J. Umm Al-Qura Univ. Eng.Archit.
https://doi.org/10.1007/s43995-024-00046-4

ORIGINAL ARTICLE

Text-independent speaker identification system using discrete wavelet transform with linear prediction coding

Othman Alrusaini[1] · **Khaled Daqrouq**[2]

Received: 30 November 2023 / Accepted: 3 January 2024

Abstract
One of the key problems of the modern day is the presentation of an identity verification system that can perform sufficient accuracy in identity verification, is resilient to assaults and noises, and can be recorded in the simplest possible method. In this study, a new speaker feature extraction which based on discrete wavelet transform (DWT) and linear prediction coding (LPC) algorithm (WLPCA) are investigated. This paper's primary objective is to evidence the performance of the new method for speaker identification by a Gaussian mixture model (GMM). The proposed method improves the recognition rate over the Mel-frequency cepstral coefficient (MFCC). Experimental evaluation of the process performance is performed on two speech databases; our recorded database and the publicly available TIMIT database. We show that the speech features derived by the newly proposed method are more suitable for GMM (91.53%), in terms of the time-consumed, by requiring less Gaussian mixtures than MFCC (85.77%). For testing the presented method in a noisy environment, Additive white Gaussian noise (AWGN) was added to the TIMIT database, where a slight improvement in the performance of the presented method (60.02%) over the MFCC (59.89%) was observed.

Keywords Speech · Wavelet · Linear prediction coding · Text-independent · Gaussian mixture model

1 Introduction

The information located in person's sound waves by using the process of automatic recognition is called speaker recognition [1]. To confirm a person's identity, the person's voice can be used to trigger such authority utilizing speaker recognition techniques. Applicable services are voice dialing, telebanking, teleshopping, remote access of computers, security control for confidential information, database access services, forensic purposes, information services, reservation services, and voice mail [2]. Speaker recognition has two classifications, speaker verification and speaker identification [3–5]. Speaker verification is described as the procedure of recognizing or refusing the identity declared by one of the speakers. Speaker identification defined as the identification of the person who speaks among the registered speakers.

Identifying a speaker is accomplished by comparing samples from an unknown test speaker with those of unknown speakers. Based on the best model match to his/her speech, the unknown person is identified as a speaker. In speaker verification, when an unknown person is claiming to be the owner of the identity, his speech is compared with the claimed person's model. Therefore, the authentication process is accepted if the match level is excellent and above a specified threshold [6]. The number of taken decisions is a fundamental difference between identification and verification. In identification, the number of registered models is the same as the number of decisions, therefore, the rise of models number can hurt the performance. In addition, acceptance or rejection are the only choices in the verification process. Hence, the verification performance does not depend on the number of models [7].

Text-dependent or text-independent speaker recognition systems are another characteristic of these systems. Text-dependent based on an individual's speech utterance of a

✉ Othman Alrusaini
oarusaini@uqu.edu.sa

Khaled Daqrouq
haleddaq@yahoo.com

[1] Department of Engineering and Applied Sciences, Applied College, Umm Al-Qura University, 24382 Makkah, Saudi Arabia

[2] Department of Electrical and Computer Engineering, King Abdulaziz University, 21589 Jeddah, Saudi Arabia

Published online: 09 February 2024

specific word or phrase. On the other hand, the speaker recognition system is identified regardless of the spoken word or phrase called text-independent [6, 8].

The feature extraction process consists of getting the signal's parameters (characteristics) to be used for the classification of the signal. Pattern recognition problems can be solved by extracting prominent features. The recognition of the speaker requires identifying the desired person from the speech signal based on a unique identification feature [6]. Speech signals are controversial because of the features they contain. Various features have been found to perform better for certain applications than others. So far, we have not found a feature that is ideal for all applications [6].

Most feature extraction methods use the Karhunen–Loeve Transform (KLT) [9, 10]. With outstanding results, these methods were applied to cases of text-independent speaker recognition [8]. The KLT transform is the best conversion in terms of minimum mean square error (MMSE) and maximum energy packing. Moreover, the most popular identification systems use Mel-frequency cepstral coefficient (MFCC) [11] and linear prediction cepstral coefficients (LPCC) [12, 13] as features. MFCC and LPCC have excellent features in speaker identification. The MFCC has the disadvantage of employing the short-time Fourier transform (STFT), which has double resolution time–frequency signal and the assumption that is fixed. Thus, it is difficult to identify explosive phonemes because of these characteristics.

