Received: May 6, 2023. Revised: July 8, 2023. 350

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Speaker Identification System Employing Multi-resolution Analysis in Conjunction with CNN

Huda W. Al-Dulaimi[1]* Ahmed Aldhahab[1] Hanaa M. Al Abboodi[1]

Department of Electrical Engineering, University of Babylon, Iraq
* Corresponding author's Email: HudaWasfi61@gmail.com

Abstract: The acoustic features extracted from the speech-signal are a critical challenge for implementing an accurate speaker identification system. In this paper, two-dimension discrete multi-wavelet transform (2D-DMWT) in conjunction with the deep learning neural networks are proposed for speaker identification. The DMWT is based on a vital sampling scheme preprocessing that uses the filter invented by Geronimo, Hardian, and Massopust, which is call GHM. The system proposed involves firstly preprocessing in which the speech-signal is resampled into 16kHz. Then, the speech-signal is divided to five different durations: 0.5 sec., 1 sec., 2 sec., 3 sec., and 5 sec. In this paper, each duration is tested separately. Second, 2D-DMWT is employed to obtain discriminant features from the speech-signal and reduce speech-signal dimensions in the feature's selection phase. Finally, neural network algorithm based on convolution neural network (CNN) is used for classification. The system proposed is tested using four databases: SALU-AC, ELSDSR, TIMIT, and RAVDESS. These databases include various speech variances, such as age, gender, etc. The results obtained by the proposed system are 95.86%, 96.59%, 89.90%, and 89.83% for 0.5sec of the SALU-AC, ELSDSR, RAVDESS, and TIMIT databases, respectively. For 1sec, the SALU-AC, ELSDSR, RAVDESS, and TIMIT databases obtained 96.30%, 97.31%, 96.05%, and 93.59%, respectively. The SALU-AC, ELSDSR, RAVDESS, and TIMIT databases achieved 96.63%, 97.76%, 96.12%, and 95.90%, respectively, over the 2sec time duration. During the time duration of 3sec, the SALU-AC, ELSDSR, and RAVDESS databases obtained 97.04%, 98%, and 97.96%, respectively. For 5sec, the SALU-AC, and ELSDSR databases attained 97.56%, and 98.30%, respectively. The results accomplished by the proposed system are outperformed those results discussed in the previous works based on the same databases.

Keywords: Speaker identification, 2D-DMWT, GHM.

1. Introduction

Speaker recognition is a biometric system that uses specific characteristics derived from voice utterances to authenticate users' individuality. Recognizing a person based on speech utterances is known as speaker recognition. Speaker recognition is a crucial task in speech processing, and it has wide applications in security. For example, speaker recognition is used to identify personal intelligent devices such as cellular phones, cars, and laptop computers. It ensures the safety of bank transactions and distant payments [1].

There are two types of speaker recognition: speaker identification (SI) and speaker verification (SV). The term "speaker verification" refers to verifying speakers' identity based on information contained in the speech signal to verify that the client is the one who claimed to be, which is a 1:1 question of confirmation. On the other hand, Speaker identification refers to determining the identity of anonymous speakers. It is a 1: N classification issue [2]. Open-set and closed-set systems are the two main categories of speaker recognition systems. A "closed-set" system only allows for a certain number of speakers to be registered on the system at any given moment. The machine can identify a specific human voice from a recording in this technique. In another respect, the term "open-set" refers to a system intended to function with any number of trained speakers, given that the anonymous speech might originate from a diverse group of unknown speakers

[3]. The closed-set speaker identification system is considered in this study by employing 2D-DMWT to extract the features of the speech signal.

As mentioned above, speaker recognition plays a vital role in security. For this reason, it has become the subject of study by many researchers. Many works have been published on this topic; some are briefly explained below. The authors in [4] suggested an emotional speaker identification method based on both machine learning (ML) and deep neural network (DNN) models. The authors extracted the features of the speech signal using a variety of approaches. Several ML and DNN models were employed to enhance the classification. The suggested system was assessed using the "RAVDESS" database, which contains eight emotions recorded by 24 speakers. After the models were assessed, DNN beat ML models in terms of accuracy. Hence, the system employing DNN was accomplished 92% of accuracy in comparison to the system using ML that achieved 88% of accuracy. The authors in [5] suggested a speaker identification system based on Random Forest. For the feature extraction phase, the authors used and compared two techniques, which are MFCC and Reconstructed Phase Space (RPS). Random Forest was used in the classification phase. They tested their approach on 38 speakers from the TIMIT database. The system demonstrated superior performance in the case of MFCC as compared to RPS. The authors proposed a speaker identification system based on a modified Support Vector Machine (SVM) as a classifier to enhance the degraded speaker identification performance for disguised voices under an extremely high-pitched condition in a neutral talking environment [6]. For the feature extraction phase, MFCC was employed. The authors used a modified Support Vector Machine (SVM) for the classification phase. Their suggested approach was assessed using different speech datasets, including an Arabic Emirati-accented database, the SUSAS English database, and the RAVDESS database. Their system achieved a good recognition rate.

A speaker identification system based on Mel-frequency cepstral coefficient (MFCC) and deep neural network (DNN) was proposed by the authors in [7]. The authors employed data augmentation methods (DA) to increase the database size. For the feature extraction step, MFCC was employed. A seven-layer Deep Neural Network (DNN7L) was applied for the classification process. The suggested approach was tested using a database of Indonesian Regional Language-301 Languages spoken in Indonesian. The database included speakers of Indonesian from a variety of various backgrounds within Indonesia. Their system accomplished a high recognition rate. In 2021, the authors in [8] proposed a hybrid speaker identification system. Their method was designed to obtain constant speech features and accomplished high recognition rates. For the feature extraction phase, the authors employed MFCC, which they modified by adding a pitch frequency coefficient. The classification phase was implemented with a feed-forward neural network (FFNN) trained using an optimized particle swarm optimization (OPSO) technique. The authors collected 250 speech samples from different speakers and evaluated their system using 10-fold cross-validation. Their system attained a recognition rate of 97.83%.

A speaker identification system based on convolutional and recurrent neural networks was proposed in [9]. The authors built the deep learning architecture by combining convolutional neural networks and recurrent neural networks using long short-term memory models. Their method was tested by using 77 different non-native speakers who read the same text in Turkish. Their approach achieved an identification rate of 98%. The authors suggested approaches for speaker identification systems based on adaptive orthogonal transformations [10]. The authors proposed an adaptive operator to extract significant features from input signals with a minimum dimension. Dynamic time wrapping (DTW) was employed for the classification phase. The authors utilized a database of ten speakers to test their technique. The highest accuracies accomplished using the system-proposed in [10] were 96.8% and 98.1% when using Fourier transform and correlation method as a compression approach, respectively.

The authors presented a speaker identification system based on a convolutional neural network [11]. The authors employed a spectrogram, a graphical representation of speech regarding raw features, to identify speakers. These features were fed to CNN, and the speakers were identified using a CNN-visual geometry group (CNN-VGG) architecture. The system proposed accomplished 98.78% accuracy based on 78 speakers, each with 10 different samples. The authors suggested a two-branch network [12] as a way to improve the performance of the speaker identification system. This network would be able to extract features from both the face and the speech signals. The VGGFace and VGGVox subnetworks were employed to extract features from face and voice, respectively. The authors employed support vector machine (SVM) as a classifier. In order to evaluate the effectiveness of their approach, the authors used the VoxCeleb1 database, which is a large collection of audio-visual recordings of human

speech that were obtained "in the wild" from YouTube. The authors found that by including information about the user's face in their system, the system performance improved to 97.2% accuracy. A speaker recognition system based on an optimization-based support vector neural network was suggested by the authors in [13]. In the feature extraction phase, the authors employed frequency-dependent characteristics such as multiple kernel weighted Mel frequency cepstral coefficient (MKMFCC), spectral kurtosis, spectral skewness, and autocorrelation. An adaptive fractional bat-based support vector neural network (AFB-based SVNN) was used for the classification phase. The authors used the ELSDSR database to analyze their system. The results of their suggested approach attained a recognition rate of 95%.

