INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 2, PP. 1–7
Manuscript received January 9, 2026; revised April 2026. doi: 10.24425/ijet.2026.157935

Robust text-independent speaker identification and verification using multi-feature fusion and student's t modelling

Musab T.S. Al-Kaltakchi, Mohanad Abd Shehab, Emad A. Hussien, and Amal Ibrahim Nasser

***Abstract*—This paper presents a text-independent speaker identification system that utilizes MFCC, LPC, prosody, and optimized multi-level DWT features for robust speaker modeling. The system is designed for multiple standard speech databases, including TIMIT, NTIMIT, SITW, and NIST2008. During training, features from each speaker are normalized to zero-mean and unit-variance, and Student-t distributions are fitted to model the statistical characteristics of each speaker. For testing, features are normalized using the corresponding speaker's training statistics, and speaker identity is predicted based on maximum log-likelihood estimation over the trained models. Experimental results confirm the superiority of the proposed system, which achieves high accuracy across multiple datasets (e.g., 98.33% on TIMIT, 89.38% on NTIMIT, 96.88% on SITW, and 100.00% on NIST2008) and consistently outperforms existing state-of-the-art methods under AWGN conditions, demonstrating significant improvements in identification accuracy and the effectiveness of multi-feature fusion and Student-t modeling.**

***Keywords*—Speaker identification and Verification, Student's t model, MFCC, LPC, Prosody, Optimized multi-level DWT, Fusion approach**

I. Introduction

SPEAKER identification's primary aim is to identify the speaker by comparing an unidentified voice sample to a list of recognized speakers in a database. "Whose voice is this?" is the question it answers. Using the distinctive physical (vocal tract shape, larynx size) and behavioral (accent, pitch, speaking style) features of a person's voice to create a unique "voiceprint" or model, this procedure is a type of biometric authentication [1].

In the work [2], the resilience to noise and the use of spectrograms as input to these spatial networks are examined. The SpectroNet model for speech-based speaker identification is presented in this work in order to reduce memory requirements (storage) and speed up training (computation). The proposed system was evaluated using Voxceleb1 and Part 1 of the RSR 2015 databases. The experimental results show a relative improvement in speaker identification of approximately 16% (accuracy: 96.21%) with spectrograms and about 10% (accuracy: 98.92%) with log Mel spectrograms when compared to existing models. When the cochleagram was used, the recognition accuracy was 99.26%.

The research article [3] introduces a novel approach to text-independent speaker recognition by integrating Mel-Frequency Cepstral Coefficients (MFCCs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks, with noise removal facilitated by Convolutional Neural Networks (CNNs). The primary objective is to improve the robustness and precision of speaker reputation systems in real-global environments where history noise is general. A CNN-based noise removal mechanism that reduces historical past interference is introduced to the advised method, which makes use of MFCC functions to capture speech timbral traits. A Bi-LSTM community, which efficaciously models temporal dependencies in speech records, techniques the ensuing denoised MFCC capabilities. Speaker popularity accuracy has drastically stepped forward, achieving 98.17% at a sign-to-noise ratio of 30 dB, in keeping with experiments conducted on publicly reachable datasets. This technique indicates how deep getting to know, noise reduction, and complex characteristic extraction can paintings collectively to provide reliable speaker recognition in noisy settings. Speaker recognition accuracy has significantly improved.

Automatic speaker identity (SI) has received reputation due to the latest tendencies in deep learning and hardware [4]. Nevertheless, there isn't a thorough analysis of current SI methods and their advantages and disadvantages. By reviewing several SI fields and outlining the future research issues that require attention, this study seeks to close that gap.

Natural language processing and image recognition are just two of the many fields in which deep learning (DL), a potent machine learning technique, finds use. Conventional learning methods limit their potential in practical tasks, despite their success. Although DL has outperformed conventional methods in speaker identification [5], the research community is ignorant of its application in this field. By examining DL approaches and algorithms pertinent to speaker identification, classifying them according to their application procedures, and

Musab T.S. Al-Kaltakchi is with Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq; (e-mail: m.t.s.al_kaltakchi@uomustansiriyah.edu.iq).

Mohanad Abd Shehab is with Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq; (e-mail: mohanadshehab@uomustansiriyah.edu.iq).

Emad A. Hussien is with Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq; (e-mail: dr.emadeng@uomustansiriyah.edu.iq).

Amal Ibrahim Nasser is with Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq; (e-mail: amalalshemmiri@uomustansiriyah.edu.iq).

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

talking about their potential future research, this paper seeks to close that gap.