Furthermore, the wavelet-transform is being studied by some researchers for extracting the speaker feature [14–16]. Wavelet transform [17, 18] has been widely used in various fields of science and engineering. The operation involves signal analysis using dilated and translated versions of a base signal called the mother wavelet. Using the wavelet analysis, we can express signals of interest by a set of coefficients (wavelet coefficients), and we can implement signal processing algorithms by adjusting these coefficients. From a mathematical standpoint, the mother wavelet scale can be a real positive value, and the translation value can be any real number [19]. In practice, however, to improve calculation efficiency, the translation and scale parameters are frequently limited to discrete lattices [20, 21].

The paper has been structured into five distinct parts. The first component of the investigation starts by presenting a comprehensive background. Subsequently, this study undertakes an exhaustive examination of scholarly literature sources in order to ascertain the current status of research pertaining to the topic. The article continues by outlining the technique used in the investigation, followed by the presentation and analysis of the results. The researchers ultimately engage in a discussion of these results within the context of the existing literature, derive conclusions, and provide suggestions about the optimal use of a text-independent speaker identification system.

2 Literature review

The subject of speaker recognition [22] began to develop in the mid-twentieth century. The first known published paper on this topic was in the 1950s [23, 24]. This research was interested in preserving the personal qualities of the speakers through the analysis of speech. As [23] pointed out the need to identify the speaker for the emergence of communication networks in early the 1950. Most of the early studies were based on text-dependent analysis to facilitate the task of identification. In 1959, [24] tried to facilitate the identification process and started to compare the formants of speech. Human experience incarnation of the speaker's first recognition has been used until now to deal with the speaker identification forensic [25]. Legal experts have used this type of approach for various analyses of criminal forensics [26, 27]. Pruzansky et al. [28, 29] used a text-dependent approach to make an automatic statistical comparison of speakers by analyzing 10 speakers, where each speaker utters a few unique words. At least to identify a speaker, it was clear that a text-dependent analysis method was needed [22]. However, for speaker verification, there were cases where a text-dependent analysis could perform better than the text-independent method [30]. The Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) approaches are currently the most popular modeling techniques. Classifiers such as artificial neural networks have been used [8, 22].

As per [31, 32] presented a technique for speaker identification using a frame linear predictive coding spectrum (FLPCS). The FLPCS technique can be used to reduce the size of a speaker's feature vector. In classification, the general regression neural network (GRNN) and the GMM were used. In a very short time, GMM can achieve a higher recognition rate with feature extraction using the FLPCS method. Avci [33] presented a discrete wavelet adaptive network model based on the fuzzy inference system (DWANFIS). The DWANFIS model has two layers: discrete wavelet and an adaptive network based on a fuzzy inference system. For the sample speakers, the classification rate was approximately 90.55%.

E. Avci and D. Avci [34] presented a genetic wavelet adaptive network model based on a fuzzy inference system (GWANFIS). The model was made up of three layers: a genetic algorithm, a wavelet, and an adaptive network based on a fuzzy inference system (ANFIS). The classification rate was approximately 91%. As a feature selection technique, Singular Value Decomposition (SVD) followed by QR Decomposition with Column Pivoting (QRcp) was proposed by Chakroborty

and Saha [35]. The method was accomplished by extracting the most salient information from the speaker's data. The proposed SVD-QRcp-based method outperforms the F-Ratio based method, and the proposed feature extraction tool outperforms the baseline MFCC and Linear Frequency Cepstral Coefficients (LFCC).

As per [36] presented feature analysis and compensator design for speaker recognition in stressed speech conditions. MFCC, linear prediction (LP) coefficients, LPCC, reflection coefficients (RC), arc-sin reflection coefficients (ARC), and log-area ratios (LAR) were six speech features that were widely used for speaker identification and were analyzed for evaluation of their characteristics under a stressed condition. To evaluate speaker identification results with different speaker features, the GMM and Vector Quantization (VQ) classifiers were used. This analysis aided in the selection of the best feature set for stressed-out speaker recognition.

