SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I1P117

Volume 13 Issue 1, 207-240, January 2026
© 2026 Seventh Sense Research Group®

Review Article

Speaker Diarization: A Comprehensive Survey of Clustering Methods, Neural Networks and Hybrid Frameworks for Robust Speaker Segmentation and Identification

Merlin Revathy S[1], Kumar S S[2]

[1,2]Department of Electronics and Communication Engineering, Noorul Islam Centre for Higher Education, Kanyakumari, India.

[1]Corresponding Author : merlinrevathy@gmail.com

Received: 09 November 2025 Revised: 11 December 2025 Accepted: 10 January 2026 Published: 14 January 2026

Abstract *- The speaker diarization involves splitting an audio recording into segments based on who is speaking and when they speak. In simple words, it helps you know, "Who spoke when?" This method of sorting voices is significant for various aspects of real life. People use it in automatic meeting notes, video subtitles, voice helpers, and legal or medical recordings. With this, you can see which person is always speaking, so it is easier to follow and understand the talk. Early systems depended on the handcrafted features, which performed well, but they faced challenges in the real world. Mostly, in the time of background noise, many speakers are talking at the same time, or speakers have similar voices. This survey paper provides an extensive review of 95 research papers published between 2023 and 2025, categorizing the existing diarization methods into three groups: supervised learning-based techniques, unsupervised learning-based techniques, and hybrid learning frameworks. Furthermore, the enhancement of diarization performance is facilitated by integrating speaker diarization methods with voice recognition applications. This paper offers a useful review by combining the most recent advancements with neural techniques, hence promoting further advancements in the direction of more effective speaker diarization.*

Keywords *- Clustering Techniques, Deep Learning, Hybrid Frameworks, Multimodal Analysis, Speaker Diarization.*

1. Introduction

In addition to transmitting linguistic information, human speech also reflects speaker identity, intent, and conversational dynamics, making it a rich source of information. Speaker diarization, which aims to identify "who spoke when" in an audio recording, is a key speech processing task that makes such analysis possible. Diarization divides a stream of audio into homogeneous sections and allocates them to different speakers, in contrast to Automated Speech Recognition (ASR), which concentrates on transcribing spoken words. Determining each speaker's temporal limits is crucial in various applications, including meeting transcription, transmission monitoring, medical equipment, customer service data analytics, and surveillance systems.

With the growth of voice-driven technologies, speaker diarization has become increasingly important. Accurate diarization enhances readability in computerized transcription systems by assigning speakers to the text. Diarization enables adaptive processing tailored to the primary talker, facilitating the separation of speech from noise in cochlear implants and hearing aids. Diarization aids in sentiment analysis and agent-customer assessment in call center analytics. Similarly, in legal and technical contexts, diarization permits investigators to track persons across recordings [1]. All these sectors share the need for reliable speaker identification and segmentation in difficult acoustic and environmental circumstances.

The methodology in the field has undergone significant changes. Hand-crafted features, clustering algorithms, and statistical modeling using Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs) were all integral components of early rule-based systems. More condensed illustrations of speaker characteristics were obtained through diarization with the introduction of i-vectors; nonetheless, these methods remained susceptible to domain mismatches, overlapping speech, and noise. An important turning point was the introduction of deep neural networks and x-vectors, which offered deeper discriminatory embeddings that increased clustering robustness and accuracy [2]. To lessen the dependency on modular pipelines, End-to-End Neural

Diarization (EEND) and transformer-based frameworks have lately sought to combine speaker assignment, embedding extraction, and segmentation into a single model.

Even with these developments, diarization is still a problem, especially in devices with limited resources like hearing aids. These devices can function dependably in noisy and extremely dynamic acoustic environments while requiring minimal latency, low power consumption, and limited memory utilization. Another recurring issue is overlapping speech, as traditional segmentation and clustering techniques have trouble with numerous speakers speaking at once. Additionally, diarization systems need to be resilient to reverberant surroundings, different recording devices, and invisible speakers, all of which are typical in real-world applications.

Deep learning-based diarization techniques that go beyond traditional pipelines have become increasingly popular in recent years. Although Generative Adversarial Networks (GANs) have been investigated for data augmentation and embedding refinement, Convolutional Neural Networks (CNNs), autoencoders, and transformers have been used to train high-level feature representations [3].

In parallel, solutions that lessen reliance on labeled data have been made available by self-supervised and unsupervised learning techniques, including Wav2Vec 2.0, HuBERT, and clustering with pseudo-label refinement. The diversity and quick growth of various techniques underscore the importance of a complete evaluation. Figure 1 shows the traditional speaker diarization system.

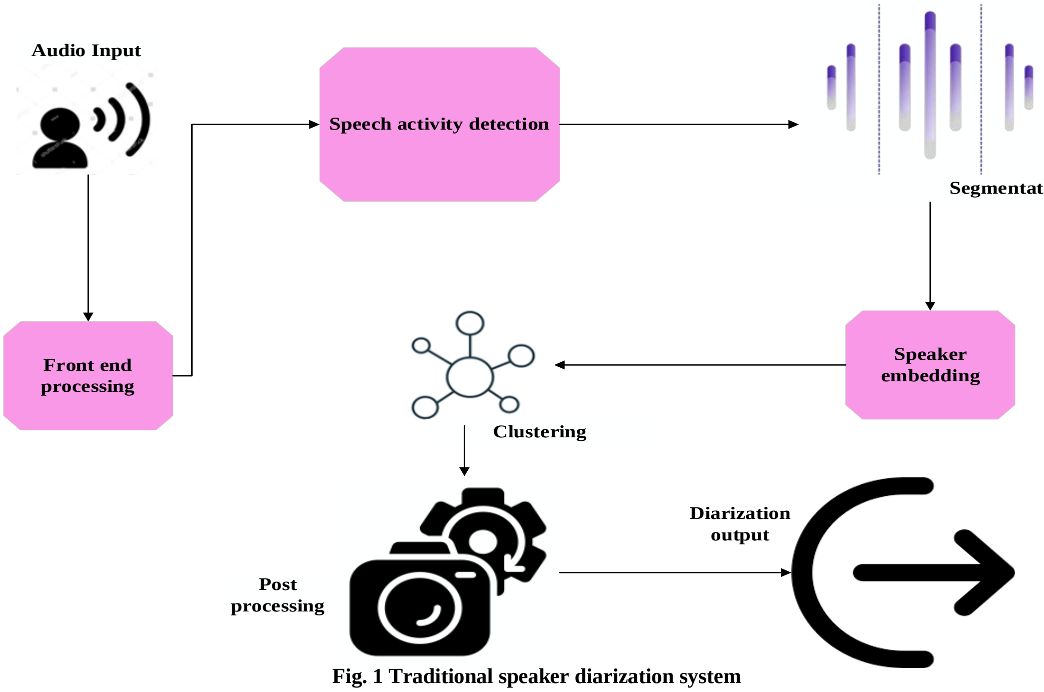


Fig. 1 Traditional speaker diarization system

As a result, this survey offers a thorough analysis of speaker diarization strategies, categorized into supervised techniques, unsupervised techniques, and hybrid learning frameworks. The report also highlights the use of diarization in assistive devices and hearing aids, which can directly affect the quality of life for those who are deaf or hard of hearing. This work aims to guide the development of diarization systems that are both resilient and flexible enough to overcome the various challenges of real-world speech processing by combining advancements made after 2023 and critically evaluating their benefits and drawbacks.

1.1. History of the Speaker Diarization

The goal of research in the 1990s, when diarization technology was still in its infancy, was to improve ASR on broadcast news recordings and air traffic control dialogues by separating the speech segments of each speaker and allowing speaker-adaptive training of acoustic models [4-10]. The Generalized Likelihood Ratio (GLR) [4] and the Bayesian Information Criterion (BIC) [11], two fundamental techniques for determining the separation between speech segments for speaker change detection and clustering, were developed at this time and quickly gained prominence in the field.

All these efforts culminated in the creation of several research consortia and challenges in the early 2000s, such as the National Institute of Standards and Technology (NIST)-organized Rich Transcription (RT) Evaluation [13] and the European Commission-funded Augmented Multiparty Interaction (AMI) Consortium [12].

Building on this basis, later research advances in speaker diarization technology have been used in a wide range of data domains, such as broadcast news [14–18], Conversational Telephone Speech (CTS) [19–22], and meeting conversations [23–27]. The feature representation of short speech segments, in an unsupervised manner, was quickly migrated to speaker-specific representation in a total variability space based on the simplified Joint Factor Analysis (JFA), referred to as i-vector [28-30] for speaker diarization systems. This representation was highly successful in speaker recognition.

In order to enhance the performance of clustering in speaker diarization, i-Vector was indeed a successor of former methods, including simply the Mel-Frequency Cepstral Coefficient (MFCC) or speaker factors (or eigen voices) [31]. It has been used together with Principal Component Analysis (PCA) [32, 33], Variational Bayesian (VB-GMM) [34], mean shift [35], and Probabilistic Linear Discriminant Analysis (PLDA) [36]. Because deep learning was introduced in the 2010s, numerous works have been performed to use the powerful modeling of neural networks to diarize the speaker.

A representative example of how speaker embeddings can be extracted is the use of neural networks, as in the d-vectors [37-39] or the x-vectors [40], which are commonly the embedding vector representations of the bottleneck layer output of a Deep Neural Network (DNN) trained for speaker recognition. These neural embeddings substituted i-vector, resulting in greater performance, training under bigger data sets with less complexity [41], and resistance to acoustic situations and speaker variation. Most recently, EEND has come under a lot of attention and has shown promising results by replacing one neural network with the individual sub-modules of conventional speaker diarization systems [42, 43]. Though this line of research is still immature, when massive data is accessible to train such powerful neural network-based models, it can offer opportunities, never heard of before, to solve speaker diarization issues, including concomitant optimization with other speech functions, especially where there is overlapping speech.

The review will be structured in the following manner: the background and motivation of speaker diarization will be presented in Section 2. Section 3 gives the speaker diarization methodology. Section 4 presents an overview of recent deep learning-based diarization methods. Section 5 presents a compilation of datasets commonly used in contemporary speaker diarization studies. Section 6 discusses the evaluation metrics and performance benchmarks employed to assess the effectiveness of speaker diarization models. Section 7 examines the challenges and limitations currently faced in speaker diarization, highlighting potential future research directions and emphasizing areas for further improvement in speaker diarization. The article concludes in Section 8, summarizing key findings and the need for continuous advancements in this crucial field.

2. Background and Motivation

Traditionally, a modular pipeline consisting of three main stages, mainly Speech Activity Detection (SAD), speaker embedding extraction, and clustering, has been used to tackle the speaker diarization challenge. Within this paradigm, speech and non-speech regions are separated by SAD, embeddings compactly represent speaker-specific attributes, and these embeddings are grouped by clustering to be assigned to specific speakers. Although this modular architecture has been widely used, the reliability of all the parts and the uniformity of their integration frequently determine how well it performs.

Over time, there has been a major evolution in the depiction of speakers in diarization systems. I-vectors, a low-dimensional, statistical assessment of speakers, were used in early methods. Even though they worked well, i-vectors had trouble with overlapping speech and a variety of acoustic circumstances. By learning discriminative characteristics straight from massive amounts of voice data, x-vectors, which are obtained from deep neural networks, improved the resilience of embeddings. The limits of modular pipelines have been lessened with the advent of EEND, which can model speech that overlaps and optimize speaker attribution and segmentation simultaneously.

Traditional diarization techniques, however, have serious drawbacks when used with hearing aids. Hearing aids must provide outcomes in an immediate form with the least amount of lag possible while operating on limited hardware. Complex modular pipelines cannot be directly deployed on these devices due to their small form sizes, low memory, and limited computing power. Furthermore, the performance of traditional systems is further deteriorated by voice overlaps and noisy settings, which are prevalent in ordinary listening scenarios.

Researchers are increasingly using methods based on deep neural networks to get over these obstacles. DNN-based diarization systems can provide low-latency inference and more precise and effective speaker representations by leveraging large-scale training data and optimized architectures in CNNs and autoencoders. There is also growing interest in hybrid frameworks that combine the representational strength of neural embeddings with the interpretability of clustering techniques, particularly for devices with limited resources, such as hearing aids. The need for a thorough analysis that assesses neural networks, clustering techniques, and hybrid approaches in the context of reliable speaker segmentation and identification is highlighted by these discoveries.

3. Methodology

Figure 2 depicts the systematic procedure used to select studies for this survey. During the identification phase, 360 records were collected, with 320 coming from electronic databases such as IEEE, Springer, Scopus, and ScienceDirect, and 40 coming from other relevant sources such as cross-references and research repositories. After eliminating duplicates, 280 unique records were selected for screening. During the screening step, the titles and abstracts were examined to determine their relevance to the topic of speaker diarization. At this point, 150 records were removed because they were irrelevant to diarization, were outside of the designated publishing time, or had duplicated content. The eligibility phase included a full-text review of 130 articles to guarantee methodological rigor, complete experimental validation, and relevance to the primary issues of clustering, neural network-based techniques, and hybrid frameworks for diarization. 35 papers were omitted because they lacked an appropriate methodological description, had incomplete results, or did not contribute to the review objectives. Finally, 95 papers met all the inclusion criteria and were used in the qualitative synthesis, which served as the foundation for this survey. These selected articles provide a complete overview of current accomplishments and issues in supervised, unsupervised, and hybrid diarization methodologies, ensuring that the evaluation is based on high-quality and relevant information.

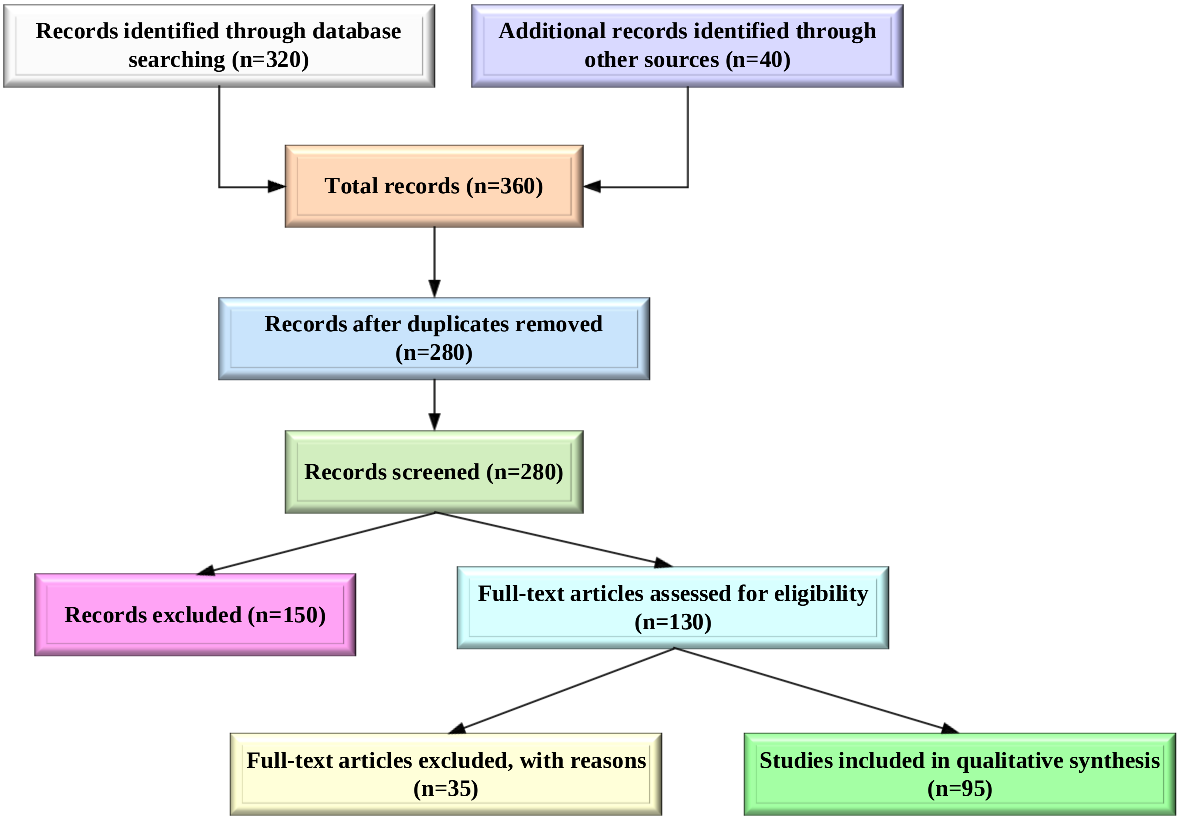

Records identified through database searching (n=320)
Additional records identified through other sources (n=40)
Total records (n=360)
Records after duplicates removed (n=280)
Records screened (n=280)
Records excluded (n=150)
Full-text articles assessed for eligibility (n=130)
Full-text articles excluded, with reasons (n=35)
Studies included in qualitative synthesis (n=95)

Fig. 2 PRISMA flow diagram for study selection process

4. Recent Deep Learning-Based Diarization Methods

This section presents various methods for speaker diarization, categorized into supervised, unsupervised, and hybrid learning techniques. These methods have been assessed to identify challenges and gaps, thereby providing a foundation for developing more effective diarization techniques to address the complexities of multi-speaker audio recordings. Neural embeddings as well as attention mechanisms are used in recent deep learning-based diarization techniques to enhance speaker clustering and segmentation. For robust representation, these methods frequently use EEND, x-vectors, or d-vectors. Compared to conventional models, they show improved accuracy in noisy situations with overlapping speech. All things considered, these techniques represent a major step forward for accurate, scalable, and real-time speaker diarization.

4.1. Supervised Learning-Based Approaches

Supervised learning methods have become more and more significant in improving the reliability of diarization tasks in recent years. These methods help models learn discriminating characteristics and enhance speaker recognition in challenging settings by utilizing annotated training data. Figure 3 shows the supervised learning-based techniques.

