IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 September 2022, accepted 20 October 2022, date of publication 26 October 2022, date of current version 31 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217226

RESEARCH ARTICLE

3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU

MOHAMMAD REZA FALAHZADEH[1], EDRIS ZAMAN FARSA[2], ALI HARIMI[3], ARASH AHMADI[4], (Senior Member, IEEE), AND AJITH ABRAHAM[5,6], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Islamic Azad University, Central Tehran Branch, Tehran 14778-93855, Iran
[2]Department of Computer Engineering, Islamic Azad University, Sanandaj Branch, Sanandaj 61349-37333, Iran
[3]Department of Electrical Engineering, Islamic Azad University, Shahrood Branch, Shahrood 36199-43189, Iran
[4]Department of Electronics, Carleton University, Ottawa, ON K1S 5B6, Canada
[5]Machine Intelligence Research Laboratories (MIR Labs), Auburn, WA 98071, USA
[6]Center for Artificial Intelligence, Innopolis University, 420500 Innopolis, Russia

Corresponding authors: Edris Zaman Farsa (edriss.zamanfarsa@gmail.com) and Ajith Abraham (ajith.abraham@ieee.org)

This work was supported by the Analytical Center for Government of the Russian Federation under Grant 70-2021-00143 dd. 01.11.2021 (IGK000000D730321P5Q0002).

ABSTRACT Due to the high level of precision and remarkable capabilities to solve the intricate problems in industry and academia, convolutional neural networks (CNNs) are presented. Speech emotion recognition is an interesting application for CNNs in the field of audio processing. In this paper, a speech emotion recognition system based on a 3D CNN is suggested to analyze and classify the emotions. In the proposed method, the three-dimensional reconstructed phase spaces of the speech signals were calculated. Then, emotion-related patterns formed in these spaces were converted into 3D tensors. Accordingly, a 3D CNN for speech emotion recognition applied to two datasets, EMO-DB and eNTERFACE05, using a speaker-independent technique achieved 90.40% and 82.20% accuracy, respectively. By employing gender recognition, the accuracy rates on EMO-DB increased to 94.42% and on eNTERFACE05 rose to 88.47%. Realization of the introduced 3D CNN on both Intel CPU and NVIDIA GPU is also explored. The results of the implemented 3D CNN without and with regard to gender recognition show that GPU-based running is faster for the EMO-DB and eNTERFACE05 datasets than CPU-based executions (using Python).

INDEX TERMS 3D convolutional neural networks (3D CNNs), speech emotion recognition, reconstructed phase space, 3D tensor.

I. INTRODUCTION

During the past years, because of the availability of big data including audio, video, image, text, etc. and progression in digital electronics devices, deep leaning has received increasing attention by researchers [1], [2], and [3]. Convolutional neural networks (CNNs) are one of the prominent and credible deep learning models due to its computational efficiency and high accuracy in comparison with other artificial intelligence algorithms. CNNs are heavy in computations and memory requirements. Therefore, running CNNs in embedded computing devices requires effective hardware/software co-design [4], [5]. The ability of CNNs in the comprehension of complex structures is an impressive feature in applications with high-dimensional data such as text processing [6], [7], face detection [8], [9], speech recognition [10], [11], character recognition [12], [13], image classification [14], [15], video classification [16], and gesture recognition [17]. Moreover, Microsoft, Instagram, Amazon, Google, and Facebook are examples of high-tech corporations which have applied CNNs in different types of services [18].

Speech emotion recognition is a challenging topic in the field of pattern recognition and processing the speech signals has received a great deal of research interest in the recent

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

decades. The purpose of emotion recognition from speech is classifying the basic emotions including sadness, happiness, fear, anger, disgust, surprise, boredom, and neutral from speech signals and can be used in a variety of applications in human-computer interaction [19], [20], and [21].

It is common to use 2D CNN models for visual tasks [22], [23], [24]. However, these networks have also been employed for audio-visual purposes. For example, in [25], a 2D CNN have been proposed for emotion recognition from speech and visual information. Regarding the benefits of 2D CNNs for image processing tasks, some research introduced feature engineering techniques to convert one-dimensional speech signal to 2D images, which allow to benefit from 2D CNNs in speech processing tasks, and in particular emotion recognition applications. In this way, spectrogram [26] and CyTex [27] are two feature engineering-based methods that have been employed to convert speech signal to images as a compatible input for 2D DCNNs. Moreover, in [28], the phase space reconstruction has been employed to represent the emotional speech in a 3D space. Then, a transformation technique has been used to convert the 3D speech patterns to 2D chaogram images for speech emotion recognition task.

3D CNNs have been successfully used for speaker verification [29], video scene understanding [30], action recognition [31], and also introduced as promising models to recognize the emotions on the base of feature extraction technique in the speech signals [32], [33]. Although feature extraction is a very common method, it still suffers from extracting of the ineffective features. Hereby, finding more practical ways is inevitable. Phase space reconstruction has been exposed to discussion for analyzing signals with nonlinear dynamics and presented as a meritorious alternative to conventional signal classification approaches [34]. It is also an effective tool for representing the one-dimensional speech signal in a multidimensional space, that should be compatible to the employed model input [35]. It motivates to extract 3D tensors based on mutual information [36] from speech signals instead of extracting features. Two famous corpuses named EMO-DB [37] and eNTERFACE05 [38] are very common and popular to explore the performance of various algorithms in recognizing speech emotions [26], [32], [33], [39], [40], [41], [42], [43], [44], [45], [46], [47].

This work proposes a 3D CNN architecture to recognize various speech emotions and its realization on both Intel CPU and NVIDIA GPU. Experiments on two public datasets known EMO-DB and eNTERFACE05 have shown highly valuable results for investigating 3D tensors using a 3D CNN model in speech emotion recognition application. To reach a better accuracy, gender recognition technique is added to the suggested method. As envisaged and on the mentioned corpuses, GPU implementations have less running times than CPU implementations.

The rest of the paper is organized as follows: Section 2 explains the recommended 3D CNN model for speech emotion recognition. In section 3, the experimental outcomes are shown and compared with other related published works. Finally, section 4 concludes the paper.

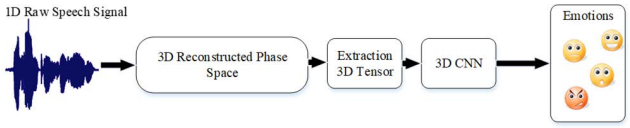

FIGURE 1. The schematic diagram of the proposed method.

II. PROPOSED METHOD FOR SPEECH EMOTION RECOGNITION

The presented method consists of two main stages. In the first stage, 3D tensors are provided using reconstructed phase space of speech signals, and in the second stage, a 3D CNN is trained based on the 3D tensors provided in the first stage and their corresponding emotion labels. As the speech signal has nonlinear and chaotic behavior, showing the correlation of the emotional speech parameters in one-dimensional space is not possible. Reconstruction of signal in the phase space is an appropriate method for studying signals in higher dimensions. In order to apply the compatible inputs for the 3D CNN network and to study the relationship between emotional parameters, speech signals have been modeled and analyzed in a 3D space. Figure 1 shows a schematic diagram of the proposed method for speech emotion recognition. As displayed, by using the reconstructed phase space, the one-dimensional signal is mapped to the three-dimensional space and then a 3D tensor is extracted to apply as the input of the 3D CNN.