In the paper [6], the I-vector model and the Gaussian Mixture Model-Universal Background Model (GMM-UBM) are compared for speaker identification. Four combinations of I-vectors and seven fusion techniques (maximum, mean, weighted sum, cumulative, interleaving, and concatenated) are tested. An Extreme Learning Machine (ELM) is also used for identification, and Speaker Identification Accuracy (SIA) is calculated for 120 speakers from the TIMIT and NIST 2008 databases under clean speech conditions. The evaluation includes Additive White Gaussian Noise (AWGN) and Non-Stationary Noise (NSN) effects. Results indicate that the I-vector method outperforms GMM-UBM in clean and AWGN conditions without a handset, while GMM-UBM shows better performance for NSN types.

In the paper [7], a brand new finite Student's-t aggregate model (SMM) for picture segmentation is introduced, emphasizing its robustness in comparison to Gaussian fashions due to its heavy-tailed distribution. Unlike present models, this method explicitly contains the spatial relationships between pixels by means of using the Dirichlet distribution and the Dirichlet regulation to enforce neighborhood spatial constraints. The new version at once addresses the Student's-t distribution for parameter estimation, averting the complexity of representing it as an infinite combination of scaled Gaussian, as seen in earlier models. Rather than using the expectancy maximization (EM) set of rules, the proposed method utilizes a gradient technique to decrease the top sure on the information's negative log-chance, optimizing model parameters correctly. Comparative numerical experiments display the version's effectiveness towards current ultra-modern finite aggregate models on simulated and actual scientific snap shots, showcasing its capability for stepped forward picture segmentation overall performance. Feature fusion is specially valuable in text-unbiased scenarios, in which the gadget need to understand a speaker irrespective of the spoken content material, as it reduces the chance of overlooking important identification cues that can most effective appear in a specific function domain [8]. Equally important is the choice of statistical model for representing the fused feature space. Gaussian Mixture Models (GMMs) have long been popular due to their flexibility and well-understood behavior [9], [10].However, their assumption of normality regularly falls short in actual-international speech, that may show off heavy tails and outliers. To triumph over this hassle, recent work has explored more strong probabilistic frameworks. The multivariate Student's t-distribution has emerged as a sturdy opportunity, seeing that its tiers of freedom parameter obviously incorporates heavier tails. By modelling each speaker's characteristic distribution with a multivariate t-density, structures grow to be more resilient to odd observations and non-Gaussian variability, in the end improving the reliability of likelihood estimates at some stage in identification [11].

In this paper, we present a textual content-independent speaker identification and verification system that integrates multi-characteristic fusion with Student's t modelling. Our method combines MFCC, LPC, prosodic, and optimized DWT functions, making use of according to-speaker normalization to beautify consistency. The fused function vectors are modeled for each speaker the usage of a multivariate Student's t-distribution, providing a principled and strong probabilistic framework for classification. This approach has clean benefits: the heavy-tailed nature of the Student's t-distribution certainly handles outliers and variability in speech alerts, providing extra robustness than traditional Gaussian fashions. As a end result, the system grants more dependable and correct speaker fashions, specifically while schooling facts is confined or noisy. Evaluated on benchmark datasets including TIMIT, NTIMIT, SITW, and NIST2008, the proposed framework demonstrates advanced accuracy, underscoring the efficacy of mixing diverse feature representations with a sturdy statistical model to strengthen the performance of speaker identification structures.

II. Proposed Method

The proposed speaker identification system methodology consists as shown in Figure 1 of four main stages: data preprocessing, feature extraction, speaker modeling, and testing.

A. Data Preprocessing

The system is designed for text-independent speaker identification and has been evaluated across multiple benchmark databases, including TIMIT, NTIMIT, SITW, and NIST2008. Each speaker has multiple training and testing utterances. To ensure consistency, all audio signals are resampled to a consistent sampling rate of 16 kHz. Although the current implementation uses the entire utterance, silence and noise segments can be optionally removed during preprocessing.

B. Feature Extraction

The system extracts a multi-level fused feature vector that includes the following in order to capture speech's spectral and temporal characteristics:

- **MFCC (Mel-frequency cepstral coefficients):** 13 coefficients capturing perceptually relevant spectral information.
- **LPC (Linear Predictive Coding):** 12 coefficients modeling vocal tract dynamics.
- **Prosody features:** 3 coefficients of the fundamental frequency (pitch), energy, and zero-crossing rate.
- **Optimized multi-level DWT (Discrete Wavelet Transform):** 12 features generated from 4 statistical descriptors (mean, standard deviation, skewness) from a three-level wavelet decomposition capturing time-frequency patterns.

