IRAQI Academic Scientific Journals

ISSN: 1813-162X (Print); 2312-7589 (Online)

Tikrit Journal of Engineering Sciences

available online at: http://www.tj-es.com

TJES Tikrit Journal of Engineering Sciences

Speaker Identification and Verification Using Convolutional Neural Network CNN

Azhar S. Abdulaziz *[a], Akram Dawood [b], Amar Daood [b]

a Computer Networks and Internet Department, College of IT, Ninevah University, Mosul, Iraq.
b Department of Computer Engineering, Engineering College, Mosul University, Mosul, Iraq.

Keywords:
Artificial intelligence; Biometrics; Digital voice communication; Deep learning; Signal processing; Speaker identification; Speaker verification.

Highlights:

- *A new online training technology called train-on-the-fly side-by-side with speaker verification feedback highly improved identification accuracy.*
- *The paper proved that split-add-noise, specifically AWGN, positively affected the accuracy of speaker identification/verification.*
- *The proposed system accurately verified the speaker, regardless of the speech context.*
- *The deep learning CNN model with MFCC features was compared with other classifiers and proved its superiority over them.*

ARTICLE INFO

Article history:

Received	01 Oct.	2023
Received in revised form	30 Oct.	2023
Accepted	17 May	2024
Final Proofreading	27 Aug.	2024
Available online	31 May	2025

Citation: Abdulaziz AS, Dawood A, Daood A. **Speaker Identification and Verification Using Convolutional Neural Network CNN**. *Tikrit Journal of Engineering Sciences* 2025; **32**(2): 1746. http://doi.org/10.25130/tjes.32.2.19

**Corresponding author:*

Azhar S. Abdulaziz

Computer Networks and Internet Department, College of IT, Ninevah University, Mosul, Iraq.

Abstract: Speaker identification and verification are important fields contributing to smart IoT, phone banking, remote login services, E-learning, and other applications. In this work, the speaker identification and verification processes have been experimentally proven to have mutual enhancement effects if they are merged together in a proper manner. Speaker identification and verification work cooperatively so that the verifier will enhance the identifier model. The first step is to identify the speaker using context – independent speech signal, and the identifier model (ID) is trained using a classification model. The model's outputs are then used to control the verification process as a next step. When the verification result is positive, the first process outcome is approved with high confidence. Otherwise, the negative verification will force the ID process to re-configure itself. The loop continues until both verification and ID agree on the speaker. A multiple Gaussian mixture GMM was used to efficiently model each person's speech features (MFCC) for using expectation maximization (EM). On the other hand, the conducted experiments showed that the one-dimensional convolutional neural network (1D-CNN) proved its superiority over other models for speaker identification. A novel approach was proposed, proving that little data can be expanded with split-add-noise and train-on-the-fly procedures. In many speaker identification approaches, the specific context was used as a keyword or a password to simplify the processing, requiring big data to achieve high accuracy. It is noteworthy that a small amount of data was enough to efficiently train the proposed model, with a verification error of around 3%, i.e., an accuracy of 97%. Meanwhile, 95% and 96% identification accuracy was achieved using two different datasets. Additionally, the suggested algorithm did not imply using any keyword or password because it is a context-independent approach.

التحقق والتعرف على المتكلم باستخدام الشبكة العصبية الالتفافية

أزهر صباح عبد العزيز [١]، أكرم عبد الموجود داود [٢]، عمار ادريس داود [٢]

[١] قسم شبكات الحاسوب والانترنيت/ كلية تكنولوجيا المعلومات / جامعة نينوى / الموصل – العراق.

[٢] قسم هندسة الحاسوب/كلية الهندسة / جامعة الموصل / الموصل – العراق.

الخلاصة

ان التعرف على المتكلم اليا والتحقق منه هما مجالان مهمان يساهمان في تطوير انترنت الأشياء الذكية وخدمة المصرف الهاتفية وخدمة الولوج وخدمة التعليم الالكتروني عن بعد وتطبيقات أخرى. ثبت تجريبيًا من خلال هذا البحث أن عمليات تحديد هوية المتكلم والتحقق منه بينها تأثيرات تعزيزية تبادلية، إذا دمجا معًا بطريقة مناسبة. حين يعمل التعرف على المتكلم والتحقق منه بشكل تكاملي، بحيث ان المتحقق سيحسن الية التعرف. الخطوة الأولى هي التعرف على المتكلم باستخدام إشارة كلام الغير معتمدة على سياق معين، ونموذج تمييز المتكلم يدرب باستخدام أي نموذج تصنيف رياضي. ان مخرجات النموذج الرياضي عندها ستستخدم في السيطرة على عملية التحقق كخطوة ثانية. عندما تكون نتيجة التحقق إيجابية، فإن مخرجات المرحلة الأولى سيتم قبولها بدرجة عالية من الوثوقية. في ماعدا ذلك، فان التحقق السلبي سيجبر الية التعرف على إعادة تشكيل اعداداتها. وتستمر هذه الدورة من العمليات الى ان يتفق كل من المتعرف والمتحقق على هوية المتكلم. ان المزيج الغاوسي المتعدد تم استخدامه لنمذجة خواص الكلام لكل شخص وبكفاءة عالية باستخدام الية تعظيم التوقع. من ناحية اخرى، اثبتت التجارب العملية افضلية الشبكة العصبية الالتفافية الأحادية مقارنة مع النماذج الأخرى في مايخص التعرف على المتكلم. ان الطريقة المبتكرة المقترحة في هذه الورقة البحثية اثبتت انه يمكن توسعة البيانات القليلة باستخدام اسلوب التدريب المجزأ المتطاير. في العديد من أساليب التعرف على المتحدث، فإن سياق محدد من الكلام يستخدم ككلمات مفتاحية أو كلمات سر لتبسيط المعالجة. إن هذه الالية تتطلب بيانات ضخمة لتحقيق دقة عالية. من الجدير بالذكر ان كمية قليلة من البيانات تكفي لتدريب النموذج الرياضي المقترح، بنسبة خطأ في التحقق حوالي ٣٪، بمعنى آخر كانت نسبة الدقة ٩٧٪. بينما كانت نسبة الدقة في التعرف على المتكلم بمقدار ٩٥٪ و٩٦٪ لنوعين مختلفين من البيانات. فضلاً عن ذلك، فإن الخوارزمية المقترحة لا تشترط استخدام كلمات مرور أو مفتاحية، وبعبارة أخرى، فإنها طريقة تعرُف غير معتمدة على سياق معين في الكلام.

الكلمات الدالة: الذكاء الاصطناعي، القياسات الحيوية، الاتصالات الصوتية الرقمية، التعلم العميق، معالجة الإشارات، تحديد هوية المتحدث، التحقق من المتحدث.