A. RECONSTRUCTED PHASE SPACE AND CREATING A 3D TENSOR OF SPEECH

Phase space reconstruction is a powerful technique for analyzing nonlinear dynamic systems with chaotic characteristics and has been presented as a pivotal alternative to conventional nonlinear signal classification methods [27]. The phase space reconstruction approach transforms a one-dimensional signal known as a vector to a d-dimensional signal called a tensor. In order to reconstruct the phase space of a system, the output signal of the system is assumed as a time series S_n, $n = 1, 2, 3, \ldots, N$. Equation (1) shows a row vector that is a single point in the reconstructed phase space.

$$\bar{s_n} = [s_n, s_{n+\tau}, s_{n+2\tau}, \ldots\ldots\ldots., s_{n+(d-1)\tau}] \tag{1}$$

where τ denotes the time delay and d indicates the dimension. All possible points of the system in the reconstructed phase space are defined by the following trajectory matrix:

$$S = \begin{pmatrix} s_1 & s_{1+\tau} & s_{1+2\tau} & \ldots & s_{1+(d-1)\tau} \\ s_2 & s_{2+\tau} & s_{2+2\tau} & \ldots & s_{2+(d-1)\tau} \\ s_3 & s_{3+\tau} & s_{3+2\tau} & \ldots & s_{3+(d-1)\tau} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ s_N & s_{N+\tau} & s_{N+2\tau} & \ldots & s_{N+(d-1)\tau} \end{pmatrix} \tag{2}$$

Each row vector s_n represents a speech element and its relation to the samples with τ delay. Major methods determine optimum time delay, τ, and dimension, d, based on mutual information and the false nearest neighbours, respectively [36]. Because the inputs to be compatible for the 3D CNN, the false nearest neighbours method is chosen to set $d = 3$. To compute the suitable value for τ, the mutual information approach is employed. The mutual information between two signals is obtained from the following equation:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p_{(X,Y)}(x,y) \times log(\frac{p_{(X,Y)}(x,y)}{p_{(X)}(x)\,p_{(Y)}(y)}) \quad (3)$$

where $p_{(X,Y)}$ is the joint probability distribution function and $p_{(X)}$ and $p_{(Y)}$ are the marginal probability distribution functions. The first minimum of mutual information is optimal [36]. While computing the τ parameter for each speech sample is time-consuming and imposes heavy computations, the first minimum of average mutual information as the optimum time delay is considered. Except the dependency of τ based on the time delay, sampling rates of the speech signals is another parameter that has effect on τ computing. By finding the appropriate values of τ and d, the reconstructed phase spaces are modelled. Setting the optimal value of τ is vital in reconstructed phase space analysis. As shown in Figure 2, the first minimum of average mutual information as the optimum time delays in the EMO-DB and eNTERFACE05 datasets are located on $\tau = 17$ and $\tau = 31$, respectively. Figure 2 (a) displays 535 mutual information of speech samples from the EMO-DB dataset, and Figure 2 (b) depicts 1166 mutual information of speech samples from the eNTERFACE05 dataset. Figure 2 (c) and Figure 2 (d) show the first minimum of average mutual information for each dataset. Each speech sample has different first minimum value in mutual information.

For a better understanding, the reconstructed phase spaces from a speaker in 7 various emotions of the EMO-DB dataset have been shown in Figure 3.

The procedure of the creating 3D tensors from the 1D raw speech signals to be compatible with the suggested 3D CNN for speech emotion recognition has been displayed in Figure 4. As shown in this figure, the appropriate values of time delays have been set to $\tau = 17$ for the EMO-DB dataset and $\tau = 31$ for the eNTERFACE05 dataset.

Typically, the reconstructed phase space of a speech signal is presented in a multidimensional space. Since the goal is to create the compatible inputs for the 3D CNN using speech samples, the 3D reconstructed phase space of each speech signal is formed and then converted into a 3D tensor. To this end, each axis of the 3D space is split into 256 segments. Consequently, the space is converted to a $256 \times 256 \times 256$ grid net. The frequency of points in each cell of the grid is calculated. Therefore, a 3D matrix considered as a 3D tensor. In other words, the output 3D tensor of this stage is a 3D histogram of points in the space. In summary, each 3D tensor can be considered as a new representation of the corresponding speech signal that is labeled with the corresponding emotion in the dataset. Finally, this 3D tensor would be applied to the 3D CNN model for training and testing. The number 256 is set arbitrarily, as it can be matched to the network input size via a resize procedure. However, it is preferable to choose a close value to the size of the network input to avoid additional computations.

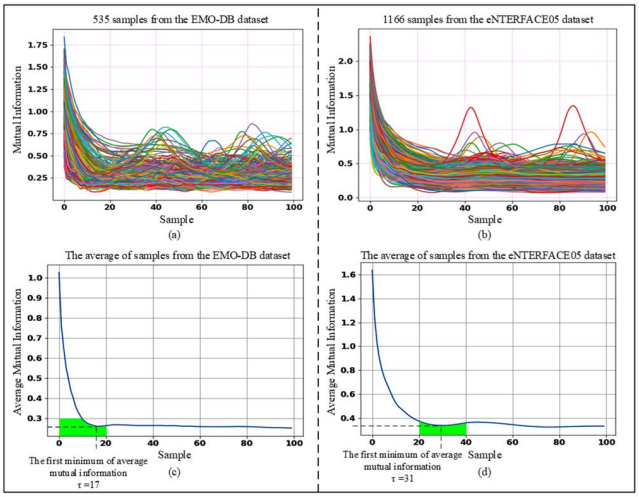

FIGURE 2. The mutual information of the speech signals for the EMO-DB and eNTERFACE05 datasets. (a) 535 mutual information of speech samples extracted from the EMO-DB dataset. (b) 1166 mutual information of speech samples extracted from the eNTERFACE05 dataset. (c) The first minimum of average mutual information for EMO-DB ($\tau = 17$) (d) The first minimum of average mutual information for eNTERFACE05 ($\tau = 31$).

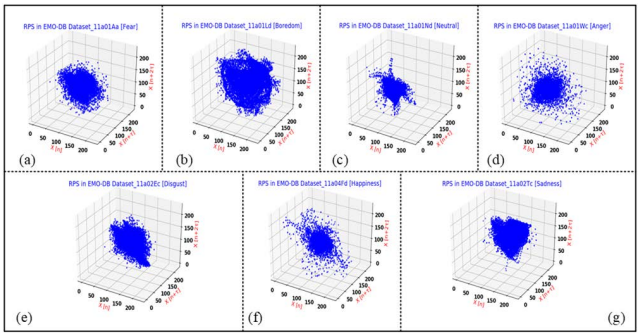

FIGURE 3. Reconstructed phase spaces of the speech signals expressed by a speaker in 7 different emotions from the EMO-DB dataset. (a) 11a01Aa (Fear), (b) 11a01Ld (Boredom), (c) 11a01Nd (Neutral), (d) 11a01Wc (Anger), (e) 11a02Ec (Disgust), (f) 11a04Fd (Happiness), and 11a02Tc (Sadness) with $d = 3$, *and* $\tau = 17$.

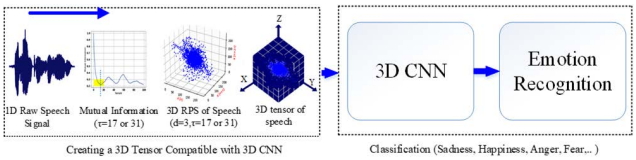

FIGURE 4. The schematic of the proposed method for speech emotion recognition based on the average mutual information on the datasets.