The authors in [14] presented a speaker identification system based on fused spectral features with hybrid machine learning (ML). For the feature extraction phase, the authors implemented several spectrum features, such as mel-frequency cepstral coefficients (MFCCs), normalized pitch frequency (NPF), spectral kurtosis, formants, and spectral skewness. Random forest-support vector machine (RF-SVM) was applied for the classification phase. The authors analyzed their system by utilizing the ELSDSR database. Their suggested method achieved a recognition rate of 98.16 %. A speaker identification system based on squeeze-and-excitation (SE) components was proposed by the authors in [15]. For the feature extraction phase, MFCC was employed. A combination of squeeze-and-excitation (SE) components with a simplified residual convolutional neural network (ResNet) was used for the classification phase. Their approach was evaluated by utilizing the TIMIT and Librispeech databases. Their suggested approach achieved accuracies of 95.8% and 93.92%, respectively, when compared to the TIMIT and Librispeech databases.

In order to reduce the dimensions of the databases used and maintain the storage required by selecting discriminant features while achieving high recognition accuracies, this paper presents a closed-set speaker identification system based on multi-resolution analysis (2D-DMWT) and CNN. The three main phases of the suggested design are preprocessing, feature extraction, and classification. During the preprocessing phase, silence is removed from the speech signal, divided into segments of 0.5 sec., 1 sec., 2 sec., 3 sec., and 5 sec., and resampled to 16kHz. 2D-DMWT is used to extract meaningful features during the feature extraction step and compress the data. The essential sampling technique preprocessing that forms the foundation of this transform utilizes the GHM filter. CNN is used during the classification step. Four different speech databases that have various speech variations, namely, SALU-AC, ELSDSR, TIMIT, and RAVDESS are used to assess the proposed system. The system-proposed accomplished an accuracy that is superior to the other approaches presented in [4-6, 13-15]. The system proposed accomplished high dimensionality reduction, which leads to less storage requirement, in comparison with the other approaches discussed in [4-6, 13-15]. The dimensionality reduction is 93.75% achieved. The robust performance of the multi-wavelet transform is closely related to the availability of multi-wavelet properties, such as multiple vanishing moments, short support, orthogonality, and symmetry. The following outline constitutes the arrangement of the content of this paper: The background of the DMWT is described in section 2. In section 3, the proposed system will be discussed. The presentation of the speech databases may be found in section 4. Section 5 provides the results of the experiment and the discussion. Section 6 summarizes the conclusion and discusses the study's future employment prospects.

2. Background: DWT and DMWT

The wavelet transform is a signal transform based on multi-resolution analysis (MRA) and is used in computer vision, signal processing, pattern recognition, and image processing [16]. The DWT is constructed using two main functions: wavelet function $\Psi(t)$ and scaling function $\Phi(t)$ Unlike DWT, the 2D-DMWT, likewise based on multi-resolution analysis, employs multiple wavelet and scaling functions. The collection of scaling operations can be represented using vector notation [17].

$$\Phi(t) = [\Phi_1(t), \Phi_2(t), \dots, \Phi_r(t)]^{\mathrm{T}} \quad (1)$$

Where $\Phi(t)$ is called a multi-scaling function, similarly, the collection of wavelet functions defines the multi-wavelet function [17].

$$\Psi(t) = [\Psi_1(t), \Psi_2(t), \dots, \Psi_r(t)]^{\mathrm{T}} \quad (2)$$

$\Psi(t)$ is called a scalar wavelet or simply wavelet when r=1. Although, in theory, r can have any size. The majority of the multi-wavelets investigated so far are for r=2. Scalar wavelet equations are similar to those for multi-wavelet two-scale wavelets and can be expressed as [18].

$$\Psi(t) = \sqrt{2}\ \textstyle\sum_{k=-\infty}^{\infty} G_k \,.\, \Phi(2t-k) \quad (3)$$

$$\Phi(t) = \sqrt{2}\ \sum_{k=-\infty}^{\infty} H_k\,.\,\Phi(2t-k) \tag{4}$$

where G_k and H_k , are the r × r filter matrices for each integer k. These filters' matrix components offer more degrees of freedom than a conventional scalar wavelet. The multi-wavelet filters can rely on their beneficial characteristics, including orthogonality, symmetry, and high order of approximation [19]. The GHM filter, which Geronimo, Hardian, and Massopust developed, is a well-known multi-wavelet filter. The GHM basis combines orthogonality, symmetry, and compact support in a way that no other scalar wavelet basis can [17]. According to Eqs. (3) and (4), the GHM two scaling and wavelet functions satisfy the two- scale dilation equations shown below:

$$\begin{bmatrix}\Phi_1(t)\\ \Phi_{2(t)}\end{bmatrix} = \sqrt{2}\ \sum_k H_k \begin{bmatrix}\Phi_1(2t-k)\\ \Phi_2(2t-k)\end{bmatrix} \tag{5}$$

$$\begin{bmatrix}\Psi_1(t)\\ \Psi_2(t)\end{bmatrix} = \sqrt{2}\ \sum_k G_k \begin{bmatrix}\Psi_1(2t-k)\\ \Psi_2(2t-k)\end{bmatrix} \tag{6}$$

where H_k in the GHM system are four scaling matrices $H_0, H_1, H_2,$ and H_3 [20].

$$H_0 = \begin{bmatrix}\frac{3}{5\sqrt{2}} & \frac{4}{5}\\ -\frac{1}{20} & -\frac{3}{10\sqrt{2}}\end{bmatrix},\ H_1 = \begin{bmatrix}\frac{3}{5\sqrt{2}} & 0\\ \frac{9}{20} & \frac{1}{\sqrt{2}}\end{bmatrix},$$
$$H_2 = \begin{bmatrix}0 & 0\\ \frac{9}{20} & -\frac{3}{10\sqrt{2}}\end{bmatrix},\ H_3 = \begin{bmatrix}0 & 0\\ -\frac{1}{20} & 0\end{bmatrix} \tag{7}$$

Also, G_k in the GHM system are four wavelet matrices $G_0, G_1, G_2,$ and G_3 [20].

$$G_0 = \begin{bmatrix}-\frac{1}{20} & -\frac{3}{10\sqrt{2}}\\ \frac{1}{10\sqrt{2}} & \frac{3}{10}\end{bmatrix},\ G_1 = \begin{bmatrix}\frac{9}{20} & -\frac{1}{\sqrt{2}}\\ -\frac{9}{10\sqrt{2}} & 0\end{bmatrix},$$
$$G_2 = \begin{bmatrix}\frac{9}{20} & -\frac{3}{10\sqrt{2}}\\ \frac{9}{10\sqrt{2}} & -\frac{3}{10}\end{bmatrix}, G_3 = \begin{bmatrix}-\frac{1}{20} & 0\\ -\frac{1}{10\sqrt{2}} & 0\end{bmatrix} \tag{8}$$

The iteration approach is outlined in Eqs. (3) and (4) to draw the wavelet and scaling function for the GHM filter. However, in this instance, two wavelet functions and two scaling functions are derived from two box functions, as seen in Fig. 1 [19].

There are three major drawbacks associated with DWT that severely limit its ability to be used in signal-processing applications. The drawbacks of DWT are being sensitive to shifts, having poor directionality, and lacking of phase information [21].