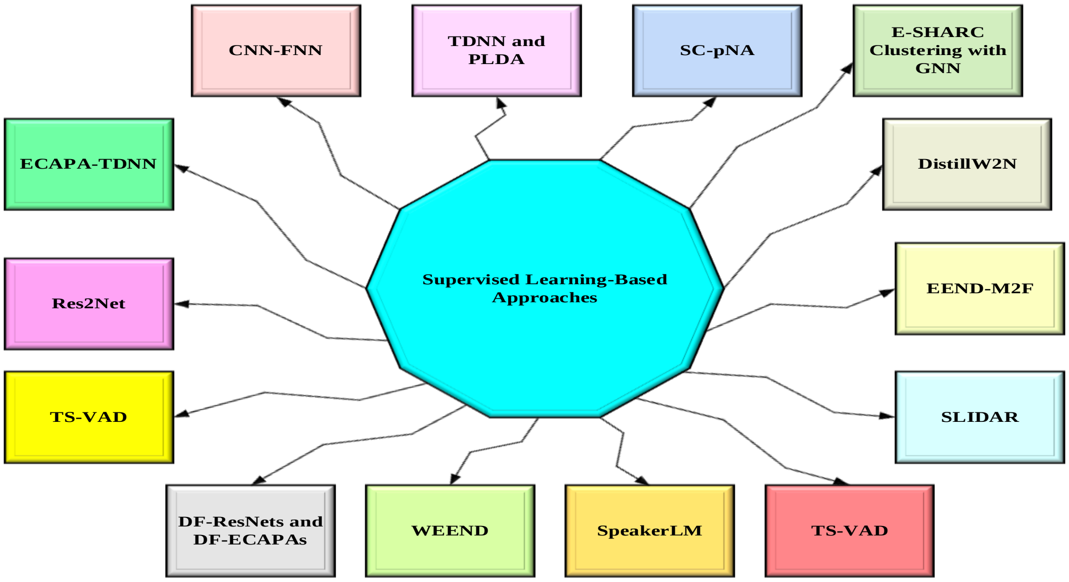


Fig. 3 Supervised learning-based approaches

4.1.1. Embedding Extractors

Shankar et.al. (2025) presented a system designed specifically for segment attribution and speaker identification in audio recordings [44]. While diarization aimed to precisely segment audio to assign specific parts to the appropriate speakers, speaker recognition required acute identification of individuals based on their distinctive voice features. The solution used Agglomerative Hierarchical Clustering to aggregate retrieved embeddings, used the Emphasized Channel Attention, Propagation, and Aggregation – Time Delay Neural Network**. (**ECAPA-TDNN) architecture for speaker embeddings, and the OpenAI Whisper model for transcription. This procedure enabled detailed segment classification and revealed complex speaker relationships. In addition to addressing current issues, this study developed a novel approach that improved accuracy and effectiveness in speaker identification and diarization. By challenging accepted methods, this approach opens new possibilities for research in the ever-evolving field of audio analysis and provides useful implications for a range of applications.

By adding multi-scale convolutional focus to a Feed-forward Network (FFN) based on Global Response Normalization (GRN), Chen et al. (2025) proposed a novel solution that enhances the Transformer model and creates a lightweight backbone architecture known as the Lightweight Simple Transformer (LST) [45]. By adding CNN's Res2Net structure to LST, further enhance it and produce the low-parameter, high-precision Speaker Verification (SV) model known as the Res2Former. By combining features at various depths at the frame level, we create and apply a Time-Frequency Adaptive Feature Fusion (TAFF) technique in Res2Former that permits fine-grained feature propagation. Holistic fusion is also used to propagate global features across the model. Multiple convergence techniques are introduced to improve performance, increasing the SV system's overall effectiveness.

To address the two main challenges of feature extracting task-relevant labels from sparse metadata for system creation and evaluation, and to deal with the unconstrained Acoustical circumstances encountered within the archive, which range from quiet studios to unfavorable noisy environments, Loweimi et al., (2025) have presented challenges, solutions, efficiency, and robustness of audio retrieval systems developed "in the wild" [46]. It systematically examines several facts of system development, such as diarization of speakers, embedding extraction, and query selection, and evaluates the difficulties, potential fixes, and usefulness of each. Results establish the flexibility and adaptability of the suggested framework for a variety of applications outside the British Broadcasting Corporation (BBC) Rewind corpus, demonstrating the efficiency and resilience of the created speaker retrieval systems.

By recognizing speeches in the classroom, Zheng et.al. (2025) have introduced a layered Residual Network (ResNets) featuring multi-scale aggregation as well as a speaker attention system that can differentiate between instructor and student speech [48]. Thus, the verbal exchange between professors and students can be used to examine the teaching approach. However, due to the uneven language environment and the disparity in speaker distribution, the current SV approach cannot be modified for use in classroom settings. To maximize the validity of voiceprint information, a deep multiple-scale aggregation ResNets model is used for the extraction of key features. To determine the variations in teachers' pronunciation patterns and voiceprint amplitudes, a speaker attention system that incorporates both channel-domain and frequency-domain information was implemented. According to experimental data, the suggested approach outperforms the state-of-the-art techniques and delivers exceptional performance with a considerable learning capacity. Using the English public dataset LibriSpeech, the suggested approach achieved a 4.00% equal error rate improvement and a 6.20% accuracy enhancement over the compared techniques.

An innovative domain-robust pre-training technique of SV using local prototypes has been presented by Gu et.al. (2025) [50]. Present a self-distillation system for local feature acquisition using a transformer-based encoder. Online clustering is used to create local prototypes using domain-agnostic pre-training. Additionally, domain-robust local characteristics are learned using domain-aware alignment. Using utterance-level supervision for fine-tuning shows how well the suggested approach performs on the CNCeleb, along with VoxCeleb benchmarks.

Two methods have been proposed by Wang et.al. (2023) to reduce the complexity of the conformer-driven system without sacrificing its functionality [51]. To replace shallow Conformer blocks, a lightweight CNN front-end featuring channel-frequency attention is first proposed. To obtain more informative speaker features for further processing, this modification was made. Second, to reduce the model size of Conformer blocks, a light FFN was utilized that utilized depth-wise separable convolution.

As a result, Kim et.al. (2024) have highlighted the importance of intermediate information processing and given a strategy toward such a paradigm [52]. A novel method is introduced for taking advantage of the multilayered nature of trained models for Automatic Speaker Verification (ASV). This method consists of two steps: pooling designs for each layer and each frame axis, as well as a layer/frame-level network. Allow a stack of level outputs to be processed directly by the convolutional design. Next, squeeze together by being level with the highest representative value, and introduce a channel focus-based scheme for assessing layer significance.

Lastly, a single vector speaker embedding is obtained by careful statistics over frame-level representations. To validate the suggested method, comparative experiments are created using a variety of pretraining models and flexible data contexts. The experimental findings show how stable the multi-layer output technique is while utilizing pretrained designs. Next, confirm that the suggested ASV backend structure, which employs layer-wise operations, is better than the traditional approach in terms of both cost and performance. The ablation analysis demonstrates how the suggested interlayer processing helps to optimize the benefit of using pretrained models.

By requiring the space of embedding to have less intra-class variance than inter-class variance regarding similarity scores, Han et.al. (2025) have introduced a score comparative learning objective that directs the training framework toward being more in line with the verification task [53]. Additionally, suggest a generalized loss function that includes numerous common training losses and regularization strategies for score comparison-based learning. The VOiCES, VoxCeleb, and CN-Celeb, along with the Common Voice datasets, are used to compare the suggested method with the traditional approaches. According to experimental data, the suggested approach can improve system performance and increase its resistance to over-fitting in speech verification tasks.

For speaker verification, Liu et.al. (2023) have proposed an effective architecture design [54]. First, a methodical investigation of the impact of network width and depth on performance reveals empirically that, for SV tasks, network depth is more significant than network width. Make a new Depth-First (DF) design for the architecture guideline based on this finding. Two new families of substantially deeper models, ResNets, DF-ResNets, and DF-ECAPAs, are created through transferring them to ResNet and ECAPA-TDNN. Two new Attentive Feature Fusion (AFF) techniques, Sequential AFF (S-AFF) as well as Parallel AFF (P-AFF), are also suggested to dynamically fuse features in a learnable manner to further improve the performance of minimal models in the low computation regime.

The newly suggested DF-ResNets and DF-ECAPAs can accomplish a significantly better trade-off between performance and complexity over the original ResNet and ECAPA-TDNN, according to experimental results based on the VoxCeleb dataset. Furthermore, by using the AFF approach, tiny models can achieve a relative EER improvement of up to 40% at a low computational cost. Lastly, a thorough comparison with a few previously published SV systems shows that, in both low and elevated computation scenarios, these suggested models provide the optimal trade-off between complexity and performance.

Chen et al. (2025) have proposed Speaker Reciprocal Points Learning (Speaker RPL), which greatly improves single-shot Open-set Speaker Identification (OpenSID) efficiency by integrating unknown sample learning through speech-synthesized unknown samples [55]. To show SpeakerRPL's versatility across different speaker foundation models, and to investigate the best model-tuning techniques, zero-shot timbre-controllable synthesis techniques, and training protocols. Extensive tests on several multilingual, mostly text-dependent recognized speakers datasets validate the framework's effectiveness in intricate home settings, achieving better few-shot open-set recognition results than several cutting-edge speaker foundation models.

Table 1. Comparative analysis of embedding extractors in supervised learning-based approaches

Author and Year	Method / Approach	Dataset	Contribution	Performance Metrix
Shankar et.al., (2025) [44]	Agglomerative Hierarchical Clustering + ECAPA-TDNN embeddings + Whisper transcription	General audio recordings	System for segment attribution and speaker identification using Agglomerative Hierarchical Clustering, ECAPA-TDNN embeddings, and Whisper transcription.	Improved accuracy and effectiveness in speaker identification and diarization.
Chen et.al., (2025) [45]	Integrates Res2Net with Transformer, adds TAFF, and holistic fusion	VoxCeleb	LST with Res2Net, Res2Former with TAFF, and holistic fusion.	Enhanced performance and efficiency for speaker verification.
Loweimi et al., (2025) [46]	Speaker retrieval framework tested on BBC Rewind archive under diverse conditions.	BBC Rewind corpus	Audio retrieval system addressing diarization, embedding extraction, and query selection under noisy/unconstrained conditions.	Robust and adaptable retrieval system in noisy environments.
Zheng et.al., (2025) [48]	NResNet	LibriSpeech	Layered ResNets with multi-scale aggregation and speaker attention for classroom speech recognition.	4.00% EER improvement and 6.20% accuracy enhancement.
Gu et.al., (2025) [50]	Domain-robust pre-training using local prototypes, self-distillation, and online clustering.	CNCeleb, VoxCeleb	Domain-robust pretraining with local prototypes and self-distillation for SV.	Improved generalization across domains.
Wang et.al., (2023) [51]	CNN-FNN	Conformer-based SV	Lightweight CNN front-end with channel-frequency attention and light FFN.	Reduced complexity with retained performance.
Kim et.al., (2024) [52]	Universal pooling of multi-layer features from pretrained models	Various pretrained models, flexible contexts	Intermediate information processing with layer/frame-level pooling and convolution-based interlayer design.	Better performance and cost-efficiency vs. traditional ASV.
Han et.al. (2025) [53]	Generalized score comparison-based learning objective with multiple loss functions.	VoxCeleb, VOiCES, CN-Celeb, Common Voice	Score comparative learning with generalized loss for verification.	Improved robustness and reduced overfitting.
Liu et.al., (2023) [54]	DF-ResNets and DF-ECAPAs	VoxCeleb	Sequential and Parallel AFF.	Up to 40% EER improvement at low computational cost.
Chen et al., (2025) [55]	SpeakerRPL	Multilingual OpenSID datasets	open-set identification with unknown sample learning.	Better few-shot open-set recognition than baseline models.
Wang and Li (2024) [68]	TS-VAD	DIHARD III, AliMeeting	Diarization with self-embedding updates.	Outperforms offline clustering; matches SOTA multi-channel diarization.

An online Target Speaker Voice Activity Detection (TS-VAD) technique for speaker diarization that is independent of prior knowledge from clustering-based systems is suggested by Wang and Li (2024) [68]. The approach develops self-embeddings to identify speaker activities, extending classic TS-VAD for real-time operation. The approach prevents permutation ambiguities and guarantees consistent performance by repeatedly updating target speaker embeddings during inference. The suggested approach performs better than offline clustering-based diarization techniques, according to experiments conducted on the DIHARD III and AliMeeting datasets. Its performance is on par with the most advanced offline diarization methods when used with multi-channel data. Table 1 shows the comparative analysis of embedding extractors in supervised learning approaches.

4.1.2. Clustering and Classification Models

By classifying speech segments and allocating them to certain speakers, classification and clustering models are essential to speaker diarization. One of the most popular methods is still the x-vector + PLDA framework, which uses probabilistic scoring and discriminative embeddings to provide reliable speaker attribution. Graph Neural Networks (GNN) and deep clustering techniques like Improved Deep Embedding Clustering (IDEC) improve cluster separability in spectral clustering using learned affinity matrices, a recent development. By simulating intricate speaker connections under various acoustic situations, these methods greatly increase the accuracy of diarization. The clustering and classification models are discussed below,

Singh and Ganapathy (2025) proposed an end-to-end supervised hierarchical clustering approach built on GNN [57]. While the GNN model was first trained using the x-vector embeddings from the pre-trained model, the embedding extractor was initialized using a pre-trained x-vector model. Lastly, the E-SHARC model performed representation learning, metric learning, and clustering with end-to-end optimization by the usage of the front-end mel-filter bank features as input and jointly optimizing the embedding extractor and the GNN clustering module. Furthermore, the E-SHARC method could estimate the speakers in the overlapping speech regions when it received extra inputs from an external overlap detector. The experimental assessment on benchmark datasets such as AMI, Voxconverse, and DISPLACE showed that the suggested E-SHARC framework used the graph-based clustering techniques to produce competitive diarization outcomes.

Singh et al. (2023) used a hierarchical structure employing GNN to accomplish supervised clustering in their innovative Supervised Hierarchical Graph Clustering technique (SHARC) for speaker diarization [59]. A single-step method for diarization is made possible by the supervision, which enables the model to update its representations and directly enhance the clustering performance. The input segment’s embeddings are viewed as regions of a graph in the proposed approach, and the edge weights are determined by the nodes' similarity scores. Additionally, suggest a method for End-To-End speaker diarization (E2E-SHARC) that updates the GNN model and the embedding extractor simultaneously. Node densities & edge existence probabilities are used in the hierarchical clustering process during inference to merge the segments till convergence. Demonstrate in the diarization trials that the suggested E2E-SHARC method outperforms the baseline systems by 53% and 44%, respectively, on benchmark datasets such as AMI and Voxconverse.

A speaker diarization system utilizing speaker embedding parameters, particularly the x-vector, was reported by Khadar et al. (2025) [60]. The system is more resilient to noise changes when auto-correlated MFCC features are used for x-vector extraction through a pre-trained delayed neural network. With PLDA scoring as the distance metric and speaker grouping achieved through agglomerative clustering, the system is very useful for identifying possible speakers, particularly in forensic applications. By combining different kinds of noise, including red, pink, and white noise, over a broad range of signal-to-noise ratios (20 dB to − 20 dB), the system's noise adaptability is extensively assessed. Furthermore, by altering the total number of speakers and speech duration, the system's performance is thoroughly evaluated, demonstrating its resilience and efficacy in practical situations.

Nareaho (2023) has provided a process of identifying "who spoke when" in an audio clip [61]. Speech technology has advanced significantly in many areas and measurements since the discovery of deep learning, and speaker diarization seems no exception. The purpose of this thesis is to examine which acoustic conditions are challenging to diarize, as well as to assess how well the state-of-the-art speaker diarization system, Pyannote, performs in more challenging acoustic environments and investigate ways to improve that performance. At first, Pyannote had trouble distinguishing between audio with a lot of reverberation and audio with significantly inferior sound quality, such as a phone call. The performance was significantly enhanced for the most challenging environments and stayed relatively constant for the easier ones by employing fine-tuning methods and a method for augmenting the data used for training. This suggests that Pyannote is resilient and capable of adjusting to notable changes in the audio signal strength.

Pande et.al. (2025) analyzed the effectiveness of various speaker diarization techniques in practical settings [62]. I-vectors used statistical models to produce concise and effective representations of speakers, which made them useful for automatic speaker recognition. They had trouble, nevertheless, when there was background noise or multiple sounds. However, X-vectors get over these restrictions. VoxCeleb and the AMI Meeting Corpus were two common datasets used to assess these two methods. Diarization Error Rate (DER) and Jaccard Error Rate (JER) were the two metrics used to assess their performance.

A new pruning approach named Spectral Clustering using P-Neighborhood maintained Affinity matrix (SC-pNA) is introduced by Raghav et.al. (2025) to produce a sparse affinity matrix [63]. By permitting a variable number of neighbors, this approach outperforms node-specific fixed neighbor selection. Additionally, since the pruning parameters are obtained directly from the affinity matrix, no extra tuning data is required. To accomplish this, SC-pNA finds two clusters in each row of the original affinity matrix and only keeps the highest percentage similarity scores from the cluster with the most similarities. After that, spectral clustering is performed, and the largest eigenvalue gap is used to determine the number of clusters. The superiority of SC-pNA, which is additionally computationally more economical than current auto-tuning techniques, is demonstrated by experimental results using the difficult DIHARD-III dataset. Table 2 shows the comparative analysis of clustering and classification Models in supervised learning-based approaches.