For each speaker, the 40-features from all training utterances are concatenated and normalized to zero mean and unit variance to reduce inter-session variability.

C. Speaker Modeling

Speaker identity is modeled using the Student-t distribution, which is robust to outliers and heavy-tailed feature distributions. For each speaker, the normalized training features are used to estimate the degrees of freedom, mean vector, and covariance matrix for each speaker. To prevent singularities, covariance matrices are regularized.

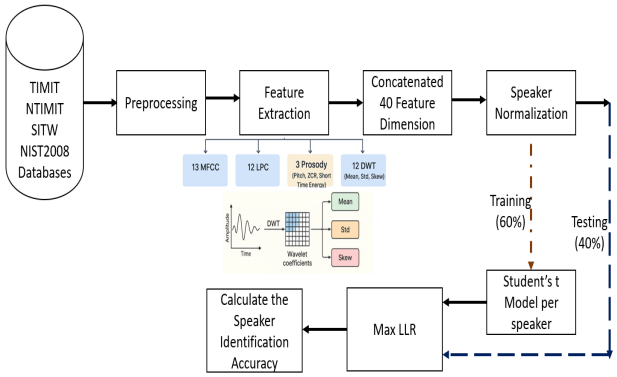

Fig. 1. Block Diagram of the Proposed Student's t Speaker Identification Model

D. Testing and Identification

The corresponding speaker's training statistics are used to normalize the features that are taken from unseen utterances during testing. The log-likelihood is calculated for every speaker model for every test feature vector. The predicted identity is chosen to be the speaker associated with the model that has the highest average log-likelihood. The percentage of correctly classified test utterances is used to calculate identification accuracy.

III. Mathematical Implementation

A. Feature Extraction

Four features are extracted from the speech signal x[n]:
a) MFCC (Mel-Frequency Cepstral Coefficients) For a frame $x_f[n]$, the MFCC coefficients are:

$$\mathrm{MFCC}_k = \sum_{m=0}^{M-1} \log\big(X_f[m]\big) \cdot \cos\left(\frac{\pi k}{M}\left(m+\tfrac{1}{2}\right)\right) \quad (1)$$

$$k = 1, 2, \ldots, \text{numMFCC}$$
$$X_f[m] = \text{FFT of the frame } x_f[n]$$
$$M = \text{number of Mel filter banks}$$
$$k = \text{MFCC coefficient index}$$

b) LPC (Linear Predictive Coding)
For each frame, LPC coefficients a_i are computed using the predication error minimization:

$$x_f[n] \approx \sum_{i=1}^{p} a_i\, x_f[n-i] \quad (2)$$

where

$$n = p+1, \ldots, N$$
$$p = \text{LPC order } \ (\text{numLPC} = 12)$$
$$N = \text{frame length}$$
$$a_i = \text{LPC coefficients}$$

c) Prosody Features

1) **Pitch (Fundamental Frequency):**

$f_0[n]$ from autocorrelation or cepstral method.

2) **Energy:**

$$E_f = \sqrt{\frac{1}{L}\sum_{n=0}^{L-1}\big(x_f[n]\big)^2} \quad (3)$$

where L is the frame length.

3) **Zero-Crossing Rate (ZCR):**

$$\mathrm{ZCR}_f = \frac{1}{2L}\sum_{n=1}^{L}\big|\mathrm{sign}[x_f[n]] - \mathrm{sign}[x_f[n-1]]\big| \quad (4)$$

d) Optimized Multi-Level DWT (3 levels)
For each frame $x_f[n]$, a three-level DWT is computed, producing the coefficients:

$$[cA_3, cD_3, cD_2, cD_1] = \mathrm{DWT}_3(x_f[n])$$

The statistical features (mean, standard deviation, skewness) are then calculated:

$$\text{feat\_DWT} = [\mu(cA_3), \sigma(cA_3), \mathrm{skew}(cA_3), \ldots, \mathrm{skew}(cD_1)]$$

e) Fused Feature Vector
All features are concatenated to form the frame-level feature vector:

$$f_t = [\mathrm{LPC}_f, \mathrm{MFCC}_f, \mathrm{Prosody}_f, \mathrm{DWT}_f] \in \mathbb{R}^d$$

where the total feature dimension is:

$$d = \text{numMFCC}(13) + \text{numLPC}(12) + 3\ (\text{prosody}) + 12\ (\text{DWT})$$

