3, k, (j, n), x3, k, (j+1, m)\right) \tag{5}$$

There is no parameter in this layer, and k is fixed. By a pooling layer, the padding dimension of features is reduced by half, thus minimizing overfitting. The second and third core blocks (L5-L8) follow the same rules as the first core block (L1–L4) and can be determined from there.

$$\sigma 13, m = \omega 13, 0, m + \sum_{j=1}^{P12} \sum_{i=1}^{D} X12, j, i \cdot \omega 13, m \tag{6}$$

Layer L12 has D neurons and is fully connected to the flattened L12, in which w13 gives a threshold, P12 denotes the number of features, and L13 gives a pair of channels. This layer provides channel combinations such as L14:

$$\sigma 14, m = \omega 14, 0, m + \sum_{j=1}^{P13} X13, j \cdot \omega 14, m \tag{7}$$

A layer is used to select valid spatial information for classifications using $\omega 14$, 0, as a threshold and P13 as the number of neurons in the layer.

3.4 Model Organization and Computational Setup

In this section, the depth analysis of the suggested framework is set up to be used in a Python environment utilizing Scikit Learning and other relevant machine learning modules. To make the model suitable for SER, adjusted several hyper-parameters. Additionally, examined this model using a range of batch sizes, learning rates, optimizers, positional sizes, and regulation variables, such as L1 and L2 with varying values. The data was divided into an 80:20 ratio, with 80% going toward model training and 20% for model testing. Instead of performing pre-processing or modification in these trials, simply predicted emotions from the unprocessed audio or speech. The model was trained and tested using a GeForce RTX 3080 NVIDIA GPU with 10 GB of memory. To save the best model the model train using the initial stop approach and set the learning rate to 0.0001 with a fall after ten iterations. For all datasets, the 64-batch size selected delivered great accuracy.

4 Experimental Assessments and Results

In this section, the proposed TCN design is empirically assessed using two standard raw speech emotion corpora include, IEMOCAP [25] and EMO-Db, [26]. Datasets are scripted, where actors read scripts to record and express different emotions. They conducted experiments and obtained high-level results to show the efficiency and strength of the proposed technique in the SER field. A detailed description of the database, model evaluation, and performance are included in upcoming sections.

4.1 Datasets

4.1.1 Interactive Emotional Dyadic Motion Captures (IEMOCAP)

The most widely used dataset in the SER field is the notoriously difficult and complex Interactive emotional dyadic motion captures (IEMOCAP) [25]. Different male and female professional actors performed scripted and spontaneous speeches for the IEMOCAP dataset. Ten professional actors—five men and five women—recorded twelve hours of audio-visual, audio, video, text transcription, and face motion data throughout five sessions, each of which had a male and a female actor. Three distinct

experts annotated the recorded script, giving each label a suitable description. The authors chose the documents for which at least two experts concurred on the same label and the total number of each class/emotion is mentioned in Table 2. Experimental evaluation and comparative analysis are often employed in the literature.

Table 2: Detailed information of EMO-DB & IEMOCAP corpus

Emotion	Total voices in EMO-DB	Total voices in IEMOCAP
Happiness	71	1636
Anger	127	1103
Sadness	62	1084
Neutral	79	1708
Fear	69	-
Boredom	81	-
Disgust	46	-

4.1.2 Berlin Emotion Database (EMO-DB)

The SER extensively used the EMO-DB [26], a Berlin emotion database recorded by ten professional actors, five of whom were men and five of whom were women. The EMO-DB has a total of 535 prepared utterances, five male and five female professional actors read aloud to convey a range of emotions, including rage, boredom, fear, melancholy, happiness, and disgust. The average duration of a dataset is 3.5 s, captured at a sample rate of 16 kHz, while the utterances were three to five seconds long. The authors assessed each emotional category and contrasted it with standard procedures further details about the number of utterances in each class are mentioned in Table 2.

