IJIE

INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

Vol. 18 No. 1 (2026) 155-169
https://publisher.uthm.edu.my/ojs/index.php/ijie

Speaker Recognition Systems: A Review of Machine Learning and Deep Learning Approaches

Fadhliyatun Farah Mahadi[1], Rafizah Mohd Hanifa[2,3]*, Shamsul Mohamad[1], Roszaini Haniffa[4]

[1] *Department of Electronic Engineering, Faculty of Electrical & Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, MALAYSIA*
[2] *Department of Information Technology, Centre for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Hab Pendidikan Tinggi Pagoh, KM1, Jalan Panchor, 84600 Panchor, Johor, MALAYSIA*
[3] *ICT for Technology Humanization (iTecH) Focus Group, Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Hab Pendidikan Tinggi Pagoh, KM1, Jalan Panchor, 84600, Johor, MALAYSIA*
[4] *Edinburgh Business School, Heriot-Watt University, SCOTLAND*

*Corresponding Author: rafizah@uthm.edu.my
DOI: https://doi.org/10.30880/ijie.2026.18.01.012

Article Info

Received: 15 December 2025
Accepted: 28 March 2025
Available online: 17 April 2026

Keywords

Speaker recognition, machine learning, deep learning, MFCC, Wav2Vec, CNN, systematic review

Abstract

Speaker recognition has emerged as a cutting-edge technology with a diverse array of applications, ranging from biometric authentication to personalized user experiences. Its primary objective is to identify or authenticate individuals based on their unique vocal characteristics. This paper presents a comprehensive analysis of the evolution of speaker recognition systems, systematically comparing traditional statistical techniques with modern deep neural network–based approaches in terms of robustness, scalability, and computational efficiency. Historically, speaker recognition systems have relied on statistical models, including Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). While these methods provide a reliable performance under controlled conditions, they suffer notable degradation in the presence of speech variability, environmental noise, and large-scale datasets. In contrast, recent deep learning frameworks, including d-vector and x-vector architectures, have achieved substantial performance gains, with reported classification accuracies exceeding 99% in benchmark studies and relative improvements of 10–30% over traditional methods. Furthermore, speaker verification research increasingly emphasizes metrics such as Equal Error Rate (EER) and minimum Detection Cost Function (minDCF), with state-of-the-art systems reporting EER values as low as 3–4% and minDCF scores below 0.03 in challenging evaluation conditions. Key challenges, including spoofing attacks, dataset bias, multilingual variability, and privacy preservation, are critically examined. Based on the reviewed findings, this study concludes that deep learning–driven speaker recognition systems offer superior performance and adaptability for real-world deployment. Future research should focus on the ethical use of data, bias mitigation,

and the development of lightweight yet secure models suitable for global and resource-constrained environments.

1. Introduction

Speaker recognition, a crucial technology in biometric authentication, has attracted significant attention due to its applications in security, biometric verification, and human-computer interaction. By analyzing vocal characteristics, speaker recognition systems can verify or identify individuals within modern intelligent systems [1]. Fundamentally, this technology involves identifying or authenticating individuals based on their unique vocal traits. The field is divided into two primary tasks: speaker verification, which compares voice samples against a claimed identity, and speaker identification, where a system determines the identity of an unknown speaker from a set of known voices. Each person's voice possesses distinctive characteristics shaped by their physiology and speaking habits, making it an effective biometric indicator. However, these vocal traits are not static; they can fluctuate due to factors such as mood, health, background noise, and the recording device used. This variability poses an ongoing challenge in developing highly accurate and robust speaker recognition systems.

Traditionally, speaker recognition relied on conventional statistical models, such as Gaussian Mixture Models (GMMs) and i-vector frameworks [2]. While these methods established a solid foundation, they often struggled to perform well under diverse real-world conditions. Variations in language, accent, background noise, and channel interference hindered the ability of these traditional models to generalize effectively. The landscape began to change with the advent of machine learning and, more recently, deep learning, which introduced new opportunities for modeling complex patterns in speech data.

Modern systems utilize advanced architectures, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), to generate robust voice models that effectively capture everyone's unique vocal characteristics [3]. A significant breakthrough in this field is the x-vector framework, which employs DNNs to produce compact yet distinctive voice representations. This model has demonstrated enhanced performance not only in controlled environments but also shows improved generalization across a diverse range of speakers and recording conditions.

Researchers are investigating methods to improve deep learning systems through advanced preprocessing techniques, feature extraction methods, and hybrid model architectures to further boost performance. However, challenges remain, particularly in developing systems that prevent fraud, maintain accuracy across various languages and dialects, and operate effectively in diverse environments [4].

This paper aims to provide a comprehensive overview of the development of speaker recognition systems, focusing specifically on the transition from traditional machine learning methods to advanced deep learning models. By reviewing various architectures, techniques, and applications, we highlight the advantages and challenges of current approaches. In doing so, we seek to identify key areas for future research and innovation, with the goal of creating speaker recognition systems that are accurate, scalable, secure, inclusive, and capable of performing effectively across a wide range of scenarios [5].

Unlike earlier reviews that focused mainly on x-vector and CNN-based systems, this study provides a broader perspective by incorporating recent self-supervised models (e.g., WavLM, HuBERT) and hybrid architectures (e.g., ECAPA-TDNN, SincNet). It also identifies research gaps related to multilingual robustness, ethical considerations, and on-device optimization, which have been underexplored in prior works.

2. Machine Learning Approaches

Traditional speaker recognition systems have historically relied on statistical and discriminative models such as GMMs, Hidden Markov Models (HMMs), and Support Vector Machines (SVMs) [1], [5]. These models operate in conjunction with handcrafted acoustic features, most notably Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients, to capture the unique characteristics of each speaker's voice. Each method offers distinct advantages: GMMs effectively model complex feature distributions, HMMs excel at capturing the temporal dynamics of speech, and SVMs provide robust classification capabilities in high-dimensional feature spaces. The following subsections explore these approaches in greater detail, examining their underlying principles, typical applications, and the limitations that have driven the evolution of modern speaker recognition techniques.

2.1 Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMMs) represent the probability distribution of acoustic features, effectively capturing speaker-specific characteristics. The GMM-Universal Background Model (UBM) has become a foundational approach in this domain. Speaker models are adapted from the universal background model using Maximum A Posteriori (MAP) adaptation. This method has proven effective, particularly when limited speaker-specific data is

available. However, GMMs often struggle with channel variability and require large amounts of data to achieve robust performance [5].