Table 2. Comparative analysis of clustering and classification models in supervised learning-based approaches

Reference	Method/Model	Dataset	Contribution	Performance Metrics
Singh and Ganapathy (2025) [57]	E-SHARC Clustering with GNN	AMI, Voxconverse, DISPLACE	End-to-end optimization with mel-filterbank features and overlap detection.	Competitive DER on-benchmark datasets
Singh et al., (2023) [59]	SHARC/E2E-SHARC	AMI, Voxconverse	Supervised clustering with GNN updating embeddings and extractor jointly.	53% and 44% DER improvement over baselines
Khadar et al., (2025) [60]	X-vector (TDNN) + PLDA + Agglomerative Clustering	Synthetic noisy data with multiple noise types	Noise-robust diarization for forensic applications.	Stable DER across SNR range (+20dB to -20dB)
Nareaho, (2023) [61]	Pyannote (fine-tuned deep network system)	Varied acoustic environments (reverberant, phone-quality)	Evaluation of Pyannote robustness and fine-tuning for difficult environments.	DER improved in reverberant/low-quality audio
Pande et.al., (2025) [62]	I-vector vs X-vector embeddings	VoxCeleb, AMI Meeting Corpus	Comparison of I-vector and X-vector in noisy conditions.	DER, JER (X-vectors outperform I-vectors)
Raghav et.al., (2025) [63]	SC-pNA	DIHARD-III	Sparse affinity matrix without tuning data, computationally efficient.	DER reduced compared to auto-tuning baselines

4.1.3. End-to-End Supervised Models

In speaker diarization, end-to-end supervised models assign speaker labels directly, eliminating the need for independent embedding and clustering steps. At the same time, UIS-RNN employs a recurrent neural network that can sequentially estimate speaker labels with temporal consistency, EEND, and its enhanced variation EEND-EDA jointly simulate multi-speaker interactions to handle overlapping speech. Conversely, Speaker-Attributed Automatic Speech Recognition (SA-ASR) combines speech recognition and diarization to provide transcripts that are already in line with speaker identities. This capability has been expanded more recently by Whisper-based systems, which ensure robustness across many domains by simultaneously executing transcription and diarization. The studies listed below provide a detailed description of these methods.

Dasari (2025) [64] found a method based on the current End-to-End Neural Diarization with Masked-attention Mask Transformers (EEND-M2F) architecture, which developed an end-to-end diarization model by adding embeddings taught by multi-view contrastive learning to the backbone. The main tasks included training the model on publicly accessible data sets, duplicating the model, and expanding the backbone as outlined.

Kalda *et.al.* (2025) examined simple post-processing methods to reduce SA-ASR timestamp mistakes brought on by lengthy silences and artifacts in sources that were segregated at the file level [65]. With performance comparable to typical approaches without any speaker embedding fine-tuning, it demonstrates the possibility of directly extracting speaker embeddings for the diarization pipeline from isolated sources. After fine-tuning the ASR system on the separated sources, their PixIT-based technique achieved a 20% improvement in Word Error Rate (WER) over the CSS-based baseline on the NOTSOFAR-1 challenge dataset. Interestingly, without utilizing any of the given domain-specific synthetic data, the system can outperform the baseline even when using the identical ASR model. These developments establish PixIT as a reliable and adaptable SA-ASR system for the real world.

Landini (2024) introduced a system called Variational Bayesian x-vector clustering (VBx**)**, which uses a Bayesian hidden Markov framework and shows strong performance on different datasets [66]. Its advantages and limitations have been evaluated on several corpora. EEND techniques have also been explored. Since these models require large training sets and there is a lack of manually annotated diarization data, an alternative approach is to generate artificial data that mimics real conversations, including speaker overlaps and turns. Training the commonly used EEND with Encoder-Decoder Attractors (EEND-EDA) on such simulated conversations achieves better results than earlier simulated mixture methods.

An online target speaker SAD method for speaker diarization is suggested by Wang and Li (2023) [67]. In contrast to clustering-based diarization, this system extends the conventional target speaker speech detection technique for real-time operation by using self-generated embeddings to detect speaker activities rather than requiring prior information on the number of speakers. Throughout inference, the model continuously modifies target speaker embeddings, providing consistent performance even in the face of permutation problems. Experimental results demonstrate that this method delivers performance comparable to advanced offline systems when applied to multi-channel data and outperforms offline clustering-based diarization on the DIHARD III and AliMeeting datasets.

Kwon et al. (2023) presented a framework for online speaker diarization that performs well in various areas [69]. Online diarization, in contrast to offline approaches, must manage real-time outputs that cannot be changed, so mistakes made early in the session may have an adverse effect on later results. This is addressed by the framework, which progressively raises the projected number of speakers and permits lowering the number if a previous rise is found to be inaccurate. The system makes use of two buffers: centroids, which represent speaker embeddings, and checkpoints, which estimate the number of speakers. A clustering-based label matching method is employed for real-time label assignment. The method reaches state-of-the-art performance on DIHARD II and III and performs well on the AMI and VoxConverse datasets despite being lightweight.

Regardless of the number of speakers in a discussion, Cheng et.al. (2023) have proposed a neural architecture that concurrently generates speaker representations according to the speaker diarization aim and recognizes each speaker's presence on a frame-by-frame basis [70]. Implemented by a network with residuals and processing data across the temporal and speaker scales, a time-speaker contextualizer and a speaker representation (referred to as a z-vector) extractor are combined into a single framework. Experiments on the CALLHOME corpus demonstrate that the model performs better than most previously suggested approaches. Tests in a more difficult scenario with two to seven simultaneous speakers reveal that the approach reduces the relative DER by 6.4% to 30.9% compared to several common baselines.

The Word-level (WEEND) using auxiliary network is a multi-task learning method that combines speaker diarization and end-to-end ASR in a single neural architecture, according to Huang et.al. (2023) [71]. In other words, speaker labels are simultaneously predicted for every word that is recognized during speech recognition. According to experimental results, WEEND can generalize to audio lengths of five minutes and performs better than the turn-based diarization baseline system in all 2-speaker short-form scenarios. When given enough in-domain training data, WEEND has been shown to produce high-quality diarized text; nevertheless, managing interactions with three or more speakers is still difficult.

SpeakerLM is a unified multimodal big language framework for Speaker Diarization and Recognition (SDR) that collaboratively executes SD with ASR in an end-to-end fashion, as demonstrated by Yin et.al. (2025) [72]. Furthermore, it integrates a versatile speaker registration method into SpeakerLM to support a variety of real-world applications, allowing SDR under various speaker registration configurations. A multi-stage training approach is used to gradually improve SpeakerLM on vast amounts of real data. Numerous tests reveal that SpeakerLM outperforms the most recent cascaded baselines when using in-domain as well as out-of-domain public SDR benchmarks, exhibiting significant data scaling capabilities and generalizability. Additionally, experimental findings demonstrate that the suggested speaker diarization process successfully guarantees a strong SDR response from SpeakerLM for a range of speaker registration circumstances and speaker registration numbers.

Cornell et.al. (2024) introduced Sliding-Window Diarization-Augmented Recognition (SLIDAR), a unique framework for ASR with joint speaker diarization [73]. Because SLIDAR can handle any number of speakers and processing inputs of any duration, it can solve the "who spoke what, when" problem simultaneously. SLIDAR uses a sliding window technique and is comprised of an End-To-End Diarization-Augmented Speech Transcription (E2E DAST) architecture that offers speaker embeddings, diarization, and transcripts locally for each window. Based on a simple encoder-decoder architecture, the E2E DAST model makes use of modern methods, including "Whisper-style" prompting and serialized output training. After clustering speaker embeddings to obtain global speaker identities, the local outputs are combined to obtain the final Speaker Diarization+ASR result. The method's efficacy in both close-talk and far-field speech settings is confirmed by experiments conducted using monaural recordings from the AMI corpus.

A lightweight single-shot Distilled Whisper-to-Normal (DistillW2N) model is proposed by Tan et.al. (2025) to transform whispered speech into regular speech [74]. The two main parts of the system are a Speech-to-Unit (S2U) encoder that extracts Hidden - Unit BERT (HuBERT) - Soft representations from normal and whispered speech and learns to reduce noise using a Voice Activity Detection (VAD) model; and a Unit-to-Speech (U2S) decoder that uses the SoundStream decoder for lightweight, high-quality speech synthesis and Style-Adaptive Layer Normalization (SALN) to combine content units with timbre units. According to experimental results, DistillW2N outperforms other Self-Supervised Learning (SSL) techniques in improving whisper intelligibility while maintaining speaker similarity. Table 3 shows the comparative analysis of end-to-end supervised models in speaker diarization.

Table 3. Comparative analysis of end-to-end supervised models in speaker diarization

Reference	Method/Model	Dataset	Contribution	Performance Metrics
Dasari (2025) [64]	EEND-M2F + Multi-view Contrastive Learning	Publicly available datasets	Extended EEND-M2F backbone with contrastive embeddings for robust diarization.	Improved generalization and accuracy on benchmark data
Kalda et.al., (2025) [65]	PixIT-based SA-ASR	NOTSOFAR-1 Challenge	Reduced SA-ASR timestamp errors; direct embedding extraction from isolated sources.	20% WER improvement over CSS-based baseline
Landini, (2024) [66]	VBx + EEND, EEND-EDA, DiaPer	AMI, simulated corpora	Proposed DiaPer outperforming EEND-EDA in overlap scenarios; VBx as Bayesian baseline.	DiaPer achieves superior DER vs EEND-EDA.
Wang and Li (2023) [67]	TS-VAD	DIHARD III, AliMeeting	Introduced online diarization with real-time target speaker tracking.	Outperforms offline clustering-based baselines.
Kwon et.al., (2023) [69]	Conservative Online Diarization	AMI, VoxConverse, DIHARD II & III	Proposed buffer-based correction to avoid catastrophic early errors.	SOTA efficiency on DIHARD datasets
Cheng et.al., (2023) [70]	Multi-target Extractor & Detector (z-vector)	CALLHOME, simulated multi-speaker	Proposed z-vector extractor with contextualization for unknown # speakers.	6.4%–30.9% relative DER reduction
Huang et.al., (2023) [71]	WEEND (Word-level EEND + ASR)	Short 2-speaker corpora	Integrated diarization and ASR at the word-level with an auxiliary network.	Outperforms turn-based diarization in 2-spk settings.
Yin et.al., (2025) [72]	SpeakerLM (Multimodal LLM-based diarization)	In-domain & out-of-domain SDR benchmarks	Unified diarization + recognition with scalable multimodal LLM.	Outperforms cascaded baselines across SDR tasks.
Cornell et.al. (2024) [73]	SLIDAR (Sliding-window Diarization-ASR)	AMI corpus	Proposed E2E diarization-augmented ASR with Whisper-style prompting.	Robust SD+ASR results across far-field speech
Tan et.al., (2025) [74]	DistillW2N (Whisper-to-Normal)	Synthetic Whisper datasets	Lightweight one-shot Whisper-to-normal voice model with SALN.	5× faster than QuickVC; intelligibility gains while preserving speaker identity

When compared to conventional clustering methods, supervised learning-based techniques have proven to be the most successful in speaker diarization because they use annotated data to acquire extremely discriminative speaker characteristics regardless of noisy and overlapping environments.

Advanced end-to-end models, such as DiaPer, improve upon earlier EEND approaches by effectively handling multi-speaker overlaps and reducing diarization errors across different datasets.

Similarly, embedding-based models, such as ECAPA-TDNN, combined with hierarchical clustering, enhance the embedding quality, enabling accurate speaker segmentation and robust speaker identification. When combined, these supervised systems exhibit exceptional accuracy, flexibility, and resilience, making them the top option for contemporary diarization tasks.

4.2. Unsupervised or Self-Supervised Approaches

By allowing models to develop directly from unlabeled or raw voice data, unsupervised and self-supervised methods in speaker diarization seek to reduce the significant reliance on huge, annotated datasets. These techniques are intended to maximize separation in intricate auditory situations, cluster segments without labels, and derive meaningful speaker representations. Self-supervised methods like HuBERT, Wav2Vec 2.0, and DINO-style contrastive models use contrastive learning or masked input prediction to collect speaker-specific and contextual information. By iteratively improving cluster assignments, clustering-based unsupervised techniques, such as DEC, IDEC, and pseudo-label refinement, enhance embedding separability. Optimization-inspired methods combine clustering pipelines with meta-heuristic algorithms, such as Memory-GWO and the Chronological Political Optimizer (CPO), to increase global search efficiency and robustness. Below is a detailed discussion of these methods. Figure 4 shows the unsupervised or self-supervised techniques.

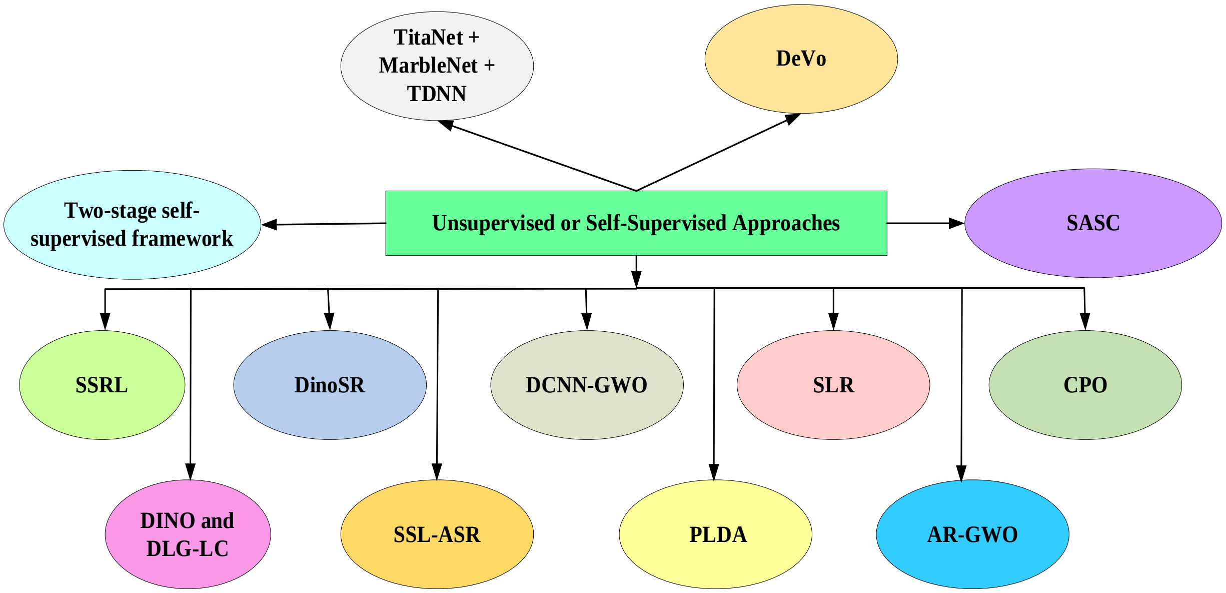

TitaNet + MarbleNet + TDNN
DeVo
Two-stage self-supervised framework
Unsupervised or Self-Supervised Approaches
SASC
SSRL
DinoSR
DCNN-GWO
SLR
CPO
DINO and DLG-LC
SSL-ASR
PLDA
AR-GWO

Fig. 4 Unsupervised or self-supervised techniques

4.2.1. Self-Supervised Embedding Extraction

Wang et al. (2024) documented progress in speaker encoders, encompassing supervised and self-supervised learning techniques, independent and extensive pretrained models, exclusive speaker embedding training, joint optimization with downstream tasks, and methodologies for interpretability [47]. Several open-source toolkits, including Kaldi, SpeechBrain, and pyannote-audio, offer speaker embedding extractor implementations.

The topologies, training methods, and robustness of these toolkits can be compared across various speech recognition and diarization tasks. A clear reference for researchers working in the field of speaker characterization and modeling, as well as for those wishing to apply speaker modeling techniques to downstream tasks.

A two-stage self-supervised framework using unlabeled data has been presented by Cai (2023) [49]. While the second step incorporates discriminative training and clustering, the first stage concentrates on representation learning. Self-supervised knowledge distillation, which is tuned to reduce label noise effects, is at the heart of the self-supervised learning through reflection approach, which further streamlines this framework. The quality of self-supervised speaker representation is greatly enhanced by this method.

Transfer learning techniques are investigated to make effective use of sparse training data by utilizing the connection between ASR and actual speaker verification. ASR-based knowledge distillation, initialization using ASR-pretrained encoders, and a speaker adaptor that transforms ASR features into speaker-specific ones are some of the methods. To identify the attackers behind conversions, the thesis also investigates voice conversion spoofing countermeasures. To put it simply, this work provides improvements in speaker representational learning, addresses data limitations, and strengthens security against voice spoofing, all of which strengthen audio applications.

Self-Supervised Reflection Learning (SSRL), a better framework that overcomes these drawbacks by permitting ongoing pseudo-label revision during training, was introduced by Cai et.al. (2025) [56]. SSRL does away with iterative training cycles by using an online clustering technique and teacher-student architecture. Use temporally consistent pseudo-label queues and noisy label modeling to deal with label noise. VoxCeleb experiments demonstrate that SSRL outperforms existing two-stage iterative methods, outperforming a 5-round method in a single trained round. The contributions of important elements such as pseudo-label queues and noisy label modeling are validated by ablation studies. Furthermore, the convergence among cluster counts and steady gains in pseudo-labeling show how well SSRL decodes unlabeled data. With the use of the innovative reflective learning paradigm, this study represents a significant breakthrough in accurate and efficient self-supervised speaker representation learning.

Prosody-Enhanced wav2vec (PE-wav2vec), a self-supervised speech model, was introduced by Liu et.al. (2024) using prosody learning [75]. In particular, the wav2vec 2.0 architecture's first transformer blocks are subjected to supervision using Linear Predictive Coding (LPC) so that the remaining signals can accomplish prosody learning. The relevant frames in a spoken utterance are represented prosodically by the embedding vectors that were taken from the first Transformer blocks of the PE-wav2vec model. The Speech Synthesis System with Self-Supervisedly Learned Prosodic Representations (S4LPR) is an acoustic model that is based on FastSpeech 2 and is designed to use these PE-wav2vec representations for Text-To-Speech (TTS). The experimental findings show that the suggested PE-wav2vec model outperforms the vanilla wav2vec 2.0 model in describing voice prosody. In addition, when compared to baseline models, the S4LPR model with PE-wav2vec representations can successfully enhance the subjective naturalness and lessen the objective distortions in synthetic speech.

According to Jafarzadeh et.al. (2024), deep learning methods have become more and more popular, including recurrent and CNN being used for Speech Emotion Recognition (SER) tasks [76]. Most of the input for these methods consists of spectra and manually constructed features. In this work, we examine the usage of two self-supervised transformer-driven algorithms, such as HuBERT and Wav2Vec2, to infer speakers' emotions. The models use raw audio inputs to derive features for the classification task. Reliable datasets, including EmoDB, AESDD, SAVEE, SHEMO, along with RAVDESS, are used to test the proposed method. The results show the effectiveness of the proposed approach on multiple datasets. Furthermore, the model has been utilized in real-world situations, such as contact center encounters, and the results demonstrate how accurately it predicts emotions.

