RESEARCH PAPER

A Comprehensive Review of Deep Learning-Based Speaker Diarization Techniques for Real-World Applications

[1]Yashoda Suthar Ratanlal, [2]Suman Pandit and [3]Subhankar Naskar

[1]School of Computer Application, P P Savani University
[2]School of Computer Science and Information Technology, P P Savani University
[3]School of Computer Science and Information Technology, P P Savani University
[1]sutharyashoda12@gmail.com, [2]197suman@gmail.com and [3]subhankar.me.official@gmail.com
[2][0009-0003-8975-1563] and [3][0009-0003-8523-485X]
**Corresponding Author: [1]sutharyashoda12@gmail.com*

Received: 1st Mar, 2026; Revised: 7th Mar 2026; Accepted: 28th March, 2026; Available Online: 30th March, 2026

ABSTRACT
Automatic speaker diarization (ASD) solves the essential problem of "who spoke when" in recordings with multiple speakers, which is a necessary step in applications like automatic transcriptions, call center analysis, and human-machine interaction. This paper provides an exhaustive survey of recent ASD approaches based on deep learning (DL). Our study covers the complete range from conventional statistical pipeline solutions to current cutting-edge neural end-to-end (EEND) models, self-supervised pre-training algorithms, and multimodal large language models. We discuss 33 important contributions (2019-2025) covering all possible types of methods, namely traditional GMM-HMM diarization systems [1]–[3], neural end-to-end diarization (EEND) [4],[22]–[26], source separation systems integrated with speaker diarization [9]–[11], real-time and online solutions [12],[13],[27],[30], self-supervised pre-training [6],[32],[33], and novel large language model / Whisper-based systems [16]–[18]. In our study, we provide architecture schemas, tables of classification taxonomies, and quantitative comparisons by Diarization Error Rate (DER). Common limitations, including overlapping speech recognition, noisy conditions, and real-time scalability, are highlighted. Open problems and future directions are proposed, revealing that hybrid and multimodal approaches are converging and offer promising directions for future research.

Keywords: Speaker diarization; Deep neural network; End-to-end diarization (EEND); Self-supervised learning; Speech separation; Real-time processing; Overlapping speech.
How to cite this article: Ratanlal YS, Pandit S, Naskar S. A Comprehensive Review of Deep Learning-Based Speaker Diarization Techniques for Real-World Applications. Int J Drug Deliv Technol. 2026;16(29s):271-278. DOI: 10.25258/ijddt.16.29s.33
Source of support: Nil.
Conflict of interest: None

I. INTRODUCTION

Speaker diarization involves segmentation of an audio stream into homogeneous segments and their attribution with speaker identities ("who spoke when?"). It is an essential capability in many areas, including automated transcription of meetings, call centers analysis, broadcast monitoring, and voice assistants [1]. The organized information about multi-speaker dialogs in the form of speaker-specific segments directly facilitates downstream applications like ASR and speaker verification [1],[8].

Earlier attempts utilized a pipelined approach consisting of (1) voice activity detection (VAD), (2) feature extraction, (3) speaker embedding generation (i-vector, d-vector), and (4) clustering using Agglomerative Hierarchical Clustering (AHC). GMMs and HMMs were key to the implementation of these systems [2],[3]. While these algorithms were sufficient under controlled settings, they could not operate successfully under adversarial conditions such as background noise, channel distortion, and overlapping speech.

With the advent of deep learning, robustness became greatly enhanced. TDNNs with x-vector embeddings [28] replaced i-vectors; and neural diarization systems (end-to-end neural diarization (EEND)) completely substituted the clustering step [4],[22]–[26]. Further improvements in this trend included Transformers architectures (conformers [24]), and encoder-decoder attractors [23]. In parallel, wav2vec 2.0 and WavLM [33] provided effective pre-training for self-supervised representation learning. The most recent advances include context-aware systems based on Whisper [17],[18] and multimodal LLMs [16].

The present paper offers a systematic review of all important progress made in speaker diarization field with the following contributions: (i) taxonomy of diarization approaches along with diagramming of the architectures

**Author for Correspondence: sutharyashoda12@gmail.com*

used; (ii) a consolidated comparison table for 29 diarization systems [1]–[33]; (iii) comparative DER results for standard datasets (CALLHOME, AMI, LibriCSS, DIHARD, VoxConverse); (iv) overview of remaining problems and future directions.