GMMs are widely used in speaker recognition and verification to represent the probability distribution of acoustic features and to capture speaker-specific characteristics through a weighted combination of multiple Gaussian components [5]–[8]. A GMM defines the likelihood of an acoustic feature vector x as:

$$p(x \mid \lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(x; \mu_k, \Sigma_k) \tag{1}$$

where Kdenotes the number of Gaussian components, $w_k$are the mixture weights satisfying $\sum_{k=1}^{K} \quad w_k = 1$, $\mu_k$represents the mean vector, and $\Sigma_k$denotes the covariance matrix of the k^{th}Gaussian component.

The GMM–Universal Background Model (GMM-UBM) framework, has become a foundational technique in speaker modeling. In this framework, a speaker-independent UBM is first trained using data from a large and diverse set of speakers. Subsequently, speaker-dependent models are derived through Maximum A Posteriori (MAP) adaptation of the UBM parameters to the target speaker's data. This approach allows robust modeling even when only limited speaker-specific data are available [5]–[8].

During verification, a likelihood-ratio test is used to compare the probability of an observed feature sequence under the speaker model versus the background model, as given by:

$$\Lambda(X) = \log p(X \mid \lambda_s) - \log p(X \mid \lambda_{\text{UBM}}) \tag{2}$$

The GMM-UBM framework demonstrates strong performance under constrained data conditions due to the statistical strength of MAP adaptation. However, its effectiveness is limited by channel variability and the need for large training datasets to estimate reliable covariance matrices [6], [8]. These limitations have motivated a shift toward deep learning-based architectures, such as i-vector and x-vector systems, which offer improved robustness through hierarchical feature extraction and end-to-end training.

2.2 Hidden Markov Models (HMM)

Hidden Markov Models (HMMs) are highly effective for modeling temporal sequences, making them well-suited for text-dependent speaker recognition tasks. They represent speech as a sequence of states with probabilistic transitions, effectively capturing the dynamic aspects of speech [5]. However, despite their strengths, HMMs are less effective in text-independent scenarios and are sensitive to variations in speaking rate and style.

Moreover, Hidden Markov Models (HMMs) demonstrate notable efficiency in text-dependent speaker recognition scenarios. They function by capturing phonetic sequences through their hidden states and transition dynamics, leveraging the stable lexical content inherent in these tasks [5]–[8]. These models are characterized by their state transition probabilities, expressed as $a_{ij}=P(q_{t+1}=s_j|q_t=s_i)$, emission probabilities $B=\{b_j(o_t)\}$, typically modeled using Gaussian Mixture Models (GMMs) or neural networks, and initial state distributions $\pi=\{\pi_i\}$ [5], [7], [8]. Fundamental algorithms in this domain include the forward algorithm, which calculates the likelihood $P(O|\lambda)$; the Viterbi algorithm, used for decoding state sequences; and the Baum-Welch algorithm, which applies the Expectation-Maximization (EM) method for parameter estimation. Nonetheless, ongoing research consistently highlights the limitations of HMMs in text-independent environments, where variable speech content disrupts state alignment, further complicated by variations in speech rate, style, and background noise [5]-[8]. Although HMMs excel in temporal modeling for narrowly defined tasks, recent advancements increasingly favor deep learning techniques, such as Deep Neural Networks (DNNs) and ResNeXt architectures, for text-independent scenarios. This preference is driven by their robustness to variability and unpredictable conditions [5], [6].

2.3 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are discriminative classifiers designed to identify the optimal hyperplane that separates different classes. In the context of speaker recognition, SVMs have been employed using Gaussian Mixture Model (GMM) supervectors as input features, effectively combining the generative modeling capabilities of GMMs with the discriminative power of SVMs [5]. This hybrid approach has significantly enhanced performance, particularly under challenging conditions. SVMs function as discriminative classifiers that optimize class separation by finding a maximum-margin hyperplane in feature in the typically using the objective the following:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad \text{subject to} \quad y_i(\mathbf{w}^\top\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, \tag{3}$$

Here w denotes the weight vector, C is the regularization parameter, and ξ_i are slack variables [5], [6]. SVMs are often combined with GMM supervectors to improve speaker classification performance. The GMM supervector is an adapted mean vector expressed as:

$$m = [\mu^{1\top}, \mu^{2\top}, \ldots, \mu^{M\top}]^\top, \tag{4}$$

This approach is derived from speaker-specific generative GMMs. This integration allows the combination of generative modeling (GMM) with discriminative classification (SVM) approaches [5]-[8]. The GMM-SVM framework significantly enhances robustness, particularly in text-independent speaker recognition tasks, by converting variable-length speech segments into fixed-length supervectors that effectively handle variations in recording duration [7]-[8]. Despite these advantages, several limitations persist. The performance of SVMs can deteriorate without proper hyperparameter tuning, and they may face computational challenges when processing large-scale datasets—especially when compared to recent deep learning–based methods [8], [10]. Kernel functions ϕ(·) and soft-margin techniques are commonly applied to manage nonlinear separable data. Although GMM-SVM systems have been foundational in speaker recognition, they are increasingly being replaced by neural network–based embedding models such as Waveform-based Language Model (WavLM) and other deep learning architectures [6], [9].

3. Deep Learning Approaches

The advent of deep learning has transformed speaker recognition by allowing models to automatically learn complex and discriminative representations directly from raw waveforms or minimally processed acoustic features [4]-[5]. Unlike traditional machine learning methods, which depend heavily on handcrafted features, deep learning architectures can jointly learn feature extraction and classification within a unified framework. This capability has significantly improved accuracy, robustness, and adaptability across real-world conditions.

3.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are well-suited for capturing local temporal and spectral structures in speech signals. By applying convolutional filters over spectrograms or other time–frequency representations, CNNs can detect speaker-specific patterns at multiple scales as shown in Fig. 1.

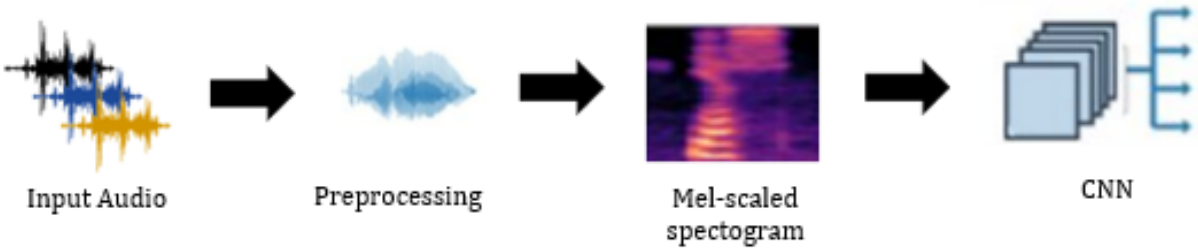

Fig. 1 *Schematic diagram representing CNN-based audio classification adapted from [10]*

Architectures such as ResNet and its variants (e.g., ResNeXt, Res2Net) have been adapted for speaker recognition, offering enhanced representation capacity through multi-scale feature modeling and residual connections [6]-[7]. These designs improve the system's ability to capture fine-grained local patterns and broader contextual cues. However, CNNs process input in relatively short receptive fields, making them less suited for capturing the long-term temporal dependencies that characterize a speaker's voice over an entire utterance. Sequence modeling architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been introduced to address this.