4.2 Experimental Evaluations

This section conducted an experimental evaluation of the proposed technique on two benchmark datasets called: IEMOCAP and EMO-DB. The data was split into an 80:20 testing and training evaluation to validate the proposed system. The test set was used to evaluate the model recognition ability on unknown data. All the datasets have different numbers of speakers, which is why split the data into an 80:20 ratio. The work used statistical parameters to investigate the proposed method to calculate the accuracy using various functions. The work used the confusion matrix to predict the true positive, false positive, and true negative false negative values and validate the model testing [27]. The letter T denotes accurate forecasts, the letter F denotes inaccurate predictions, and the sum of the letters true positives (TP) and false negatives (FN) denote the amount of positive data. The amount of negative data in the actual condition is revealed by adding the TN (true negatives) and FP (false positives) [28]. The entire recognition ratio is shown by the accuracy factor. Accuracy by itself will not suffice, factors including precision, memory, and F1-score evaluation criteria are needed. The harmonic mean of precision and recall rate is used to define the depth of the F1-score [29]. Additionally, assessed this suggested model from several angles presented the findings in terms of weighted accuracy and unweighted accuracy, which applied to both comparative analysis and the literature. Table 3 displays the training results for various high-level models and this system. Table 3 compares the performance of the proposed model's architectures in terms of weighted accuracy and unweighted accuracy (WA and UA) using various datasets.

Table 3: Performance comparison of various model architectures utilizing datasets in terms of weighted accuracy (WA) and unweighted accuracy (UA)

Database	Model	WA (%)	UA (%)
IEMOCAP	CNN	68.27	62.75
	LSTM	70.21	65.68
	Proposed model	**80.84**	**78.34**
EMO-DB	CNN	86.35	82.57
	LSTM	83.44	78.73
	Proposed model	**92.31**	**91.61**

The results of the model training for the IEMOCAP and EMO-DB datasets are displayed in Table 2. The speech signals are the model's input source. For both datasets, high precision has been attained. Modified filters in the TCN convolution and pooling layers are employed in this study to implement a novel TCN model for the SER. The approach improves generalization and accuracy, demonstrating the significance and effectiveness of the method. The authors compare the suggested model to the baseline to illustrate the improvement, robustness, and effectiveness of the system.

The suggested system's unweighted accuracy compares with other standard techniques. Few studies in the SER literature discuss sequential architecture and those that do not significantly enhance performance. The speech emotion recognition temporal convolution network is a new type of learning strategy for SER. To demonstrate the model's efficacy and accuracy, tested the proposed system using two SER datasets and compared the findings with the industry standard methods. Consequently, the proposed system achieved good weighted and unweighted accuracy compared to other architectures shown in Table 2. This model has a straightforward structure that effectively recognizes emotions from voice signals using TCN layers with a modified filter shape, core block, and max pooling.

For further analysis, obtained the confusion matrix (CM) was to clarify the misunderstanding between factual and the predicted labels with other emotions in the respective lines. The achieved CM class-wise precisions/recalls of all suggested corpora are demonstrated in the following figures.

Figs. 2 and 3 show the confusion matrixes and class-level perception of all sentiments of IEMOCAP, and EMO-Db corpora, respectively. In comparison with state-of-the-art techniques, better recognition results are shown by this model. The final productivity of the given system is important and performs better for every emotion of happiness [30]. Due to their linguistic data or information, the happy emotion with low accuracy was identified by the model. The proposed model gives the correct un-weighted accuracy and is identified and recognized by this model from the frequency pixels. The reason is that all the emotions recognized by the proposed SER model with superior identification rate. Thus, for more examinations elaborate on the CM needed for each corpus which is given in Figs. 2 and 3. The uncertainty among each other is shown in the corresponding lines, and the actual predicted values of each emotion diagonally. The class-level accuracy of the proposed system is illustrated in Fig. 4.