1.INTRODUCTION

Speaker identification is a technique used to determine an unknown speaker's identity. The vocal voice can be considered a unique characteristic to recognize speakers from their signature voices. However, there is some common confusion and ambiguity among people between speaker recognition and speech recognition. The major difference between speaker identification and speech recognition is that speaker identification can be defined as the technique of recognizing who is speaking [1, 2]. On the other hand, speech recognition recognizes what is being said. The primary focus in the present work is to recognize people or grant them access based on their voice characteristics. Each individual has unique characteristics within his voiceprint, such as style of speaking, accent type, rhythm, manner of pronunciation, intonation, and even vocabulary [3-5]. The recognition process of speaker identity can be classified into two types: speaker identification and speaker verification [6]. Speaker identification is the process of recognizing people from their voices within a set of known people by comparing the input voice with the stored reference voices as a database. On the other hand, speaker authentication is the process of granting access or authorizing a particular person based on his/her claimed identity [7,8]. The recognition method can be categorized into text-dependent and text-independent. The first analysis requires the identified person to speak a specific set of words. In this type of system, the speaker must read the same texts during the enrollment (training) and the authorizing (testing) phases [9]. The second type of speaker identification is text-independent, where the recognition process does not require specific content [10]. The recognition process can be done by reading anything, which makes the identification system much more complicated [11]. Additionally, there are two types of speaker recognition systems from a speaker standpoint: open-set and closed-set systems based on the stored references in the database [12]. Speaker identification has many advantageous applications, such as identifying a criminal solely by their voice compared with previous criminals' voices in the database, called forensic voice verification, customer service voice recognition to avoid fraud cases in banks, and access control to different services, such as telephone network services, voice dialing, mobile shopping [13, 14], and even E-learning [15]. The captured voice can be used to grant access to different types of applications to provide security layers, which is a crucial factor for any particular entry of various services. Speaker recognition and identification is one of the most common pillars of the authentication and verification process in the biometric field. Recently, the need to provide security to ensure secure entrance increases the speaker recognition process value and motivates a huge number of research to be conducted in this direction [16, 17]. Many studies have measured the outcome of a wide range of acoustic characteristics for speaker identification and the feature fusion effect. Many acoustic feature types, such as LPCC, MFCC, and PLP, have been extensively studied. The most important inference about these features is their highly integration with speaker and speech recognition algorithms. Biometric wrapping does not show any benefit to features for speaker recognition. Occasionally, merging two or more features enhances the precision of speaker identification with little risk of degrading accuracy [18]. This paragraph summarizes some of the last works that tried to tackle the speaker recognition problem. For example, the author in [19] extracted MFCC and UMRT features from ten samples of 15 persons using dependent and independent speakers' data. The extracted features were classified using the MLP classifier. They achieved an accuracy of 97% and 94% for dependent and independent speech, respectively. The authors in [20] collected Emirati-accented speech databases from 25 speakers in two modes: neutral and shouting talk. They extracted MFCC features from the gathered dataset. The recognition process was performed based on text-independence. Finally, the classification was done using Hidden Markov Models. The researchers in [21] used the VGG CNN network to learn discriminative features by adding SoftMax loss and center loss. They used the VoxCeleb dataset to convert speech audio into spectrogram images. Then, the modified VGG was tuned to perform the identification process. The work in [22] used low-dimensional feature vectors to perform the feature extraction. The authors extracted cepstral features from the TIMIT data database. GMM analysis was used to classify the extracted features. The authors in [23] used wavelet transform to extract wavelet coefficient features from recorded samples of 40 speakers. The extracted features were used to train a multilayered neural network and a generalized regression neural network. The final stage of this system was the decision-making based on the majority voting technique. The authors in [24] used ATCOSIM and Voxfore datasets to evaluate the proposed method. The suggested model used wavelet transform to convert the audio samples and then extracted MFCC from the wavelet sub-bands. SVM was trained using the extracted features to perform the classification process. They used their method in aeronautical communication applications where the noise effect was analyzed. The authors in [25] transformed the audio samples into spectrogram images. The audio samples were collected from YouTube of 5 speakers. The transformed images were used to train convolutional neural networks to perform the identification process. They benchmarked their results with the MFCC features. The trained CNN achieved 95% accuracy. The work in [26] used the Voxceleb1 dataset to suggest a two-stage attention technique to perform speaker recognition in a noisy environment. Time delay neural network and convolutional neural network were used to extract features in time and frequency domains. Three types of noise were added in the experiments to evaluate the proposed method's performance. The researchers in [27] presented sum-product networks based on deep probabilistic graphical models. They extracted spectral features from the TIMIT dataset. The extracted features were used to learn the marginal probability density function. They compared their results with GMM and CNN methods. The work in [28] proposed a multi-model i-vector system using short speech length. The researchers used two datasets: THUYG-20 and their own collected speeches in the English language. The proposed system was tested and evaluated; the results showed an improvement of 28%. The authors in [29] suggested an open-set speaker by identifying outliers against a blacklist of speakers. The outlier detection was implemented using cosine-similarity. Probabilistic linear discriminant analysis was used to learn the blacklist parameters. All the conducted experiments were conducted using the MCE dataset. The paper [30] used two datasets that comprise Arabic and Algerian Berber languages. The authors extracted MFCC features and used PCA analysis for dimensionality reduction. MPL network was used to perform the identification process. Finally, the speaker's verification was done using dynamic time wrapping (DTW). Most of the previous works have used a large size of dataset to perform the identification and verification processes to gain high accuracy. However, collecting data requires huge effort not only for the gathering process but also for the labeling process, which requires extensive manual laboring using human operators. Therefore, an identification context-independent model is proposed based on a small amount of data and boosts the accuracy of the trained model using noise addition. Therefore, a text-independent identification system is proposed. The proposed system does not depend on the content of the speech. First, MFCC features were extracted from the captured voice to get some representative and discriminative features. The analysis of GMM was used to perform the recognition process, and a noise addition technique was adopted to improve the results by adding more variations to the captured voice to accommodate various kinds of realistic circumstances. The paper structure is organized as follows: Section 1 already discusses the introduction and related works. Section 2 explains the methodology of the proposed work and speaker identification. The third section describes the implementation details of the speaker verification. Whereas Section 4 discusses and analyzes the results. The last section is dedicated to the conclusions of the paper.

2.THE METHODOLOGY

It is intended to design a system that can be trained to identify and verify speakers when sufficient data is unavailable. In addition, the proposed system should be able to verify a speaker with a totally unconstrained model so that the speaker can be identified in any speech context, i.e., no specific keyword is required, which is known as a context-independent criterion. The proposed ID-verification model is achieved by following steps:

1- A trained classification model will classify the signal predicting the speaker (x).

2- The verification step will get the log-likelihood for a person x'.

If $x = = x'$ The classification model will fix its values.

Else if $x \neq x'$ The classification model updates its parameters.

For GMM (Gaussian Mixture Model), the model changes the means and covariance matrix while using the 1D-CNN (one-dimensional convolutional neural network) filters, and the final dense layers weights are updated.

3- If both identification and verification agree, stop.