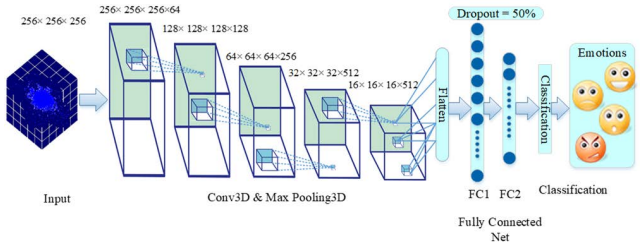

FIGURE 5. The proposed 3D CNN architecture.

B. PROPOSED 3D CNN FOR EMOTIONS CLASSIFICATION

The structure of the suggested 3D CNN model is displayed in Figure 5. This model has been inspired from the architecture of VGG16 [48], which have gained great achievements in classification of two-dimensional images. In this research, a similar architecture proposed in three-dimensional manner. Experiments have been done to check whether simplifying the model by removing some of the layers, or using additional layers, can improve the classification performance. To this end, first, a convolutional layer added before the first layer of the suggested model. This resulted in decreasing of the overall classification rate. Next, a convolutional layer added to the end of the model, just before the flatten layer. It also led to falling in the classification rate. In other experiments, the first and last convolutional layers of the model have been removed, respectively. These experiments also result in decrease in overall classification performance. Consequently, the proposed structure of the model is set similar to the VGG16 architecture [48].

This 3D CNN will train on 3D tensors obtained from reconstructed phase space representation of speech signals. In contrast to classical pattern recognition systems, CNNs usually combine the feature extraction and classification stages in an end-to-end model that perform both tasks. Hence, the 3D CNN can be considered as a model that can train the features of the 3D tensors (which are a new representation of the input speech signals) related to the target emotion labels. As shown in Figure 5, the 3D CNN uses 3D filters to analyze input data, which refers to feature extraction. Then, it assigns the weights to the outputs of the filters to emphasize or ignore informative or redundant features, respectively. These weights are computed through the training procedure by minimizing a loss function trying to link input data to the output target classes.

The input of the proposed network is a 3D tensor with the size of 256 × 256 × 256. While the smaller size of input can reduce the resolution and consequently cause loss of useful information, larger size can complicate computations. The 3D CNN consists of three types of layers, including the convolution, pooling, and fully connected layers, where each layer performs its particular task. The proposed 3D CNN architecture consists of five convolutional layers with 64, 128, 256, 512, and 512 3D filters. The kernel size of convolutional layers is 3 × 3 × 3 with the stride size of 1. There is a max pooling layer with a kernel size of 2 × 2 × 2 and the stride size of 2 after each convolutional layer. This max pooling layer is not only responsible for sampling but also reduces the features dimensions. Finally, a flattening and two fully connected layers are located afterwards. The flattening layer converts the 3D tensor into a vector. The first fully connected layer (FC1) consists of 1000 neurons and the last fully connected layer (FC2) is a classifier layer with 6 or 7 neurons corresponding to 6 or 7 emotions. The number of emotions in the EMO-DB and eNTERFACE05 datasets is 7 and 6, respectively. In the first fully connected layer and convolutional layers, rectified linear unit (ReLU) activation function is used as follows:

$$ReLU\ (x_i) = \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases} \tag{4}$$

where, x_i is the i^{th} input to the convolutional layer. Also, in the classifier layer or the last fully connected layer, softmax activation function is employed as follows:

$$softmax\ (z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where, z_i and K refer to the values and total number of the neurons from the last layer in the 3D CNN model. The parameters of the proposed 3D CNN model are shown in Table 1.

TABLE 1. The details of each layer parameters of the proposed 3D CNN model.

Layer	Feature Map	Size	Kernel Size	Stride	Activation Function
Input	1	256×256×256	-	-	-
Conv3D_1	64	256×256×256×64	3×3×3	1	ReLU
Max Pooling3D_1	64	128×128×128×64	2×2×2	2	-
Conv3D_2	128	128×128×128×128	3×3×3	1	ReLU
Max Pooling3D_2	128	64×64×64×128	2×2×2	2	-
Conv3D_3	256	64×64×64×256	3×3×3	1	ReLU
Max Pooling3D_3	256	32×32×32×256	2×2×2	2	-
Conv3D_4	512	32×32×32×512	3×3×3	1	ReLU
Max Pooling3D_4	512	16×16×16×512	2×2×2	2	-
Conv3D_5	512	16×16×16×512	3×3×3	1	ReLU
Max Pooling3D_5	512	8×8×8×512	2×2×2	2	-
Fully Connected_1	-	1000	-	-	ReLU
Fully Connected_2	-	6 OR 7	-	-	Softmax

C. SPEECH EMOTION RECOGNITION BASED ON GENDER

Because of the key role of τ in reconstructed phase spaces and mutual information analysis, an interesting idea to obtain a better accuracy lead us to compute τ based on the gender recognition. To that end, the speech signals from men and women are independently modelled in each dataset. The results show that there is a considerable difference between average of τ for men and women in the EMO-DB and eNTERFACE05 datasets. Figure 6 proposes the structure of a schematic diagram for speech emotion recognition based on gender recognition. As depicted, gender recognition has

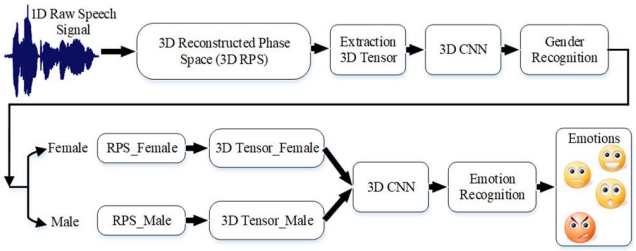

FIGURE 6. The Schematic diagram of the proposed method with considering gender recognition.

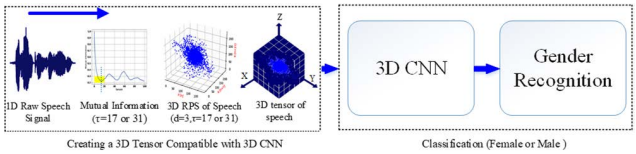

FIGURE 7. The schematic of the proposed method for gender recognition based on the average mutual information on the datasets.

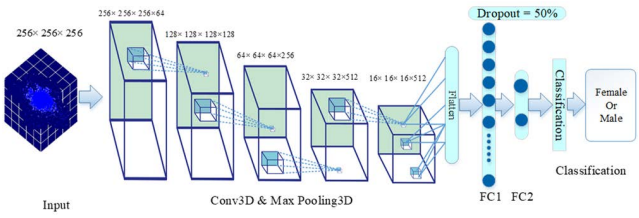

FIGURE 8. The proposed 3D CNN architecture for gender recognition.