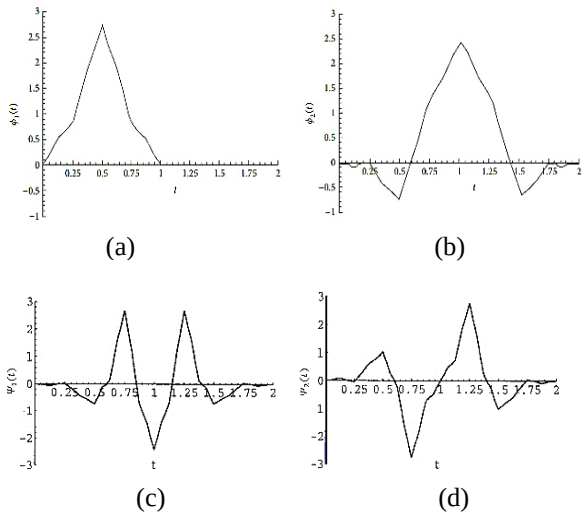

Figure. 1 GHM pair of $\Phi(t)$ and $\Psi(t)$: (a) is $\Phi_1(t)$, (b) is $\Phi_2(t)$, (c) is $\Psi_1(t)$, and (d) is $\Psi_2(t)$.

The DWT has a problem known as shift sensitivity. This means that the DWT coefficients cannot recognize between the input-signal shifts. Hence, the DWT transform is shift-sensitive when the input-signal shift produces an unreliable change in the transform coefficients. Secondly, The DWT is subjected to poor directionality since the transform coefficients reveal only several feature orientations in the spatial domain. Finally, the DWT does not provide any phase information, which is a crucial component in accurately describing a function's amplitude and local behavior.

In comparison to DWT, the drawbacks of DMWT include higher levels of complexity and longer implementation times. However, using DMWT to execute the proposed system is more reliable than utilizing DWT due to the fact that DMWT offers a number of benefits compared to DWT. In signal processing, it is common knowledge that certain features, such as short support, orthogonality, symmetry, and vanishing moments, play an essential role.

Vector inputs are required for the multi-wavelet filter bank. For some multi-wavelets, preprocessing must be followed by an adequate pre-filtering operation. However, due to desired aspects of their fundamental functions, some multi-wavelets eliminate the need for pre-filtering (and preprocessing); these multi-wavelets are known as balanced multi-wavelets [18]. There are several approaches for preprocessing, such as over sampled-scheme-preprocessing, critical-sampled- scheme-preprocessing, etc. In this paper, the critical-sampled-scheme-preprocessing based on the first order of approximation is utilized.

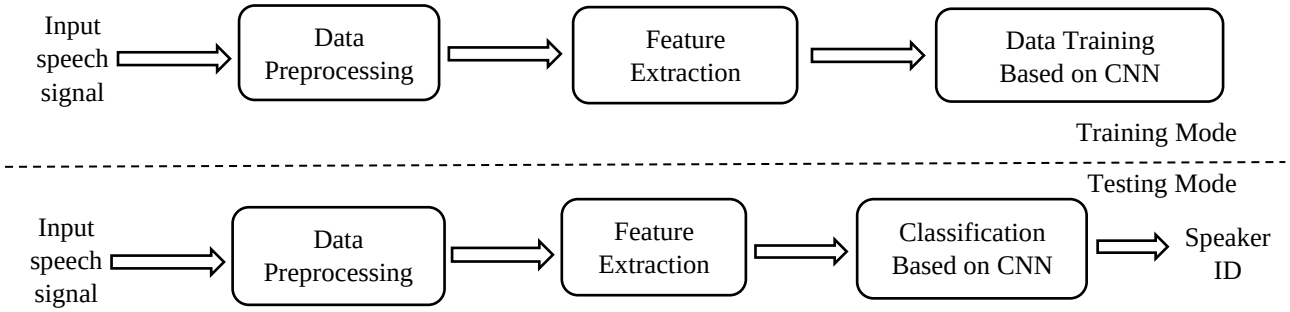

Figure. 2 The proposed system

3. System-proposed (SP)

The system-proposed (sp) is displayed and discussed in this section. The system-proposed is shown in Fig. 2, the proposed method has two modes: training and testing. Three significant steps are in any recognition system: Preprocessing, feature extraction, and classification.

Several datasets, including SALU-AC, TIMIT, ELSDSR, and RAVDESS, are utilized to evaluate the system-proposed. These databases consider various speech variations, including age, gender, noise, etc.

The three parts of the speaker recognition system will each have their role below.

1. Preprocessing

The initial phase of the proposed system is the preprocessing of the data. This step is necessary to format the data correctly. The following procedures are included as part of the preprocessing stage:

First, a 16-kilohertz resample must be performed on the speech signal before storing it on a single mono channel. In the second step, each speech signal is processed to remove silence. Silence removal is vital in speech recognition systems since it improves performance and streamlines the processing time. In the third step, divide the whole signal time of each speech sample into (0.5sec., 1sec., 2sec., 3sec., and 5sec.). Lastly, the 1D speech-signal is transformed into a 2D format so that the DMWT process may begin.

2. Feature extraction (FE)

Speech feature extraction aims to convert speech signals to coefficient vectors that contain only the information required for the identity of the provided phrase. The target of the FE phase is to obtain discriminate speech features from input speech-signal and reduce the speech-signal dimensions. Hence, it considered an intrinsic phase on every classification system. The SP employs a 2-D discrete wavelet transform (2D-DMWT) for FE.

2.1 Two-dimension discrete multi-wavelet transform (2D-DMWT)

In this work, 2D-DMWT is used for feature selection and dimensionality reduction. Multi-wavelet filter banks can only accept vector-valued input signals. Developing strategies for converting scalar input signals into their corresponding vector forms are necessary. The process of undergoing this transformation is known as preprocessing.

Preprocessing consists of two categories:

1- Oversampling scheme: Compared to the input with dimensions N×N, and where N is 2^a and a is an integer number, the dimensions of the DMWT matrix is doubled by making use of an oversampled method of preprocessing known as repeated row preprocessing.

2- Critical sampling scheme: The dimensions of the DMWT matrix are identical to those of the input matrix with dimensions N×N, where N is 2^a. An approximation-based-preprocessing technique commonly known as a critically sampled-scheme-preprocessing was used to achieve this.

A critical sampling strategy is presented in this paper as a potential solution. A first-order approximation is a method of approximation that may be used for critically sampled multi-wavelets. The following is a synopsis of how the first-order approximation-based-preprocessing operates (where every two rows produce two new rows) [20]:

a- For any odd-row:

$$new\ odd-row = (0.373615)[same\ odd-row] + (0.11086198)[next\ even-row] + (0.11086198)[previous\ even-row] \quad (9)$$

b- For every even-row:

$$new\ even-row = \left(\sqrt{2}-1\right)[same\ even\ row] \quad (10)$$

While figuring out the value of the first odd-row in Eq. (9), it is essential to remember that the value of the previous even-row is zero. Similarly, when calculating the value of the final odd-row in Eq. (9),

the value of the following even-row is zero [16].

The following procedures, as shown in Fig. 3, should be taken to compute DMWT using approximation-based-preprocessing:

1. Checking input dimensions: The length of the input matrix should be N×N, where N is 2^a. If the input matrix is not square, adding rows or columns of zeros must be executed to generate a square matrix. The input matrix in this study is a square matrix with N equal to 256 (256×256).

2. Constructing a transformation matrix H: The transformation matrix is shown in Eq. 11 [16, 20]:

$$H = \begin{bmatrix} H_0 & H_1 & H_2 & H_3 & 0 & 0 & ... \\ G_0 & G_1 & G_2 & G_3 & 0 & 0 & ... \\ 0 & 0 & H_0 & H_1 & H_2 & H_3 & ... \\ 0 & 0 & G_0 & G_1 & G_2 & G_3 & ... \\ ... & ... & ... & ... & ... & ... & ... \end{bmatrix} \quad (11)$$

where H_i and G_i are the impulse responses of the low-pass and high-pass filters, respectively. The matrices for the GHM low-pass and high-pass filters are given in Eqs. 7 and 8, respectively. These matrices should generate an N/2×N/2 transformation matrix (128×128). Then replace the GHM matrix filter coefficients with the values supplied by the following matrix.

$$H = \begin{bmatrix} H_0 & H_1 & H_2 & H_3 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & H_0 & H_1 & H_2 & H_3 & ... & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ H_2 & H_3 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & H_0 & H_1 \\ G_0 & G_1 & G_2 & G_3 & \vdots & \vdots & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & G_0 & G_1 & G_2 & G_3 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & G_0 & G_1 & G_2 & G_3 \\ G_2 & G_3 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & G_0 & G_1 \end{bmatrix} \quad (12)$$

3. Preprocessing rows: This is done by using Eqs. 9 and 10 for the odd-even-rows of the input N×N matrix, respectively, for the 1st-order approximation-based-preprocessing. The dimension of the input matrix, N×N (256×256), does not change after row preprocessing.