As per [37] demonstrated speaker identification using empirical mode decomposition (EMD), a feature extraction method, and an artificial neural network. The EMD is a nonlinear, non-stationary data analysis technique that uses adaptive multi-resolution decomposition. The proposed system's performance and training time were validated using back-propagation neural networks (BPNN) and GRNN. The experimental results showed that the GRNN outperformed the BPNN in terms of feature extraction using the EMD method.

Daqrouq [38] presented a text-independent speaker identification method using wavelet transform (WT) and neural networks. In the text-independent speaker identification system, the use of a discrete wavelet transform approximating sub-signals of the original signal via several levels instead of the original imposter had a good performance on Additive white Gaussian noise (AWGN) facing, particularly on levels 3 and 4. As per [39, 40] used a fused Mel feature set and GMM to develop a method for text-independent speaker identification. Each speaker's MFCC and Inverted Mel Frequency Cepstral Coefficient (IMFCC) features were obtained. The identification efficiency of this method was 93.88%. As per [41–43] presented a text-independent speaker identification system using an average framing linear prediction coding (AFLPC) technique. The distinguished speaker's vocal tract characteristics were extracted using the AFLPC technique during the feature extraction stage, and the size of the feature vector was optimized. The probabilistic neural network (PNN) classifier outperformed the wavelet packet (WP) and AFLPC in terms of recognition rate 97.36%.

3 Methodology

The Discrete Wavelet Transform (DWT) is a mathematical technique used to split a given signal into a set of coefficients, which correspond to various frequency bands and temporal scales. This characteristic makes it a potent instrument for extracting speech signal elements that are pertinent to speaker identification. Linear predictive coding (LPC) is a statistical technique used for modeling the correlation between previous and current values of a given signal. The use of this technique enables the anticipation of forthcoming signal values. The DWT and LPC are two complementing methodologies that may be effectively used to enhance the precision of speaker identification systems [30].

3.1 Wavelet Speaker Identification Method

By employing low pass and high pass filters, g and h generated from wavelets parents; scaling and mother functions denoted by φ and Ψ, respectively, we obtain approximation and detail coefficients of a speech signal X through the DWT and are given by:

$$a_X(j+1,k) = \big((a_X(j) * g) \downarrow 2\big)(k) = \sum_{m\epsilon Z} g_{2k-m} a_x(j,m),$$

$$\text{And } d_X(j+1,k) = \big((a_X(j) * h) \downarrow 2\big)(k) = \sum_{m\epsilon Z} h_{2k-m} a_X(j,m), \text{ respectively} \quad (1)$$

where $j \in \{1,2,\ldots,J\}, k \in \{1,2,\ldots,n_j - 1\}$,* is convolution, ↓ is decimation, and n_j represents the number of DWT coefficients at level j. We assume that

$$D_X(1) = \{d_X(1,0), d_X(1,1), \ldots, d_X(1,n_j-1)\},$$

$$D_X(2) = \{d_X(2,0), d_X(2,1), \ldots, d_X(2,n_j-1)\},\ldots$$

$$D_X(J) = \{d_X(J,0), d_X(J,1), \ldots, d_X(J,n_j-1)\}, \text{ and}$$

$$A_X(J) = \{a_X(J,0), d_X(J,1), \ldots, a_X(J,n_j-1)\}, \quad (2)$$

are the DWT sub-signals.

First of all, we divide the speech signal into windows and decompose each window separately into DWT sub-signals. And then, each sub-signal $D_X(1), D_X(2), \ldots, D_X(J)$, and $A(J)$ is divided into S frames as follows:

$$D_X(j) = \{frame_{X1}, frame_{X2}, \ldots, frame_{XS}\} \quad (3)$$

For each frame d_{X_s} LPC coefficients of a specific length are obtained as follows:

$$LPC_{D_X(j)} = \left\{lpc_{f_{X1}}, lpc_{f_{X2}}, \ldots, lpc_{f_{XS}}\right\}.$$