Else Go to step 1.

The proposed flowchart steps of the algorithm are depicted in Fig. 1. Regarding the classification model, the analysis of GMM was used to build multinormal distributions for individual speakers. Additionally, different types of classifiers were tested, other than GMM, were tested to classify the MFCC (Mel-Frequency Cepstral Coefficients) features. However, the deep learning technique of 1-D Convolutional neural network showed better accuracy than other methods.

2.1.MFCC Feature Extraction

The Mel-Frequency Cepstral analysis is a key feature used by typical speech and sound signals analysis. The MFCC is becoming the de-facto feature that abstracts almost all the important properties of the speech signal. It reduces the speech signal's dimensionality, reducing the processing power and complexity. It has been known for being robust, and they are used frequently in the field of speech recognition. Although there are different algorithms to extract those features, in this research, the feature estimation is achieved by the following sequence: The MFCC feature extraction starts with the pre-emphasize filter, which is simply LPF and is used to emphasize the low-level high variation with respect to the high-level low variation usually found in speech signals. The following equation is the z-plane (frequency domain) equation used in this filter:

$$\boldsymbol{Y(z) = \frac{1}{1-\alpha z^{-1}}\, X(z)} \qquad \textbf{(1)}$$

where $Y(z)$ represents the output of the filter, $X(z)$ stands for the input, and α=0.97 in the present case. The energy was calculated for the 20 ms frames after applying the Hamming window. A 50% overlap was applied for the ongoing speech signal consequent frames. The spectrum of the frame was calculated using the Fast Fourier Transform (FFT). The spectrum samples were filtered using Mel-scale filter banks, let be Mel(FFT(x)), and then the logarithm of the amplitude of those coefficients was taken, or log(|Mel(FFT(x))|). A discrete Cosine transform (DCT) was applied to the resultant samples, and only the first 12 were considered, as shown in Fig. 2.

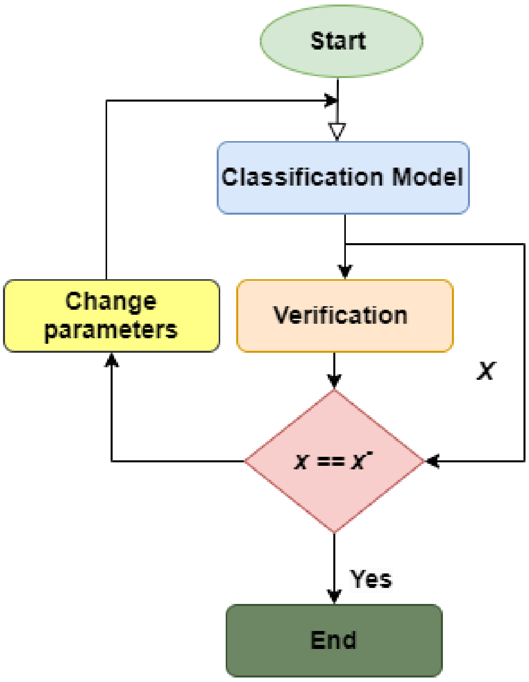

Fig. 1 The Proposed Algorithm Flowchart.

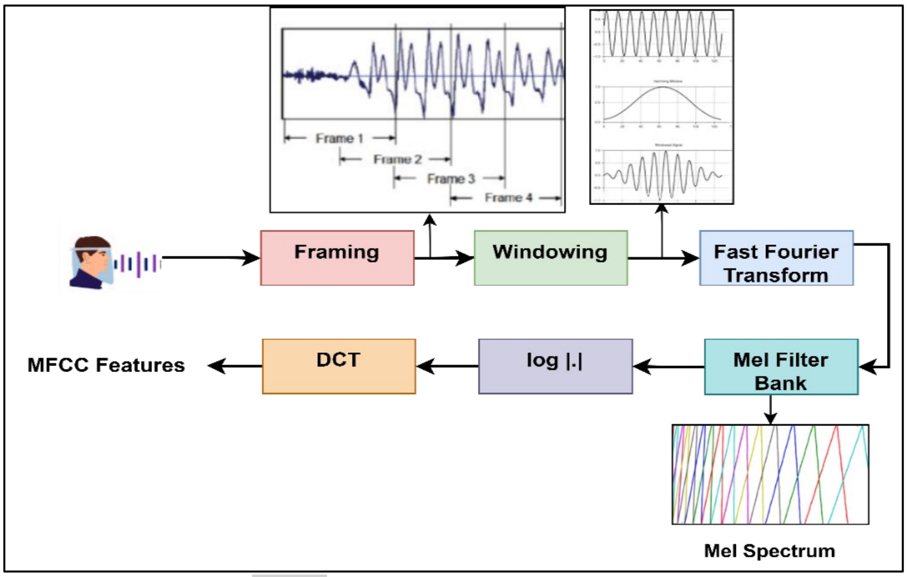

Fig. 2 MFCC Features Extraction.

The MFCC was formed by considering the first 12 coefficients with the frame's energy, forming 13 MFCC coefficients [31]. Calculating the MFCC feature is achieved by the following equation:

$$MFCC_i = \sum_{n=1}^{N_{fb}} \log|m_n| \cos\left[i \left(\frac{2n-1}{2}\right)\left(\frac{\pi}{N_{fb}}\right)\right], \; i = 1, 2, \dots, L \qquad (2)$$

where m_n is the nth Mel-filter bank output, N_{fb} is the number of filter banks, i.e., =24, and L is the number of the first 12 MFCC coefficients, i.e., =12. The first and the second derivatives were then derived from the 13 MFCCs to remove the linear filtering effect, making the overall features 39 per frame. The audio signal a human can handle was up to 20 kHz; therefore, the FFT samples were diminishing as frequency increased. Taking the first 12 samples of Mel-Scale frequency was enough because the values of the coefficients were too low after that. The energy is related to the phoneme articulation as well as the person who pronounces it. Hence, the energy sample of each speech frame is a vial of information to distinguish individuals according to their speech audio signal [31].