4. Transformation of rows: can be accomplished in the following manner:

I. Apply Eq. 12 to the N×N row preprocessed matrix via matrix multiplication.
II. To perform a permutation on the rows of the resultant N×N matrix, first, arrange the row pairs 1,2 and 5,6..., N-3, N-2 after each other in the top half of the rows of the resulting matrix, and then set the row pairs 3,4 and 7, 8, …, N-1, N below them in the next lower half of the rows.

5. Preprocess columns: To repeat the method that was utilized in preparing rows,

I. It is required to transpose the row-transformed N×N matrix that was created in step 4.
II. Repeat step 3 on the N×N matrix to obtain the N×N columns preprocessed matrix.

6. Transformation of columns: The following transformation is done to the N×N column preprocessed matrix to transform the columns:

I. Perform matrix multiplication between the resultant matrix with the transformation matrix H.
II. Conduct a permutation on the rows of the resulting N×N matrix by putting the row pairs 1, 2 and 5, 6... N-3, N-2 after each other in the top half and then placing the row pairs 3, 4 and 7, 8... N-1, N below them in the next lower half.

7. The final Transformed matrix: The following processes need to be followed to obtain the final transformed matrix:

I. After completing the column transformation stage, the created matrix must be transposed.
II. The resultant transpose matrix should then have the coefficients permuted applied to it.
III. When approximation-based-preprocessing is used, the final DMWT matrix has the exact dimensions (N×N) as the input matrix.

The approximation-based-preprocessing used to create the final DMWT matrix results in the exact dimensions (N×N) as the original input matrix (256×256), as shown in Fig. 4. After applying 2D-DMWT to 2D processed speech-signal with 256×256 dimensions, the resultant processed speech-signal matrix is partitioned into four main sub-bands, as shown in Fig. 3. Each sub-band has 128×128 dimensions. Also, each main one (sub-band) is further divided to four sub-sub-bands, each with 64×64 dimensions. Since most of the discriminate features of the speech-signal is localized in the low-low (LL) frequency sub-band, therefore, the main one (LL-sub-band) is preserved and the other sub-bands are ignored. The resultant extracted speech matrix is

Received: May 6, 2023. Revised: July 8, 2023.

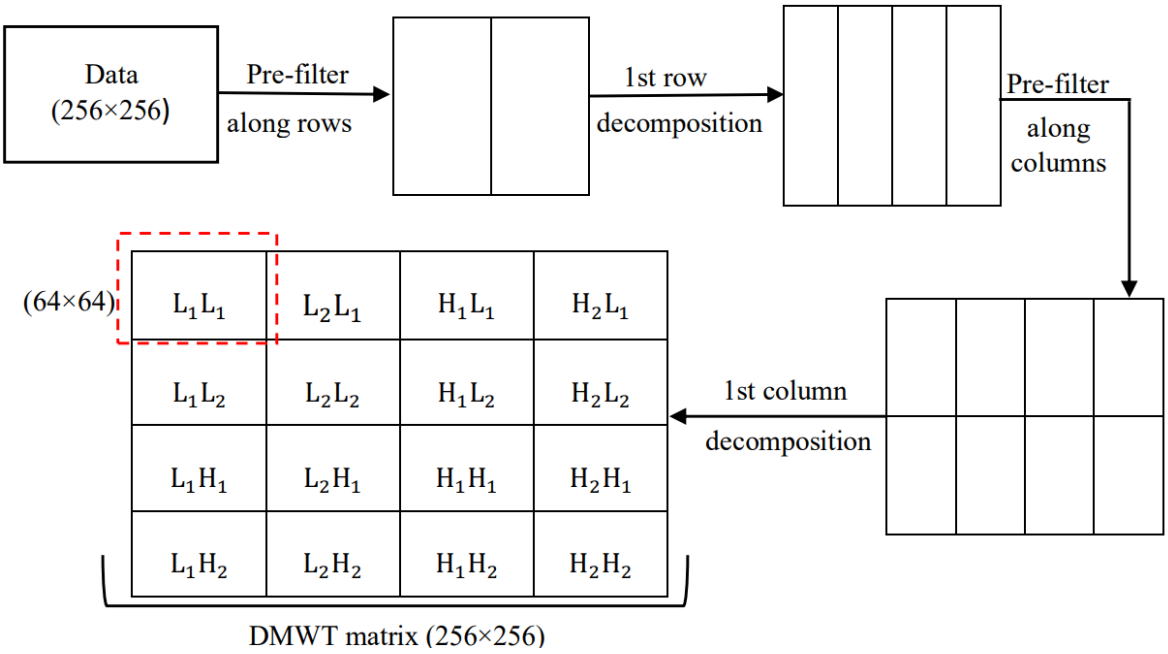

Figure. 3 Single level of decomposition of DMWT

the average matrix the four-sub-sub-bands localized in LL main sub-band.

Hence, the resultant matrix features, considered as an input to the classification phase, has 64×64 dimensions. As a results, the features extraction and dimensionality reduction are accomplished in this step.

3. Classification (convolution neural network (CNN))

Deep learning relies on neural network models, and one such model is called CNN. Convolutional neural networks (CNN) is a network that will use the convolution operation as its principal processing operation. The concepts of sparse interaction, parameter sharing, and equivariant representation are the foundations upon which CNN networks are built [22]. The convolution and subsampling layers are the first two layers included in a traditional CNN. One or more fully connected layers (FC) and an output layer follow subsequently.

The convolutional layer comprises many distinct learnable convolution kernels or filters that are used to generate a wide range of feature maps. Every feature map unit is connected to the receptive field of an earlier layer. Convolving the input with the kernels comes first in generating the new feature map. After that, a non-linear activation function is applied to the output of the convolving procedure. The pooling or subsampling layer takes as its input a tiny piece of the output generated by the convolutional layer and then down-samples it to produce a single output. The pooling layer's objective is to decrease the computational complexity and dimensionality [23]. One or more FC layers, typical in feedforward neural

Table 1. The structure of CNN layers

Layer No.	Layer Name	Detail
1	Input	64×64×1×Number of samples
2	Convolution	3×3, 24 filters, padding "same."
3	Batch Normalization	
4	relu	
5	Max Pooling	Pool size = 4×4
6	Convolution	3×3, 36 filters, padding "same"
7	Batch Normalization	
8	relu	
9	Max Pooling	Pool size = 4×4
10	Convolution	3×3, 48 filters, padding "same"
11	Batch Normalization	
12	relu	
13	Fully Connected Layer	
14	softmax	
15	Classification output	Number of classes

networks, are included in CNN's topmost layer. The FC layer takes the last pooling or convolutional layer's input and generates the CNN's final output layer. The architecture of the CNN layers that are used in this system is shown in Table 1.

The CNN consists of 15 layers. The input image for the CNN, a 4-dimensional (64x64x1xNumber of samples) matrix, is entered into the CNN to begin the classification process. A learning rate of 0.01 was used throughout the CNN model's training, which lasted for a total of 500 epochs, and the sgdm optimizer is responsible for utilizing the optimization procedure.

4. Speech databases

The system proposed is tested using databases: SALU-AC, ELSDSR, TIMIT, and RAVDESS. Each database's attributes are briefly explained:

SALU-AC/

The database SALU-AC, "Salford University Anechoic Chamber," was compiled from 110 speakers, all of which were in English. The speakers relied on reading passages from books, newspapers, and other sources [24]. Each speaker was given three speech samples, each with a different length of time (approximately 60 seconds for the first sample and around 40 seconds for the other instances). This paper employs 104 speakers (56 females and 48 males). The SALU-AC database is accessible on the website of the University of Salford (https://salford.figshare.com/).