Int. Journal of Integrated Engineering Vol. 18 No. 1 (2026) p. 155-169

3.2 Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

Building upon CNNs' ability to extract local spectral features, RNNs and LSTMs extend the modeling horizon by explicitly learning temporal dependencies across entire speech sequences. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are specifically designed to model sequential data, effectively capturing temporal dependencies in speech. These architectures have been employed to learn speaker embeddings from sequences of acoustic features, enabling the modeling of long-term temporal dependencies and consequently improving recognition accuracy, as illustrated in Fig. 2.

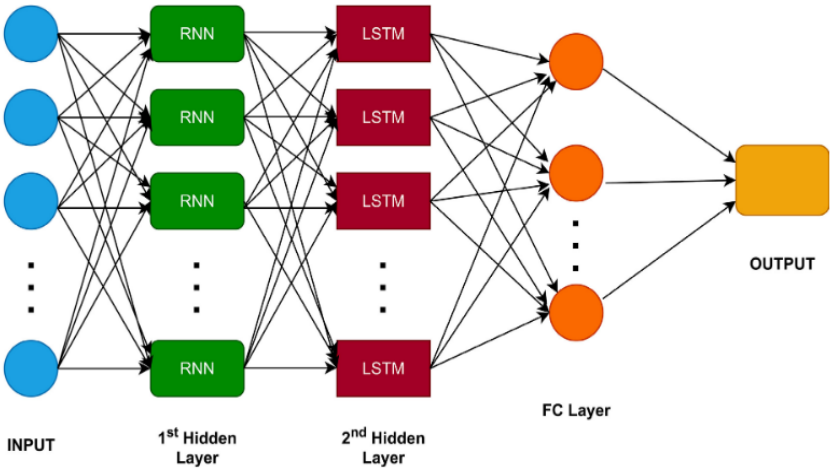

Fig. 2 *A hybrid RNN-LSTM base model architecture [11]*

For speaker recognition tasks, RNNs and LSTMs are particularly relevant, as they have proven instrumental in capturing temporal dependencies in speech [9]. One of the mathematical foundations of RNNs is the hidden state update:

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{5}$$

and the LSTM's gating mechanisms, where the cell state is updated as follows:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

Information flow is controlled by input, forget, and output gates [9]. By maintaining historical context through recurrent connections, these architectures excel at processing sequential acoustic signals such as MFCCs. This capability enables the extraction of speaker embeddings that capture long-term voice characteristics. Recent research [12]-[14] indicates a shift toward convolutional networks and transformer-based designs (such as WavLM and HuBERT), which demonstrate superior performance in speaker verification. Although RNNs and LSTMs were previously the industry standard for temporal modeling [9], self-attention methods have become widely adopted in contemporary systems [12]-[13] due to the inefficiency of RNNs and LSTMs in parallel computing and their challenges in handling very long-range dependencies. Pure transformer architectures still dominate state-of-the-art techniques [12]-[13]; however, hybrid designs that combine convolutional layers with LSTM components remain in use in some research for integrated local and temporal feature extraction. Nevertheless, the mathematical foundations of RNNs and LSTMs remain essential for understanding sequence modeling in speech processing [9],[14]. Although these recurrent architectures effectively capture sequential context, they can be computationally slow and sometimes inefficient in parallel processing, leading to interest in alternative designs such as Time Delay Neural Networks (TDNNs), which retain temporal modeling capability while improving efficiency.

3.3 Time Delay Neural Networks (TDNN) and X-Vectors

Time Delay Neural Networks (TDNNs) model temporal context in speech without relying on recurrent connections. Unlike RNNs, they enable parallel computation while effectively capturing relevant speech patterns over time. TDNNs have become the industry standard for speaker embedding extraction, especially within the x-vector framework [9]. To better understand the strength of TDNNs in speaker recognition, this section highlights six key aspects: (a) their core principles, (b) their role in feature representation, (c) their temporal modeling ability, (d) the x-vector embedding process, (e) their architectural variants, and (f) their integration with self-supervised learning methods.

(a) Core Principle
TDNNs process fixed-size temporal windows of acoustic frames to model local temporal dependencies. This approach enables reliable speaker representation while maintaining computational efficiency. Compared to traditional i-vector methods, the TDNN-based x-vector framework offers significant improvements in robustness and discriminative power [9],[14].

(b) Feature Representation
TDNNs effectively model local temporal patterns essential for speaker characteristics by applying convolutional operations over fixed-size temporal windows of acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs) [9], [12], [14]. MFCCs are commonly used as input features because they capture the short-term power spectrum of speech based on human auditory perception using the mel-frequency scale. The conversion from linear frequency f (Hz) to the mel scale m is defined as:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

The speech signal is segmented into overlapping frames and transformed into the frequency domain using the Fast Fourier Transform (FFT). A set of triangular Mel filterbanks is then applied to the magnitude spectrum to obtain the filterbank energies. These energies are converted to a logarithmic scale and decorrelated through the Discrete Cosine Transform (DCT) to yield the MFCCs:

$$c_n = \sum_{k=1}^{K} \log(S_k) \cos\left[\frac{\pi n(k-0.5)}{K}\right], \quad n = 1, 2, \ldots, L \tag{8}$$

where S_k represents the mel-filtered spectral energies.

(c) TDNN Temporal Modeling
The x-vector framework, originally proposed in [9], [13], aggregates frame-level features through a statistics pooling layer that computes both the mean and standard deviation of the frame-level activations. The resulting fixed-dimensional embeddings, known as speaker embeddings, capture distinctive speaker characteristics and demonstrate robustness against channel variability, background noise, and emotional factors. The Time-Delay Neural Network (TDNN) architecture serves as foundation of the x-vector system for speaker embedding extraction. Its design enables effective modeling of local temporal dependencies within speech signals while maintaining computational efficiency. The mathematical basis of TDNNs lies in temporal convolution, where a set of learnable filters $w[k]$ is applied to an input sequence $x[t]$ within a defined temporal context window:

$$y[t] = \sum_{k=-K}^{K} x[t+k]\, w[k] \tag{9}$$

Here, $x[t]$ denotes the input acoustic frame at time t, $w[k]$ represent the learnable weights and is $y[t]$ the resulting output. This structure enables TDNNs to capture speaker-specific information across consecutive frames, thereby forming a discriminative representation.