2.2.Expectation Maximization on GMM

In signal processing, the normality of the probability distribution of a random signal simplifies many statistical measures. However, noisy speech signal has a complicated probability density function (PDF) that cannot be described as a normal, or Gaussian-distributed PDF. Hence, the noisy speech data can be considered a mixture of many different Gaussian distributions. The Gaussian mixture is a suitable approximation where a single-dimensional Gaussian PDF fails to represent such a complicated random signal. The GMM can efficiently represent the speaker/speech variability and model the incoming features robustly [32]. In this work, the Gaussian Mixture Model (GMM) was designed to represent the noisy speech signal. The proposed GMM consisted of three mixtures of normally distributed PDF, each contributing to the mixture with a specific weight. Modeling the speech signal using the GMM is achieved by an unsupervised fitting process known as the Expectation Maximization (EM) algorithm. The EM algorithm is an iterative algorithm that converges monotonically towards maximum likelihood. The following relationships represent the proposed GMM mixtures:

$$M_1 = c_1 p(\mu_1 \Sigma_1) \qquad (3)$$

$$M_2 = c_2 p(\mu_2 \Sigma_2) \qquad (4)$$

$$M_3 = c_3 p(\mu_3 \Sigma_3) \qquad (5)$$

where c_i is the weight, μ_i is the mean, and Σ_i is the covariance of the mixture for the i^{th} mixture in the fitted GMM. Building the GMM model in the training and testing stages was accomplished by relating each feature with a certain class. Ultimately, the trained GMM produced the maximum log-likelihood probability p(ω|X). The latter likelihood cannot be estimated directly, and it is described according to the following Bayes rule:

$$p(\omega/x) = \frac{p(x/\omega)p(\omega)}{p(x)} \qquad (6)$$

where $p(\omega)$ and $p(x)$ are the class and the observation prior probabilities, respectively, $p(\omega|x)$ is the posterior probability of the class ω given the observation x, and $p(x|\omega)$ is the probability of the observation x given the class ω. For each class ω_k, the posteriori probability of the observation given the k^{th} class was computed by applying each observation to the multivariate normal distribution, which has the following formula:

$$\boldsymbol{p(x|w_k)} = \frac{\mathbf{1}}{\sqrt{(2\pi)^d|\Sigma_k|}} \boldsymbol{exp}\frac{-\mathbf{1}}{\mathbf{2}} (\boldsymbol{x} - \boldsymbol{\mu_k})^T \, \Sigma_k^{-1}(x-\mu_k) \qquad (7)$$

The GMM is designed so that k=1, 2, and 3 represent the three mixtures of the model, and x is the entry MFCC feature matrix, i.e., its dimension = 39. GMM analysis was used to recognize and identify speakers, and a deep learning method was used. The next section provides details of the implementation of both methods.

2.3.Speaker Identification Models

The identification process was started by building a Gaussian Mixture Model (GMM) for each speaker. It was based on the log-likelihood function between the input speech and the models. The Expectation maximization (EM) algorithm was based on the GMM of three mixture models. The input speech is said to come from an individual if it is more likely to match that individual's model. Increasing the mixtures in each model might increase the system's accuracy. However, it will increase the computational complexity as well [33]. Therefore, it is intended here to make suitable speaker recognition with acceptable complexity and accuracy. Figure 3 shows the training and testing methodology for the proposed identification system. In the training part, a 3-mixture GMM has been trained for each speaker individually. When it comes to testing, the log-likelihood between the input speech features and the trained GMM was measured. The candidate speaker was selected when its associated model had the maximum likelihood with the input test speech features.

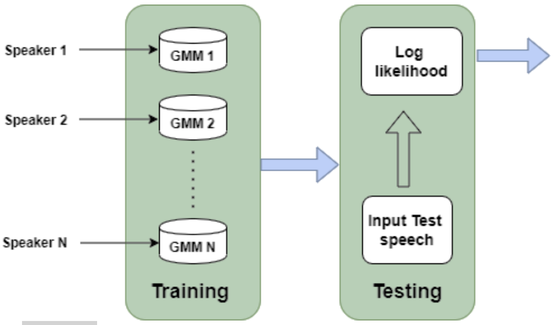

Fig. 3 Speaker Identification Using GMM.

Furthermore, different types of classifiers were trained to compare the results of the conducted experiment. To improve the identification process results, deep learning strategies were adopted by training a 1D convolutional neural network. Neural networks and deep learning have shown significant performance on different types of tasks [34, 35]. Figure 4 shows the architecture of the proposed network, consisting of 8 layers starting with the input layer and ending with the SoftMax layer to recognize six speakers. The training process was performed on more than 20,000 samples, each representing a vector of 39 features for six participants. The extracted features were divided into two segments: 80% for training and 20% for testing the trained model. The designed network was trained for 60 epochs. Figure 5 shows the training process of the trained and tested data, where the accuracy was calculated during the learning process to monitor the improvement of the learning knowledge against the epochs.

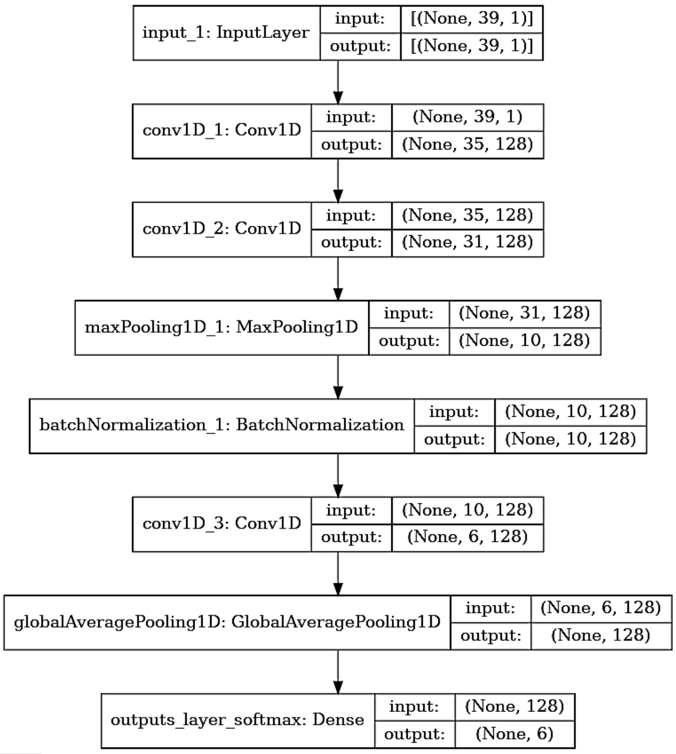

Fig. 4 Proposed 1D CNN Deep Learning Network Architecture.

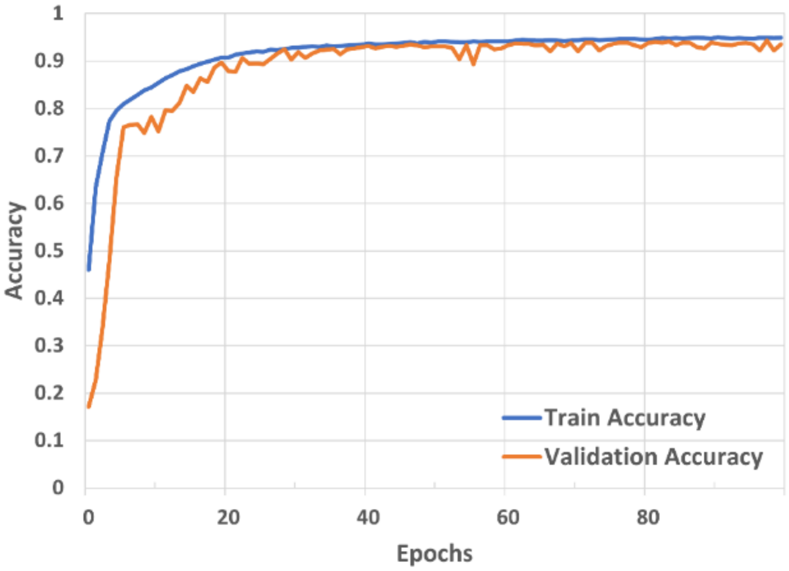

Fig. 5 Training and Validation Accuracy vs. Epochs for Deep Learning Model in Fig. 4.