TIMIT/

The TIMIT database, which stands for "Texas Instruments Massachusetts Institute of Technology," is an acoustic-phonetic continuous speech corpus containing 630 speakers of eight American English dialects. Each speaker has ten utterances that last (2-3) seconds. The speech was sampled at 16 kHz with 16 bits per sample. The TIMIT database can be downloaded from (https://www.kaggle.com/datasets/tommyngx/timit-corpus).

RAVDESS/

RAVDESS is a collection of 24 expert public speakers' emotional speeches. (12 females and 12 males). Each speaker has 60 utterances, each lasting 3-4 seconds. The database includes eight distinct emotions: disgust, indifference, astonishment, calmness, fear, anger, joy, and sadness. Visit (https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio) to access the RAVDESS database.

ELSDSR/

Twenty-two speakers were used to generate the ELSDSR database (10 females and 12 males). There are nine total utterances from each speaker. Twenty Danes, one Icelander, and one Canadian all speak English. 16 kHz sampling frequency is employed. The ELSDSR database can be downloaded from (http://www2.imm.dtu.dk/~lfen/elsdsr/).

Table 2. The recognition rate of the system-proposed

Database	Accuracy (%)				
	0.5sec	1sec	2sec	3sec	5sec
SALU-AC	95.86	96.30	96.63	97.04	97.56
ELSDSR	96.59	97.31	97.76	98	98.30
RAVDES	89.90	96.05	96.12	97.96	--
TIMIT	89.83	93.59	95.90	--	--

5. The experimental results and discussion

In this section, the experimental results of the SP are exhibited and explained. Moreover, the outcomes of the proposed method are compared with the state-of-the-art approaches [4-6, 13-15]. MATLAB was used to implement the SP. Table 2 displays the results of the proposed approach. Table 2 presents the recognition rate of all the types of speech databases according to the 2D-DMWT method. The SALU-AC and ELSDSR databases may be partitioned into (0.5sec., 1sec., 2sec., 3sec., and 5sec.). RAVDESS database is split into (0.5sec., 1sec., 2sec., and 3sec.). While TIMIT database is split into (0.5sec., 1sec., and 2sec.). Depending on the duration of time that each database provides. The SALU-AC and ELSDSR databases each have the ability to divide their length into 5sec., whilst the RAVDESS database has the ability to split its duration into 3sec. Due to the fact that each sample has a length of (3-4) seconds, whereas the samples in the TIMIT database each have a duration of (2-3) seconds, it is possible to break it up to 2 seconds. The system-proposed attained a recognition rate of 100% during evaluation using the training poses.

Figs. 4-7 illustrate the normalization process of some outcomes of the proposed method for the required amount of time as an example (1sec.) for each type of database.

As seen in Table 2, the classification rates of the system proposed are increased when the time duration of speaker sample of every database used is increased. The database's size is another factor contributing to the SP's overall performance. Figure 8 shows the histogram illustrating the effect of the signal duration on the model performance. Because the longer the test lasts, the greater the likelihood that a more considerable number of words will be said.