Vielzeuf (2024) studied the evolution of high-level information during pretraining, concentrating on the HuBERT model with less attention to its "autoencoder" behavior [77]. By analyzing the variables that affect HuBERT's top layers, the goal was to enhance both the training procedure and the top layers' performance on challenging tasks. Experiments demonstrated that these training improvements produce competitive outcomes on downstream tasks in addition to accelerating convergence.

Self-distillation with online clustering was proposed by Liu et.al. (2023) for self-supervised speech representation learning (DinoSR) [78]. To create a successful voice learning model, this system integrates online clustering, self-distillation, and masked language modeling. DinoSR first extracts contextualized embeddings from input audio using a teacher network, then uses online clustering to generate a machine-discovered phone inventory, and lastly employs discretized tokens to drive a student network. On several downstream tasks, experiments demonstrate that DinoSR performs better than prior state-of-the-art models, and a thorough analysis demonstrates the potency of the learnt discrete units.

A self-supervised learning approach for reliable SV in the absence of labeled data was presented by Han et.al. (2023) [79]. The method uses the self-distillation with no labels (DINO) architecture, which minimizes the impact of false negatives in contrastive learning and does not rely on negative pairs. A cluster-aware training method is used to increase data diversity. GMMs are used to predict the loss distribution and adaptively filter unreliable labels in a Dynamic Loss-Gate with Label Correction (DLG-LC), which is suggested because unsupervised clustering may provide wrong labels.

Comparing the VoxCeleb dataset to the most advanced self-supervised SV systems, experiments reveal notable relative Equal Error Rate (EER) increases of 22.17%, 27.94%, and 25.56% on the Vox-O, Vox-E, and Vox-H test sets. Table 4 shows the comparative analysis of self-supervised embedding extraction methods.

Table 4. Comparative analysis of self-supervised and unsupervised embedding extraction methods

References	Dataset	Method / Model	Contribution	Performance Metrics
Wang et al., (2024) [47]	Various (Kaldi, SpeechBrain, pyannote-audio toolkits)	Speaker modeling approaches	Reported progress in speaker encoders: supervised and self-supervised learning, pretrained models, embedding training, joint optimization, and interpretability.	Reference resource: qualitative comparison of robustness and topologies.
Cai, (2023) [49]	general SV + ASR datasets	Two-stage self-supervised framework	Two-stage self-supervised framework with reflection learning, ASR-based knowledge distillation, and spoofing countermeasures.	Improved representation quality; stronger spoofing resistance.
Cai et.al., (2025) [56]	VoxCeleb	SSRL with self-distillation and online clustering	SSRL with online clustering, teacher–student architecture, pseudo-label queues, and noisy label modeling.	Outperforms iterative methods; better convergence and label robustness.
Liu et.al., (2024) [75]	TTS datasets with prosody, wav2vec 2.0 pretraining	PE-wav2vec	Introduced prosody learning into wav2vec 2.0 using LPC supervision; integrated into Speech Synthesis system (S4LPR).	Improved prosody modeling; enhanced naturalness and reduced distortions in speech synthesis.
Jafarzadeh et.al. (2024) [76]	EmoDB, AESDD, SAVEE, SHEMO, RAVDESS	HuBERT & Wav2Vec2 for Emotion Recognition	Applied self-supervised embeddings to speaker emotion recognition using multiple emotion datasets.	Effective emotion prediction across multiple datasets; robust in real-world applications (e.g., contact centers).
Vielzeuf, (2024) [77]	HuBERT model, downstream tasks	HuBERT Pretraining Analysis	Investigated 'autoencoder behavior' in HuBERT during pretraining and improved training dynamics.	Competitive downstream performance; faster convergence.
Liu et.al., (2023) [78]	Multiple speech corpora	DinoSR (Self-distillation + Online Clustering)	Combined masked language modeling, self-distillation, and online clustering for speech representation.	Outperforms prior SOTA on several downstream tasks; strong discrete unit learning.
Han et.al. (2023) [79]	VoxCeleb (Vox-O, Vox-E, Vox-H)	Cluster-Aware DINO + DLG-LC	Extended DINO framework with cluster-aware training and dynamic loss-gate label correction.	Relative EER improvements: 22.17% (Vox-O), 27.94% (Vox-E), 25.56% (Vox-H).

4.2.2. Clustering-Based Unsupervised Methods

Dissen *et.al.* (2024) expanded on the earlier research on speaker diarization, using fully unsupervised deep learning [58]. The study specifically focused on estimating secondary hyperparameters of the model without annotations and producing high-quality neural speaker representations without any annotated data. An encoder trained in a self-supervised manner using pairs of adjacent segments presumed to be of the same speaker represents the speaker embeddings. Next, the PLDA, a trained encoder model, self-generates pseudolabels to learn a clustering stopping threshold and a similarity score between various segments of the same call. Additionally, the quality of the embeddings was further improved by iteratively retraining the embedding model with updated pseudo-labels, which improves performance.

The pseudo-label refinement (SLR) approach was introduced by Mahmood and Kang (2024) to solve the issue of inaccurate pseudo-labels [80]. This method determines a soft, refined label as a linear combination of the recently assigned label and the expected label, using data from the current and previous epoch clusters. The hierarchical clustering is applied to furnish more precise hard labels than the traditional method of maximum value when translating the cluster labels of the preceding epoch to the label space of the present epoch. It is experimentally demonstrated that the addition of the SLR algorithm to the base Re-ID model leads to a significant increase in the mean Average Precision (mAP) in Unsupervised Domain Adaptation (UDA) of person Re-Identification (Re-ID) tasks, such as real-to-synthetic, synthetic-to-real, and real-to-real. These findings show the effectiveness of the SLR algorithm in enhancing the performance of self-supervised machines in learning.

To increase the fairness and resilience of end-to-end ASR, Veliche and Fung (2023) have proposed a privacy-preserving method that does not require the use of speaker or utterance embeddings, zip codes, or metadata during training [81]. Unsupervised acoustic cluster formation is achieved by extracting utterance-level embeddings from a public dataset used to train a speaker recognition algorithm.

Cluster IDs are used as extra features during training, rather than depending on speaker utterance embeddings. This method has significant advantages for managing various dialects and enhances performance overall across demographic groupings.

Nazir *et.al.* (2023) suggested two speech analysis methods to address prosodic errors (differences in pronunciation) and phonemic errors (confusing similar phonemes) [82]. In the first approach, phonemic problems are detected using deep CNN features and a clustering algorithm, and phoneme classification is done using classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes. 94% accuracy was attained in experiments on six Arabic phoneme pairings that are commonly mispronounced. To detect prosodic faults, the second approach introduces an unsupervised Phone Variation Model (PVM), in which every phone is modelled with many pronunciation variations according to skill levels. This method produced feedback specific to each phone's variance and achieved 97% accuracy using a dataset of 28 Arabic phones.

Table 5. Comparative analysis of clustering-based and unsupervised methods

Reference	Dataset	Method / Model	Contribution	Performance Metrics
Dissen et.al., (2024) [58]	Call recordings	Self-Supervised Speaker Embeddings + PLDA	Unsupervised diarization with pseudo-label retraining.	Improved clustering threshold & similarity score, enhanced embeddings
Mahmood and Kang (2024) [80]	Person Re-ID dataset	Pseudo-label Refinement (SLR)	Proposed SLR approach using soft refined labels across epochs with hierarchical clustering for label precision.	significant improvement over baseline Re-ID
Veliche and Fung (2023) [81]	Public dataset for speaker recognition	Unsupervised Clustering for ASR Fairness	Introduced privacy-preserving ASR by clustering acoustic embeddings instead of speaker metadata.	Better fairness & resilience in end-to-end ASR, improved across dialects & demographics
Nazir et.al., (2023) [82]	Arabic phoneme dataset (6 pairs, 28 phones)	Deep Feature Clustering + PVM	Developed clustering with CNN features and PVM to detect phonemic/prosodic pronunciation errors.	94% accuracy (phonemic errors), 97% accuracy (prosodic errors)
Kou et.al., (2023) [83]	Audio signals, handwritten digits, object images, bioinformatics	SASC	Proposed SASC that integrates global and local correlations into representation learning.	Strong grouping effect, robust clustering performance
Ollagnier et.al., (2023) [84]	MLMA dataset (French & English tweets)	Multi-view Spectral Clustering for Hate Speech	Built an unsupervised pipeline for fine-grained detection of hate speech targets using multilingual embeddings.	Outperformed baselines in 69 runs (hate speech clustering)
Ma et.al., (2023) [85]	LibriSpeech test set	MonoBERT & PolyBERT	Proposed unit discovery models using context-independent and context-dependent phoneme units.	Higher accuracy than SSL baselines, no frequent re-clustering needed

Kou et.al. (2023) introduced a brand-new unsupervised technique named Structure-Aware Subspace Clustering (SASC) [83]. To accurately represent the intrinsic structure, SASC concurrently considers both global and local correlation structures. To obtain accurate data affinity, it also incorporates the collected structure into representation learning. It strengthens the resilience and application of subspace clustering and effectively encourages an all-around grouping effect. The success of the suggested technique is shown through experiments on a variety of benchmark datasets, such as audio signals, handwritten digits, object images, and bioinformatics.

Ollagnier et.al. (2023) suggested a comprehensive unsupervised pipeline that identifies the protected qualities being discriminated against as well as the targeted victims (individuals or groups) [84]. Their contributions are as follows: (1) employing clustering approaches for fine-grained identification of hate speech target populations; (2) comparing various abusive behaviours using numerous data perspectives; and (3) doing in-depth content analysis on manually generated hate speech targets. The pipeline far surpassed current clustering techniques in 69 runs on the Multilingual Hate Speech (MLMA) dataset of French and English tweets. The method combines the Multi-view Spectral Clustering (MvSC) algorithm with multilingual pre-trained language models, such as the multilingual Universal Sentence Encoder and the multilingual BERT.

An unsupervised approach is proposed by Ma et.al. (2023) to enhance the goals of Self-Supervised Learning (SSL) [85]. For pre-training, two models are presented: MonoBERT and PolyBERT. These models employ context-independent or context-dependent phoneme-based units, respectively. These models perform better than other SSL techniques on the LibriSpeech test set without the need for frequent re-clustering or re-training. Target improvement models that have been pre-trained using labeled data perform worse than models that use context-dependent units. Additionally, the study shows how the unit discovery process has been gradually improved. Table 5 shows the comparative analysis of clustering-based and unsupervised models.

4.2.3. Optimization-Inspired Unsupervised Models

Revathy and Kumar (2025) used speaker diarization, where the CPO for the speaker was used to improve the DEC's parameters [86]. Combining the chronological notion with the political optimizer yields the suggested CPO. Speaker labeling, or labeling the speakers, was made easier by speaker diarization. By matching the identified speaker with the new speech signal while taking the congruence coefficient into account, the pertinent speech was retrieved. Using the suggested CPO, a speech improvement generative adversarial network is used to improve speech. With the lowest diarization error of 0.251, the greatest F-measure of 0.910, the lowest false acceptance rate of 0.398, and the highest perceptual rating of speech quality of 0.325, the suggested model fared better than the others.

An Adaptive Randomised Grey Wolf Optimisation (AR-GWO) is suggested by Jadda and Prabha (2023) for correctly adjusting the tuning factor, also known as the tuned tuning factor [87]. The traditional GWO has been improved with the proposed AR-GWO. Training is completed first, then the tuning factor is obtained for each pertinent input signal during the testing phase. The adaptive Wiener filter is then used to provide the appropriately adjusted tuning factor into FW-NN as input to the Empirical Mode Decomposition (EMD) to decompose the spectral signal. The result of this process is a denoised, improved speech signal. Specifically, the selected AR-GWO model outperforms the current GA, ABC, PSO, FF, and GWO techniques in terms of calculation time by 34.07%, 43.57%, 28.86%, 38.88%, and 16.03%, respectively.

Table 6. Comparative analysis of optimization-inspired unsupervised models

Reference	Dataset	Method / Model	Contribution	Performance Metrics
Revathy & Kumar, (2025) [86]	Speech datasets	DEC + CPO	Integrated CPO with DEC for optimized clustering and applied GAN-based speech enhancement.	Achieved the lowest DER (0.251), highest F-measure (0.910), and improved perceptual speech quality rating (0.325).
Jadda and Prabha (2023) [87]	Speech signals (benchmark noisy speech datasets, unspecified)	AR-GWO + FW-NN + Adaptive Wiener Filter	Proposed AR-GWO for optimized tuning factor in fuzzy wavelet neural networks with Wiener filtering.	Outperformed GA, ABC, PSO, FF, and GWO in speech denoising; reduced computation time by up to 43.57%.
Falahzadeh et.al., (2023) [88]	SER datasets	DCNN-GWO for SER	Introduced Chaogram-based 3D tensor input for SER with VGG DCNN and optimized using GWO.	Achieved strong performance on EMO-DB and eNTERFACE05 datasets, enhancing SER applications.

By transforming a speech signal to a 3D tensor, Falahzadeh et.al. (2023) have demonstrated a pre-trained DCNN algorithm for SER and offer suitable input to these networks [88]. First, speech samples are then recreated in a 3D phase space utilizing a reconstructed phase space. According to studies, the speaker's significant emotional traits are contained in the patterns that form in this area. Three channels that resembled RGB images were produced by projecting these patterns onto a new speech signal structure known as Chaogram to give an input that can be compatible with DCNN.

The intricacies of Chaogram photos were then emphasized using image-enhancing techniques. Then, Chaogram high-level features and associated emotion classes are learned using the Visual Geometry Group (VGG) and DCNN, which have already been pre-trained on the extensive ImageNet dataset. Lastly, the suggested model undergoes transfer learning, and its performance is refined using the provided datasets. Table 6 shows the comparative analysis of optimization-inspired unsupervised models.

4.2.4. Emerging Hybrid Models

The context-aware fine-tuning method is a novel strategy introduced by Shon et.al. (2023) [89]. To encode the entire segment in a context embedding vector, apply a context module to the final layer of a pre-trained model. Introduce an auxiliary loss during the fine-tuning phase to encourage this context embedding vector to resemble the context vectors of neighboring segments. In contrast to conventional fine-tuned models, this entails a negligible overhead and enables the model to generate predictions without input to these surrounding parts during inference time.

For low-resource voice creation, Huang et al. (2024) have introduced a VoiceTuner, which utilizes an effective fine-tuning technique and self-supervised pre-training [90]. To be more precise, they use a large unlabeled dataset to pre-train VoiceTuner-SSL using pre-defined functions, which can be adjusted in downstream tasks, to alleviate the lack of data. Secondly, they develop a multiscale transform adapter to further lower the high training expense in full fine-tuning, which effectively updates only about 1% parameters to create a plug-and-play module. When compared to rival baseline models, VoiceTuner obtains state-of-the-art outcomes in rich-resource TTS evaluation, and experimental results show that VoiceTuner-SSL displays robust acoustic continuations.

Regarding an End-To-End (E2E) ASR, Yue et.al. (2024) have proposed a network architecture featuring lightweight adapters to modify a pre-trained SSL model [91]. Each SSL network layer gains an adapter, which follows training on the downstream ASR task, while the pre-trained SSL network layer’s parameters stay the same. All the pre-trained settings are carried over, preventing the disastrous forgetting issue. At the exact time, they proposed lightweight adapters to enable the network to swiftly adjust to ASR tasks. With up to 17.5% corresponding WER decrease on the 10-minute LibriSpeech split, the suggested adapters-based fine-tuning consistently surpasses full-fledged instruction in low-resource scenarios, according to experiments conducted using the Wall Street Journal (WSJ) and LibriSpeech datasets. The adapter-based adaptation also demonstrates competitive performance in resource-constrained scenarios, further validating the adapters' efficacy.

Lai (2024) suggested a method for sarcastic speech recognition that makes use of Waveform-based Language Model (WavLM) and Parameter-Efficient Fine-Tuning (PEFT) [92]. The study assesses Low-Rank Adaptation's (LoRA) efficacy in comparison to other PEFT techniques and traditional fine-tuning. The findings demonstrate that LoRA achieves improved F1 score (harmonic mean of precision and recall), recall, and precision while notably lowering parameter requirements. While also highlighting related difficulties, these results demonstrate the possibility of integrating LoRA with SSL to enhance natural language comprehension and human-computer interaction.

A Denoising Vocoder (DeVo) method was introduced by Irvin et.al. (2023) [93]. In this method, a vocoder learns to directly synthesize clean speech by accepting noisy representations. To find pertinent features, use rich representation from SSL speech models. Perform an adversarial training of the vocoder using the optimal SSL setup after conducting a candidate search over 15 possible SSL front ends. Provide a causal version that can operate on audio streamed with a latency of 10 ms and a slight decrease in performance.

For speaker diarization, Ahmed et.al. (2025) have introduced the Neuro-TM Diarizer, a deep learning framework based on Neural Tita-Net with Marbel-Net Diarizer [94]. To improve diarization performance in intricate acoustic settings, it combines neural diarization, adaptive beamforming, and noise reduction. The suggested multimodal framework uses Tita-Net to generate speaker embeddings and Marble-Net to detect voice activity. Time-delay neural networks are then used for neural diarization and speaker identification. Three metrics, such as DER, false alarm rate, and missed detection rate, are used to compare the suggested Neuro-TM Diarizer with clustering-based approaches on two standard datasets, VoxConverse and VoxCeleb.

The results of the empirical analysis show that the suggested strategy performs better than clustering-based techniques, achieving 6.89% and 6.93% DER on the VoxConverse and VoxCeleb datasets, respectively. Furthermore, when compared to clustering-based methods, the Neuro-TM Diarizer increased DER by 12.60% on VoxConverse as well as 14.01% on VoxCeleb. The suggested system supports practical uses in audio archiving, speaker authentication, and speech transcription.