Then are averaged as follows:

$$wlpca_{D_x(j)} = \frac{\sum_S (LPC_{D_X(j)})}{S} \tag{4}$$

Then the feature extraction vector of one window is considered as:

$$WLPCA = \left\{ wlpca_{D_x(1)}, wlpca_{D_x(2)}, \ldots, wlpca_{D_x(J)} \right\} \tag{5}$$

The feature extraction matrix contained *WLPCA* of all widows is fed to the GMM classifier**.** Consider F is a feature matrix extracted by *WLPCA* for speaker X, its mixture density is defined as

$$p(F/\lambda_X) = \sum_{i=1}^{M} p_i^X b_i^X(F) \tag{6}$$

The density $b_i^X(F)$ is a linear weighted combination of M unimodal Gaussian densities.

$$b_i^X(F) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i^X\right|^{1/2}} * \exp\{-1/2(F-\mu_i^X)\prime\left(\Sigma_i^X\right)^{-1}(F-\mu_i^X)\}s \tag{7}$$

where μ_i^X is the mean vector, and Σ_i^X is the covariance matrix. Then the mixture weights p_i^X satisfy the constraint.

$$\sum_{i=1}^{M} p_i^X = 1 \tag{8}$$

The parameters for a speaker X's density model are extracted as

$$\lambda_X = \left\{ p_i^X, \mu_i^X, \Sigma_i^X \right\}, i = 1, \ldots, M \tag{9}$$

The iterative Expectation–Maximization (EM) algorithm is used to estimate the maximum likelihood speaker model parameters. In this study, a simple maximum-likelihood classifier was used for identification. For a reference database speakers $\updownarrow = \{1, 2,..., Y\}$ represented by models $\lambda_1, \lambda_2, \ldots, \lambda_Y$, we have to recognize the speaker model of the maximum posterior probability for the input feature vector sequence, $F = \{f_1, f_2, \ldots, f_T\}$ [44–46].

Table 2 The effect of the number of DWT levels on the recognition rate

	DWT levels		
	3	4	5
Recognition rate[2]	0.9286	0.9390	0.9190

[2]Number of Gaussians = 10, Window length = 400, Number of LPC coefficients = 12

4 Results and discussions

4.1 Recorded database

Speech signals with a spectral frequency of 4000 Hz and a sampling frequency of 8000 Hz were recorded using a PC sound card. The recordings involved 50 people. Each participant recorded a minimum of 20 different Arabic utterances. The speakers ranged in age from 20 to 45 years old, with 28 men and 22 women speaking. The recording procedure was carried out in a typical university office setting. This database was used in the first stage of our investigation. All experiments were conducted using the text-independent speaker identification system. The normalized and silence removed signals were given to the WLPCA (discrete wavelet transform using a linear prediction coding algorithm) for feature vectors extraction. Then the GMM was utilized to model the feature vectors obtained for each speaker. Half of the speaker's signals were used for training, with the other half reserved for testing. All speakers in our database were used for algorithm evaluation. This section's two main objectives are to investigate the best WLPCA parameters and compare WLPCA as a new method with the well-known feature extraction method, MFCC. The experiments were conducted concerning the recognition rate.

In the first experiment, the system was run for six several LPC coefficient vector lengths 5, 10, 12, 15, 20, and 30. Four decomposition levels of the DWT with window length 400 and 10 Gaussian mixtures we applied. Table 1 summarizes the results of this experiment for the DWT. The presented results are conducted in terms of recognition rate calculated as the ratio of the number of times the speakers are correctly recognized by the total number of test signals. The best result was observed for 12 LPC coefficients.

Table 1 The effect of the number of LPC coefficients on the recognition rate

	Number of LPC coefficients					
	5	10	12	15	20	30
Recognition rate[1]	0.9357	0.9357	0.9390	0.9214	0.8976	0.8643

[1]Number of Gaussians = 10, Window length = 400, DWT Level = 4

In the next experiment, the process was repeated for different DWT decomposition levels 3, 4, and 5, with 12 LPC coefficients. Window length was 400, and Gaussian mixtures were set to 10. Table 2 presents the results of this experiment. The best recognition rate (0.939) of the presented method was observed for 4 DWT decomposition levels.