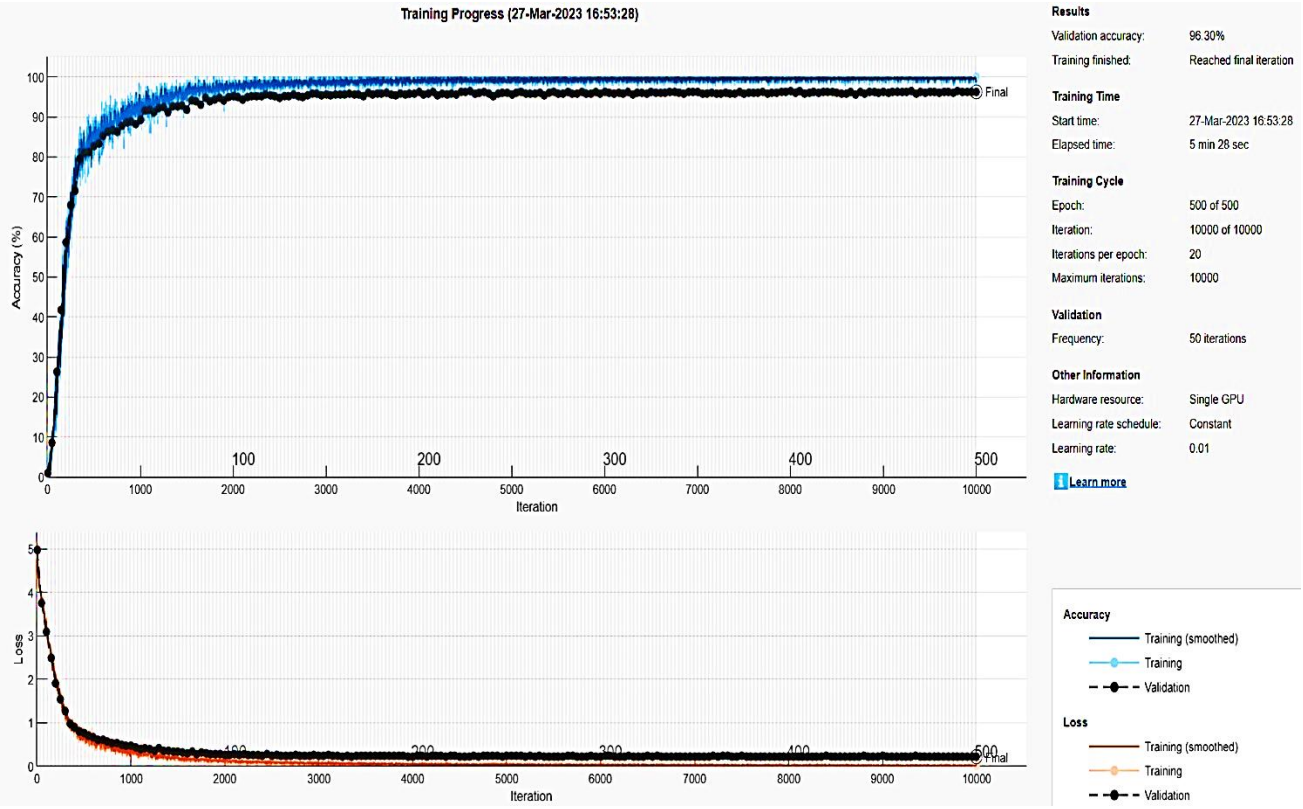

Figure. 4 The normalization process of the SALU-AC database

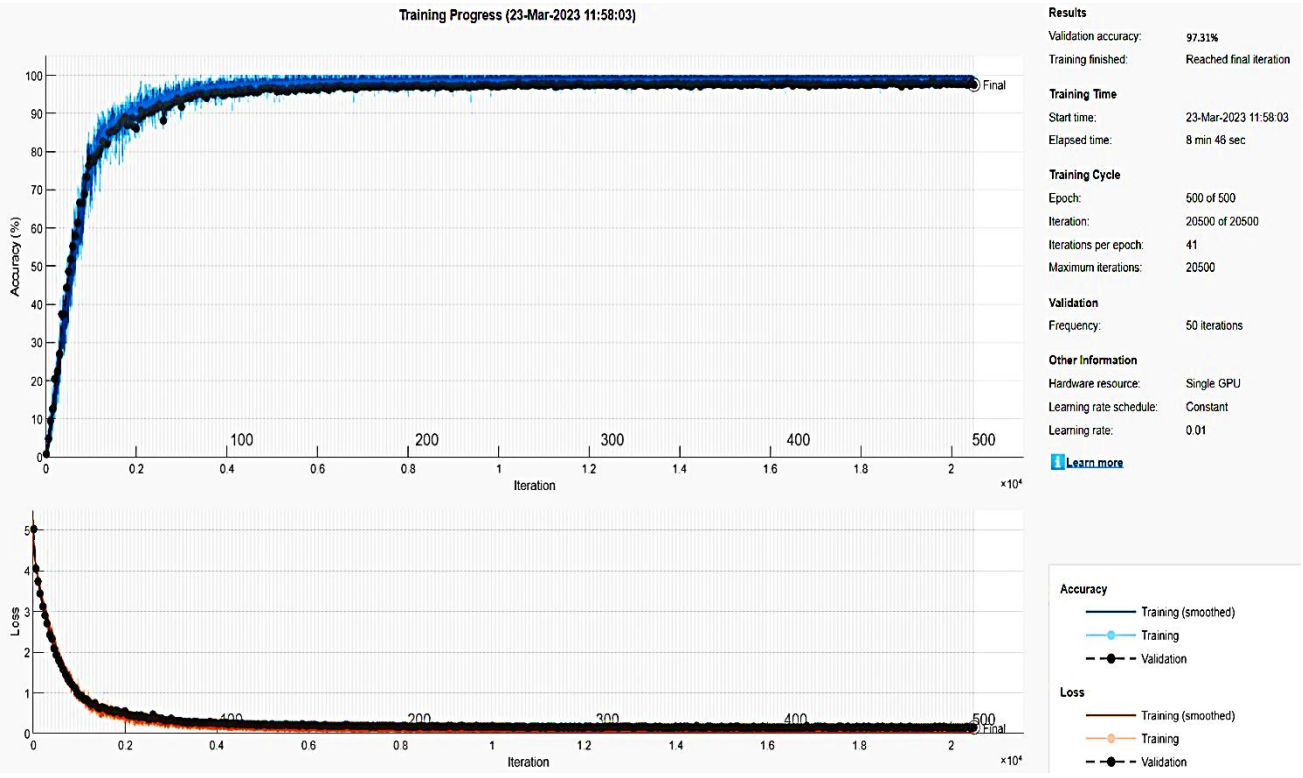

Figure. 5 The normalization process of the ELSDSR database

Fig. 9 shows a histogram of accuracies according to the length of time for all database types.

The outcomes of the proposed method are compared with those obtained in the previous works [4-6, 13-15] based on the same databases; namely, TIMIT, RAVDESS, and ELSDSR. Comparisons will be made