B. Per-Speaker Normalization

For each speaker s, the feature matrix $F_s \in \mathbb{R}^{N_s \times d}$ is normalized as:

$$F_s^{\text{norm}} = \frac{F_s - \mu_s}{\sigma_s} \quad (5)$$

where

$$\mu_s = \frac{1}{N_s}\sum_{i=1}^{N_s} f_i, \quad \sigma_s = \sqrt{\frac{1}{N_s-1}\sum_{i=1}^{N_s}(f_i-\mu_s)^2}$$

C. Speaker Modeling (Student-t Distribution)

Each speaker is modeled as a multivariate Student-t distribution:

$$p(f \mid \mu_s, \Sigma_s, \nu_s) = \frac{\Gamma\left(\frac{\nu_s+d}{2}\right)}{\Gamma\left(\frac{\nu_s}{2}\right)(\nu_s\pi)^{d/2}|\Sigma_s|^{1/2}} \left[1 + \frac{1}{\nu_s}(f-\mu_s)^T\Sigma_s^{-1}(f-\mu_s)\right]^{-\frac{\nu_s+d}{2}} \quad (6)$$

where

$$
\begin{aligned}
f &\in \mathbb{R}^d \quad \text{(feature vector)},\\
\mu_s &\in \mathbb{R}^d \quad \text{(mean of speaker } s\text{)},\\
\Sigma_s &\in \mathbb{R}^{d\times d} \quad \text{(covariance matrix)},\\
\nu_s &= \text{degree of freedom (5 in our case)},\\
d &= \text{feature dimension}.
\end{aligned}
$$

D. Speaker Identification

The log-likelihood of speaker s over N_f frames is:

$$\log L_s = \frac{1}{N_f}\sum_{i=1}^{N_f}\log p(f_i \mid \mu_s, \Sigma_s, \nu_s) \tag{7}$$

The predicted speaker is:

$$\hat{s} = \arg\max_s(\log L_s) \tag{8}$$

The overall speaker identification accuracy is:

$$\text{Accuracy} = \frac{\#\text{correct}}{\text{total}} \times 100\% \tag{9}$$

IV. The Datasets

A. TIMIT acoustic-phonetic continuous speech corpus

TIMIT corpus which is the widely adopted benchmark in speech processing research , [1], [12] comprises of 630 speakers representing eight primary dialects of American English. To ensure comparability with other studies [13], [14], a subset of 120 speakers was selected from dialect regions 1 and 4. Each speaker contributes ten utterances, of which six were allocated for training and four for testing. All 1,200 utterances were standardized to a fixed length of 129,250 samples (equivalent to 8 seconds) through truncation or concatenation as necessary.

B. The NTIMIT Corpus

The NTIMIT corpus was utilized to evaluate the system's robustness under channel-degraded conditions. Derived from the original TIMIT database, NTIMIT was created by re-transmitting the clean, studio-quality recordings through the telephone network, introducing characteristic narrowband filtering (0.3-3.4 kHz), codec artifacts, and real-world line noise [15]. To maintain a consistent experimental framework, the same 120 speakers from dialect regions 1 and 4 were selected. The audio was upsampled to 16 kHz for uniformity, and the identical utterance partitioning scheme was applied: six utterances per speaker for training and the remaining four for testing, each standardized to a fixed length of 8 seconds. The controlled pair of datasets provided by this direct derivation from TIMIT allows for an accurate evaluation of performance degradation that can be attributed exclusively to telephone-channel effects.

C. The Speakers in the Wild (SITW) recognition challenge

The Speakers in the Wild (SITW) database [16] was designed to advance speaker recognition under realistic and challenging conditions, featuring diverse acoustic environments such as clean interviews, outdoor settings, stadium events, and red-carpet recordings. A subset of 120 speakers was selected for this investigation. In order to guarantee single-speaker utterances, the target speaker was isolated using Goldwave and Audacity software, even though the dataset contains multi-speaker segments. Every speech file that resulted was concatenated or segmented to a predetermined length of 129,250 samples (8 seconds). To maintain alignment with the experimental methodology used to other benchmark corpora, a consistent partitioning scheme was used, with six segments used for training and four for testing.

D. NIST2008 speaker recognition evaluation training set

The NIST 2008 database [17] consists of multilingual speech recordings collected from telephone and microphone channels, featuring both native and bilingual English speakers. A subset of 120 microphone recordings in English was chosen for consistency with the TIMIT and SITW benchmarks. To standardize sampling conditions, the original audio signals were upsampled from 8 kHz to 16 kHz. To preserve only single-speaker content, interviewer segments were removed. The data from each speaker was divided into four testing and six training utterances, each of which was concatenated or trimmed to a set length of eight seconds.