Table 3 contains the results of the recognition rates calculated for different window lengths. The system parameters were set as follows: the DWT level is 4, the number of the LPC coefficients is12 and the number of Gaussian mixtures is 10. By analyzing the results tabulated in Table 3, we can notice that the best performance is at window length 400 (relatively small window size). The reason behind this is the ability to divide the signal into more windows. So, we can gain more features over the same length of the signal. On the other hand, a smaller window than 400 could be of less benefit.

In the next part of this study, we compare WLPCA with other published well-known feature extraction methods. A AFLPC method with wavelet packet (WPLPCF) at WP level 2 [41, 42], the conventional LPC [47], AFLPC [41, 42] and MFCC [48]. With 10 Gaussian mixtures, the proposed method gave the best recognition rate as summarized in Table 4.

The results of WP were auspicious as much as DWT. The limitation for WP is the feature extraction vector length in comparison to DWT. However, it is difficult to increase the WP level as the window length is constant. In contrast to WP, DWT has more flexibility in terms of decomposition level. In the case of WP at level two, the feature extraction vector's length is near to the length of the vector obtained by DWT at level 4.

The proposed method produced better results than the MFCC method overall suggested Gaussian mixtures. The reason behind that is that the LPC coefficients in our method are obtained through different frequency passbands. The decomposition offers that into several DWT levels. Additionally, the averaging over the frames over each sub-signal can help gain more accurate results.

Table 4 Recognition rate comparisons between several feature extraction methods[4]

Feature extraction methods	Number of Gaussian Mixtures			
	3	5	7	10
WLPCA	0.9048	0.9238	0.9333	0.9390
WPLPCF	0.9022	0.9280	0.9304	0.9380
LPC	0.7810	0.8119	0.8381	0.8505
AFLPC	0.7833	0.8310	0.8476	0.8105
MFCC	0.8786	0.9071	0.9071	0.9038

[4]Window length = 400, number of extracted features using LPC = 12, DWT Level = 4

4.2 TIMIT database

A real standard database could be an essential tool for proving out a recognition algorithm. For the generalization of our new method, a standard common database is required. Without a doubt, the TIMIT database has been one of the most widely used standard databases [49]. There are 630 speakers of the same dialect in this database. that are divided into 438 males and 192 females. 10 utterances were recorded for each speaker with an 8000Hz sampling frequency, a wideband microphone was used in a clean environment.

In the next experiment, WLPCA and MFCC are tested using the TIMIT database. The signals were preprocessed by silence removing algorithm. GMM was trained using 8 utterances out of the 10 for each class (speaker). The remaining two utterances were used for testing. The TIMIT database recognition rate was calculated by running experiments on 50 randomly selected speaker sets from a total of 630 speakers and averaging the resulting recognition rates. Figure 1 illustrates the results of the TIMIT database. The experiment was performed for 5, 10, 20, and 30 Gaussian mixtures. 10 random sets were taken for calculating 10 average recognition rates for each of these Gaussian mixture numbers. As shown in the figure, the WLPCA has better performance, particularly with a small number of Gaussian mixtures.

In contrast to WLPCA, MFCC behaved better with a high number of Gaussian mixtures. Nevertheless, the running time is increased to the extent that the number of Gaussian mixtures is increased. For instance, MFCC with 30 Gaussian mixtures needs 7.5 min to get the max recognition rate for a set of 50 speakers. Considering that, WLPCA needs 3.8 min to get the max recognition rate. So, we can state that WLPCA requires a smaller number of Gaussian mixtures and requires less running time than MFCC.

Cross-validation continuously processes recognition performance by excluding a few instances (10% for tenfold cross-validation) to be used as the test set during the training process [15pdz, 19pdz]. We ran a tenfold cross-validation test to compare our algorithm to the validation technique. In addition, a 20-fold cross-validation test was carried out.