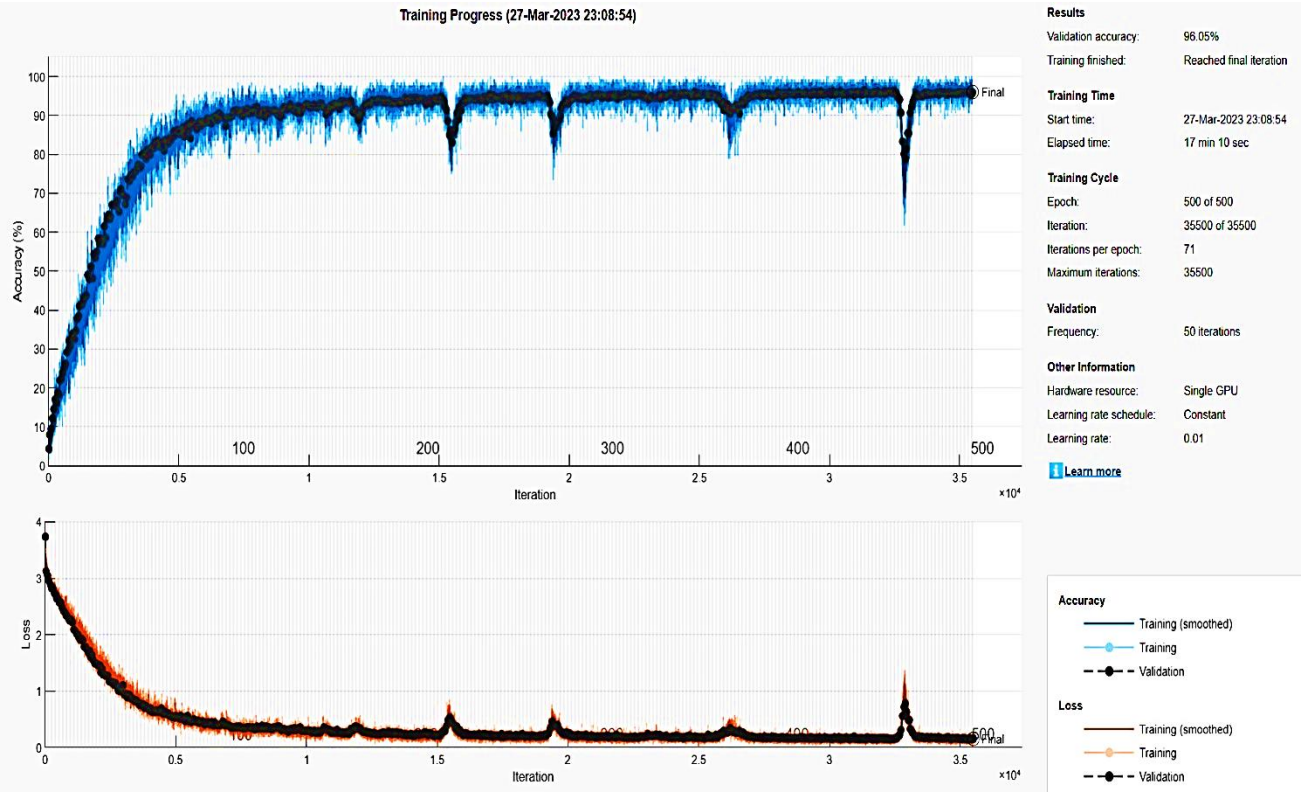

Figure. 6 The normalization process of the RAVDESS database

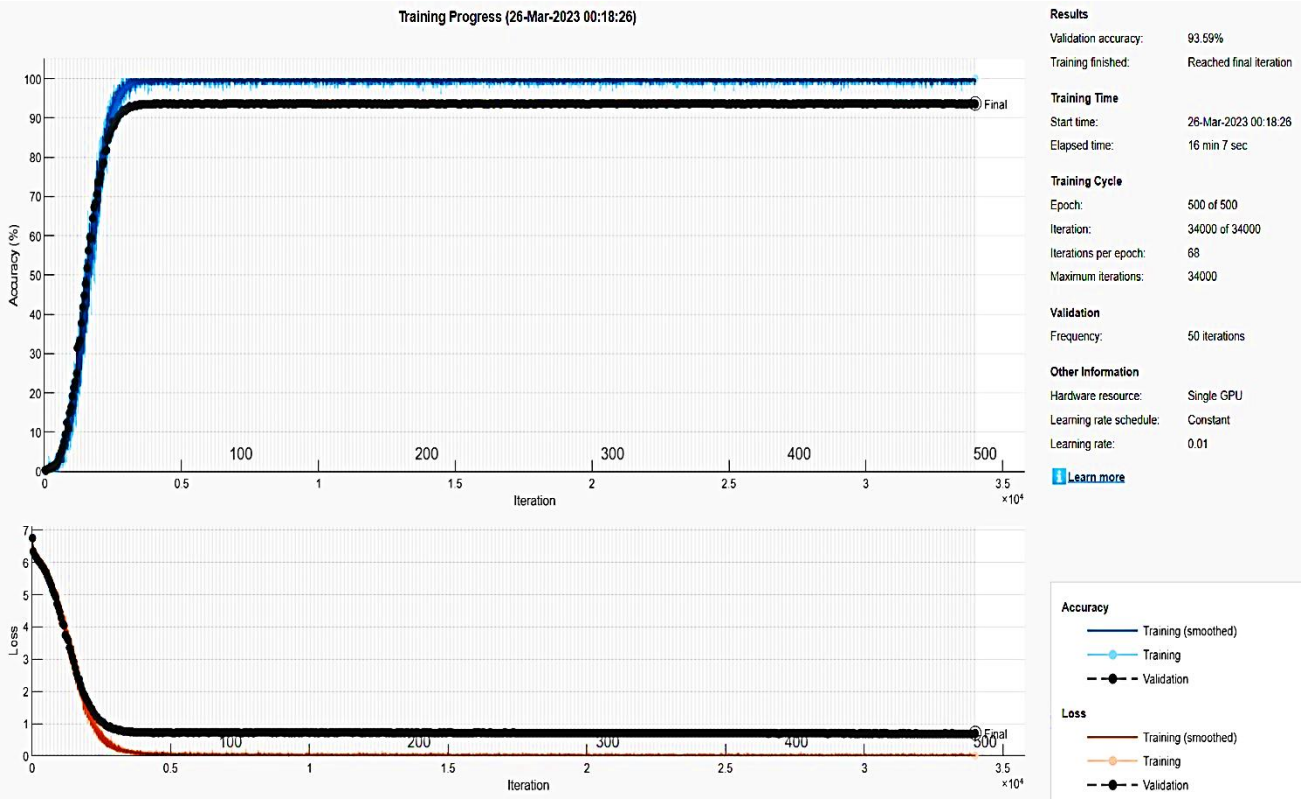

Figure. 7 The normalization process of the TIMIT database

between the approaches used in the feature extraction and classification. In order to establish which one of these approaches is the most efficient in terms of obtaining a high recognition rate.

The result of the system-proposed based on the RAVDESS database is shown in Table 3. As seen in Table 3, a comparison is made between the proposed system and the technique described in [4]. The

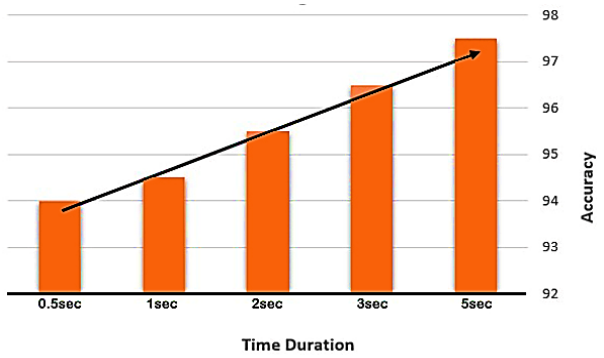

Figure. 8 illustrates the histogram of the accuracies based on the time duration based on the SALU-AC database

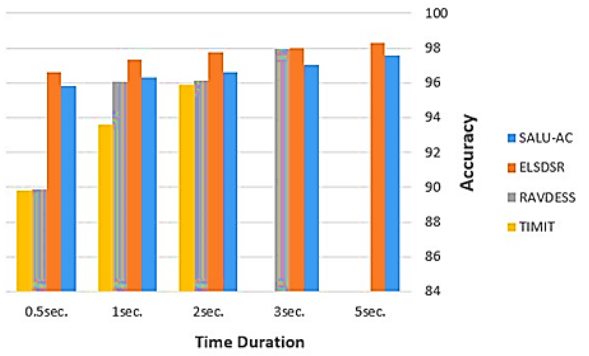

Figure. 9 illustrates a histogram of the accuracies based on the length of time

authors in [4] used hybrid techniques to extract the features from the speech signal depending on time, frequency, and cepstral domains.

Although hybrid approaches were applied in [4], the proposed system, 2D DMWT-CNN, achieved greater recognition rates. The use of the proposed 2D-DMWT approach was what ultimately made it feasible to extract the features in an efficient manner. This is a result of the fact that DMWT has a variety of benefits that make it possible for them to extract features in a manner that is more distinguishable.

The recognition rate of the SP based on the TIMIT database is shown in Table 3. As seen in Table 3, the result of the SP outperforms the result accomplished by [5]. The approach presented in [5] was evaluated using 38 speakers of the TIMIT database. So, 38 speakers from the TIMIT database were selected to evaluate the proposed method in order to have a fair comparison.

Table 3 shows a comparison between the SP and the method provided in [6] based on the RAVDESS database. In [6], The authors selected only 12 samples for each speaker. Therefore, 12 samples were chosen for each speaker to analyze the proposed approach for a fair comparison.

The comparison between the SP and the approach described in [13] is also shown in Table 3, which is

Table 3. The recognition rates of the SP compared to the previous works

Method	Database	Rates
2D DMWT-CNN (proposed) hybrid techniques-MLP [4]	RAVDESS	**98.26** 92
2D DMWT-CNN (proposed) Random Forest using MFCC features [5]	TIMIT	**97.46** 97
2D DMWT-CNN (proposed) MFCC-Modified SVM [6]	RAVDESS	**95.22** 93.01
2D DMWT-CNN (proposed) AFB-SVNN [13] hybrid techniques-RF-SVM [14]	ELSDSR	**98.42** 95 98.16
2D DMWT-CNN (proposed) MFCC-SECNN [15]	TIMIT	**95.90** 95.83

based on the ELSDSR database. The authors in [13] used combination methods in order to extract features from the speech signal based on the frequency domain. These techniques included multiple kernel weighted Mel frequency cepstral coefficient (MKMFCC), spectral kurtosis, spectral skewness, and autocorrelation. Even though hybrid techniques were used in [13], the proposed approach, 2D DMWT-CNN, was still able to attain higher recognition rates.

The result of the system-proposed based on the ELSDSR database is shown in Table 3. As seen in Table 3, a comparison is made between the proposed system and the technique described in [14]. The authors in [14] utilized hybrid techniques to extract speech signal features based on spectral features. Although hybrid approaches were employed in [14], 2D DMWT-CNN achieved better recognition rates.

The recognition rate of the SP based on the TIMIT database is presented in Table 3. As can be shown in Table 3, the outcome achieved by the SP performs better than the one that was achieved by [15].

Regarding Table 3, it is abundantly clear that the integration of 2D-DMWT with CNN lead to accomplish high recognition rates in comparison to those accuracies obtained in [4-6, 13-15]. This due to that discrete multi-wavelet contain some properties, such as compact support, orthogonality, symmetry, and high-order vanish moments. On the other hand, a multi-wavelet system has the ability to instantaneously offer perfect reconstruction while

maintaining length (orthogonality), good performance at the boundaries (via linear-phase symmetry), and a high order of approximation as a result of the system's capacity to preserve linear-phase balance (vanishing moments). As a consequence of this, the use of multi-wavelets offers the opportunity to improve the system performance in any signal-processing application.

According to the results shown in Table 3, the performance of the suggested approach was superior to those approaches discussed in [4-6, 13-15]. The high performance accomplished by employing the proposed algorithm in comparison with the other algorithms presented in [4-6, 13-15] is due to the combination of the resilience techniques that were employed in the preprocessing phases (silence elimination, signal duration splitting, and signal resampling). These techniques effect directly on the system proposed performance. Using the suggested 2D-DMWT algorithms allowed for the successful extraction of the features. DMWT has a number of advantages, some of which include qualities that are regarded fundamental in signal processing, such as short support, orthogonality, symmetry, and a large number of vanishing moments. This results in a high level of system performance since this approach has benefits that enable them to extract features in a more distinct manner. As a consequence, this leads to a high level of overall system performance. And the laborious search into which parameters best suit the CNN structure.

The dimensionality reduction can be measured by:

$$DR = \left(1 - \frac{R}{I}\right) \times 100\% \quad (13)$$

Where R and I are referred to the resultant (output) and the input matrix dimensions of every speaker sample in every database used, respectively. For all databases, the DR of every speaker sample is:

$$\left(1 - \frac{64 \times 64}{256 \times 256}\right) \times 100\% = 93.75\%$$

Therefore, the reduction in dimensions will improve the storage requirement to store the extracted feature. Whereas, for the methods described in [4-6, 13-15], the dimensions of the output matrix are the same as those of the input matrix.