Int. Journal of Integrated Engineering Vol. 18 No. 1 (2026) p. 155-169

(d) X-vector Embedding Process

The x-vector framework, originally proposed by [9], integrates a Time Delay Neural Network (TDNN) with a statistics pooling layer that aggregates frame-level representations into fixed-length embeddings by computing their mean and standard deviation. This process can be summarized as follows:

(i) Frame-level modeling: Time-Delay Neural Network (TDNN) layers extract local temporal features.
(ii) Statistics pooling: Aggregates variable-length frame representations into a fixed-length vector.
(iii) Fully connected layers: Transform pooled features into compact speaker embeddings that are used for classification or verification.

(e) Architectural Variants

Enhanced variants, such as Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) is an enhanced version of the traditional TDNN designed for speaker recognition. It introduces channel attention mechanisms and residual connections to focus on the most informative features in the speech signal. In simpler terms, ECAPA-TDNN refines how the model processes and aggregates speech features over time, allowing it to capture subtle speaker characteristics more effectively [12]. ECAPA-TDNN leads to improved robustness and accuracy, especially in real-world, noisy conditions.

(f) Integration with Self-Supervised Learning

Recent studies show that TDNN x-vector systems benefit from self-supervised pretraining using models such as Wav2Vec 2.0 and WavLM [12]–[14]. These pretrained models enhance representation quality, particularly in low-resource or multilingual settings. Nevertheless, alternative feature-based systems, such as CQCC-GMM, remain advantageous for anti-spoofing tasks [15]. This distinction highlights that TDNN x-vector systems are primarily optimized for speaker discrimination rather than spoof detection.

Fig. 3 illustrates the TDNN x-vector architecture used for extracting speaker embeddings. The process begins with raw audio input, represented as a waveform, which is processed by a Time Delay Neural Network (TDNN) composed of multiple layers that capture local temporal dependencies from fixed-size windows of speech frames. The frame-level outputs are then aggregated through a statistics pooling layer, where the mean and standard deviation are computed to summarize the entire utterance. This pooled representation is subsequently transformed into a fixed-dimensional speaker embedding, depicted as a vertical bar, providing a compact and discriminative representation for speaker verification and identification tasks.

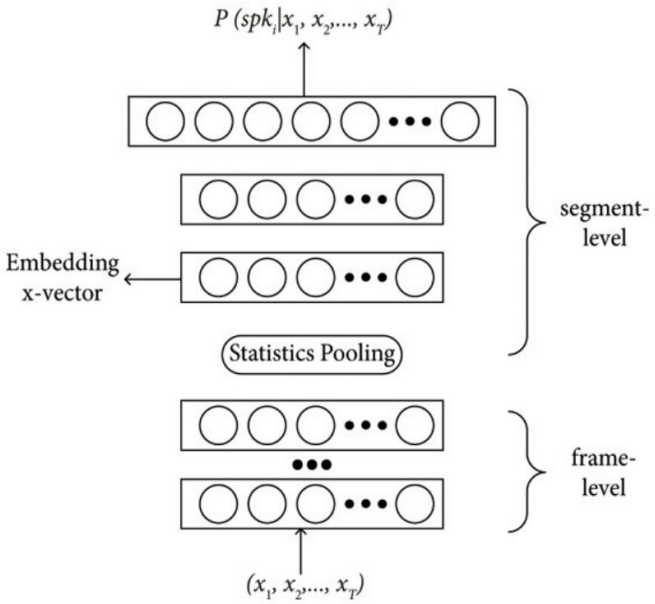

Fig. 3 *x-vector TDNN embedding architecture [7]*

Although TDNN-based x-vector systems remain highly competitive, their performance can be further enhanced through large-scale self-supervised pretraining, as demonstrated by models such as Wav2Vec 2.0 and WavLM.

3.4 Self-Supervised Models: Wav2Vec2 and WavLM

Extending beyond supervised architectures like TDNNs, self-supervised learning approaches exploit vast amounts of unlabeled data to learn general-purpose speech representations. Self-supervised learning models, such as Wav2Vec2 and WavLM, have been adapted for speaker recognition tasks [12]. These models are pre-trained on extensive amounts of unlabeled data to learn general speech representations, which can subsequently be fine-tuned for specific speaker recognition applications [13]. Wav2Vec2 has shown competitive performance in speaker verification tasks. WavLM enhances Wav2Vec2 by integrating additional training objectives and architectural improvements, resulting in further advancements in speaker recognition performance [14].

4. Datasets and Features

The success of a speaker recognition system hinges on the careful selection of datasets and feature extraction techniques. These foundational elements determine how effectively the model can identify unique speaker characteristics under varying conditions. In the following subsections, a comprehensive overview of the datasets and features utilized in this study is provided to ensure robust and accurate speaker recognition.

4.1 Dataset

Research on speaker recognition is essential in the fields of biometric authentication, forensic applications, and human-computer interaction. The National Institute of Standards and Technology (NIST) in the United States has played a pivotal role in developing and evaluating speaker recognition systems [15]. For many years, NIST has led the Speaker Recognition Evaluation (SRE) series, organizing competitions to assess and advance speaker recognition technologies while fostering innovation in the field. These evaluations serve as a standard benchmark for comparing the performance of various systems under controlled experimental conditions, making them highly influential [16].

Additionally, the VoxCeleb Speaker Recognition Challenge (VoxSRC) was established to encourage further research using the VoxCeleb dataset. Participants in VoxSRC utilize real audio, which typically includes background noise, various recording devices, and diverse speech patterns, in their open-set speaker identification and verification tasks. This work highlights the challenges faced by speaker identification systems in real-world scenarios [17]. Although the VoxCeleb dataset has been extensively used and proven effective, studies indicate that performance on this dataset is approaching its peak [14], [17]. Consequently, there is increasing emphasis on exploring more challenging conditions, such as speech across multiple genres beyond interviews, and remotely acquired far-field audio recordings.

This situation, which is inadequately represented in current benchmarks, introduces new challenges for improving speaker identification techniques. The review covers standard datasets with speaker labels, as well as specialized collections such as Gigaspeech and Wenetspeech, providing a comprehensive overview of the resources available for speaker identification. Although these datasets sometimes contain obvious speaker identifiers, they remain valuable for unsupervised or self-supervised learning methods and for developing speaker representations, especially in less formal or unstructured data collection environments. Incorporating these types of resources underscores the need for flexible and scalable training solutions that do not rely heavily on manually labeled data.

Finally, speaker identification has evolved from a purely supervised discipline focused on assessment to a more comprehensive field that addresses real-world complexities, diversity, and the need for continuous learning. The emergence of resources such as VoxCeleb, along with a diverse array of data types, highlights a shift in research toward improved generalizability and broader applicability. Fig. 4 illustrates the key differences between two major types of speaker recognition datasets: NIST SRE and VoxCeleb.