V. Simulation Results and Discussion

Table I shows the results for the proposed method as well as compared with different features combinations with various feature dimensions. In addition, Table II shows the comparative between the prposed method and other state of the art work. All datasets showed the best results for the suggested model, which concatenated LPC, MFCC, prosody, and DWT features with Student's t-modeling: TIMIT (98.33%), NTIMIT (89.38%), SITW (96.88%), and NIST2008 (100%). Since the same feature set under a GMM produced lower accuracy, especially on noisy data like NTIMIT, where performance dropped by 7.72%, comparative analysis highlights the significance of the Student's t-distribution. The need for complete feature integration is further highlighted by feature ablation studies, which show that with fewer combinations, performance significantly declined, particularly on NTIMIT, where accuracy dropped to 13.75–35.83% from 89.38% for the full model. These findings are supported by the variation in performance across databases. Clean, high-quality recordings that allow for the best feature extraction are reflected in the high accuracy on TIMIT (98.33%). The reduction in NTIMIT (89.38%) emphasizes the importance of spectral distortion and narrowband channel effects. Using the relative noise resistance of prosodic and DWT features, the model's strong performance on SITW (96.88%) demonstrates its applicability in real-world scenarios. Longer utterances provide richer feature statistics, and the model is naturally suited to handle channel variability and heavy-tailed distributions, which is

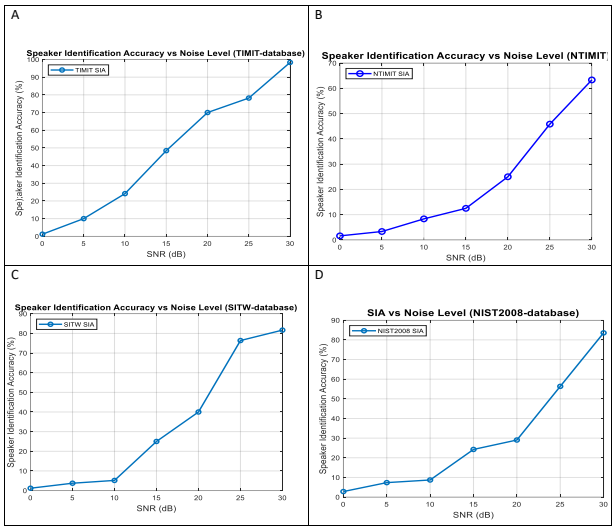

Fig. 2. Speaker identification accuracy. vs. SNR for the proposed system on the A- TIMIT B- NTIMIT C-SITW D-NIST2008 databases

probably why the optimal result on NIST2008 (100%) was obtained. Fig. 2 illustrates the speaker identification accuracy (SIA%) as a function of Signal-to-Noise Ratio (SNR) on the TIMIT, NTIMIT, SITW, and NIST 2008 databases. It explains the impact of additive noise on system performance. The degradation in identification accuracy with decreasing SNR highlights the challenge of noisy environments.

The system's resilience is further demonstrated by the SIA (Accuracy) vs. SNR plot, which demonstrates that even in the presence of high noise, the model retains non-zero accuracy even though accuracy declines with decreasing SNR. This robustness is explained by the multi-feature fusion framework's prosodic and wavelet features' ability to withstand noise as well as the Student's t-model's ability to withstand noise-induced outlier values. Figure 3 shows the Detection Error Tradeoff (DET) curve for the speaker identification system assessed on the TIMIT database, as well as the Receiver Operating Characteristic (ROC) curve for the speaker verification system on the TIMIT database. This curve illustrates the trade-off between the True Acceptance Rate (TAR) and False Acceptance Rate (FAR). It shows how the True Acceptance Rate (TAR) and the False Acceptance Rate (FAR) are traded off.

Figures 4, 5, and 6 present the performance evaluation of the proposed multi-feature fusion model on the NIST2008, NTIMIT, and SITW, databases, comprising the ROC and DET curves.