6. Conclusion

A speaker identification system based on the conjunction of 2D-DMWT and CNN was presented in this paper. This was successfully contributed to achieve dimensionality reduction, discriminant feature extraction, and data compaction. The system proposed consisted of three primary phases. The speech databases passed through various preprocessing techniques to improve speech representation and achieve greater data reduction in preprocessing phase. 2D-DMWT approach was used based on a critical sampling scheme preprocessing that employs the GHM filter in the FE phase. To accomplish the dimensionality reduction, the only frequency sub-band (LL-sub-band) of the 2D-DMWT was preserved. The resultant matrix was the average matrix among the four sub-sub-bands with 64×64 dimensions. CNN was used for classification purposes. CNN was built with 15 layers. The system proposed is assessed using four databases; namely, SALU-AC, TIMIT, ELSDSR, and RAVDESS. That have various speech variations, such as gender, age, etc. The system proposed accomplished 93.75% dimensionality reduction, which led to less storage requirement in comparison to other approaches presented in [4-6, 13-15]. The system's high performance that is being accomplished is due to the fact that multi-wavelets combine symmetry, orthogonality, and short support. On the other hand, a multi-wavelet system may deliver a good performance at the boundaries while still providing a high degree of approximation (vanishing moments). As seen in Tables 2 and 3, the SP achieved high recognition rates in comparison to those obtained by [4-6, 13-15].

In further work, the proposed system will combine 2D-DMWT with principal component analysis (PCA) and 2D-DMWT with Mel-frequency Cepstral Coefficients as a feature extraction phase to achieve high system's accuracy.

Conflicts of interest (Mandatory)

"The authors declare no conflict of interest."

Author contributions (Mandatory)

Conceptualization, Huda W. Al-Dulaimi; methodology, Ahmed Aldhahab, and Hanaa M. Al Abboodi; software, Huda W. Al-Dulaimi, and Ahmed Aldhahab; validation, Huda W. Al-Dulaimi; formal analysis, Huda W. Al-Dulaimi, Ahmed Aldhahab, and Hanaa M. Al Abboodi; investigation, Huda W. Al-Dulaimi; resources, Huda W. Al-Dulaimi; data curation, Huda W. Al-Dulaimi, Ahmed Aldhahab, and Hanaa M. Al Abboodi; writing—original draft preparation, Huda W. Al-Dulaimi; writing—review and editing, Huda W. Al-Dulaimi, Ahmed Aldhahab, and Hanaa M. Al Abboodi; visualization, Huda W. Al-Dulaimi; supervision, Huda W. Al-Dulaimi, and

Ahmed Aldhahab; project administration, Huda W. Al-Dulaimi, Ahmed Aldhahab, and Hanaa M. Al Abboodi; funding acquisition, Huda W. Al-Dulaimi", etc. Authorship must be limited to those who have contributed substantially to the work reported.

References

[1] Z. Bai and X. L. Zhang, "Speaker Recognition Based on Deep Learning: An Overview", *Neural Networks*, Vol. 140, pp. 65-99, 2021.

[2] X. Yuan, G. Li, J. Han, D. Wang, and T. Zhi, "Overview of the development of speaker recognition", *Journal of Physics: Conference Series ICETIS 2021*, Vol. 1827, pp. 1-6, 2021.

[3] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities", *IEEE Access*, Vol. 9, pp. 79236- 79263, 2021.

[4] T. J. Sefara and T. B. Mokgonyane, "Emotional Speaker Recognition based on Machine and Deep Learning", In: *Proc. of 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Kimberley, South Africa, pp. 1-8, 2020.

[5] K. K. Nawas, M. K. Barik, and A. N. Khan, "Speaker Recognition using Random Forest", In: *Proc. of ITM Web Conf.*, Vol. 37, pp. 1-5, 2021.

[6] N. A. A. Hindawi, I. Shahin, and A. B. Nassif, "Speaker Identification for Disguised Voices Based on Modified SVM Classifier", In: *Proc. of 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, Monastir, Tunisia, pp. 687-691, 2021.

[7] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and D. Setiadi, "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network", *Journal of King Saud University –Computer and Information Sciences*, Vol. 34, Issue. 7, pp. 4375-4384, 2022.

[8] R. T. A. Hassani, D. C. Atilla, and Ç. Aydin, "Development of High Accuracy Classifier for the Speaker Recognition System", *Applied Bionics and Biomechanics*, Vol. 2021, pp. 1-10, 2021.

[9] S. Kadyrov, C. Turan, A. Amirzhanov, and C. Ozdemir, "Speaker Recognition from Spectrogram Images", In: *Proc. of 2021 IEEE Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, pp. 1-4, 2021.

[10] F. Abakarim and A. Abenaou, "Comparative study to realize an automatic speaker recognition system", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 12, No. 1, pp. 376-382, 2022.

[11] S. Dwijayanti, A. Y. Putri, and B. Y. Suprapto, "Speaker Identification Using a Convolutional Neural Network", *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, Vol. 6, No. 1, pp. 140 - 145, 2022.

[12] S. H. Shah, M. S. Saeed, S. Nawaz, and M. H. Yousaf, "Speaker Recognition in Realistic Scenario Using Multimodal Data", In: *Proc. of 2023 3rd International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, pp. 209-213, 2023.

[13] V. Srinivas and C. H. Santhirani, "Optimization-Based Support Vector Neural network for Speaker Recognition", *The Computer Journal*, Vol. 63, Issue. 1, pp. 151–167, 2020.

[14] V. Karthikeyan and S. P. Suja, "Hybrid machine learning classification scheme for speaker identification", *Journal For Forensic Sciences*, Vol. 67, No. 3, pp.1033–1048, 2022.

[15] M. Qi, Y. Yu, Y. Tang, Q. Deng, F. Mai, and N. Zhaxi, "Deep CNN with SE Block for Speaker Recognition", In: *Proc. of 2020 Information Communication Technologies Conference (ICTC)*, Nanjing, China, pp. 240-244, 2020.

[16] A. Aldhahab and W. B. Mikhael, "Face Recognition Employing DMWT Followed by FastICA", *Circuits, System, and Signal Processing*, Vol. 37, No. 5, pp. 2045–2073, 2018.

[17] A. Aldhahab and W. B. Mikhael, "A facial recognition method based on DMW transformed partitioned images", In: *Proc. of 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, USA, pp. 1352-1355, 2017.

[18]A. H. Kattoush, W. A. Mahmoud, A. Mashagbah, and A. Ghodayyah, "Multi-wavelet Computed Radon-Based Ofdm Trasceiver Designed and Symulation under Different Channel Conditions", *Journal of Information and Computing Science*, Vol. 5, No. 2, pp. 133–145, 2010.

[19] H. H. Wang, J. Wang, and W. Wang, "MULTISPECTRAL IMAGE FUSION APPROACH BASED ON GHM MULTIWAVELET TRANSFORM", In: *Proc. of 2005 International Conference on Machine Learning and Cybernetics*, Guangzhou, China, pp. 5043–5049, 2005.

[20] W. A. Mahmoud, Z. J. M. Saleh, and N. K. Wafi, "The Determination of Critical-Sampling Scheme of Preprocessing for Multiwavelets

Decomposition as 1st and 2nd Orders of Approximations", *Al-Khwarizmi Engineering Journal*, Vol. 1, No. 1, pp. 26-37, 2005.

[21] F. C. A. Fernandes, R. L. C. van Spaendonck, and C. S. Burrus, "A New Framework for Complex Wavelet Transforms", *IEEE Transactions On Signal Processing*, Vol. 51, No. 7, pp. 1825-1837, 2003.

[22] S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, "Text-independent speaker identification using deep learning model of convolution neural network", *International Journal of Machine Learning and Computing*, Vol. 9, No. 2, pp. 143–148, 2019.

[23] T. Bezdan and N. B. Džakula, "CONVOLUTIONAL NEURAL NETWORK LAYERS AND ARCHITECTURES", In: *Proc. of International Scientific Conference On Information Technology And Data Related Research (SINTEZA)*, pp. 445–451, 2019.

[24] H. M. A. Abboodi, "Binaural sound source localization using machine learning with spiking neural networks features extraction", *Salford: University of Salford*, 2019.