Table 3 The effect of window length on the recognition rate

	Windows length							
	100	200	300	400	500	600	700	1000
Recognition rate[3]	0.9133	0.9371	0.9181	0.9390	0.9124	0.8952	0.8952	0.8762

[3]Number of Gaussians = 10, DWT levels = 4, Number of LPC coefficients = 12

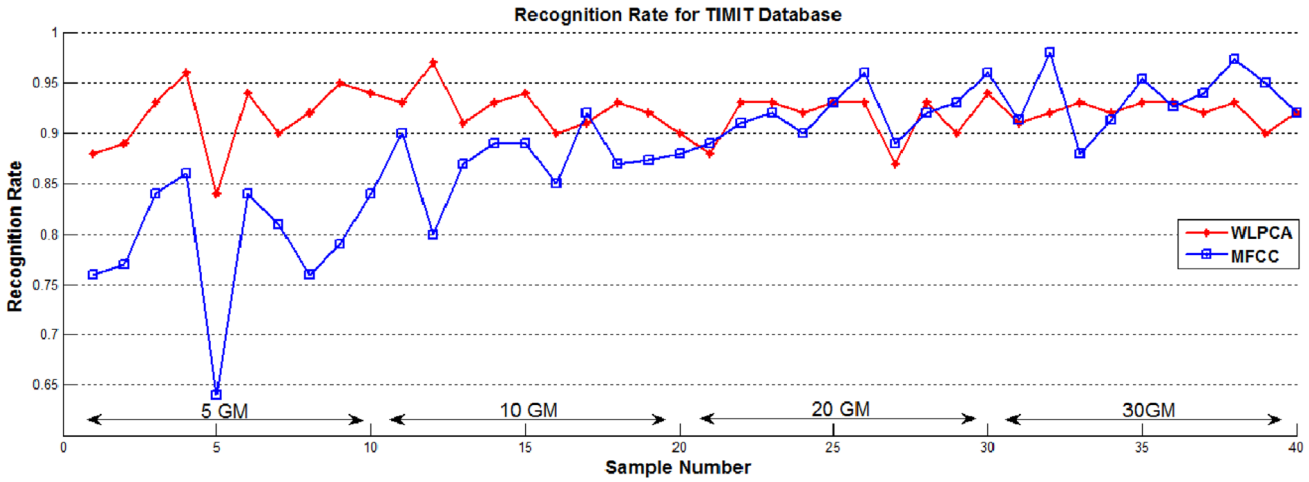

Fig. 1 The results of the TIMIT database

Table 5 Tenfold and 20-fold cross-validation tests results

Method	10-Fold		20-Fold		Avg
	5 GM	30 GM	5 GM	30 GM	
WLPCA	0.9050	0.9200	0.9150	0.9210	0.9153
MFCC	0.7820	0.9236	0.7910	0.9340	0.8577

Table 6 The recognition rate for 5 and 10 Gaussian mixtures

Method	0 dB SNR		10 dB SNR		Avg
	5 GM	30 GM	5 GM	30 GM	
WLPCA	0.5306	0.3751	0.7390	0.7561	0.6002
MFCC	0.4220	0.5133	0.6450	0.8152	0.5989

Table 5 shows tenfold cross-validation and 20-fold cross-validation tests results. 10 random sets were taken for calculating each recognition rate for five and thirty Gaussian mixture numbers. As shown in the table, that our proposed method is stable and not sensitive to the validation technique.

In the next experiment, WLPCA and MFCC were tested in a noisy environment. Additive white Gaussian noise (AWGN) was added to the TIMIT database signals. TIMIT has about 59 dB signal-to-noise ratio (SNR). We conducted the identification experiment over 0 dB and 10 dB SNR. The analysis was performed for 5 and 30 Gaussian mixtures. 10 random sets were taken for calculating each recognition rate for each of these Gaussian mixture numbers in 0 dB and 10 dB SNR environments. Table 6 summarizes the recognition rate for 5 and 30 Gaussian mixtures in 0 dB and 10 dB SNR environments. The results show that our method performs slightly better than MFCC method.