Plotting True Acceptance Rate (TAR) against False Acceptance Rate (FAR) yields a ROC curve that sharply curves toward the top-left corner, indicating the system's high discriminative power. This suggests high accuracy, and the significant area under the curve (AUC) demonstrates that a highly discriminative representation is produced by combining MFCC, LPC, prosodic, and DWT features, allowing for efficient speaker separation at different thresholds. The DET curve

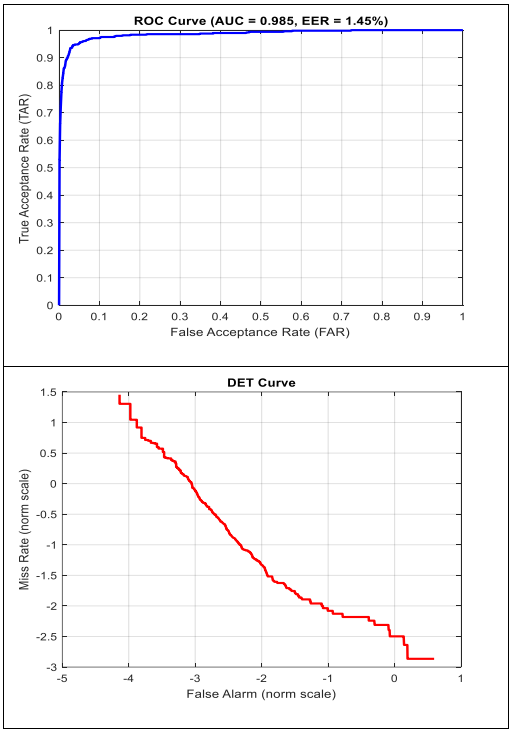

Fig. 3. ROC and DET curves of the proposed model on the TIMIT database

provides comprehensive information about performance at low error rates that are essential for real-world applications by plotting Miss Rate against False Alarm Rate on a logarithmic scale. In comparison to Gaussian models, the heavy-tailed Student's t-distribution, which lessens sensitivity to outliers and atypical variations, supports the model's robustness, as seen by its trajectory toward the bottom-left.

VI. Conclusions

This paper has presented a robust framework for text-independent speaker identification based on multi-feature fusion and Student's t modelling. A highly discriminative representation of a speaker's identity is produced by the suggested system's efficient integration of spectral, prosodic, and time-frequency features (MFCC, LPC, prosody, and optimized DWT). An important development was the addition of per-speaker normalization and the multivariate Student's t-distribution, which significantly increased robustness against noise, inter-session variability, and the heavy-tailed nature of speech feature data. Comprehensive testing on the benchmark datasets TIMIT, NTIMIT, SITW, and NIST 2008 shows that the suggested method performs better. Significantly, the system's accuracy was 98.33% on TIMIT, 89.38% on the noisy NTIMIT, 96.88% on SITW, and 100% on NIST2008. Moreover, thorough testing in Additive White Gaussian Noise (AWGN) conditions verified that the suggested system continuously surpasses the most advanced techniques, resulting in

TABLE I
COMPARATIVE PERFORMANCE EVALUATION OF THE PROPOSED AND STATE-OF-THE-ART MODELS.

Feature Type	Modelling	SIA TIMIT	SIA NTIMIT	SIA SITW	SIA NIST2008
Proposed Method 40 FD	Student's t	98.33 %	89.38 %	96.88 %	100.00 %
[LPC, MFCC, prosody, and Optimized DWT] 40 FD	GMM Mixture 256	97.29 %%	81.66%	87.92 %	97.50%
LPC + MFCC + Prosody + DWT] 40 FD Without Speaker normalization	Student's t	96.67%	30.83%	91.04%	92.29%
LPC + Prosody 15 feature dimension	Student's t	86.25 %	13.75 %	81.25%	85.83%
LPC + MFCC + Prosody 28 feature dimension	Student's t	= 96.25 %	27.50 %	91.04 %	92.08 %
MFCC + Prosody 16	Student's t	95.42 %	27.29 %	90.21 %	90.42 %
[MFCC + Delta + Delta-Delta + Prosody + DWT] 54 FD	Student's t	93.96%	33.33%	89.38%	93.12%
%[MFCC + Delta + Delta-Delta + Prosody] 42 FD	Student's t	95.83%	35.83%	90.21%	93.12%
MFCC 13 FD	Student's t	94.38%	30.42%	91.46%	94.38%