A Federated Learning approach that can recognize those participating in a conversation despite the need for a sizable audio library for training has been presented by Bhuyan et.al. (2024) [95]. For the Federated Learning model, which relies on the cosine similarity between speaker embeddings, an unsupervised online update approach is suggested. Additionally, the suggested diarization method uses BIC and Hotelling's t-squared statistic to handle the speaker change detection problem through unsupervised segmentation techniques. This innovative method lessens the severity of this trade-off that exists between missed detection and false detection rates by biasing speaker change detection around identified quasi-silences. Additionally, unsupervised speech segment clustering lowers the computational burden associated with frame-by-frame speaker identification. The outcomes show how successful the suggested training approach is when non-IID voice data is present. Along with lowering the computational overhead, it also demonstrates a significant improvement in the decrease of missing and false detections during the segmentation step. The approach is appropriate, enabling real-time speaker diarization within an autonomous IoT audio network due to its increased accuracy and decreased processing cost. Table 7 shows the comparative analysis of emerging hybrid models in speech diarization.

Table 7. Comparative analysis of emerging hybrid models in speech diarization

Reference	Dataset	Method / Model	Contribution	Performance metrics
Shon et.al., (2023) [89]	general speech datasets	Context-Aware Fine-Tuning	Added context module and auxiliary loss to SSL models for richer embeddings.	Improved SA, NER, and ASR tasks on SLUE and Librilight benchmarks with negligible overhead.
Huang et.al., (2024) [90]	Large unlabeled dataset	VoiceTuner (SSL + Efficient Fine-Tuning)	Developed a multiscale transform adapter updating ~1% parameters for voice generation.	Achieved SOTA in TTS evaluations; robust acoustic continuation in low-resource settings.
Yue et.al., (2024) [91]	LibriSpeech (10-min split), Wall Street Journal (WSJ)	Lightweight Adapters for SSL-ASR	Integrated adapters into each SSL layer for ASR fine-tuning without catastrophic forgetting.	Reduced WER by up to 17.5% on LibriSpeech; competitive in resource-constrained scenarios.
Lai, (2024) [92]	VoxCeleb	PEFT + WavLM (LoRA)	Applied LoRA to WavLM for parameter-efficient sarcasm detection in speech.	Outperformed conventional fine-tuning in F1, recall, and precision with reduced parameters.
Irvin et.al., (2023) [93]	Candidate search across 15 SSL speech model front ends (unspecified datasets)	DeVo (Denoising Vocoder)	Used SSL embeddings with an adversarial vocoder to synthesize clean speech from noisy input.	Causal version runs with 10 ms latency; effective denoising with minor performance loss.
Ahmed et.al., (2025) [94]	VoxConverse, VoxCeleb	Neuro-TM Diarizer (TitaNet + MarbleNet + TDNN)	Combined neural diarization, adaptive beamforming, and noise reduction.	Achieved DER of 6.89% (VoxConverse) and 6.93% (VoxCeleb); improved DER by 12–14% vs clustering methods.
Bhuyan et.al. (2024) [95]	tested on IoT-style conversational audio (non-IID voice data)	Federated Learning for IoT Diarization	Proposed unsupervised FL diarization with Bayesian segmentation and cosine similarity embeddings.	Improved detection accuracy in non-IID IoT data; reduced computation cost, enabling real-time use.

Unsupervised and self-supervised methods have drawn a lot of interest in speaker diarization because they can learn from unlabeled audio without the need for tagged data. Clustering-based models, such as spectral clustering and agglomerative hierarchical clustering, are well known for their ability to accurately group speaker embeddings among unsupervised techniques. Particularly, spectral clustering performs exceptionally well by encapsulating complex interactions in high-dimensional speaker models, allowing for reliable speaker separation even in situations when speech overlaps. Conversely, self-supervised models use vast volumes of unlabeled audio for pre-training speakercharacteristic representations that can subsequently be optimized for diarization tasks. Unsupervised clustering methods, which provide a compromise between accuracy and scalability in a variety of real-world audio situations, continue to be the recommended option when labeled data is limited.

5. Dataset Explanations for Speaker Diarization

In the field of speaker diarization, various datasets offer distinct characteristics that are useful for training, evaluating, and refining. These datasets also differ based on audio, number of speakers, and background noise levels. The following are the various datasets that provide strong benchmarks for creating diarization methods in real-time.

5.1. AMI Meeting Corpus

The AMI meeting corpus is a popular dataset for speaker diarization. It consists of 100 hours of recordings of meetings with various audio channels. It is the better choice for the diarization tasks because of the spontaneous conversation, speaker overlaps, and natural turn-taking. Moreover, the recordings are stored in both microphones and room-level audio. This provides clean and challenging conditions for the researchers. They can easily get access to the dataset through the AMI project website.

5.2. CALLHOME American English Corpus

The CALLHOME dataset consists of 120 telephone conversations of native English speakers, with durations of up to half an hour. The Linguistic Data Consortium (LDC) created American English Speech, which is made up of 120 unscripted 30-minute phone conversations between native English speakers. All the calls came from North America; of the 120 calls, 90 were made to different locations outside of the continent, while the other 30 were made inside.

Many participants referred to close friends or relatives. The software tools required to uncompress the voice data are included in this corpus, together with speech data files and documentation on their format and contents. An accompanying lexicon (LDC97L20) and corresponding transcripts and documentation (LDC97T14) are available separately.

5.3. CHIME-5

Twenty dinner parties make up the dataset, which was captured using six Kinect2 devices positioned in various locations. Four sample-synchronized microphones are arranged in a linear array on each Kinect device. Two of these parties are in the development set, sixteen are in the training set, and the remaining parties are in the evaluation set.

Since the same speech is captured on many channels and in numerous places, we choose a subset of the utterances at random for training. We call these audio subsets "100k" or "400k," depending on how many utterances were chosen. To produce relatively "clean" speech, each speaker is additionally recorded using a pair of wearable binaural microphones.

5.4. VoxConverse

A sizable audio-visual dataset created specifically for speaker detection is called VoxCeleb. Over 1,000 YouTube speech clips are included in the collection. The speakers are diverse in terms of their ages, occupations, and ethnicities. Speaker sex and identification metadata are included in the dataset. Researchers have enhanced the age information by comparing the speaker profiles with online sources, even though age information is not included in the original dataset.

5.5. NIST Rich Transcription Datasets

In April 2002, the first of the RT evaluation series, the RT 2002 Evaluation (RT-02), went into effect. Speech-To-Text (STT) tasks for CTS, meeting room speech, and broadcast news speech were all covered. One Multilingual Dialogue Evaluation (MDE) task, speaker diarization for broadcast news and CTS in English, was also included in the evaluation.

STT tasks were the focus of the RT Spring 2003 Evaluation (RT-03S), which was conducted in March and April of 2003. In addition to the many data domains, the STT tasks also extended beyond English to other languages. The meeting data domain was transferred to another RT evaluation, so it was not included in the RT-03S STT tasks, as it was in the RT-02 STT tasks.

5.6. DiPCo Dataset

The DiPCo dataset is a speech data corpus that mimics a "dinner party" situation that occurs in a typical home setting. Multiple groups of four Amazon employee volunteers were recorded naturally in English at a dining table to create the corpus. A single-channel close-talk microphone and five far-field seven-microphone array devices placed at various points throughout the recording room were used to record the participants. The audio recordings and human-labeled transcripts of ten sessions, each lasting between fifteen and forty-five minutes, are included in the dataset. The corpus was developed as a public research and benchmarking data set with the goal of advancing the field of distant speech processing and noise robustness.

5.7. AliMeeting

The AliMeeting Mandarin corpus contains recordings of actual meetings, including near-field speech captured by each participant's headset microphone and far-field speech captured by an 8-channel microphone array. According to the Multi-channel Multi-party Meeting Transcription (M2MeT) challenge arrangement, the dataset's 118.75 hours of speech data are split into 104.75 hours for training (Train), 4 hours for evaluation (Eval), and 10 hours for testing (Test). There are 212, 8, and 20 meeting sessions in the Train, Eval, and Test sets, respectively. Each session lasts between 15 and 30 minutes and involves two to four people in conversation. AliMeeting covers a wide range of topics related to real-world meetings, such as varied speaker overlap ratios, different meeting spaces, and variable numbers of attendees.

5.8. MIXER 6 Speech Dataset

The Linguistic Data Consortium (LDC) created Mixer 6 Speech, which includes 15,863 hours of audio recordings of 594 different native English speakers' conversations over the phone, transcripts, and interviews. LDC gathered this information between 2009 and 2010 as part of phase 6 of the Mixer project, which focused on native American English speakers in the Philadelphia region. LDC collected the speech data in this release at its Philadelphia Human Subjects Collection facilities. Like previous LDC telephone research (such as Switchboard-2 Phase III Audio-LDC2002S06), the telephone collecting procedure involved connecting recruited speakers via a robot operator to engage in informal talks for up to ten minutes.

5.9. DIHARD III/IV Dataset

In the discipline of speaker diarization, the DIHARD III and IV datasets are frequently employed as benchmarks that are intended to assess how well diarization systems perform in demanding real-world scenarios. These datasets comprise a wide range of audio recordings gathered from various sources, including broadcasts, meetings, interviews, and social media, with differing speaker counts, overlapping speech, and background noise levels. The purpose of DIHARD III and IV is to evaluate a system's performance under challenging situations, such as conversational overlap and brief utterances that are typical of natural speech. These datasets, which include standardized recordings along with annotations, enable researchers to make reliable model comparisons and advance the creation of reliable diarization techniques that function well in a variety of acoustic settings.

5.10. Hearing Aid-Specific Corpora Dataset

Specialized datasets designed to investigate and enhance speech processing in contexts relevant to hearing aid users are known as hearing aid-specific corpora in speaker diarization. Typically, these corpora are recordings made in real-world listening environments with background noise and reverberation, with numerous speakers speaking at once, including busy rooms, meetings, or outdoor areas. These datasets are primarily used to test and create diarization techniques that can reliably distinguish and separate speakers in difficult acoustic environments, assisting hearing aids in improving speech intelligibility and then the user experience. These corpora facilitate research that directly supports assistance with listening by concentrating on hearing aid settings. Below is a summary table for these datasets, which is presented in Table 8.

Table 8. Summary table of datasets for speaker diarization

Dataset Name	Type	Size (Audio)	Focus Area
AMI Meeting Corpus	Audio/Multichannel	100+ hours	Multi-speaker meetings, overlapping speech
CALLHOME American English	Audio (Telephone)	120 Conversations	Two-speaker telephone calls with informal dialogue
CHiME-5	Audio/Multichannel	Real dinner parties (multiple sessions)	Distant mic speech in home environments
VoxConverse	Audio	Broadcast clips (hundreds)	Multi-speaker real-world audio with background noise
NIST RT Datasets	Audio	Varies by year	Telephone, broadcast, and meeting speech
DiPCo	Audio	Multi-party recordings	Personal, casual conversation with overlaps
AliMeeting	Audio/Multichannel	50+ hours	Mandarin meetings, near/far-field audio
MIXER 6	Audio	1,800 hours (interviews, calls)	Speaker diarization, clustering, and ASR tasks
DIHARD III/IV	Audio/Multichannel	About 5–6 hours in DIHARD III, whereas DIHARD IV adds roughly 10–12 hours	Managing short utterances, overlapping speech, many speakers, and a variety of real-world audio situations.
Hearing aid-specific corpora	Audio	1-5 hours	A noisy, multi-speaker real-world setting

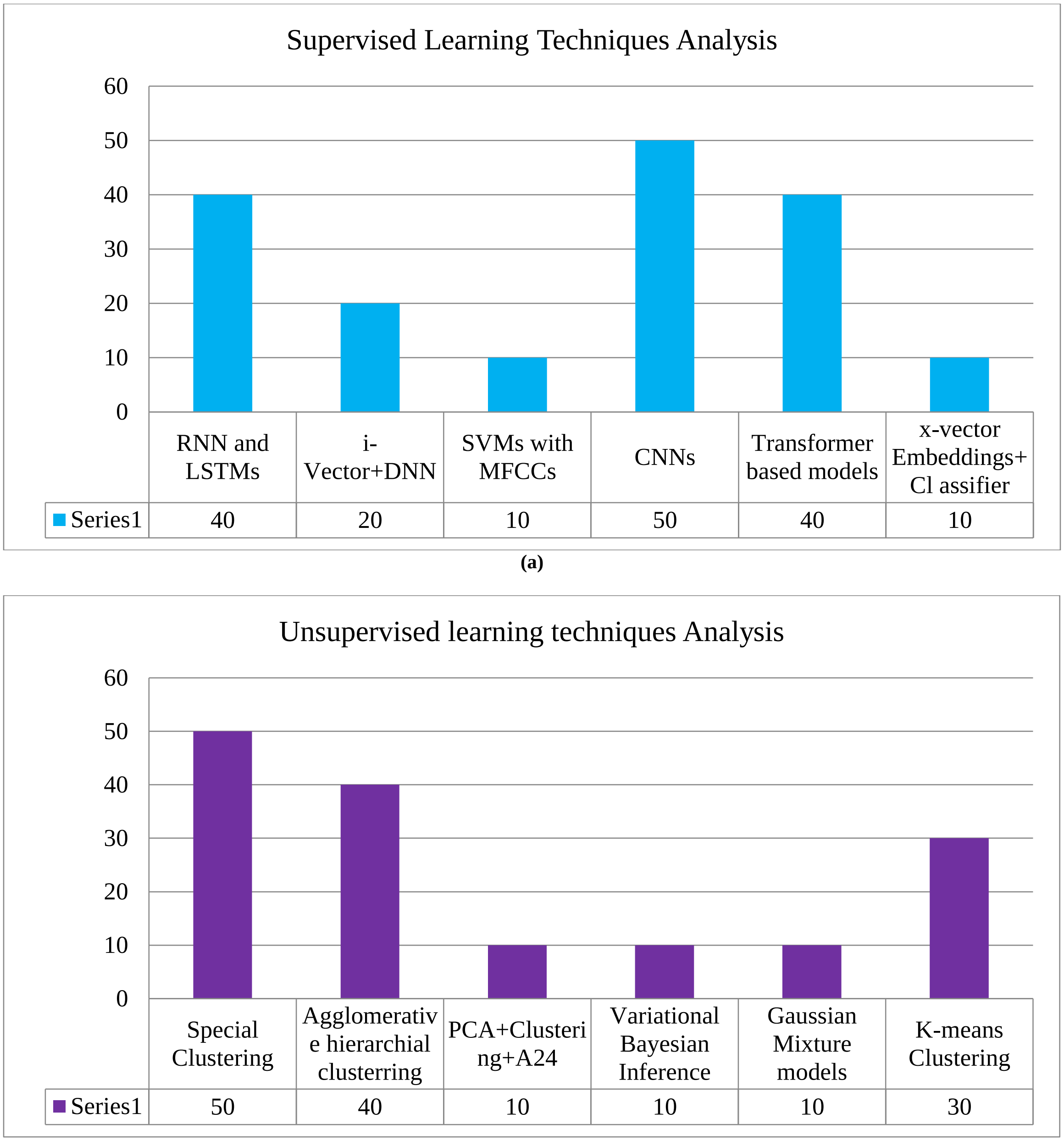
Each dataset offers unique attributes, from real-world variations and technical complexities to specific challenges that are essential for developing accurate and robust Speaker diarization.

6. Analysis and Discussion

This section presents a comprehensive analysis and discussion of speaker diarization techniques, specifically focusing on methodologies employed in the detection and classification of speaker diarization using Supervised Learning Techniques, Unsupervised Learning Techniques, and Hybrid Methods. A review of various research papers has been conducted to categorize and analyze their approaches, publication years, and performance evaluation metrics. This analysis provides a systematic overview of the current state of research in speaker diarization, highlighting key trends, challenges, and potential areas for further development.

6.1. Analysis Based on Techniques

This section examines the various research techniques used in speech diarization, focusing on approaches using Supervised Learning Techniques and Unsupervised Learning Techniques. A thorough review of the literature reveals a variety of techniques, which can be categorized into several key groups. Figures 5(a) and 5 b) illustrate the distribution of research techniques used in speaker diarization.



(a)

(b)

Fig. 5 Supervised and Unsupervised learning techniques for speaker diarization

In supervised speaker diarization, the CNN is used to extract the features of the speaker from spectrograms; meanwhile, the RNN and LSTM handle the speech flow efficiently. The lengthy audio sequences are handled by using the transformers without any challenges. Moreover, some methods rely on i-vectors with DNNs or x-vectors with simple classifiers to accurately identify speakers. Finally, the SVM with MFCC features is also used for the simpler tasks.

In terms of speaker diarization, unsupervised learning plays an important role in finding "who spoke when" without the use of any training data. When compared to the various methods, spectral clustering is the widely used technique because of its effectiveness. Agglomerative Hierarchical Clustering is also used for merging the smaller clusters into larger ones based on the same features. Then, k-means clustering divides the data into a fixed number of clusters. More methods, such as Variational Bayesian Inference and GMM, offer many ways for designing the speaker effectively. In addition, PCA combined with clustering decreases the feature dimensionality before using the clustering algorithms.

6.2. Publication Year Analysis

The analysis of publication trends in speaker diarization reveals key shifts in research focus over recent years. The year-by-year publication trends in speaker diarization used in this review are presented in Table 9. Specifically, 20 papers from 2023, 12 from 2024, and 60 from 2025 were included and examined. The inclusion of these works demonstrates the importance of recent developments as well as the expansion of study in this field. While the reduction in 2024 indicates a brief pause, perhaps due to improvements in current methods or changes in priorities in speech processing research, the moderate count in 2023 shows the stable foundation of studies that led to diarization procedures.

The significant rise in 2025, however, indicates a renewed interest and inventiveness in this area, which is likely being fueled by advances in deep learning, integration with large language models, and an increasing demand for applications such as intelligent virtual assistants, conversational AI, and meeting transcription. This review provides thorough coverage of both classic works and the most recent cutting-edge contributions in speaker diarization, encompassing these year-by-year publications.

Figure 6 illustrates a steady and increasing interest in speaker diarization overall, which is indicative of the growing significance of reliable, automated systems in applications such as meeting transcription, broadcast monitoring, and conversational analysis.