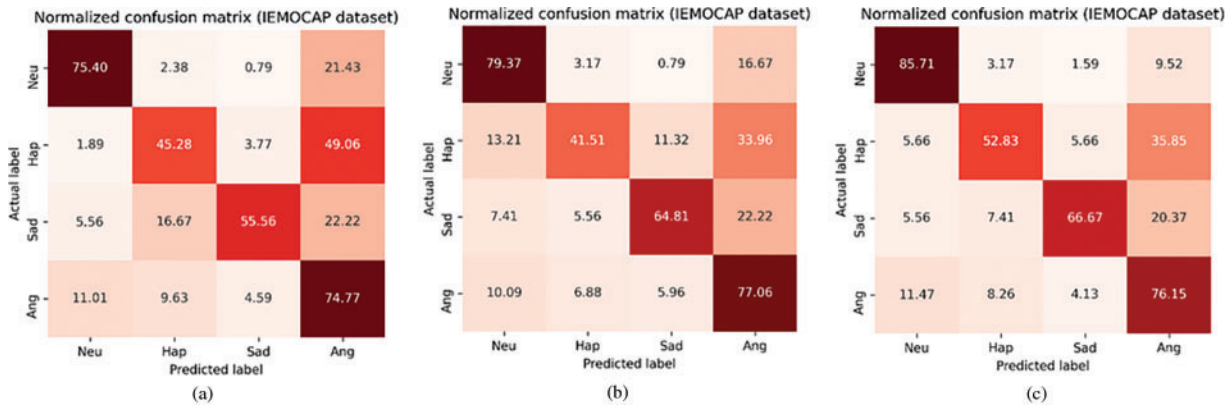

Figure 2: The IEMOCAP dataset's confusion matrix, (a) CNN model with a UA of 62.75%, (b) LSTM model with a UA of 65.68%, and (c) the suggested model with a UA of 78.34%

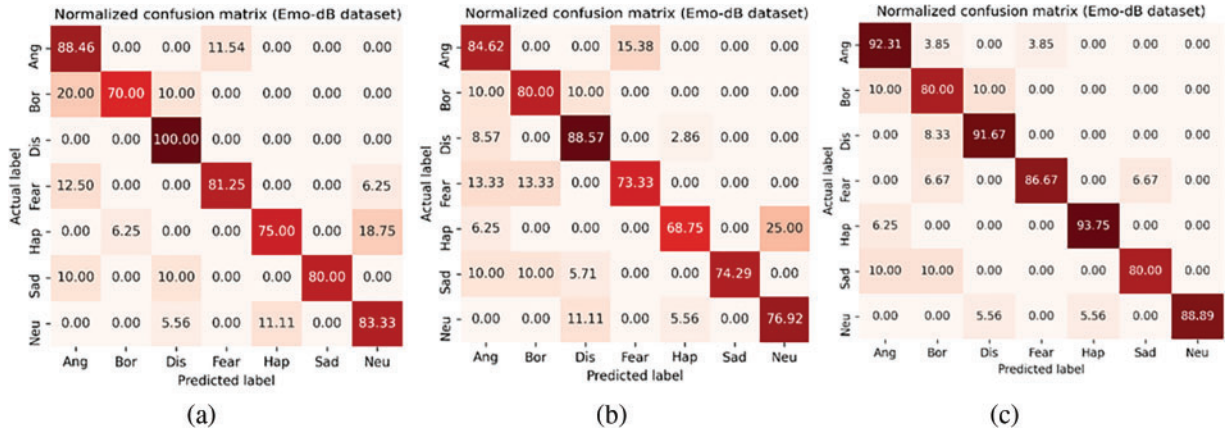

Figure 3: The EMO-DB dataset's confusion matrix, (a) CNN model with a UA of 82.57%, (b) LSTM model with a UA of 78.73%, and (c) the suggested model with a UA of 91.61%

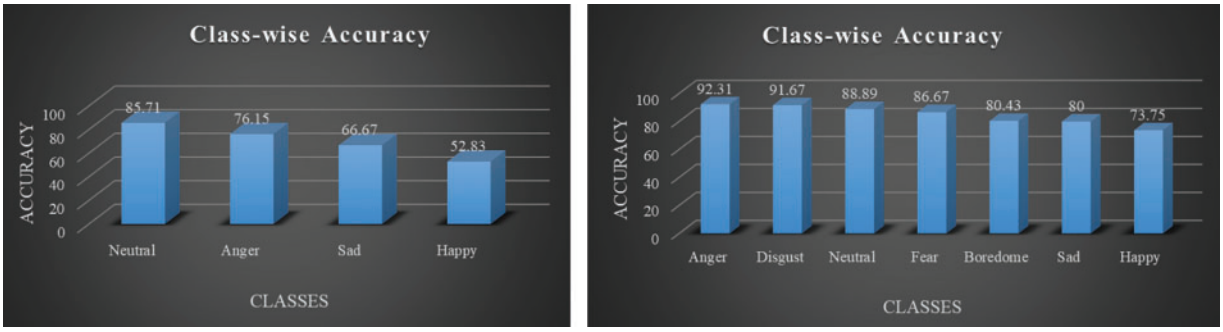

Figure 4: Class-wise accuracy of the proposed system using IEMOCAP and EMO-DB corpora