Int. Journal of Integrated Engineering Vol. 18 No. 1 (2026) p. 155-169

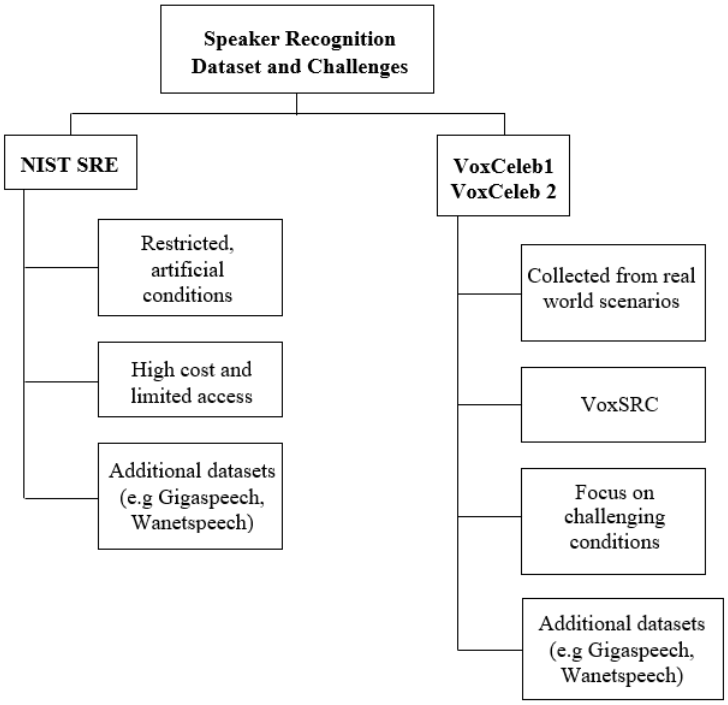

Fig. 4 *Comparison between the two main types of speaker recognition datasets: NIST SRE, and VoxCeleb*

4.2 Features

Traditional features have achieved significant success; however, they often struggle to generalize across diverse acoustic conditions and to capture the complex nonlinear patterns inherent in speech. This progress was initially demonstrated by the emergence of d-vectors, followed by the development of x-vectors. Both approaches utilize supervised learning to transform variable-length speech segments into fixed-length embeddings optimized for speaker discrimination, representing significant advances in the field.

The x-vector model is one of the most prevalent and robust architectures in the field of speaker recognition. It gained widespread acceptance through open-source implementations of efficient speech processing tools. The framework begins by analyzing features at the frame level using sophisticated time-delay neural networks (TDNNs). It then aggregates temporal information through a statistical pooling layer, culminating in a fully connected layer that transforms the segments into a dense, low-dimensional embedding space. To enhance discriminative power, these embeddings are typically refined using margin-based loss functions, such as softmax margin enhancement or angular margin loss. This approach effectively maximizes inter-speaker separability while minimizing intra-speaker variability, significantly improving the accuracy and reliability of speaker recognition and verification systems.

In scenarios with limited labeled data, combining supervised and unsupervised learning strategies has gained significant attention. Techniques such as contrastive learning, cluster-based pseudo-labeling, and masked prediction tasks—drawing inspiration from models like wav2vec and HuBERT—enable the extraction of rich speaker representations from large amounts of unlabeled speech. This approach is particularly beneficial for resource-poor languages and practical contexts where annotation costs are high, thereby improving the applicability and effectiveness of speaker recognition technologies.

The field of speaker representation learning has gradually evolved from traditional, manually designed, rule-based feature extraction methods to advanced deep learning paradigms that leverage sophisticated neural architectures, innovative loss functions, and refined training strategies. This progression has significantly improved the accuracy and robustness of speaker identification systems [19]. However, the field still faces critical challenges, including enhancing adaptability across various domains, ensuring effective generalization to unseen speakers and scenarios, and developing transparent, interpretable, and privacy-preserving embedding models [20].

Fig. 5 shows the conceptual framework of speaker representation learning, which focuses on extracting reliable features from speech while minimizing the effects of noise, emotion, and recording conditions. Earlier methods relied on handcrafted features, such as MFCCs, PLP, and i-vectors. In contrast, deep learning approaches now utilize learned representations, including d-vectors and x-vectors, through advanced models like ECAPA-

TDNN, ResNet, and Conformers. This shift marks the transition from manual feature engineering to automated, data-driven learning using deep neural networks.

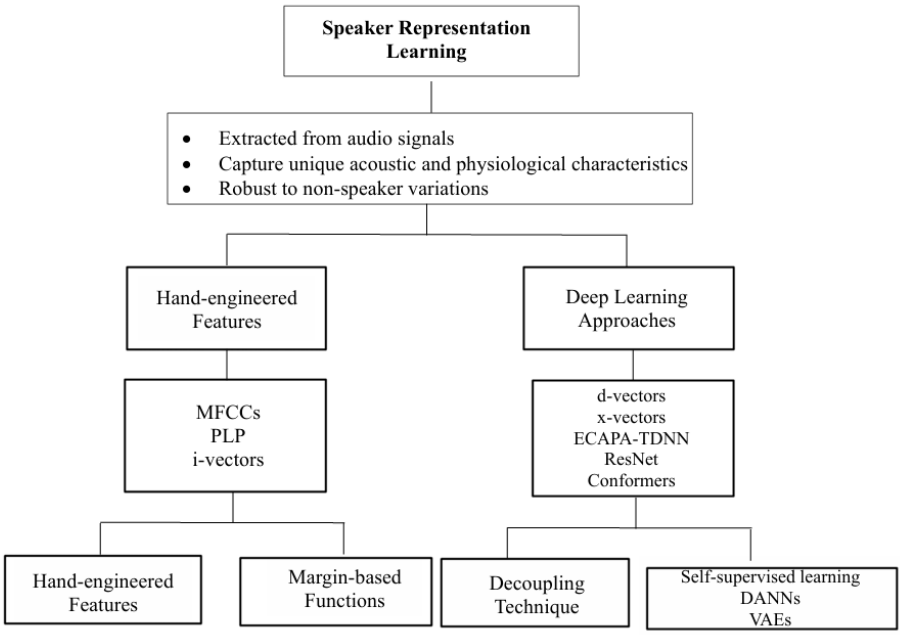

Fig. 5 *Conceptual framework of speaker representation learning, illustrating the evolution from hand-engineered features to deep learning-based approaches*

The goal is to develop discriminative and robust representations that are invariant to non-speaker variations such as noise, emotion, or recording conditions. The diagram categorizes speaker representation methods into two main approaches: hand-engineered features and deep learning-based techniques. Traditional methods rely on features like MFCCs, PLP, and i-vectors, often combined with margin-based loss functions to enhance discriminability. In contrast, deep learning approaches utilize representations such as d-vectors, x-vectors, and advanced architectures like ECAPA-TDNN, ResNet, and Conformers. These methods frequently incorporate decoupling techniques and self-supervised learning strategies, such as Domain-Adversarial Neural Networks (DANNs) and Variational Autoencoders (VAEs)—to improve robustness and generalization. This distinction reflects the evolution of speaker recognition systems from manual feature engineering to automated representation learning through deep architectures.