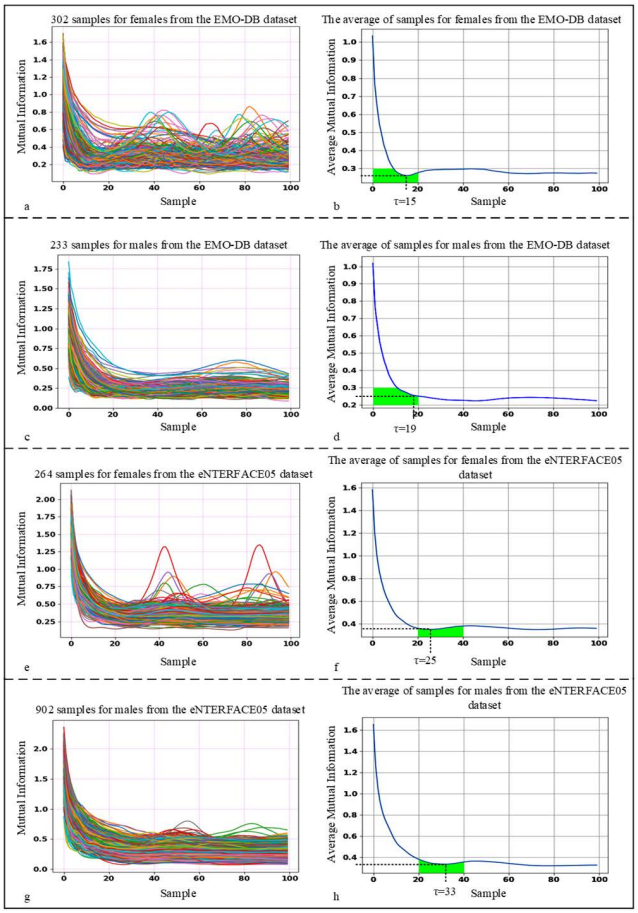

FIGURE 9. The mutual information of the speech signals for females and males in the EMO-DB and eNTERFACE05 datasets. (a), (b) 302 mutual information of speech samples for females from EMO-DB with the first minimum of average mutual information $\tau = 15$. (c), (d) 233 mutual information of speech samples for males from EMO-DB with the first minimum of average mutual information $\tau = 19$. (e), (f) 264 mutual information of speech samples for females from eNTERFACE05 with the first minimum of average mutual information $\tau = 25$. (g), (h) 902 mutual information of speech samples for males from eNTERFACE05 with the first minimum of average mutual information $\tau = 33$.

been added to the proposed method. To this end, the proposed 3D CNN should classify the input speech signals based on genders into two categories, including female and male. Then, the reconstructed phase spaces and 3D tensors have been extracted from females' and males' speech signals, separately.

As illustrated in Figure 7, a 3D tensor made from a 1D raw speech signal has been applied to the 3D CNN for classifying females or males. In this stage, the optimal time delays have been put on $\tau = 17$ for the EMO-DB dataset and $\tau = 15$ for the eNTERFACE05 dataset similar to the same way used in this work for emotion recognition without considering genders.

The block diagram of the proposed 3D CNN model for gender recognition is shown in Figure 8. The parameters of this 3D CNN are the same described in Table 1. The first fully connected layer (FC1) consists of 1000 neurons and the last fully connected layer (FC2) is a classifier layer with 2 neurons corresponding to female or male.

The first minimum of overall mutual information as the optimum time delay for female and male speech samples in the EMO-DB and eNTERFACE05 datasets are shown in Figure 9. Figure 9 (a) exhibits 302 mutual information of speech samples from females in the EMO-DB dataset, and Figure 9 (b) shows the corresponding first minimum of average mutual information $\tau = 15$. Figure 9 (c) displays 233 mutual information of speech samples from males in the EMO-DB dataset, and Figure 9 (d) shows the corresponding $\tau = 19$. By the similar way and as illustrated in Figure 9 (e)-(h), the first minimum of average mutual information for females and males in the eNTERFACE05 dataset are located on $\tau = 25$ and $\tau = 33$, respectively. In this dataset, 264 mutual information of speech samples from women and 902 mutual information of speech samples from men are considered.

Figure 10 shows a general overview of the suggested 3D CNN for the speech emotion recognition with gender designation. By obtaining the optimal time delays based on the gender recognition, the 3D tensors from females and males are employed to the propounded 3D CNN as inputs to classify various emotions in each dataset.

III. EXPERIMENTAL RESULTS AND COMPARISON

The recommended approach has been evaluated on two public datasets titled EMO-DB [37] and eNTERFACE05 [38].

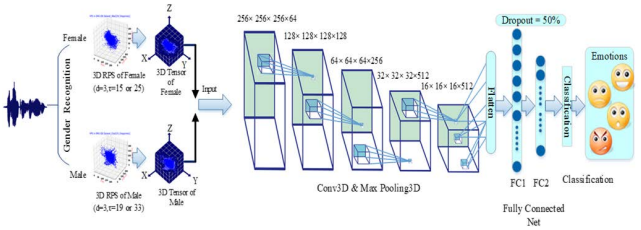

FIGURE 10. The overview of the proposed 3D CNN based on gender recognition for speech emotion recognition.

The EMO-DB database is an emotional speech dataset contains 535 utterances by 10 German actors (5 women and 5 men) in 7 emotions (sadness, fear, happiness, boredom, anger, disgust, and neutral). The eNTERFACE05 database is an emotional audio-visual dataset which is created by 42 people from 14 different nationalities. All the participants spoke English and most of them were male (19% women and 81% men). It includes 6 basic emotions (surprise, fear, happiness, disgust, sadness, and anger). This dataset contains 1166 video samples.

In this work, workstation hardware with specifications of Intel Core i7-7500U CPU, NVIDIA GeForce GTX 960M (4GB) GPU, and RAM-16GB DDR4 is used to design and evaluate the proposed 3D CNN model. Python was utilized for all implementations which are conducted in the Spyder platform under the Anaconda environment. Skedm library was applied for creating phase space reconstruction. Keras library that is running on the top of TensorFlow framework was employed to design the 3D CNN model. CUDA Toolkit v8 and cuDNN v6 were utilized libraries for GPU executions in the TensorFlow framework. A dropout technique with a rate of 0.5 has been adopted to the fully connected layers. This technique avoids the risk of overfitting by temporarily removing the neurons from each layer. The choice of which neurons to drop is random. The proposed 3D CNN was trained on the training set, with categorical cross-entropy as the loss function and Adam as the optimizer algorithm with learning rate of 1e-4 ($lr = 0.0001, beta\_1 = 0.9, beta\_2 = 0.999$). Our tests are performed on EMO-DB and eNTERFACE05 datasets based on the 3D CNN architecture using cross-validation strategy titled speaker-independent. In speaker-independent strategy, the test-runs are executed by applying leave-one-speaker-out (LOSO) and leave-one-speakers-group-out (LOSGO) schemes. When there are a few people in a dataset, the LOSO technique is chosen and if there are many people in a dataset, the LOSGO technique is selected. Thus, the LOSO scheme is applied for EMO-DB and the LOSGO scheme is employed for eNTERFACE05 [41].

As discussed, the gender recognition has been used in this research. For evaluating the accuracy of recognizing females and males, k-fold cross-validation technique has been employed. In this technique, the database is divided into K sections and in each iteration, $K-1$ sections are selected for training and one section is chosen as a test. This process is repeated until all samples participate in the training and testing. Although defining K parameter is arbitrary, it is commonly considered to 10 in which 9 parts of the datasets are given for training and 1 remained part is assigned for testing. Table 2 and Table 3 show the average accuracy of gender recognition for $d = 3$ and different time delays τ. The highest accuracy is available on $\tau = 17$ for EMO-DB by 99.06% and $\tau = 31$ for eNTERFACE05 by 98.28%, as indicated in both tables. These results have been attained in 10 iterations by randomizing the datasets and prove that the presented method for females' and males' speech signals recognition is remarkably reliable. The symbols $\pm$ refer to the standard deviation.

TABLE 2. Accuracy of the gender recognition for EMO-DB (535 Samples).