The idea assessed the suggested SER approach with two datasets, noted the unweighted (UA) accuracy, and matched it with recent approaches. During the experimentation of IEMOCAP and EMO-DB, the highest outcome of the proposed system achieved was shown in the confusion

matrices. The emotion of happiness identified compared to state-of-the-art approaches, and the whole identification outcomes of the given prototype are more countable than the baseline approaches. A better identification value and high precision due to their architecture show the model's effectiveness and robustness for real-world issues. Simple structure and user-friendly working features are used by the model, which can be real-time usage to examine human behavior.

Table 3 demonstrates the fair comparison of the suggested recognition method that is highly exceptional and more meaningful than the baseline using the same speech corpora. The current SOTA unweighted (UA) accuracy for the IEMOCAP dataset is 62.75% and 65.68% in the literature using 1D-CNN models. Nevertheless, this method increased it to 13.35% then the whole unweighted accuracy rate was boosted up to 78.34% by utilizing the suggested system. Furthermore, enhanced up to 9.4%, un-weighted accuracy with the EMO-DB dataset. Finally, this model achieved better-unweighted accuracy than the given baseline and good outcomes for all emotions (as seen confusion matrix and tables).

5 Discussion and Model Comparison

In this study, the TCN, core blocks, and an updated pooling policy are considered a large contribution to the SER system. A new and modified core block and TCN layer were used to identify the deep frequency characteristic of the speech signals. To the best of our knowledge, to recognize the emotions in the speech data, this type of core block, layer, filter, and frequency characteristic selection is lacking by the recent SER systems.

Authors investigated the SER literature and found limitations such as the rate of recognition and complexity of the model. Scientists have established several methods applying conventional, low-level, and high-level methods to increase the level of accuracy and decrease the overall model complexity, but the existing accuracy is still too low, and the modal cost is too high to address real-world issues. The article focuses on the accuracy issue and recommends a novel TCN model for emotion recognition using an updated TCN layer, core block, filter shape, and a pooling policy to identify the emotional state of speakers in their speech signals. As an input, a speech signal was used by this model to perform a temporal convolutional operation on it and extract the deep TCN characteristics from the frequency pixels. With the utilization of the deep frequency characteristics, Speech emotions with high accuracy are identified by this model. In the proposed model, three core blocks are used, and every block consists of three convolutions layers and one pooling layer. Due to this simple and novel structure, the time accuracy is increased, and a valuable and reliable SER system was designed, as proved in the experimentation. Table 4 illustrates the efficiency and utilization of the proposed system, its results, and model performance compared with the baseline models. The authors compared the performance of the proposed model in term of IEMOCAP with [1,31–38] and in term of Emo-Db with [1,30,38–40]. For comparative analysis, a detailed overview is illustrated in Tables 4 and 5.

The above tables represent the relative/comparative assessment of the proposed technique with a baseline utilizing similar corpora. The provided tables display the suggested system's outperforming result, which is much greater than that of other systems and demonstrates the efficiency of this method. The presented method demonstrates the latest achievement of deep learning in this field, which accurately identified emotions using straightforward architecture. The suggested SER model outperforms the baseline model in terms of recognition rate and time complexity. Tables 4 and 5 provide a thorough demonstration and the experimental findings. The study developed a simple TCN: Speech Emotion Recognition Using Temporal Convolution Network for the recognition with superior precision to reduce the dimension of data, which is appropriate for observing real-world products.

Table 4: Performance (%) comparison of the suggested model on the IEMOCAP compared to innovative techniques

Database	Method (Refs.)	WA (%)	UA (%)
IEMOCAP	MLT-Net (2021) [1]	73.01	-
	Efficient-SER (2017) [31]	68.80	59.40
	APL-SER (2018) [32]	71.80	68.10
	Caps-Net (2019) [33]	72.73	59.71
	HSF-DNN (2020) [34]	57.10	58.30
	DCNN-SER (2020) [35]	64.30	-
	ICRNN (2021) [36]	-	64.50
	C-Graph (2021) [37]	64.19	60.31
	3D-CNN (2018) [38]	-	64.74
	Proposed model	**80.84**	**78.34**