3.SPEAKER VERIFICATION

The designed Speaker Verifier (SV) system is based on the reality of having a GMM for each speaker. The first step was to build a mixture model for a specific speaker. Later, a new entry mixture model was built on the fly from part of the input speech. The rest of the input speech was considered an input used to measure the likelihood between it and the trained model and the entry model. The final decision could be that the features would likely come from the same person if the log-likelihood were higher with the stored model rather than the entry model. This fact comes from the reality that although part of the speech will be modeled, the other noisy part will be more likely to be similar to the long-time trained model than the short-time trained one. The following block diagram shows the proposed SV system's training and testing steps. For the training procedure, the target speaker trained a 3-state Gaussian mixture model (GMM). The training speech should be specific to that speaker and for around 3 minutes of his/her speech. The system extracted Mel-Feature Cepstral Coefficients (MFCC) for each 20 ms of speech. Those features will build the mixture model. The testing part of the system used the already stored GMM from training and the GMM of part of the entry test speech with additive white Gaussian noise (AWGN). AWGN had zero mean and unity standard deviation and was defined as follows:

$$\boldsymbol{n(x)\sim N(0,1)} \qquad \textbf{(8)}$$

The other part of that test speech was added to AWGN similarly, and then the system extracted MFCC_entry. Then, the MFCC_entry was the input to the log-likelihood function to measure its likelihood to the stored trained model and the new entry GMM model. The proposed SV system block diagram is illustrated in Fig. 6, where the input speech signal was context-independent, and the output was the speaker verification and ID result.

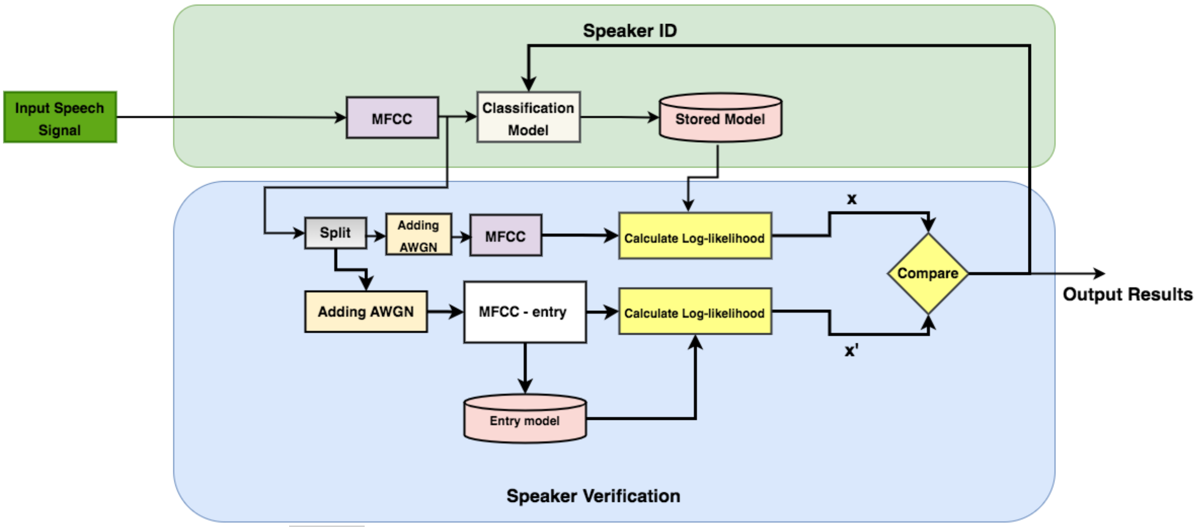

Fig. 6 The Proposed Speaker ID/Verification System.

The cooperation between the speaker ID and the speaker verifier systems was based on the feedback from the verifier to the ID model. The context-independent classifier, i.e., a part of the ID system (upper part in Fig. 6), is supposed to be trained to classify the incoming MFCC features. The result was the identified person, let it be a person (A), for example. For the speaker verification, which is the lower part in Fig. 6, a noisy speech signal was used to extract new MFCC features, fed to the on-the-fly entry model. The entry model measured the log-likelihood between the noisy MFCC and the entry classifier model to produce x'. Meanwhile, another log-likelihood, x, metric was calculated from the stored model, in the ID system, and the noisy MFCC. The verifier output is as stated in the following:

$$\boldsymbol{output} = \mathbf{max}\,(\boldsymbol{x}, \boldsymbol{x}') \qquad \textbf{(9)}$$

If the output from Eq. (9) confirms that the incoming signal is from the person (A), then the weights of the classifier model in the ID system are fixed to affirm that its output class should be stuck at the person (A) parameters. Otherwise, the weights of the classifier model take new training values.

4.RESULTS AND DISCUSSION

As discussed earlier, the proposed methodology is divided into two complementary parts: identification and verification processes. The speech signals of six speakers were collected from the GNU online database and 16 speakers from PDAs corpora, where each speaker had limited time to speak. The small amount of data was selected on purpose so that the ability of the designed model could be tested with such limited data. The data limitation was the key challenge in the proposed approach. The collected data were used to perform speaker identification/verification by building a GMM distribution for individual speakers. Furthermore, other classifiers were trained to select the best one. To compare the results, an assessment was performed to evaluate the performance of all methods. The 1D convolutional neural network proved its superiority over all of them. Table 1 shows the accuracy, precision, recall, and F1-score for those different classification models used in the proposed speaker ID/verification model. Because it had a limited amount of data, the designed approach had to increase the verification authenticity in a different approach. The proposed solution for data limitation was to expose the model to a wider variety of signals. Therefore, an additive white Gaussian noise was added to another version of the input speech, and an entry model was built on the fly. The outcomes showed that adding AWGN in building the entry GMM and feature extraction steps offered better results. Therefore, the two-step noise addition was crucial to boosting the performance. Figure 7 shows how the additive noise improved the results. It also shows the overall false acceptance error results. A false acceptance means the verifier model accepts the speaker to be 'X,' while it is not.