TABLE II
COMPARISONS WITH THE STATE OF THE ART IN TERMS OF SIA

Authors	Database	System approach	Clean speech	Noisy speech at AWGN (30dB)
Proposed work	TIMIT	Fusion based Student's t	98.33 %	98%
Proposed work	NTIMIT	Fusion based Student's t	89.38 %	64%
Proposed work	SITW	Fusion based Student's t	96.88 %	82%
Proposed work	NIST2008	Fusion based Student's t	100.00 %	84%
Al-Kaltakchi et al. [18]	TIMIT	Fusion based GMM-UBM	95%	75.83%
Al-Kaltakchi et al. [18]	SITW	Fusion based GMM-UBM	82.5%	78.33%
Al-Kaltakchi et al. [18]	NIST2008	Fusion based GMM-UBM	95.83%	26.67%
Al-Kaltakchi et al. [19]	TIMIT	I-vector	96.67%	74.16
Al-Kaltakchi et al. [19]	SITW	I-vector	85.83%	84.17%
Al-Kaltakchi et al. [19]	NIST2008	I-vector	96.67%	81.67
Kumar et al. [14]	TIMIT	GMM	93.88%	

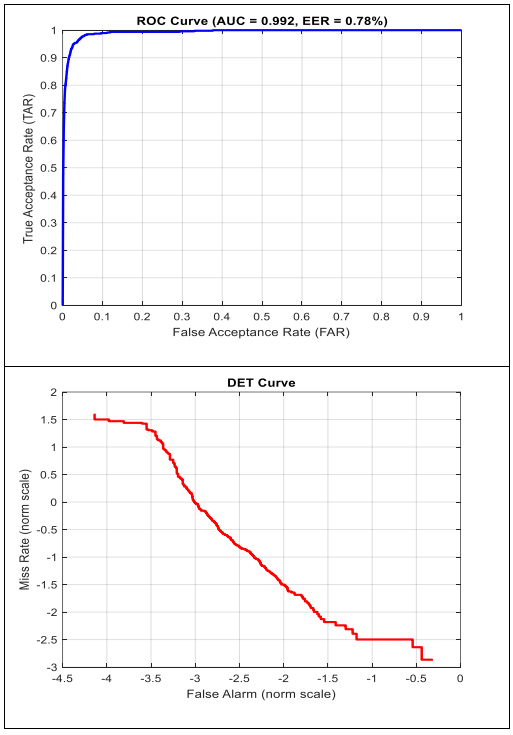

Fig. 4. ROC and DET curves of the proposed model on the NIST2008 database

notable improvements in identification accuracy. Comparative research verified that, especially under difficult acoustic conditions, the full feature fusion approach in conjunction with the Student's t-model consistently performs better than both conventional GMMs and models that use subsets of features. In summary, the effectiveness of combining complementary feature representations with a strong probabilistic model to improve speaker identification is confirmed by this work. Future work will explore the integration of deep learning-based features and the application of this framework to large-scale speaker verification tasks.

ACKNOWLEDGMENT

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

REFERENCES

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5871484/

[2] B. Saritha, M. A. Laskar, A. M. Kirupakaran, R. H. Laskar, M. Choudhury, and N. Shome, "Deep learning-based end-to-end speaker identification using time–frequency representation of speech signal," *Circuits, Systems, and Signal Processing*, vol. 43, no. 3, pp. 1839–1861, 2024. [Online]. Available: https://doi.org/10.1007/s00034-023-02542-9

[3] M. Tiwari and D. K. Verma, "Enhanced text-independent speaker recognition using mfcc, bi-lstm, and cnn-based noise removal techniques," *International Journal of Speech Technology*, vol. 27, no. 4, pp. 1013–1026, 2024. [Online]. Available: https://doi.org/10.1007/s10772-024-10150-4