Table 5: Performance (%) comparison of the suggested model on the Emo-Db compared to innovative techniques

Database	Method (Refs.)	WA (%)	UA (%)
EMO-DB	MLT-Net (2021) [1]	90.01	-
	RBF-SER (2020) [30]	-	85.57
	3D-CNN (2018) [38]	-	82.82
	SA-RNN (2021a) [39]	85.95	82.06
	PSW-SER (2019) [40]	86.44	84.53
	Proposed model	**92.31**	**91.61**

6 Conclusion

In the field of emotion recognition, using speech signals has many challenges choose an optimal cue for emotion recognition to improve the overall recognition rate through a robust and significant approach. Nowadays, researchers are working to address the limitations such as optimal features selection, best model configuration, optimizer, and learning strategy to increase the baseline models' accuracy through a novel and lightweight high-level approach. According to these challenges, proposed a TCN-based SER system utilizing a temporal convolutional network to learn spatial and sequential cues through a raw audio file and efficiently recognize each emotion accordingly.

The suggested technique reveals considerable enhancements in the standard performance and learns extra strong cues from speech waves using temporal convolution layers. The overall procedures of the proposed system are split into two sections: 1) to deal with the temporal dimension, propose the core block, and 2) use the dense layer to fuse spatial features between signals and emotions recognized. The importance of spatial information and temporal dependencies is also examined through three baselines, which are described in Table 3 (Section 4). The authors assessed the technique over two standard IEMOCAP and EMO-Db speech corpora and achieved 78.34% and 91.61%

unweighted precision. Using raw speech waves, the proposed system demonstrated excellent generality and recognition rates. The suggested system can apply to health care centers, online communication, safety control system, call centers, and many more.

In the future, researchers can further explore the 1D-TCN model for speaker recognition, speaker diarization, automatic speaker recognition, and human behavior assessment techniques to address real-world issues through a real-time application. Furthermore, researchers can investigate the proposed TCN-based deep learning model for big data in a future long train on real data to recognize the state of speakers.

Acknowledgement: Sejong University's 2021 Faculty Research Fund and the National Research Foundation of Korea supported this work through Grant NRF-2020R1F1A1060659, a project funded through the Korean government's Ministry of Science and ICT.

Funding Statement: The authors received no specific funding for this study.

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

References

[1] Mustaqeem and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach science direct," *Expert Systems with Applications,* vol. 167, no. 1, pp. 114177–114189, 2021. https://www.sciencedirect.com/science/article/abs/pii/S0957417420309131

[2] M. R. Falahzadeh, F. Farokhi, A. Harimi and R. Sabbaghi-Nadooshan, "Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 449–492, 2023.

[3] Z. Shiqing, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional LSTM," *IEEE Transactions on Affective Computing*, vol. 13, pp. 680–688, 2019.

[4] Mustaqeem and S. Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *CMC-Computers Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[5] A. A. Abdelhamid, E-S. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader *et al.,* "Robust speech emotion recognition using CNN + LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022.

[6] S. Hammal, N. Bourahla and N. Laouami, "Neural-network based prediction of inelastic response spectra," *Civil Engineering Journal*, vol. 6, no. 6, pp. 1124–1135, 2020.

[7] C. Jensen, M. Kotaish, A. Chopra, K. A. Jacob, T. Widekar *et al.,* "Piloting a methodology for sustainability education: Project examples and exploratory action research highlights," *Emerging Sciences*, vol. 3, no. 5, pp. 312–326, 2019.

[8] P. S. Kumar, H. S. Shekhawat, and S. R. M. Prasanna, "Attention gated tensor neural network architectures for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 71, pp. 103173, 2021.

[9] M. Ishaq, G. Son and S. Kwon, "Utterance-level speech emotion recognition using parallel convolutional neural network with self-attention module," in *Proc. of Int. Conf. on Next Generation Computing*, Jiju, South Korea, pp. 109–113, 2021.

[10] K. Liu, D. Wang, D. Wu, and J. Feng, "Speech emotion recognition via multi-level attention network," *IEEE Signal Processing Letters*, vol. 29, pp. 2278–2282, 2022.

[11] K. C. Wang, "Time-frequency feature representation using multi-resolution texture analysis and acoustic activity detector for real-life speech emotion recognition," *Sensors*, vol. 15, no. 1, pp. 1458–1478, 2015.