Table 1 Speaker Identification and Verification for GNU Online Dataset

	Model	Accuracy	Recall	Precision	F1-Score
RF	Random Forest Classifier	94.36%	0.9436	0.944	0.9436
DTC	Decision Tree Classifier	89.07%	0.8907	0.8911	0.8908
KNN	K Neighbors Classifier	87.8%	0.878	0.8785	0.8773
LDA	Linear Discriminant Analysis	85.79%	0.8579	0.8577	0.8575
LR	Logistic Regression	85.68%	0.8568	0.8574	0.8566
SVM	SVM - Linear Kernel	75.47%	0.7547	0.7661	0.7446
NB	Naive Bayes	71.64%	0.7164	0.7566	0.7227
ADA	Ada Boost Classifier	61.99%	0.6199	0.6661	0.621
GMM	Gaussian Mixture Model	77.08%	0.8665	0.5211	0.5926
1D-CNN	1-dimensional Convolutional Neural Network	95%	0.9417	0.945	0.94

Table 2 Speaker Identification and Verification for PDAs Corpora.

	Model	Accuracy	Recall	Precession	F1-Score
RF	Random Forest Classifier	91.64%	0.4976	0.4948	0.4929
KNN	K Neighbors Classifier	80.73%	0.4133	0.4131	0.4108
DTC	Decision Tree Classifier	65.36%	0.3524	0.3532	0.3526
LDA	Linear Discriminant Analysis	79.34%	0.2635	0.2595	0.2449
ADA	Ada Boost Classifier	75.17%	0.2519	0.2498	0.2268
LR	Logistic Regression	74.88%	0.2175	0.2079	0.1911
NB	Naive Bayes	71.73%	0.2046	0.1972	0.1815
SVM	SVM - Linear Kernel	23.56%	0.1330	0.1897	0.0818
GMM	Gaussian Mixture Model	78.84%	0.7865	0.6134	0.6261
1D-CNN	1-dimensional Convolutional Neural Network	96%	0.9336	0.9556	0.9367

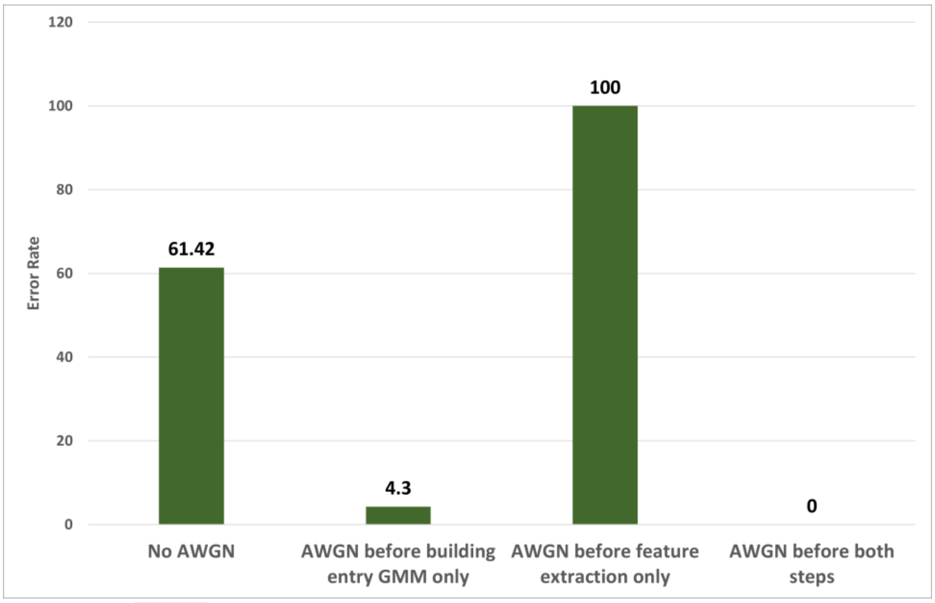

Fig. 7 Speaker Verification False Acceptance Error Rate.

Figure 7 confirms that adding noise and building on-the-fly entry models significantly influenced the speaker ID/verification accuracy. Each human individual has almost a unique vocal tract specification. However, the human voice varies for the same person due to different situations, like being tired, sleepy, sick, or angry. The additive white noise means that all frequency bands were somehow affected, while the person's voice variations may alter certain frequency bands only. Therefore, AWGN will imitate those possible variations of the person's voice biometric property. Figure 8 illustrates the AWGN spectrum compared to a speech power spectral density (PSD). It is noted that the AWGN impacted all frequency bands of the speech spectrum.

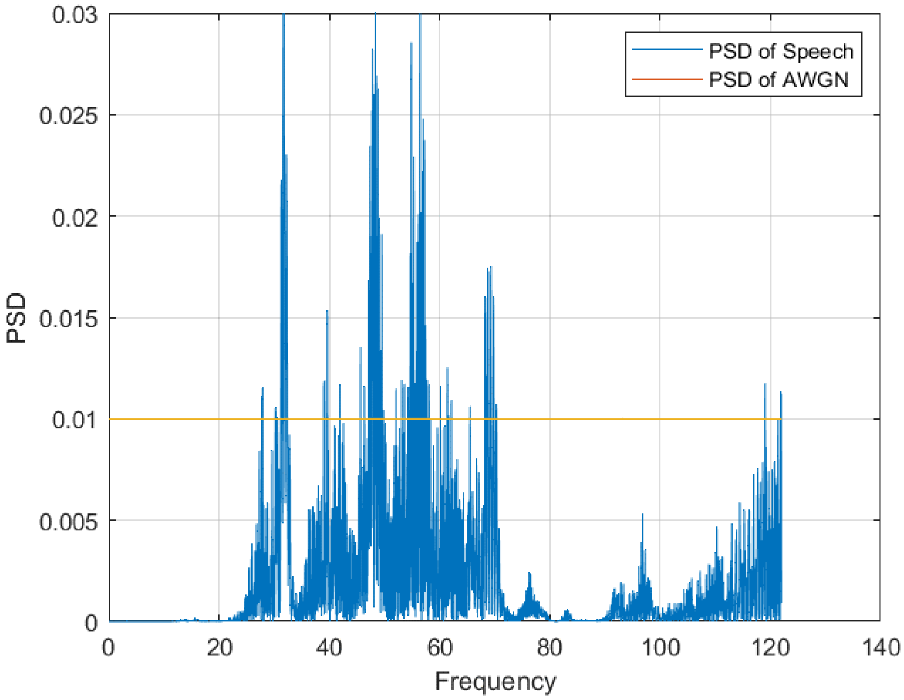

Fig. 8 The PSD of AWGN Compared to a PSD of Speech Signal.

The individual voice variations were considered in the proposed approach by adding noise to the speech signal. On the other hand, if the model was built using only noisy speech, it should be rejected, assuming that the individual variations have exceeded the individuality limits. The overall false rejection error results are shown in Fig. 9. The false rejection means the verifier model rejected the speaker as 'X' when it is actually 'X.'