Time Delay	$\tau = 1$	$\tau = 2$	$\tau = 10$	$\tau = 15$	$\tau = 17$	$\tau = 25$	$\tau = 50$
Average Error	16 ± 1	15 ± 1	11 ± 1	8 ± 1	5 ± 1	22 ± 1	47 ± 1
Average Accuracy	97.01% ± 0.2	97.20% ± 0.2	97.94% ± 0.2	98.50% ± 0.2	99.06% ± 0.2	95.89% ± 0.2	91.21% ± 0.2

TABLE 3. Accuracy of the gender recognition for eNTERFACE05 (1166 Samples).

Time Delay	$\tau = 1$	$\tau = 2$	$\tau = 10$	$\tau = 15$	$\tau = 17$	$\tau = 31$	$\tau = 50$
Average Error	42 ± 3	38 ± 2	35 ± 2	25 ± 2	27 ± 2	22±2	120 ± 7
Average Accuracy	96.74% ± 0.3	97.05% ± 0.2	97.27% ± 0.2	98.06% ± 0.2	97.90% ± 0.2	98.28% ± 0.3	90.70% ± 0.5

TABLE 4. Speaker-independent average speech emotion recognition accuracy with the SGD optimizer.

Method	Without Gender Recognition	With Gender Recognition
EMO-DB	76.81% ± 0.6	87.79% ± 0.5
eNTERFACE05	71.48% ± 0.8	80.38% ± 0.6

The average accuracy of the speech emotion recognition employing speaker-independent without and with gender recognition explained in the proposed method section has been demonstrated in Table 4. The used optimizer was stochastic gradient descent (SGD).

The datasets used in this research contain a limited number of samples, far fewer than the minimum requirements for a desirable training of a CNN, making it necessary to increase the number of samples. To address this problem, it is common to employ data augmentation techniques to increase the size of the dataset [26]. Data augmentation refers to any technique that increases the amount of data using original data. In speech emotion recognition, it can be performed by splitting each speech sample into several shorter segments. In the data augmentation procedure, all samples were split into 315ms segments [26], which are greater than the minimum required length of 250ms for emotion recognition, recommended by [40]. All new samples were labeled as the corresponding emotion of the original sample. These result in 11629 segments obtained from 535 EMO-DB utterances and 25712 segments achieved from 1166 video samples of

TABLE 5. Speaker-independent average speech emotion recognition accuracy with the SGD optimizer and data augmentation.

Method	Without Gender Recognition + Data Augmentation	With Gender Recognition + Data Augmentation
EMO-DB	87.13% ± 0.6	92.68% ± 0.5
eNTERFACE05	80.46% ± 0.8	86.73% ± 0.6

TABLE 6. Speaker-independent average speech emotion recognition accuracy by various optimizers and data augmentation without gender recognition.

Optimizer	SGD	RMSprop	Adam	Adadelta	Adagrad	Adamax	Ftrl
EMO-DB	87.13% ± 0.6	88.11% ± 0.6	89.34% ± 0.6	87.35% ± 0.6	87.46% ± 0.6	87.23% ± 0.6	89.01% ± 0.6
eNTERFACE05	80.46% ± 0.8	80.91% ± 0.8	81.48% ± 0.8	80.90% ± 0.8	80.97% ± 0.8	80.44% ± 0.8	81.14% ± 0.8

TABLE 7. Speaker-independent average speech emotion recognition accuracy by various optimizers and data augmentation with gender recognition.

Optimizer	SGD	RMSprop	Adam	Adadelta	Adagrad	Adamax	Ftrl
EMO-DB	92.68% ± 0.5	93.58% ± 0.5	94.19% ± 0.5	92.91% ± 0.5	93.15% ± 0.5	93.08% ± 0.5	93.92% ± 0.5
eNTERFACE05	86.73% ± 0.6	87.35% ± 0.6	88.24% ± 0.6	87.80% ± 0.6	87.91% ± 0.6	87.65% ± 0.6	87.94% ± 0.6

TABLE 8. Speaker-independent average speech emotion recognition accuracy by employing dropout technique without gender recognition.

Method	No Technique	Data Augmentation	Adam optimizer	Dropout Technique
EMO-DB	76.81% ± 0.6	87.13% ± 0.6	89.34% ± 0.6	90.40% ± 0.2
eNTERFACE05	71.48% ± 0.8	80.46% ± 0.8	81.48% ± 0.8	82.20% ± 0.2

TABLE 9. Speaker-independent average speech emotion recognition accuracy by employing dropout technique with gender recognition.

Method	No Technique	Data Augmentation	Adam optimizer	Dropout Technique
EMO-DB	87.79% ± 0.5	92.68% ± 0.5	94.19% ± 0.5	94.42% ± 0.2
eNTERFACE05	80.38% ± 0.6	86.73% ± 0.6	88.24% ± 0.6	88.47% ± 0.2

eNTERFACE05. This data augmentation approach, in addition to a dropout procedure, can effectively increase the test accuracy rate by reducing the risk of overfitting. Furthermore, a proper optimization technique helps to gain the highest possible accuracy rate using the proposed 3D model. Accordingly, the average recognition accuracy has considerably risen on the presented datasets for speaker-independent strategy. Table 5 demonstrates the average accuracy of the speech emotion recognition with the SGD optimizer and data augmentation.

There are various algorithms such as SGD [49], RMSprop [50], Adam [51], Adamax [51], Adadelta [52], Adagrad [53], and Ftrl [54] are used to minimize and optimize errors. All the mentioned algorithms are evaluated in the proposed method with data augmentation technique. As shown in Table 6 and Table 7, the average recognition accuracy of the speech emotions without and with considering gender recognition has been compared by employing different optimizers with data augmentation. The results prove the best possible solution is achieved by the Adam optimizer.

TABLE 10. Comparison of the speaker-independent average emotion recognition accuracy (%) of the proposed 3D CNN models with other published works.

References	Average Recognition Accuracy (%)	
	EMO-DB	eNTERFACE05
[26]	87.31	79.25
[32]	Not Reported	72.33
[33]	82.82	Not Reported
[39]	91.78	Not Reported
[40]	81.90	61.10
[41]	85.60	72.40
[42]	Not Reported	72.95
[43]	85.20	Not Reported
[44]	90.09	Not Reported
[45]	92.45	Not Reported
[46]	Not Reported	89.60
[47]	82.73	Not Reported
Proposed 3D CNN without Gender Recognition	90.40	82.20
Proposed 3D CNN with Gender Recognition	94.42	88.47

TABLE 11. Comparison of the speedups of the CPU and GPU executions.

Hardware Platform	Model	Dataset	Runtime
Intel CPU	3D CNN without Gender Recognition	EMO-DB	8:03 Hours
NVIDIA GPU		EMO-DB	6:04 Hours
Intel CPU		eNTERFACE05	9:31 Hours
NVIDIA GPU		eNTERFACE05	7:14 Hours
Intel CPU	3D CNN with Gender Recognition	EMO-DB	12:20 Hours
NVIDIA GPU		EMO-DB	9:30 Hours
Intel CPU		eNTERFACE05	14:30 Hours
NVIDIA GPU		eNTERFACE05	11:03 Hours

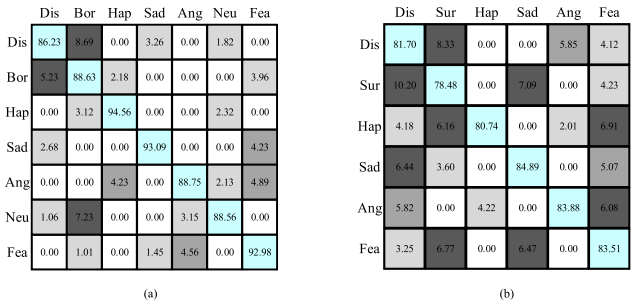

FIGURE 11. Confusion matrixes of the proposed 3D CNN without gender recognition. (a) An average accuracy of 90.40% on the EMO-DB dataset. (b) An average accuracy of 82.20% on the eNTERFACE05 dataset.