[12] E. Lieskovská, M. Jakubec, R. Jarina and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, pp. 1163, 2021.

[13] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson *et al.,* "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," *Neurocomputing*, vol. 448, pp. 238–248, 2021.

[14] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[15] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou *et al.,* "Speech emotion classification using attention-based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[16] H. Zhang, Z. Wang, D. Liu and L. Systems, "A comprehensive review of stability analysis of continuous-time recurrent neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 7, pp. 1229–1262, 2014.

[17] M. Mustaqeem and S. Kwon, "Speech emotion recognition based on deep networks: A review," in *Proceedings of the Korea Information Processing Society Conference*, Korea Information Processing Society, pp. 331–334, 2021.

[18] B. Abbaschian, D. Sierra-Sosa and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 4, pp. 1249, 2021.

[19] M. B. Akçay and K. J. Oğuz, "Speech emotion recognition: Motional models, databases, features, pre-processing methods, supporting modalities, and classifiers," *Speech Communications*, vol. 116, pp. 56–76, 2020.

[20] S. Abdullah, S. Ameen, M. A. Sadeeq, S. Zeebaree and T. Trends, "Multimodal emotion recognition using deep learning," *Applied Science and Technology Trends*, vol. 2, pp. 52–58, 2021.

[21] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch and M. R. Wrobel, "Emotion recognition and its applications," *Human-Computer Systems Interaction: Backgrounds and Applications 3*, vol. 1, pp. 51–62, 2014.

[22] A. Talabani, H. Sellahewa and S. A. Jassim, "Emotion recognition from speech: Tools and challenges," *Mobile Multimedia/Image Processing, Security, and Applications*, vol. 9497, pp. 94970N, 2015.

[23] Z. Zhang, B. Wu and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Int. Conf. on Acoustics, Speech, and Signal Processing of the IEEE*, Brighton, United Kingdom, pp. 6705–6709, 2019.

[24] W. Dai, C. Dai, S. Qu, J. Li and S. Das, "Deep convolutional neural networks for raw waveforms," in *Int. Conf. on Acoustics, Speech, and Signal Processing of the IEEE*, Hilton, New Orleans, pp. 421–425, 2017.

[25] C. Busso, B. Murtaza, L. Chi-Chun, K. Abe, M. Emily *et al.,* "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resource and Evaluations*, vol. 42, pp. 335–359, 2008.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[27] X. Deng, Q. Liu, Y. Deng and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for the classification problem," *Information Sciences*, vol. 340, pp. 250–261, 2016.

[28] J. Xu, Y. Zhang and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences*, vol. 507, pp. 772–794, 2020.

[29] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomic*, vol. 21, no. 1, pp. 1–13, 2020.

[30] 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.

[31] A. Satt, S. Rozenberg and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. of Interspeech-2017*, Stockholm, Sweden, pp. 1089–1093, 2017.

[32] P. Li, Y. Song, I. V. McLoughlin, W. Guo and L. -R. Dai, "An attention pooling based representation learning method for speech emotion recognition," *University of Kent Journal*, vol. 1, pp. 12–33, 2018.

[33] L. Wu, S. Liu, Y. Cao, X. Li, J. Yu *et al.,* "Speech emotion recognition using capsule networks," in *Int. Conf. on Acoustics, Speech, and Signal Processing of the IEEE*, Brighton, United Kingdom, pp. 6695–6699, 2021.

[34] Z. Yao, Z. Wang, W. Liu, Y. Liu and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Communications*, vol. 120, pp. 11–19, 2020.

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

[36] P. Meyer, Z. Xu and T. Fingscheidt, "Improving convolutional recurrent neural networks for speech emotion recognition," in *Proc. of Spoken Language Technology of the IEEE*, Virtual, Shenzhen, China, pp. 365–372, 2021.

[37] A. Shirian and T. Guha, "Compact graph architecture for speech emotion recognition," in *Int. Conf. on Acoustics, Speech, and Signal Processing of the IEEE*, Brighton, UK, vol. 1, pp. 6284–6288, 2021.

[38] 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, pp. 1440–1444, 2018.

[39] D. Li, J. Liu, Z. Yang, L. Sun and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, pp. 114683, 2021.

[40] L. Abdel-Hamid, "Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communications*, vol. 122, pp. 19–30, 2020.