5 Conclusion

This research has a new speaker feature extraction method based on a discrete wavelet transform using a linear prediction coding algorithm (WLPCA) is proposed. It leads to some improvement in the recognition rate over MFCC. The proposed method was tested regarding the linear prediction coding coefficients length, the discrete wavelet-transform level, the window length, and the Gaussian mixture number. Experimental evaluation of the method performance was performed on two speech databases; our recorded database and the publicly available TIMIT database. The presented work compares the proposed method's performance with the MFCC method in the computation of the feature extraction in speaker recognition. It showed that the speech features derived by the newly proposed method resulted in a more suitable representation, in terms of the time-consumed, by requiring less Gaussian mixtures in comparison to MFCC. It also showed a slight improvement in the implementation of the newly proposed method over the MFCC in a noisy environment. The idea behind that is that the LPC coefficients in our method are obtained through different frequency passbands. The decomposition offers that into several DWT levels. Additionally, the averaging over the frames through each sub-signal can help gain more accurate results.

Author contributions Conceptualization, OA and KD; methodology, OA and KD; software, OA and KD.; validation, OA and KD; formal analysis, OA and KD; investigation, OA and KD; resources, KD and OA; data curation, OA and KD; writing—original draft preparation, OA and KD; writing—review and editing, OA and KD.; visualization, OA; supervision, KD and OA.; project administration, KD; All authors have read and agreed to the published version of the manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The datasets used in this article are open source and available.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Furui S (2018) Digital speech processing, synthesis, and recognition. CRC Press
2. Nehra N, Sangwan P, Kumar D (2021) Artificial neural networks: a comprehensive review, Handbook of Machine Learning for Computational Optimization, pp 203–227
3. Zhang Z, Geiger J, Pohjalainen J, Mousa AE-D, Jin W, Schuller B (2018) Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans Intelligent Syst Technol (TIST) 9(5):1–28
4. Loweimi E, Cvetkovic Z, Bell P, Renals S (2021) Speech acoustic modelling using raw source and filter components. In: Presented at the Interspeech 2021: the 22nd annual conference of the international speech communication association
5. Hanifa RM, Isa K, Mohamad S (2021) A review on speaker recognition: technology and challenges. Comput Electr Eng 90:107005
6. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R (2017) Speaker identification features extraction methods: a systematic review. Expert Syst Appl 90:250–271
7. Sinha K, Hameed RS, Paul P, Singh KP (2021) Voice-Based Speaker Identification and Verification. In: Handbook of Research on Knowledge and Organization Systems in Library and Information Science: IGI Global, pp 288–316
8. Bai Z, Zhang X-L (2021) Speaker recognition based on deep learning: An overview. Neural Netw 140:65–99
9. Almaadeed N, Aggoun A, Amira A (2015) Speaker identification using multimodal neural networks and wavelet analysis. Iet Biometrics 4(1):18–28
10. Krobba A, Debyeche M, Selouani S-A (2020) Mixture linear prediction Gammatone Cepstral features for robust speaker verification under transmission channel noise. Multimedia Tools Appl 79:18679–18693
11. Abdul ZK, Al-Talabani AK (2022) Mel Frequency Cepstral Coefficient and its applications: a review, IEEE Access
12. Almaadeed N, Aggoun A, Amira A (2016) Text-independent speaker identification using vowel formants. J Signal Process Syst 82:345–356
13. Do HD (2022) Exploiting signal linear trend in the time domain to enhance speech feature. IEEE Access 10:117886–117899
14. Akujuobi CM (2022) Wavelets and wavelet transform systems and their applications: a digital signal processing approach. Springer Nature
15. Szeliski R (2022) Computer vision: algorithms and applications. Springer Nature
16. Tang Y et al (2022) Attention based gender and nationality information exploration for speaker identification. Digital Signal Processing 123:103449
17. Yildirim Ö (2018) A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med 96:189–202
18. Kumar S, Kumar R, Agarwal RP, Samet B (2020) A study of fractional Lotka-Volterra population model using Haar wavelet and Adams-Bashforth-Moulton methods. Math Methods Appl Sci 43(8):5564–5578
19. Duan Y, Liu F, Jiao L, Zhao P, Zhang L (2017) SAR image segmentation based on convolutional-wavelet neural network and Markov random field. Pattern Recogn 64:255–267
20. P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.
21. Lin C, Gao W, Guo M-F (2019) Discrete wavelet transform-based triggering method for single-phase earth fault in power distribution systems. IEEE Trans Power Delivery 34(5):2058–2068
22. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E (2021) A comprehensive review of speech emotion recognition systems. IEEE access 9:47795–47814
23. I. Pollack, J. M. Pickett, and W. H. Sumby, "On the identification of speakers by voice," *the Journal of the Acoustical Society of America,* vol. 26, no. 3, pp. 403–406, 1954.
24. Shearme J, Holmes J (1959) An experiment concerning the recognition of voices. Lang Speech 2(3):123–131
25. Yule G (2022) The study of language. Cambridge university press
26. Johnson K, Sjerps MJ (2021) Speaker normalization in speech perception, The handbook of speech perception, pp 145–176
27. Ahmed AI, Chiverton JP, Ndzi DL, Becerra VM (2019) Speaker recognition using PCA-based feature transformation. Speech Commun 110:33–46
28. Pruzansky S (1963) Pattern-matching procedure for automatic talker recognition. J Acoustical Soc Am 35(3):354–358
29. Pruzansky S, Mathews MV (1964) Talker-recognition procedure based on analysis of variance. J Acoustical Soc Am 36(11):2041–2047
30. Ketabi S, Rashidi S, Fallah A (2023) Text-dependent speaker verification using discrete wavelet transform based on linear prediction coding. Biomed Signal Process Control 86:105218
31. Wu J-D, Lin B-F (2009) Speaker identification based on the frame linear predictive coding spectrum technique. Expert Syst Appl 36(4):8056–8063
32. Jahangir R, Teh YW, Nweke HF, Mujtaba G, Al-Garadi MA, Ali I (2021) Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. Expert Syst Appl 171:114591
33. Avci D (2009) An expert system for speaker identification using adaptive wavelet sure entropy. Expert Syst Appl 36(3):6295–6300
34. Avci E, Avci D (2009) The speaker identification by using genetic wavelet adaptive network based fuzzy inference system. Expert Syst Appl 36(6):9928–9940
35. Chakroborty S, Saha G (2010) Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification. Speech Commun 52(9):693–709