Finally, the dropout technique is used for reaching a better training, avoids overfitting, and increases the recognition rates. Table 8 and Table 9 represent the average accuracy rates of the speech emotion recognition when dropout technique is employed. These tables show the highest recognition accuracy rates with the lowest tolerance have been obtained just after applying dropout in the fully connected layers.

Figure 11 shows the confusion matrixes of the presented 3D CNN for speech emotion recognition experiments on the datasets without specifying gender recognition. In the confusion matrix, while each row represents the goal emotion,

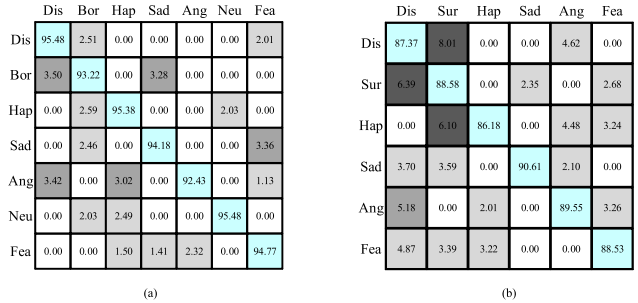

FIGURE 12. Confusion matrixes of the proposed 3D CNN with gender recognition. (a) An average accuracy of 94.42% on the EMO-DB dataset. (b) An average accuracy of 88.47% on the eNTERFACE05 dataset.

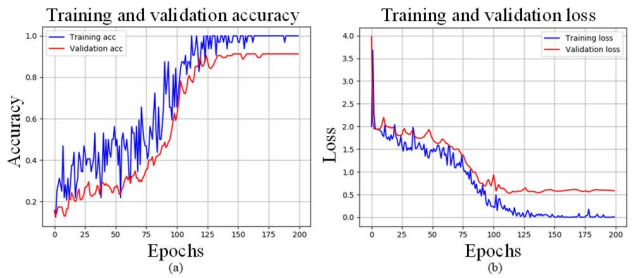

FIGURE 13. Accuracy and loss parameters on the EMO-DB database without gender recognition. (a) Training and validation accuracy per epoch. (b) Training and validation loss per epoch.

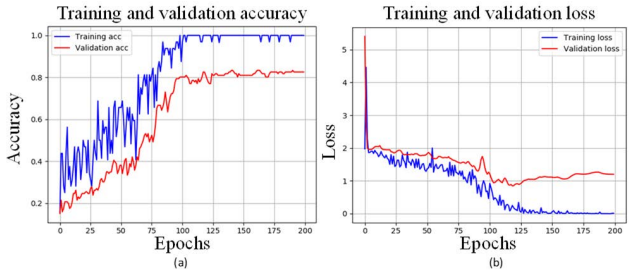

FIGURE 14. Accuracy and loss parameters on the eNTERFACE05 database without gender recognition. (a) Training and validation accuracy per epoch. (b) Training and validation loss per epoch.

columns demonstrate the recognized emotions. The diagonal line of the matrix shows the recognition rate of each emotion. It is clear from Figure 11 (a) that happiness with the recognition accuracy of 94.56% has the highest accuracy and disgust with the recognition accuracy of 86.23% has the lowest accuracy. The average recognition accuracy of 90.40% has been achieved on EMO-DB dataset. Similarly, and as understood from Figure 11 (b), sadness has the best recognition accuracy of 84.89% and surprise has the worst recognition accuracy of 78.48%.

The average recognition rate of 82.20% was obtained on eNTERFACE05 dataset. Comparing Figure 11 (a) and Figure 11 (b) reveals that the most emotional misclassification rates are 8.69% between disgust and boredom for the EMO-DB dataset and 10.20% between surprise and disgust for the eNTERFACE05 dataset.

Figure 12 demonstrates the confusion matrixes of the presented 3D CNN for speech emotion recognition experiments on the datasets with specifying gender recognition. As realized from Figure 12 (a), the maximum recognition accuracy of 95.48% has been dedicated to disgust and the minimum recognition accuracy of 92.43% has been allocated to anger. The average recognition accuracy of 94.42% has been obtained on the EMO-DB dataset. By similar way from Figure 12 (b), the most recognition accuracy of 90.61% is devoted to sadness and the least recognition accuracy of 86.18% is allotted to happiness. The average recognition rate of 88.47% has been obtained on eNTERFACE05 dataset. As can be seen from Figure 12 (a) and Figure 12 (b), the foremost emotional misclassification rates are 3.50% between boredom and disgust for the EMO-DB database and 8.01% between disgust and surprise for the eNTERFACE05 database.

Concerning Figure 11 and Figure 12, the number of zero values for gender-based speech emotion recognition is higher than the one for speech emotion recognition without considering gender. Hence, the larger number of zero value cells in the matrixes from Figure 12, proves that the classification task has been more accurately done. The accuracy and loss factors help to authenticate the consequences of our work. The superb training and validation accuracy are acquired when converging upwards, and the supreme training and validation loss are gained while converging downwards. Figure 13 describes the accuracy and loss parameters for speech emotion recognition on the EMO-DB database in speaker-independent experiments without defining gender recognition over 200 iterations. The rise in training and validation accuracy has been presented in Figure 13 (a). The fall in training and validation loss has been shown in Figure 13 (b). Figure 14 demonstrates the accuracy and loss parameters for speech emotion recognition on the eNTERFAC05 database in speaker-independent experiments without determining gender recognition over 200 iterations. Figure 14 (a) shows the training and validation accuracy increasing, while Figure 14 (b) exhibits the training and validation loss decreasing. The analogous analysis is expected for explaining Figure 15 and Figure 16 which illustrates the accuracy and loss parameters with considering gender recognition. As comprehended from Figure 13, Figure 14, Figure 15, and Figure 16, the training and validation accuracy of the proposed 3D CNN with regard to genders for EMO-DB and eNTERFAC05 are superior to the state without regard to genders. Furthermore, the training and validation loss from the datasets in the suggested gender-based 3D CNN are notably lower than the status without gender consideration.

For showing the differences when gender recognition is added to the proposed method, two bar charts have been drawn. The graphs have categorized the average accuracy of the speech emotion recognition without and with considering

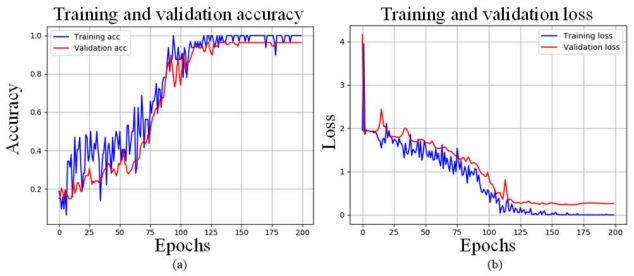

FIGURE 15. Accuracy and loss parameters on the EMO-DB database with gender recognition. (a) Training and validation accuracy per epoch. (b) Training and validation loss per epoch.