5. Performance and Comparison

A comparative analysis of recent speaker recognition studies is conducted in this paper, examining the variations in feature extraction methods, model architectures, dataset types, speaker sample sizes, and performance outcomes. Table 1 highlights the advancements in speaker recognition research, summarizing key developments across feature extraction methods, model architecture, datasets, number of speakers, and reported performance outcomes.

Table 1 *Progress in speaker recognition: techniques, datasets, and performance*

Author(s)	Year	Feature Extraction	Model/Architecture	Dataset	No. of Speakers	Result/ Performance
Fasounaki et al. [19]	2021	MFCCs	CNN	LibriSpeech	251	99.5% Accuracy
Sreekanth [20]	2024	WavLM (self-supervised)	TI-SV + password verification	English	>100	minDCF: 0.029 (Track 1), 0.142 (Track 2)
Nasr et al. [21]	2021	MFCC + Spectrograms	DNN	Free-Spoken Digit Dataset	6	93% Accuracy
Kohler & Imtiaz [22]	2025	MFCC + PCA	SVM (RBF Kernel)	LibriSpeech	245	96.91% Accuracy

Int. Journal of Integrated Engineering Vol. 18 No. 1 (2026) p. 155-169

Author(s)	Year	Feature Extraction	Model/Architecture	Dataset	No. of Speakers	Result/ Performance
Kim et al. [23]	2021	Spectrograms	ACNN	VoxCeleb	>1000	86.51% Accuracy, EER: 5.68%
Tsai et al. [24]	2021	MFCCs	GMM + HMM	Self-generated (Multilanguag e)	5	93.31%, 82.42%, EER: 3.38%
Barai et al. [25]	n.d.	MFCCs	VQ + UBM + GMM	Self-generated	100	98% Accuracy
Keshavarzi et al. [26]	2024	Monosyllabic Test Words	COPT	Self-generated	31	85.2% – 96.6% Accuracy
Chen et al. [27]	2024	WavLM	SPRL	Qualcomm	50	84.12% Accuracy
Vasquez-Serrano et al. [28]	2023	MFCCs	GMM + ANN	USC-TIMIT	9	RMSE: Male 0.55–0.78, Female 0.86–1.20
Liu et al. [29]	2023	Domain Adaptation	DANN	Not specified	-	95% Accuracy
Mavaddati [30]	2024	MFCC + ResNet	CNN + RNN	Multilang	50,000	98.78% Accuracy
Alsobhani et al. [31]	2021	-	CNN + SVM	Self-generated	30	97.06% Accuracy
Hassanzadeh et al. [32]	2024	MFCC + LPC	1D-CNN + LSTM	Multilang	630	95.22% Accuracy
Jahangir et al. [33]	2021	MFCC, LPC, PLP	DNN, CNN	VoxCeleb	1251	Not specified
Anidjar et al. [34]	2024	MFCC + Wav2Vec2	Wav2Vec2 + CNNs	VoxCeleb + LibriSpeech	1172	99.59% Accuracy
Rani & Kumar [35]	2024	LDAVGG16, VGG19, ResNet	Ensemble	English	5	0.63 Error Rate
Shome et al. [36]	2024	Raw audio	SincNet	LibriSpeech	251	Not specified
S.Bini et al. [37]	2023	-	ResNet-8	English	500	Not specified
Pham et al. [38]	2023	MFCCs + ECAPA-TDNN	ECAPA-TDNN	Vietnamese	1000	EER: 5.48%, Accuracy: 41.61%
Noishin [39]	2022	MFCC	CNN + LSTM	TIMIT + LibriSpeech	630	97.54% accuracy
Haris Isyanto [40]	2022	DWT+MFCC	CNN	Indonesian	10	99.00% accuracy
Giovanni [41]	2023	MFCC	CNN	DEMoS dataset	58	90.15%
Rekha S. Kotwal [42]	2024	Wavelet+CNN	CNN+LSTM	Emotional Speech & Song Dataset	24	81.22%
Hrithik [43]	2021	MFCC+GFCC	CNN	Ryerson Audio-Visual	24	92.59%
F. Bahmaninezhad et al. [44]	2021	Speaker embedding space	Domain adaptation in speaker embeddings	Various in-the-wild datasets	-	Improved robustness in domain adaptation
Abeer Ali [45]	2022	Mel spectrograms + MFCC	DNN + CNN	Mozilla Common Voice dataset	-	98.57%

Based on the above table, we present a structured examination of recent advances in speaker recognition and verification from 2021 to 2025 in the following sections. Our review considers the field from multiple perspectives, including comparative evaluations of contemporary systems, the evolution of feature extraction methods, the progression of model architectures, the influence of dataset diversity and scale, and the performance metrics used to assess system effectiveness. This integrated approach offers a comprehensive understanding of how recent innovations have influenced current capabilities and future directions in speaker recognition research.

5.1 Feature Extraction Techniques

Mel-Frequency Cepstral Coefficients (MFCCs) continue to be the most frequently adopted feature representation ([19], [21], [22], [24], [25], [30], [32]- [34], [37], [39] – [41], [43]) offering a reliable and compact characterization of speech signals. Many studies further enhanced MFCCs by combining them with techniques such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), and spectrograms to capture richer acoustic dynamics ([21], [32]-[33], [35], [42]). Notably, the field is experiencing a paradigm shift towards self-supervised feature learning. Models leveraging WavLM and Wav2Vec2 embeddings have demonstrated improved generalizability in multilingual and low-resource conditions. For instance, Anidjar et al. [34] reported state-of-the-art 99.59% accuracy using MFCC and Wav2Vec2 on a combined VoxCeleb and LibriSpeech dataset, while Chen et al. [27] and Sreekanth [20] utilized WavLM-based embeddings for speaker verification tasks, achieving low detection cost and high accuracy.

5.2 Model Architectures

Convolutional Neural Networks (CNNs) and their variants dominate the architectural landscape due to their ability to extract robust features from spectro-temporal representations ([19], [22], [30], [32], [37], [40]). Several hybrid models that combine CNNs with Long Short-Term Memory (LSTM) or Recurrent Neural Network (RNN) layers have demonstrated high accuracy by effectively capturing temporal dynamics ([32], [42]). Traditional machine learning classifiers, such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Support Vector Machines (SVM), remain competitive when combined with robust feature engineering ([22], [24], [25]). For example, Kohler and Imtiaz [22] employed an MFCC plus PCA feature pipeline with an SVM using a radial basis function (RBF) kernel, achieving 96.91% accuracy on the LibriSpeech dataset. Deep ensemble strategies are also emerging, as demonstrated by Rani and Kumar [35], who employed multiple CNN backbones (LDAVGG16, VGG19, ResNet) and achieved an error rate of 0.63, indicating high accuracy even with small datasets.