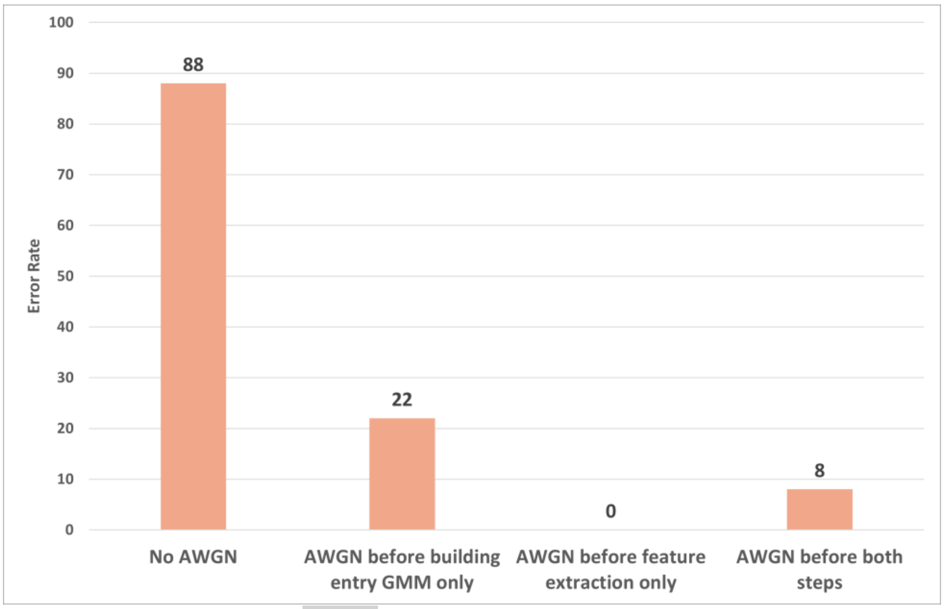

Fig. 9 False Rejection Ratio.

The present work has been designed to study the effect of the split-add-noise strategy on the ID/verification error. The false acceptance, false rejection, and average error were measured in the presence and absence of the AWGN. The lowest overall error of about 3%, as shown in Fig. 10, was achieved when adding noise and implementing the train-on-the-fly procedure. Hence, the average accuracy was around 97% in this case.

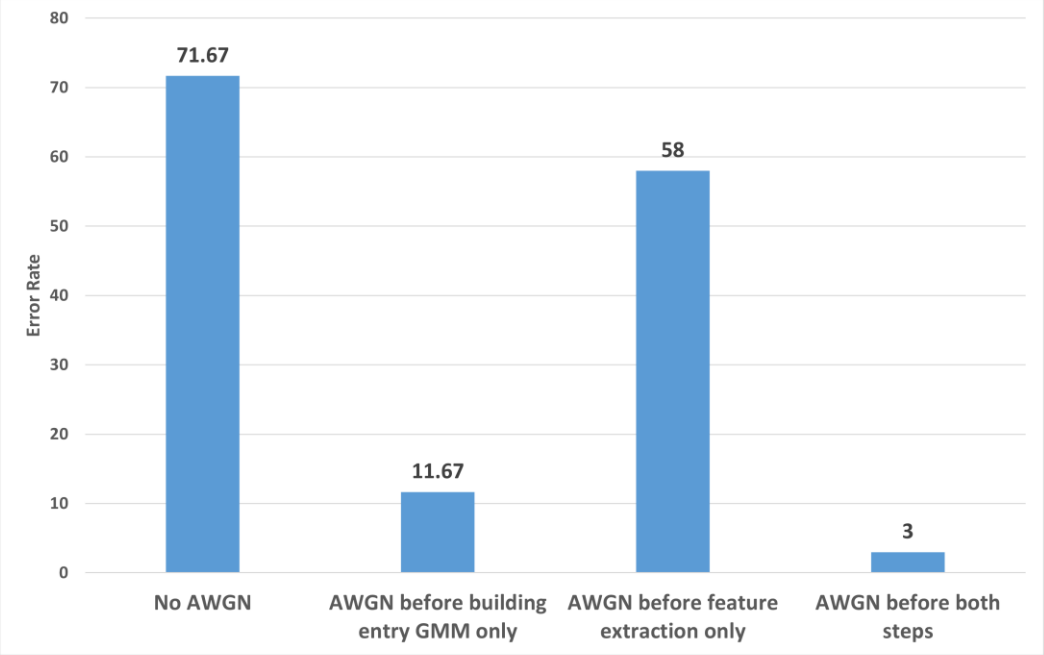

Fig. 10 The Overall Error Rate for the Proposed Model.

When the proposed approach is compared to other methodologies in previous studies, it can be noticed that the AWGN noise addition and on-the-fly modeling are superior. Table 3 shows the accuracy of different algorithms proposed by other researchers.

Table 3 The Accuracy of Speaker Identification/Verification from Different Studies.

Reference	Model and Feature types	Accuracy
[19]	MLP classifier MFCC and UMRT features	97% and 94% for dependent and independent, respectively
[25]	CNN Model	95%
[28]	Multi-model i-vector with MFCC features	91.8%
[36]	Deep Neural Network mask & voice VGG	87%
[37]	Dilated CNN, Log-MelSpectrum, excitation features	91.34%
Our approach	GMM and CNN with MFCC features	97% Context-Independent Speech

5.CONCLUSIONS

After investigating different models and approaches with limited data, it was found that the best scenario was when the speaker ID and verification work cooperatively together, and the speech variations were considered by adding noise. As the MFCC can represent a speaker's vocal tract transfer function, a mathematical model for those features yielded good results. Unfortunately, the error rate of the speaker recognition system was around 71.56% on average for two different datasets. However, the proposed split-add-noise and the train-on-the-fly side-by-side with speaker verification highly improved the accuracy to more than 97%. Where on-the-fly model building from part of speech plus random noise has added more confidence to the verification phase. The speaker was accurately verified by the proposed system, regardless of the speech context. On the other hand, different identification approaches have been applied to different speech datasets to compare them to the present approach. The comparison confirmed the superiority of the proposed methodology in terms of accuracy, specifically with a small amount of data. Context-independent speaker identification and verification is more attractive as it does not direct the user to say specific keywords or passwords. The freedom given by this approach can be used in many different fields, such as smart city and voice banking.

ACKNOWLEDGEMENTS

The authors are grateful for the support toward this research by the College of Engineering Department at Mosul university.

NOMENCLATURE

MFCC	Mel-Feature Cepstral Coefficient
GMM	Gaussian Mixture Model
ID	Identifier
AWGN	Additive White Gaussian Noise
ms	Milli Seconds
1D-CNN	One-dimensional Convolutional Neural Network
FFT	Fast Fourier Transform
Greek Symbols	
μ	Mean
Σ	Covariance Matrix
α	Filter Parameter

REFERENCES

[1] Xue Y. **Multi-Label Training for Text-Independent Speaker Identification.** *arXiv preprint arXiv* 2022 Nov 14.

[2] Mohammadi M, Mohammadi HRS. **Weighted X-Vectors for Robust Text-Independent Speaker Verification with Multiple Enrollment Utterances.** *Circuits, Systems, and Signal Processing* 2022; **41**(5):2825-2844.