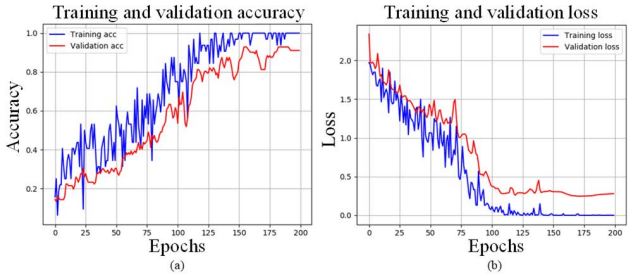

FIGURE 16. Accuracy and loss parameters on the eNTERFACE05 database with gender recognition. (a) Training and validation accuracy per epoch. (b) Training and validation loss per epoch.

gender recognition on each dataset. Figure 17 depicts the average rate for EMO-DB and Figure 18 illustrates the average rate for eNTERFACE05. In Figure 17, the magenta colour is specified to the average rate without gender recognition orientation and the cyan colour is defined to the average rate with gender recognition direction. In Figure 18, the green colour shows the average rate without recognizing genders and the blue colour displays the average rate with recognizing genders. As perceived from comparing Figure 17 and Figure 18, the improvements of average accuracy by employing gender recognition technique on the eNTERFACE05 dataset are more sensible than the effects on the EMO-DB dataset. For example, the amelioration of average accuracy for happiness and sadness emotions on EMO-DB is negligible; however, the enhancements of average accuracy for those emotions on eNTERFACE05 are completely visible.

Table 10 compares the speaker-independent average speech emotion recognition accuracy of the proposed 3D CNN without and with designating gender recognition technique in this work with the related publications and proves that our results are greatly substantial. [26] discussed an approach comprising a combination of a deep convolutional neural network with a discriminant temporal pyramid matching strategy for automatic affective feature learning to recognize speech emotions with the accuracy rates of 87.31% on EMO-DB and 79.25% on eNTERFACE05. [32] presents a 3DCNN including two convolutional layers and one fully connected layer applying k-means clustering and spectro-

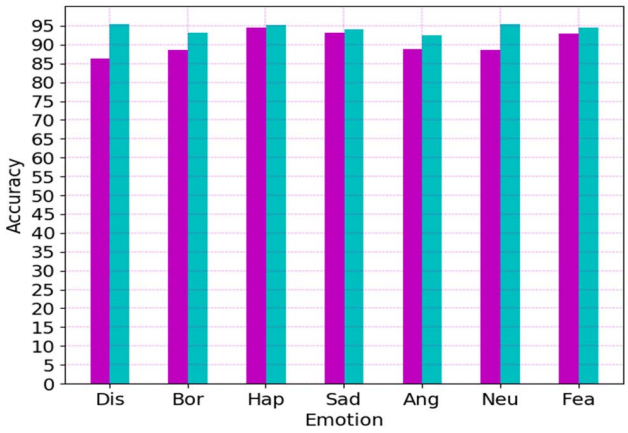

FIGURE 17. Average accuracy of the speech emotions on EMO-DB. The magenta colour refers to the average rate without gender recognition and the cyan colour refers to the average rate with gender recognition.

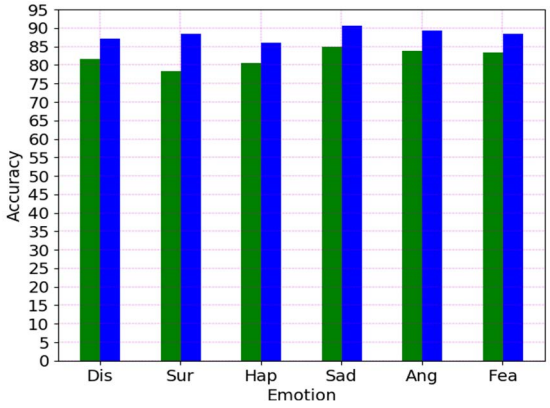

FIGURE 18. Average accuracy of the speech emotions on eNTERFACE05. The green colour refers to the average rate without recognizing genders and blue colour refers to the average rate with recognizing genders.

grams techniques in parallel for speech emotion recognition on the eNTERFACE05 dataset, obtaining the accuracy rate of 72.33%. In [33], a 3D attention-based convolutional recurrent neural network for speech emotion recognition on the EMO-DB database with the accuracy rate of 82.82% by extracting features known static, deltas, and delta-deltas from the speech signals employing as the input to the 3D convolutional network has been introduced. [39] suggested a hybrid CNN involving 1D CNN and 2D CNN. In this method, a 1D CNN and a 2D CNN were designed and then merged together. Moreover, transfer learning was employed to speed up the training process in the merged CNN. The emotion recognition rate on the EMO-DB was 91.78%. In [40], a generalized discriminant analysis (GerDA) on the base of deep neural networks was recommended for acoustic emotion recognition. The accuracy rates on the EMO-DB and eNTERFACE05 corpuses were obtained 81.90% and 61.10%, respectively. As explained from [41], two standard toolkits, frame-level by means of hidden Markov model and supra-segmental modeling using openEAR, have been applied for emotion

recognition task. The accuracy of 85.60% for EMO-DB and 72.40% for eNTERFACE05 in supra-segmental modeling was much better than the accuracy rates on the both corpora in frame-level modeling. In [42], BAUM-1 was presented as a new spontaneous audio-visual Turkish database and a multi-modal affective recognition algorithm according to apex frame selection was utilized. The experiments on the BAUM-1s and eNTERFACE05 datasets for audio emotion recognition were 29.41% and 72.95%. [43] proposed to learn emotion-salient features using semi-CNN with the recognition accuracy of 85.20% on the EMO-DB database. [44] introduced a method to utilization of the shuffle box cryptographic structure for feature generation and iterative neighborhood component analysis for feature selection to recognize the emotions from speech with the accuracy rate of 90.09% on the EMO-DB dataset. [45] discussed a bagged ensemble of support vector machines with a Gaussian kernel for the purpose of recognizing speech emotions. The accuracy rate on the EMO-DB corpus was 92.45%. [46] explained an approach for speech emotion recognition that combines attention-based long short-term memory (LSTM) recurrent neural networks with frame-level speech features. The accuracy rate of the emotion recognition on the eNTERFACE05 database was 89.60%. Due to the complexity of the method used in [46], its accuracy is slightly better than our work. Finally, [47] described a deep neural network trained by multi-conditioning and data augmentation employing Generative noise model to address the resilience of the speech emotion recognition with the accuracy rate of 82.73% on the EMO-DB dataset. According to the comparison of this research with the state of the arts as understood from Table 10, our suggested methods are remarkably worthwhile.

The comparison of the speedups between CPU- and GPU-based 3D CNN model executions is presented in Table 11. It shows the GPU speeds up the 3D CNN model without gender recognition by $\approx 1.33\times$ and $\approx 1.30\times$ faster for the EMO-DB and eNTERFACE05 datasets than the CPU-based running in our work. Besides and for the 3D CNN with gender recognition, the GPU implementations present shorter execution times by $\approx 1.31\times$ and $\approx 1.30\times$ faster for the EMO-DB and eNTERFACE05 dataPbases than the CPU-based implementations.

IV. CONCLUSION

This research presents a 3D CNN for speech emotion recognition application. The suggested 3D CNN directly employ the generated 3D tensors from the reconstructed phase space of speech signals. The results on EMO-DB and eNTERFACE05 datasets show that the proposed 3D tensors contain essential emotional cues of the speakers and consequently the 3D CNN can effectively and accurately classify the corresponding emotions. Employing gender recognition technique to the proposed 3D CNN conducts to the noteworthy accuracy rates on the datasets. Finally, a GPU has been applied to expedite the 3D CNN on the datasets by providing lower runtimes than the CPU executions.