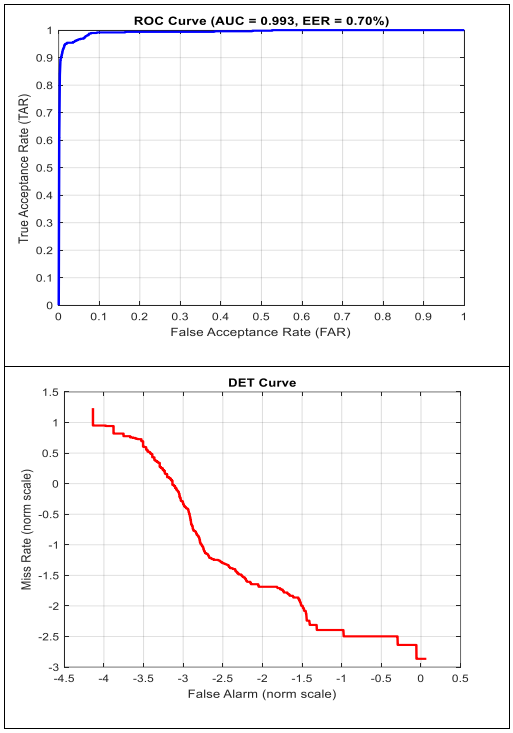

Fig. 6. ROC and DET curves of the proposed model on the SITW database

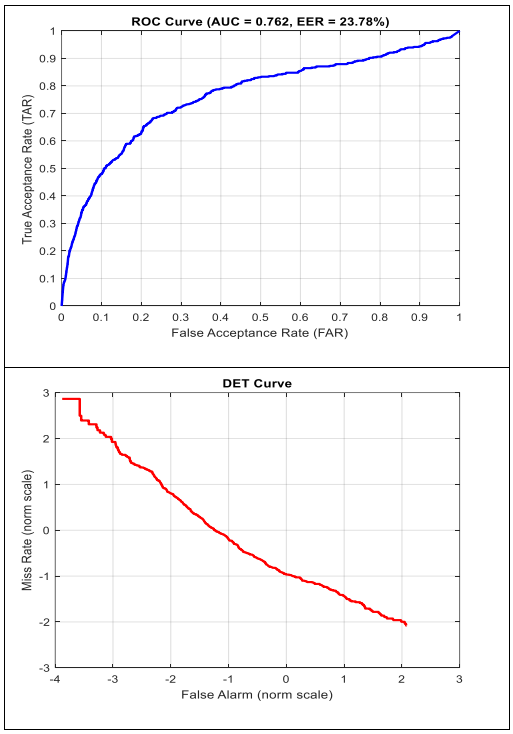

Fig. 5. ROC and DET curves of the proposed model on the NTIMIT database

[4] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.114591

[5] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *Proceedings of the 8th international conference on signal processing systems*, 2016, pp. 142–147. [Online]. Available: https://doi.org/10.1145/3015166.3015210

[6] M. T. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Comparison of i-vector and gmm-ubm approaches to speaker identification with timit and nist 2008 databases in challenging environments," in *2017 25th European signal processing conference (EUSIPCO)*. IEEE, 2017, pp. 533–537. [Online]. Available: https://doi.org/10.23919/EUSIPCO.2017.8081264

[7] T. M. Nguyen and Q. J. Wu, "Robust student's-t mixture model with spatial constraints and its application in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2011.

[8] S. O. Sadjadi *et al.*, "The 2021 nist speaker recognition evaluation," in *Proc. Odyssey 2022: The Speaker and Language Recognition Workshop*, 2022, pp. 200–207. [Online]. Available: https://doi.org/10.48550/arXiv.2204.10242

[9] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA: Springer, 2015. [Online]. Available: https://doi.org/10.1007/978-1-4899-7488-4_196

[10] J. Thienpondt, B. Desplanques, and K. Demuynck, "Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization," *arXiv preprint arXiv:2007.07689*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2007.07689

[11] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, and G. Pandey, "Application of student's t-test, analysis of variance, and covariance," *Annals of cardiac anaesthesia*, vol. 22, no. 4, pp. 407–411, 2019. [Online]. Available: https://doi.org/10.4103/aca.ACA_94_19

[12] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993. [Online]. Available: https://cir.nii.ac.jp/crid/1881146593179904768

[13] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Study of statistical robust closed set speaker identification with feature and score-based fusion," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Palma de Mallorca, 2016, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/7551807/

[14] R. S. S. Kumari, S. S. Nidhyananthan *et al.*, "Fused mel feature sets based text-independent speaker identification using gaussian mixture model," *Procedia Engineering*, vol. 30, pp. 319–326, 2012. [Online]. Available: https://doi.org/10.1016/j.proeng.2012.01.867

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus cd-rom: Ntimit corpus," Linguistic Data Consortium, Philadelphia, 1993.

[16] "Speakers in the wild (sitw) database," http://www.speech.sri.com/projects/sitw/.

[17] "Nist 2008 speaker recognition evaluation database," https://catalog.ldc.upenn.edu/.

[18] M. T. S. Al-Kaltakchi, W. L. Woo, S. Dlay, and J. A. Chambers, "Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, p. 80, 2017. [Online]. Available: https://doi.org/10.1186/s13634-017-0515-7

[19] M. T. Al-Kaltakchi, M. A. Abdullah, W. L. Woo, and S. S. Dlay, "Combined i-vector and extreme learning machine approach for robust speaker identification and evaluation with sitw 2016, nist 2008, timit databases," *Circuits, Systems, and Signal Processing*, vol. 40, no. 10, pp. 4903–4923, 2021. [Online]. Available: https://doi.org/10.1007/s00034-021-01697-7