Table 9. Year-wise Publication Trends in Speaker Diarization

Year	**2023**	**2024**	**2025**
Number of publications	20	12	60

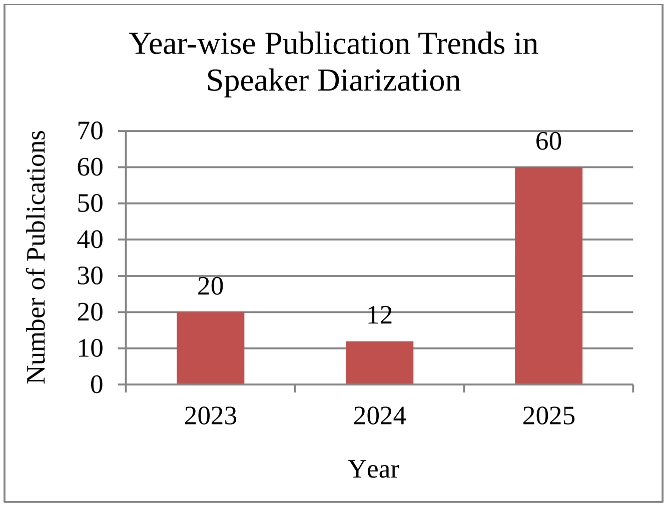


Fig. 6 Year-wise publication trends in speaker diarization (2023-2025)

6.3. Toolset Analysis for Speaker Diarization

This section examines the software tools commonly employed in speaker diarization and speech processing research, leveraging advanced machine learning and deep learning techniques. Figure 7 presents an analysis of the tools utilized in various research papers on this topic. The analysis reveals that the following software tools are used in this speaker diarization study,

6.3.1. Python

The most widely used software tool, featured in 52% of the research papers in this review. Python is widely adopted due to its vast ecosystem of libraries and frameworks, such as PyAnnote-audio, SpeechBrain, TensorFlow, and Librosa, which facilitate speech recognition, feature extraction, and deep learning.

6.3.2. Pyannote-Audio

This tool, which is mainly used for speaker diarization and offers pre-trained pipelines and diarization task metrics, is used in 21% of the research papers in this review.

6.3.3. Speech Brain

This tool is used in 15% of the studies in this review, facilitates end-to-end speech processing, and is used for tasks like speaker categorization, segmentation, and embedding.

6.3.4. Kaldi

Utilized in 7% of studies in this review for speech signal processing and feature extraction, mainly in existing machine learning diarization methods. It is best in MFCC extraction, i-vector training, and cluster algorithms.

6.3.5. OpenCV

A tool used in 6% of research in this review for audio preprocessing and feature extraction, which includes MFCCs, spectrograms, and silence removal, all of which are required for segmentation and speaker verification.

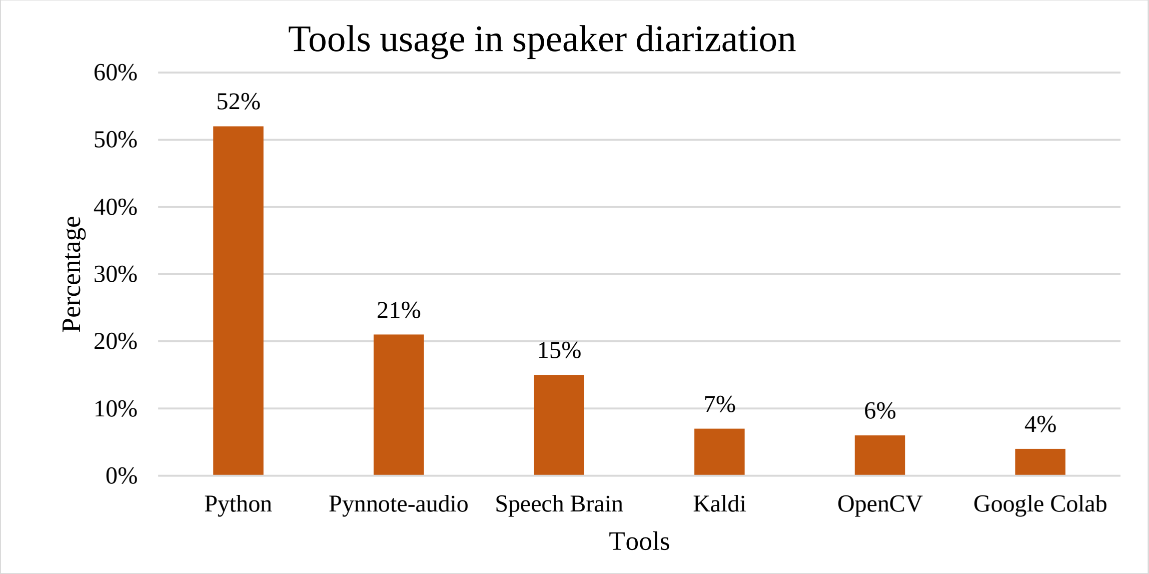


Fig. 7 Percentage usage of the tools in speaker diarization

6.3.6. Google Colab

Found in 4% of studies in this review as a cloud-based platform for running speaker diarization models on GPUs, providing accessibility and computational power for researchers working on audio datasets at a large scale.

This toolset analysis highlights that Python, with its robust ecosystem, is the leading software tool used in the majority of speaker diarization research. Researchers favor Python for its extensive libraries and frameworks, making it a highly versatile and efficient choice for speaker diarization tasks.

TensorFlow/Keras and PyTorch, as prominent deep learning frameworks, are crucial for model development, while tools like Kaldi and LIUM spk diarization play significant roles in audio processing and speaker identification. The use of cloud platforms, such as Google Colab, also demonstrates the increasing demand for accessible, high-performance computing environments.

6.4. Performance Analysis

This section presents a comprehensive analysis and discussion of speaker diarization techniques, specifically focusing on methodologies employed in the detection and classification of speaker diarization using Supervised Learning Techniques, Unsupervised Learning Techniques, and Hybrid Methods. Since this is a review paper, no new experiments were conducted; however, compiled and analyzed performance metrics from reviewed studies to highlight comparative trends. This paper is a review article that does not present novel experimental results, but rather captures and discusses the performance evaluation metrics used in the surveyed research papers. The metrics are important for benchmarking deep learning model performance in different surveillance tasks.

6.4.1. Key Performance Metrics

Diarization Error Rate

The common statistic used to assess diarization systems is called DER. It calculates the overall percentage of a recording where the output of the system differs from the reference annotation.

Jaccard Error Rate

The overlap among system as well as reference speaker segments is compared using JER, an alternative to DER. It gauges how closely anticipated speaker regions match actual speaker regions using the Jaccard similarity index.

Equal Error Rate

The False Acceptance Rate (FAR) as well as the False Rejection Rate (FRR) are identical at the Equal Error Rate (EER) point on the Detection Error Tradeoff (DET) curve.

Word Error Rate

Since the WER is one of the most widely used metrics for assessing the effectiveness of ASR and machine translation, it was employed to gauge the accuracy of automatic transcriptions.

Minimum Detection Cost Function

It is based on the Minimum Detection Cost Function (DCF), which weighs the prior probability of the target speaker and the probability of misses (false rejections) and false alarms (false acceptances) according to their application-dependent costs.

Accuracy

Accuracy is computed as the ratio of precise forecasts to all forecasts.

Precision, Recall, and F1-Score

Precision is the proportion of true positive detections to all the positive predictions. Recall is the ratio of true positive detections to all the actual positives. F1-score is the harmonic mean of precision and recall and gives a trade-off between the two.

6.5. Performance Metrics Analysis

This section provides a comprehensive evaluation of speaker diarization techniques using various performance metrics. This analysis offers insights into the evaluation practices in speaker diarization research.

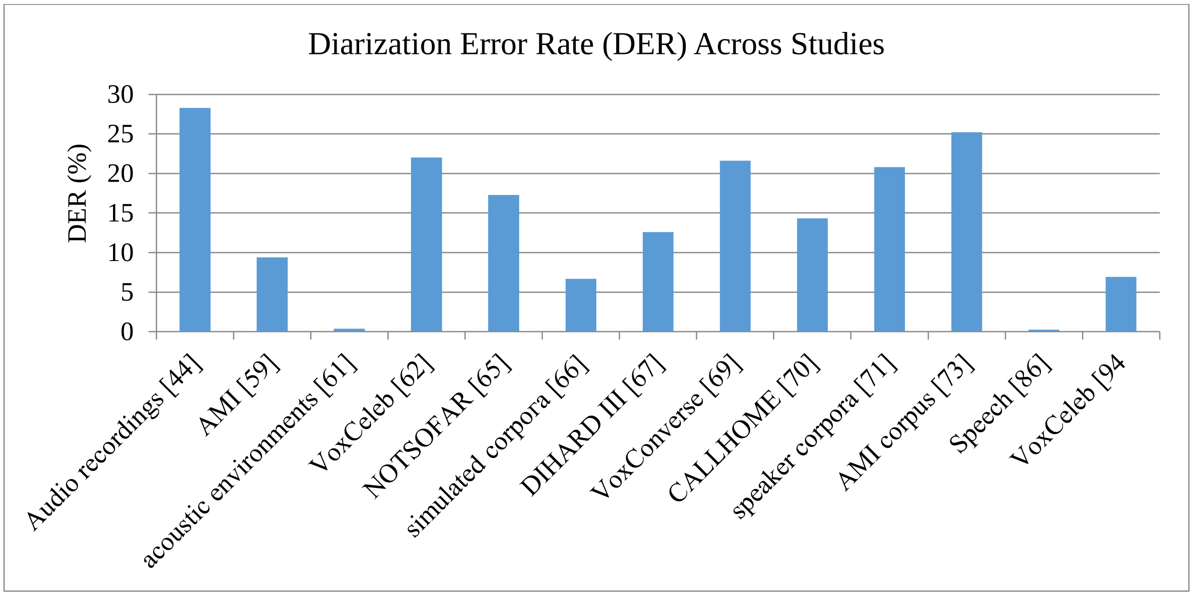
Understanding the most used metrics can assist researchers in designing and assessing their models more effectively. The overall performance analysis of the existing speaker diarization techniques is detailed in Table 10.

Table 10. Overall performance analysis of the existing speaker diarization techniques

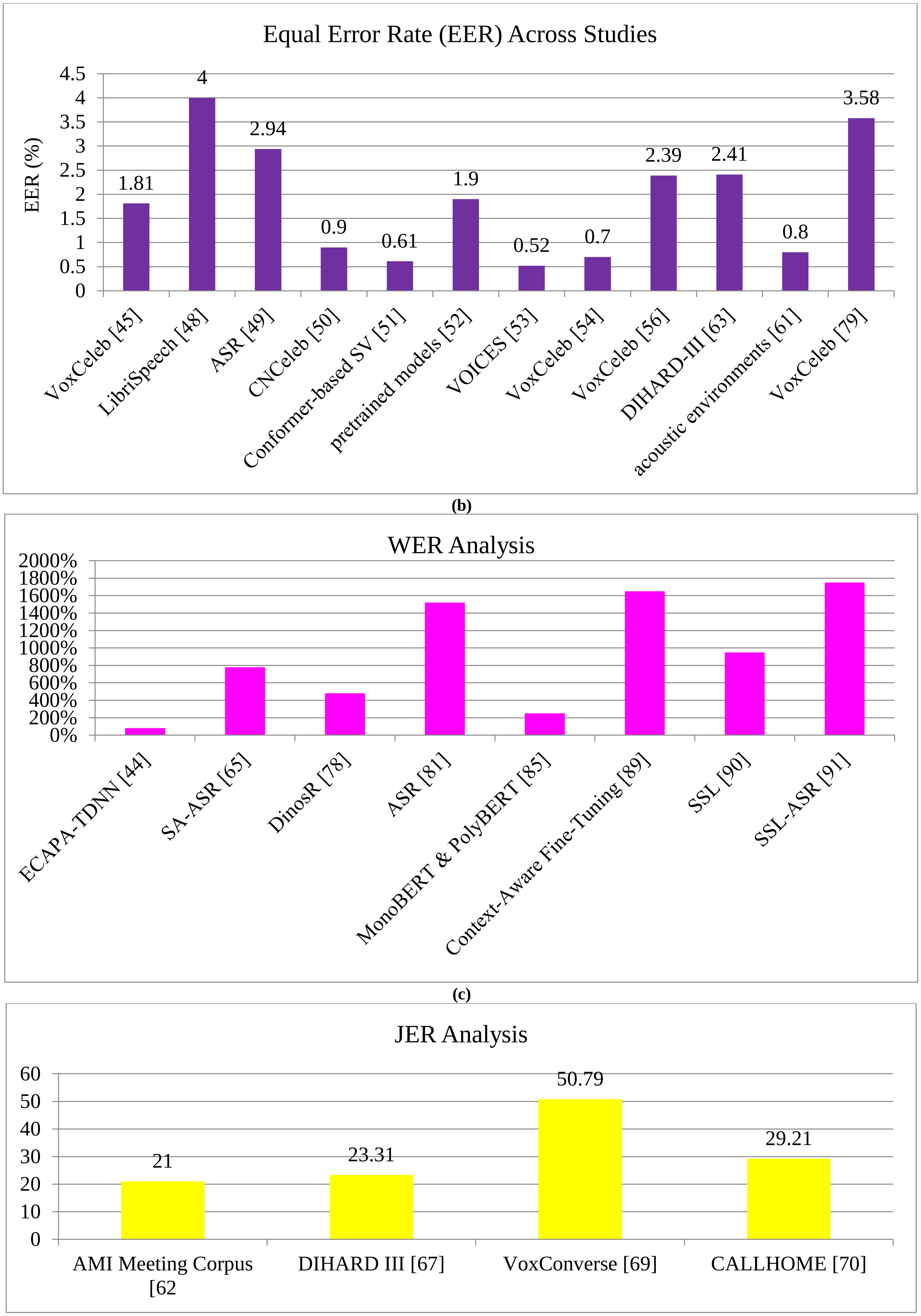
Author	WER (%)	DER (%)	EER (%)	JER (%)	Accuracy (%)	Min DCF	F1-score (%)	Recall (%)	Precision (%)
Shankar et.al., (2025) [44]	0.6	28.3	-	-	-	-	-	-	-
Chen et.al., (2025) [45]	-	-	1.81	-	-	0.17	-	-	-
Zheng et.al., (2025) [48]	-	-	4.00	-	6.20	-	-	-	-
Cai, (2023) [49]	-	-	2.94	-	65.26	0.508	-	-	-
Gu et.al., (2025) [50]	-	-	0.90	-	-	0.113	-	-	-
Wang et.al., (2023) [51]	-	-	0.61	-	-	-	-	-	-
Kim et.al., (2024) [52]	-	-	1.90	-	-	1.92	-	-	-
Han et.al. (2025) [53]	-	-	0.92	-	-	0.0653	-	-	-
Liu et.al., (2023) [54]	-	-	0.7	-	-	-	-	-	-
Cai et.al., (2025) [56]	-	-	2.39	-	79.26	0.163	-	-	-
Singh et al., (2023) [59]	-	9.4	-	-	-	-	-	-	-
Nareaho, (2023) [61]	-	0.39	0.8	-	-	-	-	-	-
Pande et.al., (2025) [62]	-	22	-	21	-	-	-	-	-
Raghav et.al., (2025) [63]	-	-	2.41	-	-	-	-	-	-
Kalda et.al., (2025) [65]	7.8	17.3	-	-	-	-	-	-	-
Landini, (2024) [66]	-	6.69	-	-	-	-	-	52.9	62.5
Wang and Li (2023) [67]	-	12.61	-	23.31	-	-	-	-	-
Kwon et.al., (2023) [69]	-	21.6	-	50.79	-	-	-	-	-
Cheng et.al., (2023) [70]	-	14.31	-	29.21	-	-	-	-	-
Huang et.al., (2023) [71]	-	20.8	-	-	-	-	-	-	-

Cornell et.al. (2024) [73]	-	25.2	-	-	-	-	-	-	-
Jafarzadeh et.al. (2024) [76]	-	-	-	-	97.67	-	-	-	-
Liu et.al., (2023) [78]	4.71	-	-	-	96.69	-	88.83	-	-
Han et.al. (2023) [79]	-	-	3.585	-	-	0.353	-	-	-
Veliche and Fung (2023) [81]	15.25	-	-	-	-	-	-	-	-
Nazir et.al., (2023) [82]	-	-	-	-	97	-	-	94	95
Ma et.al., (2023) [85]	2.5	-	-	-	96.66	-	-	-	-
Revathy and Kumar (2025) [86]	-	0.251	-	-	-	-	91	-	-
Shon et.al., (2023) [89]	16.73	-	-	-	-	-	-	-	-
Huang et.al., (2024) [90]	9.6	-	-	-	-	-	-	-	-
Yue et.al., (2024) [91]	17.5	-	-	-	-	-	-	-	-
Ahmed et.al., (2025) [94]	-	6.93	-	-	96.42	-	94.99	95.43	94.56
Bhuyan et.al. (2024) [95]	-	-	-	-	86.6	-	82	-	-

Experimental findings frequently contrast speech diarization performance with that of deep learning and clustering techniques. This comparison helps to emphasize the advantages and possible improvements that diarization brings to speaker attribution and conversation analysis. Figure 8 evaluates the current performance of diarization in terms of WER, DER, JER, Min DCF, Accuracy, memory consumption, recall, precision, and F1-score.