ACKNOWLEDGMENT

The authors would like to give their very great thankfulness to Dr. John McAllister from the Institute of Electronics, Communications and Information Technology (ECIT), Queen's University of Belfast, U.K., for his precious technical and scientific suggestions during making this research work. His generous attention is so much appreciated.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[2] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. Hoboken, NJ, USA: Wiley, 2018, p. e1253.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] J. Turner, J. Cano, V. Radu, E. J. Crowley, M. O'Boyle, and A. Storkey, "Characterising across-stack optimisations for deep convolutional neural networks," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Sep. 2018, pp. 101–110.

[5] M. P. Véstias, "A survey of convolutional neural networks on edge with reconfigurable computing," *Algorithms*, vol. 12, p. 154, Jul. 2019.

[6] J. Wang, Y. Li, J. Shan, J. Bao, C. Zong, and L. Zhao, "Large-scale text classification using scope-based convolutional neural network: A deep learning approach," *IEEE Access*, vol. 7, pp. 171548–171558, 2019.

[7] J. Wen, X. Zhou, P. Zhong, and Y. Xue, "Convolutional neural network based text steganalysis," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 460–464, Mar. 2019.

[8] X. Li, Z. Yang, and H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," *IEEE Access*, vol. 8, pp. 174922–174930, 2020.

[9] R. Qi, R.-S. Jia, Q.-C. Mao, H.-M. Sun, and L.-Q. Zuo, "Face detection method based on cascaded convolutional networks," *IEEE Access*, vol. 7, pp. 110740–110748, 2019.

[10] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[11] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.

[12] D. Das, D. R. Nayak, R. Dash, B. Majhi, and Y. Zhang, "H-WordNet: A holistic convolutional neural network approach for handwritten word recognition," *IET Image Process.*, vol. 14, no. 9, pp. 1794–1805, Jul. 2020.

[13] D. Mellouli, T. M. Hamdani, J. J. Sanchez-Medina, M. B. Ayed, and A. M. Alimi, "Morphological convolutional neural network architecture for digit recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2876–2885, Sep. 2019.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[15] M. Ha, Y. Byun, J. Kim, J. Lee, Y. Lee, and S. Lee, "Selective deep convolutional neural network for low cost distorted image classification," *IEEE Access*, vol. 7, pp. 133030–133042, 2019.

[16] M. Jubran, A. Abbas, A. Chadha, and Y. Andreopoulos, "Rate-accuracy trade-off in video classification with deep convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 145–154, Jan. 2020.

[17] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 9, pp. 1806–1819, Sep. 2019.

[18] A. Deshpande. (2018). *A Beginner's Guide To Understanding Convolutional Neural Networks*. [Online]. Available: https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

[19] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[20] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
[21] P. Song, Y. Jin, C. Zha, and L. Zhao, "Speech emotion recognition method based on hidden factor analysis," *Electron. Lett.*, vol. 51, no. 1, pp. 112–114, 2015.
[22] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.
[23] C. Shi, C. Tan, and L. Wang, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," *IEEE Access*, vol. 9, pp. 39255–39274, 2021.
[24] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, Apr. 2021.
[25] N.-C. Ristea, L. C. Dutu, and A. Radoi, "Emotion recognition system from speech and visual information based on convolutional neural networks," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–6.
[26] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Oct. 2017.
[27] A. Bakhshi, A. Harimi, and S. Chalup, "CyTex: Transforming speech to textured images for speech emotion recognition," *Speech Commun.*, vol. 139, pp. 62–75, Apr. 2022.
[28] 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, Syst., Signal Process.*, pp. 1–44, Aug. 2022.
[29] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3D convolutional neural networks," 2017, *arXiv:1705.09422*.
[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[31] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[32] N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using K-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, May 2019.
[33] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[34] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts, and J. Ye, "Statistical models of reconstructed phase spaces for signal classification," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2178–2186, Jun. 2006.
[35] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, May 2012.
[36] S. Wallot and D. Mønster, "Calculation of average mutual information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in MATLAB," *Frontiers Psychol.*, vol. 9, p. 1679, Sep. 2018.
[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.
[38] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Atlanta, GA, USA, 2006, p. 8.
[39] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Process.*, vol. 12, no. 6, pp. 713–721, 2018.
[40] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.
[41] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2009, pp. 552–557.
[42] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 300–313, Jul./Sep. 2016.
[43] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 801–804.
[44] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106547.
[45] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.
[46] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1675–1685, Jul. 2019.
[47] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7194–7198.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[49] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *J. Mach. Learn. Res.*, vol. 23, no. 3, pp. 1139–1147, 2013. [Online]. Available: http://jmlr.org/proceedings/papers/v28/sutskever13.html
[50] [Online]. Available: https://keras.io/api/optimizers/
[51] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.
[52] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
[53] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.
[54] H. B. Mcmahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica, "Ad click prediction: A view from the trenches," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 1222–1230.

MOHAMMAD REZA FALAHZADEH received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Iran, in 2007, 2010, and 2020, respectively. He is currently an Assistant Professor at Islamic Azad University, Iran. He has designed and manufactured more than 15 smart innovation devices at Iran's electricity industry. His research interests include signal processing, image processing, speech processing, deep learning networks, and smart microcontroller systems.

EDRIS ZAMAN FARSA received the B.Sc. and M.Sc. degrees in electronics engineering from Islamic Azad University, Arak Branch and Kermanshah Science and Research Branch, Iran, in 2009 and 2014, respectively. His main research interests include hardware/software co-design of signal processing systems, FPGAs, neuromorphic computing, and embedded machine learning. He has served as a Reviewer for some IEEE journals like IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, and IEEE conferences such as ISCAS and MWSCAS.

ALI HARIMI received the B.Sc. degree in electronics engineering from the Sadjad University of Technology, Iran, in 2006, the M.Sc. degree in electronics from the Shahrood University of Technology, Iran, in 2009, and the Ph.D. degree in telecommunication systems from Semnan University, Iran, in 2015. Currently, he is an Assistant Professor at Islamic Azad University, Shahrood Branch, Iran. His research interests include signal/image processing, computer vision, and machine learning, and in particular affective computing.

ARASH AHMADI (Senior Member, IEEE) received the B.Sc. degree in electronics engineering from the Sharif University of Technology, Tehran, Iran, in 1993, the M.Sc. degree from Tarbiat Modares University, Tehran, in 1997, and the Ph.D. degree in electronics from the University of Southampton, Southampton, U.K., in 2008. He was a Faculty Member with Razi University, Kermanshah, Iran. From 2008 to 2010, he was a Fellow Researcher with the University of Southampton. Then, he was an Associate Professor with the Electrical Engineering Department, Razi University, and a Visiting Scholar with the University of Windsor, Windsor, ON, Canada. Now, he is a Faculty Member with Carleton University, Ottawa, Canada. His research interests include neuromorphic, bio-inspired computing, memristors, hardware implementation of signal processing systems, hardware security, and the IoT.

AJITH ABRAHAM (Senior Member, IEEE) received the Master of Science degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. The Network with HQ, Seattle, USA, is currently more than 1,500 scientific members from over 105 countries. As an Investigator/a Co-Investigator, he has won research grants worth over more than 100 Million U.S.$. Currently, he holds two university professorial appointments. He works as a Professor in artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in artificial intelligence at UCSI, Malaysia. He works in a multidisciplinary environment. He has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as per Google Scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), from 2008 to 2021, and served as a Distinguished Lecturer for the IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and serves/served on the editorial board for over 15 international journals indexed by Thomson ISI.