5.3 Dataset Diversity and Scalability

Benchmark datasets such as LibriSpeech and VoxCeleb are prominently featured across multiple studies ([19], [22], [33]-[34], [35]), with speaker counts ranging from 245 to over 1,200. The VoxCeleb dataset, used by Jahangir et al. [33], Kim et al. [23], and Anidjar et al. [34], is notable for its large speaker diversity and real-world variability. Self-generated datasets were commonly employed for language-specific or small-scale experiments ([24]-[26], [31]-[32]), enabling precise control over acoustic conditions. Mavaddati [30] trained a CNN + RNN model on MFCCs and ResNet features using 50,000 multilingual speakers and achieved 98.78% accuracy, highlighting the impact of large-scale training.

5.4 Performance Metrics

A speaker recognition system's performance is typically evaluated using three primary metrics: Accuracy, Equal Error Rate (EER), and Minimum Detection Cost Function (minDCF). Accuracy quantifies the proportion of correctly classified samples among all test samples, offering a general measure of system reliability in identification tasks. In contrast, EER and minDCF are more informative for speaker verification, as they assess the trade-off between false acceptance and false rejection rates. The EER indicates the point where both error rates are equal, the lower the EER, the more balanced and robust system. Meanwhile, minDCF, introduced in the NIST Speaker Recognition Evaluations (SRE), integrates error rates with cost and prior probabilities, providing a more realistic assessment of system performance under specific operational constraints.

After establishing these metrics, top-performing systems can be compared as follows: accuracy remains the dominant evaluation criterion, with leading systems such as Fasounaki et al. [19] (99.5%), Anidjar et al. [34] (99.59%), and Haris Isyanto [40] (99.00%) achieving outstanding results. However, speaker verification studies have increasingly emphasized EER and minDCF. For example, Sreekanth [20] achieved a minDCF of 0.029 on Track 1 using WavLM-based embeddings, while Tsai et al. [24] reported an EER of 3.38% using a GMM + HMM framework.

6. Trends and Future Directions

Emerging trends indicate a strong interest in self-supervised learning (e.g., WavLM, Wav2Vec2), end-to-end deep speaker embeddings (e.g., ECAPA-TDNN [38]), and ensemble learning techniques. As datasets become more diverse and multilingual, robust domain adaptation and transfer learning strategies are essential. Hybrid architecture combining classical and deep models, such as CNN + SVM [31], or DNN + CNN [45], continue to yield strong results. Additionally, models using raw waveform inputs like Sinc-based Convolutional Neural Network (SincNet) [36] and novel feature combinations such as DWT + MFCC [40] are showing promising results in niche applications.

SincNet is a type of deep learning model designed to process speech directly from raw audio waveforms. Unlike traditional models that rely on manually crafted features such as MFCCs or spectrograms, SincNet learns directly from the raw sound itself. What makes it unique is the way it processes audio using sinc-based filters — mathematical filters that mimic how humans perceive sound frequencies. Instead of treating audio like random numbers, SincNet focuses on the frequency components that carry useful information, such as tone, pitch, and timbre. Thus, it makes SincNet particularly effective in several areas — identifying who is speaking (speaker recognition), recognizing the words being spoken (speech recognition), and understanding the emotional tone or stress in a person's voice (emotion recognition).

Building upon such advancements, recent studies have also explored hybrid extraction techniques, such as combining Discrete Wavelet Transform (DWT) and Mel-Frequency Cepstral Coefficients (MFCC). This approach leverages the DWT's ability to capture time–frequency information and the effectiveness in modeling the spectral envelope, yielding more robust and discriminative representations of speech. Consequently, the DWT + MFCC fusion has shown improved accuracy in speaker recognition, emotion detection, and language identification, particularly in noisy conditions.

7. Conclusion

This review has explored recent developments in speaker recognition and verification from 2021 to 2025, focusing on innovations in feature extraction, model architectures, dataset utilization, and performance evaluation. The field has clearly evolved from handcrafted acoustic features toward self-supervised and representation learning approaches, supported by increasingly diverse and large-scale datasets. Advances in deep neural architecture, often enhanced with hybrid and ensemble strategies, have led to substantial improvements in accuracy, robustness, and scalability.

Despite these gains, challenges persist in addressing cross-lingual variability, domain mismatch, and adversarial robustness, particularly in real-world and low-resource settings. Future research is expected to emphasize multimodal integration, model interpretability, and energy-efficient training and inference to promote sustainability. By consolidating current trends and identifying open research gaps, this review provides a solid foundation for researchers and practitioners to navigate the evolving landscape of speaker recognition technologies.

Beyond technical progress, speaker recognition also brings crucial ethical and societal considerations. Because voice data is inherently personal, concerns about privacy, consent, and data security must be addressed. Transparency in data collection and the mitigation of biases related to language, gender, and accent diversity are essential to ensure fairness and inclusivity. From an application perspective, these technologies are now embedded in security systems, smart devices, and assistive tools, offering both convenience and accessibility. However, responsible deployment requires striking a balance between performance and ethical safeguards to prevent misuse, protect user rights, and promote equitable access across diverse populations.

Acknowledgement

This research was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/UTHM/02/3).

Conflict of Interest

Authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Farah, Rafizah;* ***literature search and selection****: Farah, Shamsul;* ***analysis and interpretation of literature****: Farah, Rafizah, Shamsul;* ***draft manuscript preparation****: Rafizah, Shamsul;* ***critical revision of the manuscript****: Farah, Rafizah. Roszaini;* ***helps review and proofread the manuscript****.*

References

[1] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," Comput. Electr. Eng., vol. 90, p. 107005, 2021, doi: 10.1016/j.compeleceng.2021.107005.

[2] R. Sharma et al., "Milestones in speaker recognition," Artif. Intell. Rev., vol. 57, no. 3, 2024, doi: 10.1007/s10462-023-10688-w.

[3] Z. Bai and X. Zhang, "Speaker recognition based on deep learning: An overview," Neural Netw., vol. 140, pp. 65–99, 2021, doi: 10.1016/j.neunet.2021.03.004.

[4] Z. Bai and X.-L. Zhang, "Advances in deep learning for speaker recognition," Neural Netw., vol. 152, pp. 123–145, 2023.