36. Rituerto-González E, Mínguez-Sánchez A, Gallardo-Antolín A, Peláez-Moreno C (2019) Data augmentation for speaker identification under stress conditions to combat gender-based violence. Appl Sci 9(11):2298
37. Kabir MM, Mridha MF, Shin J, Jahan I, Ohi AQ (2021) A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. IEEE Access 9:79236–79263
38. Daqrouq K (2011) Wavelet entropy and neural network for text-independent speaker identification. Eng Appl Artif Intell 24(5):796–802
39. Kumari RSS, Nidhyananthan SS (2012) Fused MEL feature sets based text-independent speaker identification using Gaussian mixture model. Proc Eng 30:319–326
40. Kp B (2020) ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score. Multimedia Tools Appl 79:28859–28883
41. Daqrouq K, Al Azzawi KY (2012) Average framing linear prediction coding with wavelet transform for text-independent speaker identification system, Comput Electr Eng 38(6): 1467–1479
42. Daqrouq K, Morfeq A, Ajour M, Alkhateeb A (2013) Wavelet LPC with neural network for speaker identification system. WSEAS Trans Signal process 9:216–226
43. Hidayat S, Tajuddin M, Yusuf SAA, Qudsi J, Jaya NN (2022) Wavelet detail coefficient as a novel wavelet-mfcc features in text-dependent speaker recognition system. IIUM Eng J 23(1):68–81
44. Nautsch A et al (2019) Preserving privacy in speaker and speech characterisation. Comput Speech Lang 58:441–480
45. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: A systematic review. IEEE access 7:19143–19165
46. Ullah F, Israr M, Jan A, Ahmad AM, Dullah I, Ullah F (2020) Development of a novel system for speaker verification, In: 2020 International conference on intelligent engineering and management (ICIEM), pp 12–16: IEEE
47. Li P, Hu F, Li F, Xu Y (2014) Speaker identification using linear predictive cepstral coefficients and general regression neural network, In: Proceedings of the 33rd Chinese Control Conference, pp. 4952–4956: IEEE.
48. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52(1):12–40
49. Al-Kaltakchi MT, Al-Nima RRO, Abdullah MA, Abdullah HN (2019) Thorough evaluation of TIMIT database speaker identification performance under noise with and without the G. 712 type handset. Int J Speech Technol 22:851–863