(a)



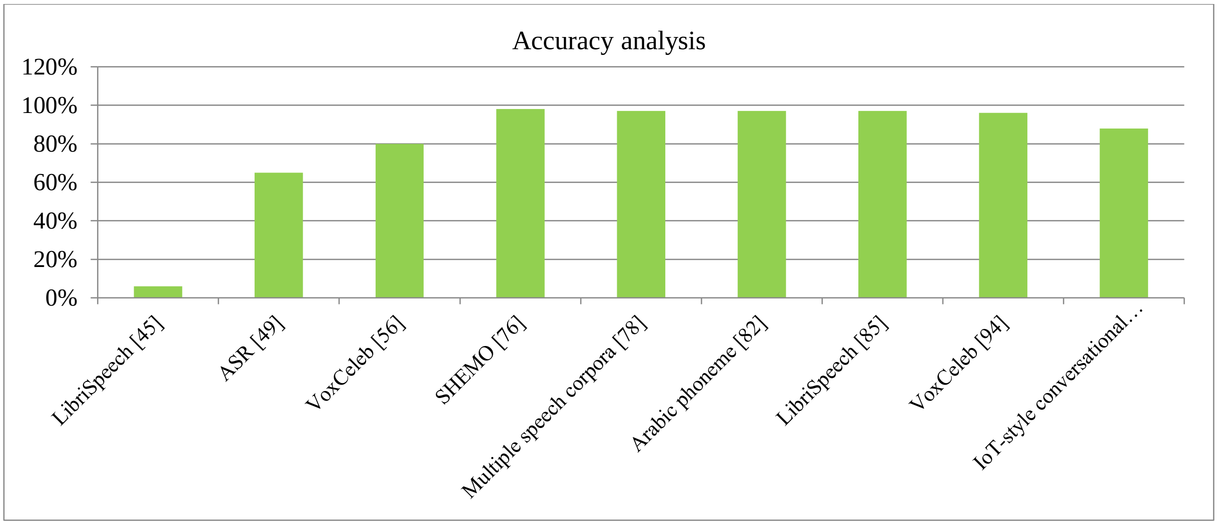
(b)

(c)

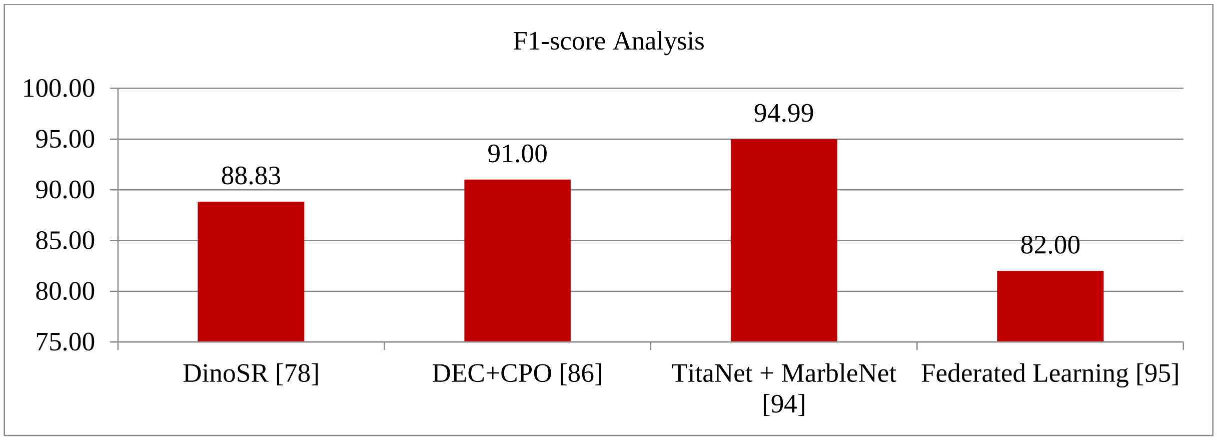
(d)

Fig. 8 DER, EER, WER, JER analysis

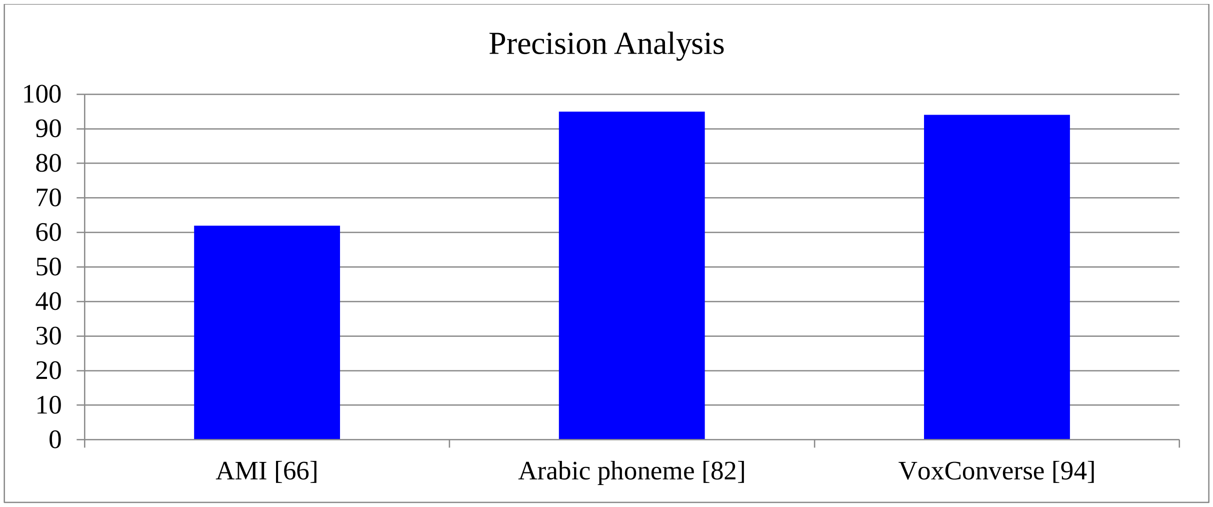
The DER performance of the researched papers is examined. The lowest DER was achieved by Revathy and Kumar (2025) [86] and Nareaho (2023) [61] models, which obtained a reduced DER of 0.25% and 0.39%. The highest DER was achieved by Shankar et.al. (2025) [44] and Cornell et.al. (2024) [73] models, which obtained 28.30% and 25.20% respectively. The moderate DER was achieved by Ahmed et.al. (2025) [94], Landini (2024) [66], Wang and Li (2023) [67], and Cheng et.al. (2023) [70] models obtained 6.93%, 6.69%, 12.61%, and 14.31% showing balanced but improvable results. The DER performance of the existing techniques is detailed in Figure 8 (a). The EER performance of the researched papers is examined. The lowest EER was achieved by Wang et.al. (2023) [51], Liu et.al. (2023) [54], Nareaho (2023) [61], and Gu et.al. (2025) [50] models, which obtained 0.61%, 0.70%, 0.80%, and 0.90%. The highest EER was achieved by Zheng et.al. (2025) [48], Han et.al. (2023) [79], Cai (2023) [49], Raghav et.al. (2025) [63], and Cai et.al. (2025) [56] models, which obtained 4%, 3.58%,2.94%, 2.41%, and 2.39%. The moderate EER was achieved by Chen et.al. (2025) [45], Kim et.al. (2024) [52], Cai et.al. (2025) [56], and Raghav et.al. (2025) [63] models obtained 1.81%, 1.90%, 2.39% and 2.41%. The EER performance of the existing techniques is detailed in Figure 8 (b). The WER performance of the researched papers is examined. The lowest WER was achieved by Shankar et.al. (2025) [44], and Ma et.al. (2023) [85] models obtained 0.6% and 2.5% respectively. The highest WER was achieved by Yue et.al. (2024) [91], Shon et.al. (2023) [89], and Veliche and Fung (2023) [81] models, which obtained 17.5%, 16.73%, and 15.25% respectively. The moderate WER was achieved by Kalda et.al. (2025) [65], Liu et.al. (2023) [78], and Huang et.al. (2024) [90] models, which obtained 7.8%, 4.71%, and 9.6% respectively. The WER performance of the existing techniques is detailed in Figure 8 (c). The JER performance of the researched papers is examined. The lowest JER was achieved by Pande et al. (2025) [62], which obtained 21%. The highest JER was achieved by the Kwon et.al. (2023) [69] model, which obtained a 50.79%. The moderate JER was achieved by Wang and Li (2023) [67], and Cheng et.al. (2023) [70] models obtained 23.31% and 29.21%. The JER performance of the existing techniques is detailed in Figure 8 (d).



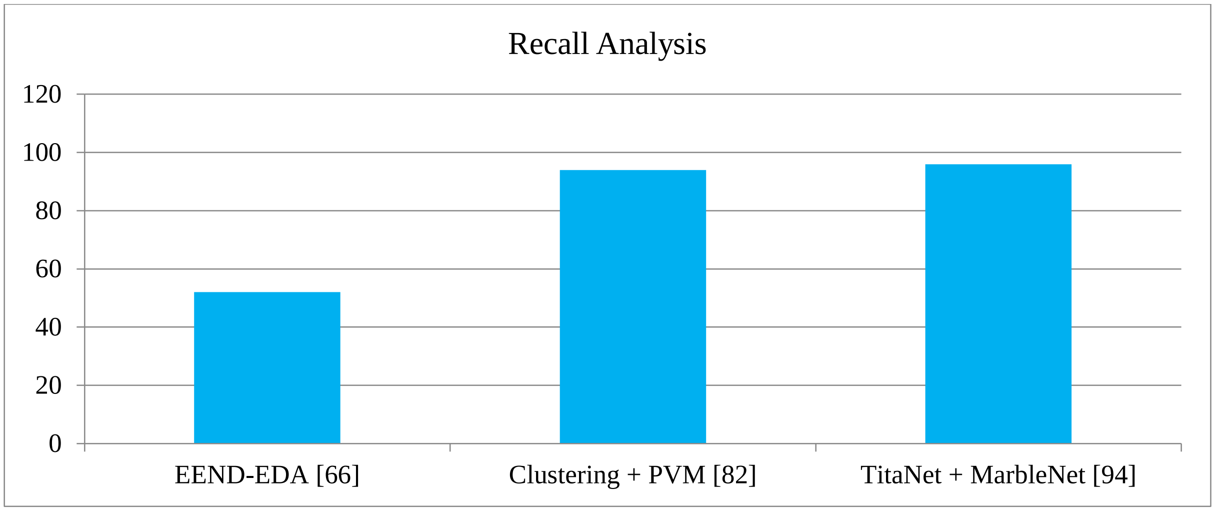
(a)



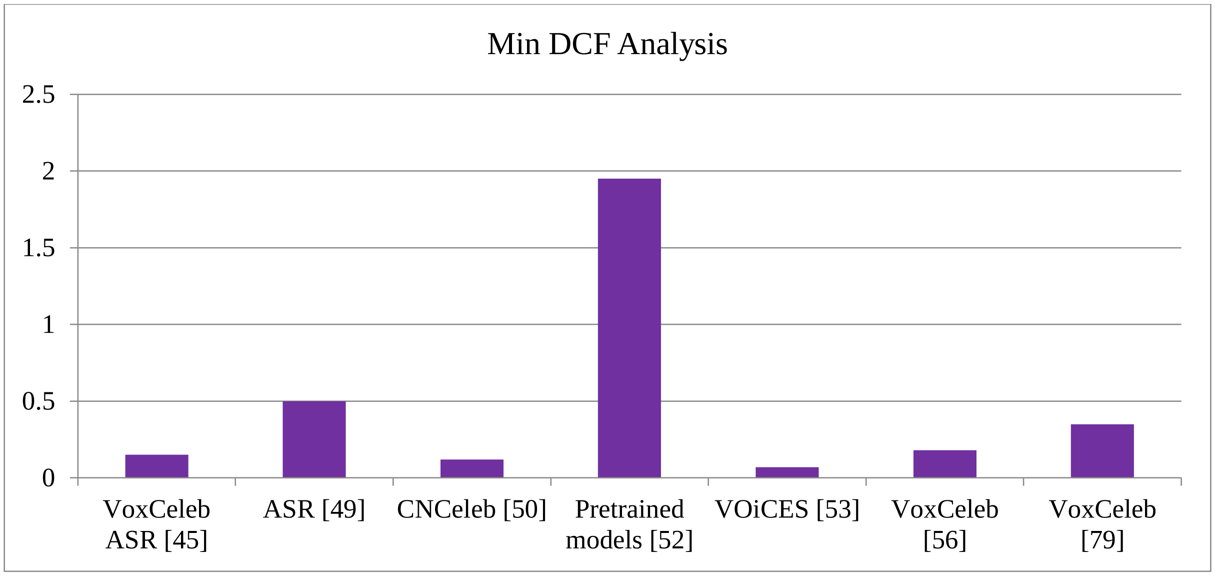
(b)



(c)



(d)



(e)

Fig. 9 Accuracy, F1-score, precision, recall, and Min DCF analysis

The accuracy performance of the researched papers is examined. The lowest accuracy was achieved by the Zheng et.al. (2025) [48] model, which obtained 6.20%. The highest accuracy was achieved by Jafarzadeh et.al. (2024) [76], Liu et.al. (2023) [78], Nazir et.al. (2023) [82], Ma et.al. (2023) [85], and Ahmed et.al. (2025) [94] models, which obtained 97.67%, 96.69%,97%, 96.66%, and 96.42%. The moderate accuracy was achieved by Cai (2023) [49], Cai et.al. (2025) [56], and Bhuyan et.al. (2024) [95] models, which obtained 65.26%, 79.26%, and 86.6% respectively. The accuracy performance of the existing techniques is detailed in Figure 9 (a). The F1-score performance of the researched papers is examined. The lowest F1-score was achieved by Bhuyan et.al. (2024) [95], and Liu et.al. (2023) [78] models obtained 82 % and 88.83% respectively. The highest F1-score was achieved by the Ahmed et al. (2025) [94] model, which obtained 94.99%. The moderate F1-score was achieved by Revathy and Kumar (2025) [86] models, which obtained 91%. The F1-score performance of the existing techniques is detailed in Figure 9 (b). The precision performance of the researched papers is examined. The lowest precision was achieved by Landini (2024) [66] models, which obtained 62.5%. The highest precision was achieved by the Nazir et.al. (2023) [82] model, which obtained 95%. The moderate F1-score was achieved by Ahmed et.al. (2025) [94], who obtained a 94.56%. The F1-score performance of the existing techniques is detailed in Figure 9 (c). The precision performance of the researched papers is examined. The lowest recall was achieved by Landini (2024) [66] models, which obtained 52.9%. The highest recall was achieved by the Ahmed et.al. (2025) [94] model, which obtained 95.43%. The moderate recall was achieved by [82] models, which obtained a 94%. The recall performance of the existing techniques is detailed in Figure 9 (d). The minimum DCF performance of the researched papers is examined. The lowest min DCF was achieved by Han et.al. (2025) [53], and Gu et.al. (2025) [50] models obtained a reduced Min DCF of 0.0653% and 0.113%. The highest Min DCF was achieved by the Kim et.al. (2024) [52] model, which obtained a 1.92%. The moderate min DCF was achieved by Chen et.al. (2025) [45], Cai (2023) [49], Cai et.al. (2025) [56], and Han et.al. (2023) [79] models, which obtained 0.17%, 0.508%, 0.163%, and 0.353%. The Min DCF performance of the existing techniques is detailed in Figure 9(e).

7. Challenges and Future Directions

The major challenge in the speaker diarization is the overlapping of the speech, which happens when two or more speakers talk at the same time. This overlapping is difficult to separate using the traditional models. The next challenge is detecting the speaker who spoke for a shorter time, such as 2 or 3 words. Next, the system gets confused with the persons with the same voice and mixes them up. Poor audio quality, echo in the room, and background noises can affect the system's ability to recognize the voices. If people speak various languages, it is difficult to define them correctly. Another level of challenge is the domain mismatch. Systems trained on phone calls may not work well on meeting recordings. Some meetings require speaker diarization instantly to record a live meeting. High-quality training data with proper labels is required; creating the data takes a lot of time and effort.

In the future, speaker diarization makes the system better at handling real-time situations, when two people speak at the same time or speak for a very short time. The system deals with background noise, different accents, and changes in recording conditions, which is very crucial. Future work may find more efficient tools that find the sound more effectively, as many current systems still struggle. Moreover, there is a growing effort to make these systems lighter and faster to use in mobile phones or smart devices. In addition, combining speaker diarization with other tasks, such as speech recognition or emotion detection, could make systems more useful in applications like virtual meetings, customer service, or healthcare. Overall, the goal is to make speaker diarization systems more accurate, faster, and smarter for everyday use.

8. Conclusion

Speaker diarization is now a crucial part of contemporary speech processing systems, allowing machines to recognize "who spoke when" in a wide range of practical applications, including voice-based assistants, meeting transcription, subtitle creation, and customer service analytics. Deep learning-driven frameworks that tackle some of the long-standing issues, such as overlapping speech, background noise, and speaker similarity, have quickly replaced traditional clustering-based approaches in recent years. This assessment categorized 95 research publications into two groups: supervised and unsupervised learning-based approaches. The papers were published between 2023 and 2025. By training end-to-end neural architectures, supervised techniques—which often utilize extensive labeled datasets—have proven to be highly accurate and resilient. On the other hand, by utilizing self-supervised learning, clustering, and representation learning techniques, unsupervised approaches have demonstrated significant promise in situations when labeled data is limited. When combined, these methods enhance the system's flexibility across various acoustic settings, languages, and domains. The integration of speaker diarization with speech recognition systems is another important topic covered in this review. Effective diarization improves downstream tasks like SA sentiment analysis, conversation summary, and multimodal interaction systems in addition to making automatic transcripts easier to read. Building real-time, speaker-aware AI assistants and transcription tools that can manage the intricacies of realistic conversations has become easier thanks to this connection. It is clear from the reviewed literature that considerable advancements have been made in the areas of diarization accuracy, robustness, and scalability. However, several obstacles still exist. In multi-speaker settings where natural breaks happen, overlapping speech identification remains a significant barrier. Likewise, it remains challenging to distinguish between speakers who share very similar vocal traits. Models that strike a compromise between accuracy, low latency, and computing efficiency are also necessary for real-time diarization. Furthermore, domain adaptation continues to be a problem since models developed on a single dataset frequently perform poorly when used in real-world settings with disparate backgrounds, languages, and accents. The review identifies several encouraging trends and prospects, despite these obstacles. New directions in diarization research are being made possible by transformer-based models, self-supervised representation learning, and multimodal fusion (e.g., merging audio and video cues). Similarly, larger adoption in edge devices and sensitive applications is being made possible by lightweight model architectures and privacy-preserving approaches. Enhancing overlapping speech handling, domain adaptation, and real-time low-latency processing will be the key goals of future speaker diarization research. Accuracy in complicated contexts will be further improved by the multimodal integration of audio-visual cues and context-aware models. Furthermore, defined standards and privacy-preserving techniques will promote scalable and moral implementation in practical applications.