[3] Nagakrishnan R, Revathi A. **Generic Speech-Based Person Authentication System with Genuine and Spoofed Utterances: Different Feature Sets and Models.** *Multimedia Tools and Applications* 2022; **1**:1-30.

[4] Gaurav, Bhardwaj S, Agarwal R. **An Efficient Speaker Identification Framework Based on Mask R-CNN Classifier Parameter Optimized Using Hosted Cuckoo Optimization (HCO).** *Journal of Ambient Intelligence and Humanized Computing* 2022; **5**:1-3.

[5] Shareef SRS, Al-Irhayim YFM. **Comparison Between Features Extraction Techniques for Impairments Arabic Speech.** *Al-Rafidain Engineering Journal* 2022; **27**(2):190-197.

[6] Monir M et al. **Cancelable Speaker Identification Based on Cepstral Coefficients and Comb Filters.** *International Journal of Speech Technology* 2022; **25**(2):471-492.

[7] Karthikeyan V. **Adaptive Boosted Random Forest-Support Vector Machine Based Classification Scheme for Speaker Identification.** *Applied Soft Computing* 2022; **131**:109826.

[8] Hamsa S et al. **Speaker Identification from Emotional and Noisy Speech Using Learned Voice Segregation and Speech VGG.** *Expert Systems with Applications* 2023; **224**:119871.

[9] AL-Shakarchy ND, Obayes HK, Abdullah ZN. **Person Identification Based on Voice Biometric Using Deep Neural Network.** *International Journal of Information Technology* 2023; **15**(2):789-795.

[10] Jahangir R et al. **Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network.** *IEEE Access* 2020; **8**:32187-32202.

[11] Shafik A et al. **Speaker Identification Based on Radon Transform and CNNs in the Presence of Different Types of Interference for Robotic Applications.** *Applied Acoustics* 2021; **177**:107665.

[12] Sidorov M et al. **Survey of Automated Speaker Identification Methods.** *Proceedings of the 9th International Conference on Intelligent Environments* 2013; 236-239.

[13] Dey N. **Applied Speech Processing: Algorithms and Case Studies.** *Academic Press*; 2021.

[14] Dawood A et al. **Simulation of Multimedia Data Transmission Over WSN Based on MATLAB/SIMULINK.** *International*

Journal of Computing and Digital Systems 2023; **14**(1):147-157.

[15] Shi Y, Huang Q, Hain T. **H-VECTORS: Improving the Robustness in Utterance-Level Speaker Embeddings Using a Hierarchical Attention Model.** *Neural Networks* 2021; **42**:329-339.

[16] Alshaykha AM. **E-Learning Visual Design Elements of User Experience Perspective.** *Tikrit Journal of Engineering Sciences* 2022; **29**(1):111-118.

[17] Mokgonyane TB et al. **A Cross-Platform Interface for Automatic Speaker Identification and Verification.** *Proceedings of the International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems* 2021; 1-6.

[18] Wang X et al. **A Network Model of Speaker Identification with New Feature Extraction Methods and Asymmetric BLSTM.** *Neurocomputing* 2020; **403**:167-181.

[19] Lawson A et al. **Survey and Evaluation of Acoustic Features for Speaker Recognition.** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* 2011; 5444-5447.

[20] Antony A, Gopikakumari R. **Speaker Identification Based on Combination of MFCC and UMRT Based Features.** *Procedia Computer Science* 2018; **143**:250-257.

[21] Shahin I, Nassif AB, Bahutair M. **Emirati-Accented Speaker Identification in Each of Neutral and Shouted Talking Environments.** *International Journal of Speech Technology* 2018; **21**:265-278.

[22] Yadav S, Rai A. **Learning Discriminative Features for Speaker Identification and Verification.** *Proceedings of Interspeech* 2018; 2237-2241.

[23] Chakroun R, Frikha M. **Improved Text-Independent Speaker Identification and Verification with Gaussian Mixture Models.** *12th International Conference on Knowledge Science, Engineering and Management* 2019; 3-10.

[24] Pawar MD, Kokate R. **A Robust Wavelet Based Decomposition and Multilayer Neural Network for Speaker Identification.** *7th Innovations in Electronics and Communication Engineering* 2019; 197-209.

[25] Sekkate S, Khalil M, Adib A. **Speaker Identification for OFDM-Based Aeronautical Communication System.** *Circuits, Systems, and Signal Processing* 2019; **38**(8):3743-3761.

[26] Bunrit S et al. **Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network.** *International Journal of Machine Learning and Computing* 2019; **9**(2):143-148.

[27] Shi Y, Huang Q, Hain T. **Improving Noise Robustness in Speaker Identification Using a Two-Stage Attention Model.** *arXiv preprint arXiv* 2019 Sep 24.

[28] Nicolson A, Paliwal KK. **Sum-Product Networks for Robust Automatic Speaker Identification.** *arXiv preprint arXiv* 2019 Oct 26.

[29] Tiwari V et al. **Speaker Identification Using Multi-Modal I-Vector Approach for Varying Length Speech in Voice Interactive Systems.** *Cognitive Systems Research* 2019; **57**:66-77.

[30] Wilkinghoff K. **On Open-Set Speaker Identification With I-Vectors.** *Proceedings of Odyssey* 2020; 408-414.

[31] Roumiassa F, Chelali FZ. **Speaker Identification and Verification System for Arabic and Berber Language.** *1st International Conference on Communications, Control Systems and Signal Processing* 2020; 242-247.

[32] Benesty J, Sondhi MM, Huang Y (Eds). **Springer Handbook of Speech Processing.** *Springer*; 2008; **1**.

[33] Dawood AAM, Abdulaziz AS, Mohammed AJ. **RLC-Based Image Compression Using Wavelet Decomposition with Zero-Setting of Unnecessary Sub-Bands.** *Journal of Engineering Science and Technology* 2022; **17**(1):391-403.

[34] Reynolds DA. **Gaussian Mixture Models.** *Encyclopedia of Biometrics* 2009; 659-663.

[35] Mahmood MS, Al Dabagh NB. **Improving IoT Security Using Lightweight Based Deep Learning Protection Model.** *Tikrit Journal of Engineering Sciences* 2023; **30**(1):119-129.

[36] Divya V, Kumar SS, Usha S, Hemamalini S, Krishnan G. **Improving EEG Electrode Sensitivity with Graphene Nano Powder and Neural Network for Schizophrenia Diagnosis.** *Tikrit Journal of Engineering Sciences* 2023; **30**(1):84-93.

[37] Hamsa S, Shahin I, Iraqi Y, Damiani E, Nassif AB, Werghi N. **Speaker Identification from Emotional and Noisy Speech Using Learned Voice Segregation and Speech VGG.** *Expert*

Systems with Applications 2023; **224**:119871.

[38] Pentapati HK. **Enhancement in Speaker Identification Through Feature Fusion Using Advanced Dilated Convolution Neural Network.** *International Journal of Electrical and Computer Engineering Systems* 2023; **14**(3):301-310.