[5] R. Jahangir et al., "Text-independent speaker identification through feature fusion and deep neural network," IEEE Access, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[6] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net structures for speaker verification," arXiv preprint arXiv:2007.02480, 2020, doi: 10.48550/arxiv.2007.02480.

[7] D. Sztahó, G. Szaszák, and A. Beke, "Deep learning methods in speaker recognition: A review," Period. Polytech. Electr. Eng. Comput. Sci., vol. 65, no. 4, pp. 310–328, 2022.

[8] N. Shome et al., "Speaker recognition through deep learning techniques: A comprehensive review," Period. Polytech. Electr. Eng. Comput. Sci., vol. 67, no. 3, pp. 300–336, 2023, doi: 10.3311/PPee.20971.

[9] S. Wang et al., "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning," IEEE/ACM Trans. Audio Speech Lang. Process., 2024, doi: 10.1109/TASLP.2024.3492793.

[10] H. Sinha, V. Awasthi, and P. Ajmera, "Audio classification using braided convolutional neural networks," *IET Signal Processing*, vol. 14, no. 9, pp. 679–687, 2020, doi: 10.1049/iet-spr.2019.0381.

[11] S. Faru, A. Waititu, and L. Nderu, "A hybrid neural network model based on transfer learning for forecasting forex market," J. Data Anal. Inf. Process., vol. 11, no. 2, pp. 103–120, 2023, doi: 10.4236/jdaip.2023.112007.

[12] S. Chen, S. Wang, and J. Li, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE J. Sel. Topics Signal Process., vol. 16, no. 6, pp. 1505–1518, 2022.

[13] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 3451–3460, 2021.

[14] S. Chen et al., "Harnessing the power of Wav2Vec2 and CNNs for robust speaker identification on the VoxCeleb and LibriSpeech datasets," Expert Syst. Appl., vol. 205, p. 124671, 2024, doi: 10.1016/j.eswa.2024.124671.

[15] M. Todisco, H. Delgado, and N. Evans, "ASVspoof 2021: Advancing speaker verification anti-spoofing," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 3760–3775, 2022.

[16] Y. Fan et al., "CN-Celeb: An extensive Chinese speaker recognition dataset," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 2765–2778, 2022.

[17] M. V. Villasana, J. D. Méndez, and J. C. Pérez, "Ethics and privacy in AI-based speaker recognition systems," Ethics Inf. Technol., vol. 26, no. 1, pp. 45–63, 2024.

[18] E. A. Albadawy, T. H. Soliman, and K. M. Badran, "Lightweight deep learning architectures for on-device speaker verification," Appl. Sci., vol. 13, no. 2, p. 948, 2023.

[19] M. Fasounaki et al., "CNN-based text-independent automatic speaker identification using short utterances," Proc. Int. Conf., 2021, p. 413.

[20] S. Sreekanth, "Self-supervised speaker representation learning using WavLM with text-dependent speaker verification," J. Speech Technol., vol. 32, pp. 1–12, 2024.

[21] M. B. Nasr et al., "Text-independent speaker recognition using deep neural networks with MFCC and spectrogram features," Proc. Int. Conf., 2021.

[22] J. Kohler and S. Imtiaz, "Speaker recognition using MFCC-PCA features with RBF-SVM classifier," J. Signal Process. Syst., vol. 13, pp. 1–15, 2025.

[23] Y. Kim et al., "Attention-based CNN for speaker recognition in large datasets," J. Audio Speech Process., vol. 3, no. 2, pp. 86–95, 2021.

[24] V. L. Tsai et al., "Multilingual speaker recognition using GMM-HMM hybrid systems," Expert Syst. Appl., vol. 171, p. 114591, 2021, doi: 10.1016/j.eswa.2021.114591.

[25] B. Barai, T. Chakraborty, and N. Das, "Closed-set speaker identification using VQ and GMM-based models," Speech Commun., 2021, doi: 10.1007/s10772-021-09899-9.

[26] M. Keshavarzi et al., "Cognitive optimization for speaker recognition using monosyllabic words," J. Intell. Syst., vol. 31, pp. 1–5, 2024.

[27] S. Chen et al., "Speaker representation learning using WavLM features," IEEE Access, vol. 32, pp. 1–12, 2024.

[28] J. Vasquez-Serrano et al., "GMM-ANN hybrid systems for speaker recognition," Int. J. Speech Technol., vol. 11, pp. 1–15, 2023.

[29] Y. Liu et al., "Domain adversarial neural networks for speaker recognition," IEEE Access, 2023.

[30] H. Mavaddati, "Multilingual speaker recognition using CNN-RNN architectures," J. Audio Speech Technol., 2024.

[31] A. Alsobhani et al., "CNN-SVM hybrid for speaker recognition," Proc. Int. Conf. Speech Commun., 2021.

[32] A. Hassanzadeh et al., "1D-CNN-LSTM architecture for multilingual speaker recognition," J. Intell. Syst., vol. 12, pp. 1–15, 2024.

[33] R. Jahangir et al., "Text-independent speaker identification through feature fusion and deep neural network," IEEE Access, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[34] O. H. Anidjar et al., "Wav2Vec2-CNN hybrid for speaker recognition," Expert Syst. Appl., vol. 205, p. 124671, 2024, doi: 10.1016/j.eswa.2024.124671.

[35] P. Rani and S. Kumar, "Ensemble deep learning for speaker recognition," Appl. Soft Comput., 2024.

[36] N. Shome et al., "Raw audio analysis using SincNet for speaker recognition," IEEE Access, 2024.

[37] S. Bini et al., "ResNet-8 architecture for speaker recognition," J. Audio Speech Technol., 2023.

[38] V. T. Pham et al., "ECAPA-TDNN for Vietnamese speaker recognition," IEEE Access, 2023.

[39] A. Noishin, "CNN-LSTM hybrid for speaker recognition," IEEE Access, vol. 10, pp. 1–12, 2022.

[40] H. Iryanto, "DWT-MFCC features with CNN for Indonesian speaker recognition," Indones. J. Electr. Eng., vol. 20, no. 1, pp. 1–10, 2022.

[41] G. Giovanni, "CNN-based emotion-dependent speaker recognition," Speech Commun., vol. 145, pp. 1–12, 2023.

[42] R. S. Kotwal, "Wavelet-CNN-LSTM for emotional speaker recognition," IEEE Trans. Affect. Comput., 2024.

[43] I. Hritnik, "Hybrid MFCC-GFCC features for speaker recognition," Multimed. Tools Appl., vol. 80, pp. 1–15, 2021.

[44] F. Bahmaninezhad et al., "Domain adaptation in speaker embeddings," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 1–12, 2021.

[45] A. Ali, "DNN-CNN with mel-MFCC fusion for speaker recognition," Appl. Sci., vol. 12, no. 5, p. 2345, 2022.