References

[1] Gongfan Chen et al., “Meet2Mitigate: An LLM-Powered Framework for Real-Time Issue Identification and Mitigation from Construction Meeting Discourse,” *Advanced Engineering Informatics*, vol. 64, pp. 1-42, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Józef Kotus, and Grzegorz Szwoch, “Separation of Simultaneous Speakers with Acoustic Vector Sensor,” *Sensors*, vol. 25, no. 5, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Linzhao Jia et al., “Enhanced Speaker-Turn Aware Hierarchical Model for Automated Classroom Dialogue Act Classification,” *Expert Systems with Applications*, vol. 296, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[4] Giulio Bertamini et al., “Automated Segmentation of Child-Clinician Speech in Naturalistic Clinical Contexts,” *Research in Developmental Disabilities*, vol. 157, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Dmitrii Korzh et al., “Certification of Speaker Recognition Models to Additive Perturbations,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 17, pp. 17947-17956, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rana Zeeshan, John Bogue, and Mamoona Naveed Asghar, “Relative Applicability of Diverse Automatic Speech Recognition Platforms for Transcription of Psychiatric Treatment Sessions,” *IEEE Access*, vol. 13, pp. 117343-117354, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ryan Duke, and Alex Doboli, “Dialogic: A Multi-Modal Framework for Automated Team Behavior Modeling Based on Speech Acquisition,” *Multimodal Technologies and Interaction*, vol. 9, no. 3, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Nassim Asbai, Hadjer Bounazou, and Sihem Zitouni, “A Novel Approach to Deriving Adaboost Classifier Weights using Squared Loss Function for Overlapping Speech Detection,” *Multimedia Tools and Applications*, vol. 84, pp. 38545-38572, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yueran Pan “Assessing the Expressive Language Levels of Autistic Children in Home Intervention,” *IEEE Transactions on Computational Social Systems*, vol. 12, no. 5, pp. 3647-3659, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Eric Ettore et al., “Childhood Trauma Affects Speech and Language Measures in Patients with Major Depressive Disorder during Clinical Interviews,” *Journal of Affective Disorders*, vol. 388, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Joonas Kalda et al., “Enhancing Pixit: Robust Methods for Real-World Speech Separation and Applications,” *SSRN*, pp. 1-36, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mitchell A. Klusty et al., “Toward Automated Clinical Transcriptions,” *AMIA Summits on Translational Science Proceedings*, 2025. [Google Scholar] [Publisher Link]

[13] Hasan Almgotir Kadhim, Lok Woo, and Satnam Dlay, “Novel Algorithm for Speech Segregation by Optimized K-Means of Statistical Properties of Clustered Features,” *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, China, pp. 286-291, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kalidindi Lakshmi Divya et al., *Query-Facilitating Multidomain Summarization with BART Model and Accessible Query Interface*, 1st ed., Hybrid and Advanced Technologies, CRC Press, pp. 321-326, 2025. [Google Scholar] [Publisher Link]

[15] Jule Pohlhausen, Francesco Nespoli, and Jörg Bitzer, “Towards Privacy-Preserving Conversation Analysis in Everyday Life: Exploring the Privacy-Utility Trade-Off,” *Computer Speech & Language*, vol. 95, pp. 1-15, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[16] Igor Abramovski et al., “Summary of the NOTSOFAR-1 Challenge: Highlights and Learnings,” *Computer Speech & Language*, vol. 93, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ian Kwok et al., “New Frontiers in Artificial Intelligence: A Multimodal Communication Model,” *Journal of Pain and Symptom Management*, vol. 69, no. 5, pp. e454-e455, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Naela Fauzul Muna, and Mukhammad Andri Setiawan, “Automated Framework for Communication Development in Autism Spectrum Disorder Using Whisper ASR and GPT-4o LLM,” *JTP-Jurnal of Educational Technology*, vol. 27, no. 1, pp. 137-149, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Sadhana Singh, Lotika Singh, and Nandita Satsangee, “Automated Assessment of Classroom Interaction Based on Verbal Dynamics: A Deep Learning Approach,” *SN Computer Science*, vol. 6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Liqaa Fadil, Alia K. Abdul Hassan, and Hiba B. Alwan, “Employing Chroma-Gram Techniques for Audio Source Separation in Human-Computer Interaction,” *AIP Conference Proceedings*, vol. 3264, no. 1, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tong Li et al., “Trimodal-Persona: Leveraging Text, Audio, and Video for Big Five Personality Scoring In Psychological Interviews,” *Research Gate*, pp. 1-8, 2025.[Google Scholar]

[22] S. Mhammad, and S. Molodyakov, “Developing and Analyzing an Algorithm for Separate Speech Recording of Multiple Speakers,” *International Journal of Open Information Technologies*, vol. 13, no. 5, pp. 41-48, 2025. [Google Scholar] [Publisher Link]

[23] Thilo von Neumann et al., “Word Error Rate Definitions and Algorithms for Long-Form Multi-talker Speech Recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3174-3188, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[24] Pinyan Li et al., “Enhancing Speaker Recognition with CRET Model: A Fusion of CONV2D, RESNET and ECAPA-TDNN,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Haoyuan Wang et al., “An Evaluation Framework for Ambient Digital Scribing Tools in Clinical Applications,” *NPJ Digital Medicine*, vol. 8, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Yuta Hirano et al., “Toward Fast Meeting Transcription: NAIST System for CHiME-8 NOTSOFAR-1 Task and Its Analysis,” *Computer Speech & Language*, vol. 95, pp. 1-13, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[27] Srikanth Madikeri et al., “Autocrime-Open Multimodal Platform for Combating Organized Crime,” *Forensic Science International: Digital Investigation*, vol. 54, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Marisha Speights et al., “Transforming Child Speech Data into Clinical-Grade Artificial Intelligence Pipelines for Speech-Language Impairment Detection,” *The Journal of the Acoustical Society of America*, vol. 157, no. 4, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Kengatharaiyer Sarveswaran et al., “A Brief Overview of the First Workshop on Challenges in Processing South Asian Languages (chipsal),” *Proceedings of the First Workshop on Challenges in Processing South Asian Languages*, Abu Dhabi, UAE, pp. 1-8, 2025. [Google Scholar] [Publisher Link]

[30] Ho Seok Ahnet al., “Social Human-Robot Interaction of Human-care Service Robots,” *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, Chicago IL USA, pp. 385-386, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[31] Zhenzhen Weng et al., “Artificial Intelligence–Powered 3D Analysis of Video-Based Caregiver-Child Interactions,” *Science Advances*, vol. 11, no. 8, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Padma Jyothi Uppalapati, Madhavi Dabbiru, and Venkata Rao Kasukurthi, “AI-Driven Mock Interview Assessment: Leveraging Generative Language Models for Automated Evaluation,” *International Journal of Machine Learning and Cybernetics*, vol. 16, pp. 10057-10079, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] Hatem Zehir, Toufik Hafs, and Sara Daas, “Unifying Heartbeats and Vocal Waves: An Approach to Multimodal Biometric Identification at the Score Level,” *Arabian Journal for Science and Engineering*, vol. 50, pp. 19535-19554, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Farhan Samir et al., “A Comparative Approach for Auditing Multilingual Phonetic Transcript Archives,” *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 595-612, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[35] Lauren Harrington et al., “Variability in Performance Across four Generations of Automatic Speaker Recognition Systems,” *IAFPA 2025*, Rotterdam, The Netherlands, pp. 3993-3997, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Jose Luis Medellin-Garibay, and J.C. Cuevas-Tello, *Artificial Neural Networks for Speaker*, Intelligent Sustainable Systems: Selected Papers of WorldS4 2024, Volume 3, pp. 1-511, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[37] Federico Pardo, Óscar Cánovas, and Félix J. García Clemente, “Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps,” *Applied Sciences*, vol. 15, no. 12, pp. 1-41, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[38] Steffen T. Eberhardt et al., “Development and Validation of Large Language Model Rating Scales for Automatically Transcribed Psychological Therapy Sessions,” *Scientific Reports*, vol. 15, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[39] Rita Francese et al., “DeepTald: A System for Supporting Schizophrenia-Related Language and Thought Disorders Detection with NLP Models and Explanations,” *Multimedia Tools and Applications*, vol. 84, pp. 46273-46306, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Andrii M. Striuk, and Vladyslav V. Hordiienko, “Research and Development of a Subtitle Management System using Artificial Intelligence,” *CEUR Workshop Proceedings*, pp. 415-427, 2025. [Google Scholar] [Publisher Link]

[41] Alymzhan Toleu et al., “Speaker Change Detection with Pre-trained Large Audio Model,” *Asian Conference on Intelligent Information and Database Systems*, Kitakyushu, Japan, pp. 262-274, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[42] Soubhik Barari, and Tyler Simko, "The Promise of Text, Audio, and Video Data for the Study of us Local Politics and Federalism," *Publius: The Journal of Federalism*, vol. 55, no. 2, pp. 223-252, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[43] Bang Zeng, and Ming Li, "USEF-TSE: Universal Speaker Embedding Free Target Speaker Extraction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2110-2124, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[44] Ravi D. Shankar, R. B. Manjula, and Rajashekhar C. Biradar, "Revolutionizing Speaker Recognition and Diarization: A Novel Methodology in Speech Analysis," *SN Computer Science*, vol. 6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[45] Defu Chen et al., "Res2Former: Integrating Res2Net and Transformer for a Highly Efficient Speaker Verification System," *Electronics*, vol. 14, no. 12, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[46] Erfan Loweimi et al., "Speaker Retrieval in the Wild: Challenges, Effectiveness and Robustness," *arXiv preprint*, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[47] Shuai Wang et al., "Overview of Speaker Modeling and its Applications: From the Lens of Deep Speaker Representation Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4971-4998, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Qiuyu Zheng et al., "NResNet: Nested Residual Network Based on Channel and Frequency Domain Attention Mechanism for Speaker Verification in Classroom," *Multimedia Tools and Applications*, vol. 84, pp. 14235-14251, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[49] Danwei Cai, "*Speaker Representation Learning Under Self-Supervised and Knowledge Transfer Setting*," Doctoral Thesis, Duke University, pp. 1-24, 2023. [Google Scholar]

[50] Qing Gu et al., "A Domain Robust Pre-Training Method with Local Prototypes for Speaker Verification," *Interspeech*, pp. 1-5, 2025. [Google Scholar] [Publisher Link]

[51] Hao Wang, Xiaobing Lin, and Jiashu Zhang, "A Lightweight CNN-Conformer Model for Automatic Speaker Verification," *IEEE Signal Processing Letters*, vol. 31, pp. 56-60, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[52] Jin Sob Kim et al., "Universal Pooling Method of Multi-Layer Features from Pretrained Models for Speaker Verification," *arXiv preprint*, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[53] Min Hyun Han, Sung Hwan Mun, and Nam Soo Kim, "Generalized Score Comparison-Based Learning Objective for Deep Speaker Embedding," *IEEE Access*, vol. 13, pp. 51194-51207, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[54] Bei Liu, Zhengyang Chen, and Yanmin Qian, "Depth-First Neural Architecture with Attentive Feature Fusion for Efficient Speaker Verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1825-1838, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[55] Zhiyong Chen et al., "Open-Set Speaker Identification through Efficient Few-shot Tuning with Speaker Reciprocal Points and Unknown Samples," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3347-3362, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[56] Danwei Cai et al., "Self-Supervised Reflective Learning through Self-Distillation and Online Clustering for Speaker Representation Learning," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1535-1550, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[57] Prachi Singh, and Sriram Ganapathy, "End-to-End Supervised Hierarchical Graph Clustering for Speaker Diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 448-457, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[58] Y. Dissen, S. Harpaz, and J. Keshet, "Label-Free Speaker Diarization Using Self-Supervised Speaker Embeddings," *SSRN*, 2025. [Google Scholar]

[59] Prachi Singh, Amrit Kaul, and Sriram Ganapathy, "Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[60] K.V. Aljinu Khadar, R.K. Sunil Kumar, and V.V. Sameer, "Speaker Diarization based on X Vector Extracted from Time-Delay Neural Networks (TDNN) using Agglomerative Hierarchical Clustering in Noisy Environment," *International Journal of Speech Technology*, vol. 28, pp. 13-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[61] Mathias Näreaho, "Speaker Diarization in Challenging Environments using Deep Networks: An Evaluation of a State-of-the-Art System," *DiVA Portal*, 2023. [Google Scholar] [Publisher Link]

[62] Vinod K. Pande, Vijay K. Kale, and Sangramsing N. Kayte, "Feature Extraction Using I-Vector and X-Vector Methods for Speaker Diarization," *ICTACT Journal on Soft Computing*, vol. 15, no. 4, pp. 3717-3721, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[63] Nikhil Raghav et al., "Self-Tuning Spectral Clustering for Speaker Diarization," *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[64] Rahul Dasari, "*Speaker Diarization Using Multi-View Contrastive Learning Embeddings*," Technical Report, pp. 1-6, 2025. [Google Scholar] [Publisher Link]

[65] Joonas Kalda et al., “Design Choices for PixIT-based Speaker-Attributed ASR: Team ToTaTo at the NOTSOFAR-1 Challenge,” *Computer Speech & Language*, vol. 95, pp. 1-16, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[66] Federico Landini, “From Modular to End-to-End Speaker Diarization,” *arXiv preprint*, pp. 1-138, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[67] Weiqing Wang, and Ming Li, “End-to-End Online Speaker Diarization with Target Speaker Tracking,” *arXiv preprint*, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[68] Weiqing Wang, and Ming Li, “Online Neural Speaker Diarization with Target Speaker Tracking *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 5078-5091, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[69] Youngki Kwon et al., “Absolute Decision Corrupts Absolutely: Conservative Online Speaker Diarisation,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[70] Chin-Yi Cheng et al., “Multi-Target Extractor and Detector for Unknown-Number Speaker Diarization,” *IEEE Signal Processing Letters*, vol. 30, pp. 638-642, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[71] Yiling Huan et al., “Towards Word-Level End-to-end Neural Speaker Diarization with Auxiliary Network,” *arXiv preprint*, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[72] Han Yin et al., “SpeakerLM: End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models,” *arXiv preprint*, pp. 1-9, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[73] Samuele Cornell et al., “One Model to Rule Them All? Towards End-To-End Joint Speaker Diarization and Speech Recognition,” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 11856-11860, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[74] Tianyi Tan et al., “DistillW2N: A Lightweight One-Shot Whisper to Normal Voice Conversion Model Using Distillation of Self-Supervised Features,” *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[75] Zhao-Ci Liu et al., “PE-Wav2vec: A Prosody-Enhanced Speech Model for Self-Supervised Prosody Learning in TTS,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4199-4210, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[76] Pourya Jafarzadeh, Amir Mohammad Rostami, and Padideh Choobdar, “Speaker Emotion Recognition: Leveraging Self-Supervised Models for Feature Extraction Using Wav2Vec2 and HuBERT,” *arXiv preprint*, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[77] Valentin Vielzeuf, “Investigating the 'Autoencoder Behavior' in Speech Self-Supervised Models: A Focus on HuBERT's Pretraining,” *arXiv preprint*, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[78] Alexander H. Liu et al., “Dinosr: Self-Distillation and Online Clustering for Self-Supervised Speech Representation Learning,” *Advances in Neural Information Processing Systems*, pp. 1-17, 2023. [Google Scholar] [Publisher Link]

[79] Bing Han, Zhengyang Chen, and Yanmin Qian, “Self-Supervised Learning with Cluster-Aware-Dino for High-Performance Robust Speaker Verification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 529-541, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[80] Zia-ur-Rehmana, Arif Mahmood, and Wenxiong Kang, “Pseudo-Label Refinement for Improving Self-Supervised Learning Systems,” *arXiv preprint*, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[81] Irina-Elena Veliche, and Pascale Fung, “Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[82] Faria Nazir et al., “A Computer-Aided Speech Analytics Approach for Pronunciation Feedback using Deep Feature Clustering,” *Multimedia Systems*, vol. 29, pp. 1699-1715, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[83] Simin Kou et al., “Structure-Aware Subspace Clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10569-10582, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[84] Anaïs Ollagnier, Elena Cabrio, and Serena Villata, “Unsupervised Fine-Grained Hate Speech Target Community Detection and Characterisation on Social Media,” *Social Network Analysis and Mining*, vol. 13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[85] Ziyang Ma et al., “Pushing the Limits of Unsupervised Unit Discovery for SSL Speech Representation,” *arXiv preprint*, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[86] S. Merlin Revathy, and S.S. Kumar, “Optimized Deep Embedded Clustering-Based Speaker Diarization with Speech Enhancement,” *Circuits, Systems, and Signal Processing*, vol. 44, pp. 5044-5074, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[87] Amarendra Jadda, and Inty Santi Prabha, “Adaptive Weiner Filtering with AR-GWO based Optimized Fuzzy Wavelet Neural Network for Enhanced Speech Enhancement,” *Multimedia Tools and Applications*, vol. 82, pp. 24101-24125, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[88] Mohammad Reza Falahzadeh et al., "Deep Convolutional Neural Network and Gray Wolf Optimization Algorithm for Speech Emotion Recognition," *Circuits, Systems, and Signal Processing*, vol. 42, pp. 449-492, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[89] Suwon Shon et al., "Context-Aware Fine-Tuning of Self-Supervised Speech Models," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[90] Rongjie Huang et al., "VoiceTuner: Self-Supervised Pre-Training and Efficient Fine-tuning For Voice Generation," *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10630-10639, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[91] Xianghu Yue et al., "Adapting Pre-Trained Self-Supervised Learning Model for Speech Recognition with Light-Weight Adapters," *Electronics*, vol. 13, no. 1, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[92] Weixi Lai, "*Parameter-Efficient Fine-Tuning for Sarcasm Detection in Speech Using the Self-Supervised Pre-Trained Model WavLM*," Master's Thesis, University of Groningen, pp. 1-39, 2024. [Google Scholar] [Publisher Link]

[93] Bryce Irvin et al., "Self-Supervised Learning for Speech Enhancement Through Synthesis," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[94] Muzamil Ahmed et al., "An Enhanced Deep Learning Approach for Speaker Diarization using TitaNet, MarbelNet and Time Delay Network," *Scientific Reports*, vol. 15, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[95] Amit Kumar Bhuyan, Hrishikesh Dutta, and Subir Biswas, "Unsupervised Speaker Diarization in Distributed IoT Networks Using Federated Learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 2, pp. 1934-1946, 2025. [CrossRef] [Google Scholar] [Publisher Link]