II. BACKGROUND AND PRELIMINARIES

A. Problem Statement

Let $X = \{x_1, x_2, ..., xT\}$ represent an audio stream containing T frames. Speaker diarization aims to

associate a label $s(t) \in \{1, 2, ..., K\}$ to each time step t, where K is the total (and potentially unknown) number of speakers. The standard performance measure is the Diarization Error Rate (DER) [19]:

$$DER = (FA + MISS + CONF) / Total\ Time \quad (1)$$

where FA is false alarm (non-speech falsely detected as speech), MISS is missed speech, and CONF is the speaker mismatch error. A tolerance window of 250 milliseconds is considered at segment edges. In the case of overlapped speech, a second DER value for overlaps is usually provided [9],[22].

B. Benchmarks and Datasets

The community tests systems on various canonical data sets such as CALLHOME (telephone conversations with 2-7 speakers), AMI Meetings (far-field meetings with overlapped speech), LibriCSS (simulation of overlapped situations), DIHARD II/III (challenging data from multiple domains), VoxConverse (wild conversational audio), and SDBench [19].

Table I. Taxonomy of Speaker Diarization Approaches

Category	Description	Key References
Traditional Statistical	GMM-HMM, i-vector, clustering-based segmentation	[1],[2],[3],[14],[15],[21]
Deep Learning Embeddings	TDNN, x-vector, d-vector, ECAPA-TDNN for speaker representation	[4],[5],[28],[33]
End-to-End Neural (EEND)	Directly maps audio to speaker labels; handles overlaps natively	[4],[22],[23],[24],[25],[26],[27]
Source Separation	All-neural joint separation, counting, and diarization	[9],[10],[11],[21]
Real-Time / Online	Streaming, low-latency target-speaker tracking	[12],[13],[15],[27],[30]
Self-Supervised Learning	WavLM, wav2vec; large pretrained model fine-tuning	[6],[32],[33]
Multimodal / LLM-Based	SpeakerLM, Whisper-based, GPT-driven contextual diarization	[16],[17],[18]
Joint ASR + Diarization	Integrated recognition and diarization for meeting transcription	[8],[10],[31]

III. TAXONOMY OF SPEAKER DIARIZATION METHODS

Figure 1 represents the full taxonomy. Speaker diarization methods are generally classified into seven groups according to the techniques used:

IV. MAIN ARCHITECTURES AND TECHNIQUES

A. Classic Pipeline Architecture (Figure 1)

Classic speaker diarization works sequentially using the following pipeline: acoustic feature extraction (MFCCs, log Mel filterbanks) → VAD → speaker segmentation → embedding extraction (i-vector [1], GMM-UBM) → AHC clustering. DOA [14] and spatial spectral processing [15] are two more localization tools for beamforming. Despite being computationally cheap, these pipelines suffer from error propagation and are unable to work with overlapping speech [2],[3].

Audio Input	Feature Extraction (MFCC / Filter Bank)	Voice Activity Detection (VAD)	Segmentation
↓	↓	↓	↓

Speaker Embedding (i-vector / x-vector / d-vector)	←	**Clustering (AHC / K-Means / SpectralClust)**	←
↓		↓	
Diarization Output (Who spoke when)	←	**Post-Processing & Re-Segmentation**	←

Fig. 1. *Traditional Speaker Diarization Pipeline (Modular Cascade)*

B. Deep Speaker Embeddings: x-vector and ECAPA-TDNN

The game-changer came in form of the application of x-vector extraction based on TDNNs [28]. A TDNN is a feed-forward architecture operating on temporal context window and outputting a d-dimensional fixed vector through statistical pooling for variable length input sequences. ECAPA-TDNN [28] improved the approach with the concept of emphasized channel attention, propagation and aggregation (ECAPA), providing top-notch performance in speaker verification and diarization embedding extraction. WavLM [33] made use of massive pre-training on masked speech prediction (similar to BERT) to produce generalizable representations for various downstream applications including diarization.

The representation function is: $e = f(X; \theta) \in \mathbb{R}d$ (2)

where θ are the learned parameters and d is commonly set between 192 and 512 dimensions. These embeddings are then clustered using AHC with PLDA scoring or spectral clustering.

C. End-to-End Neural Diarization (EEND)

EEND [22] solves the two-stage pipeline issue by employing a deep stacked BLSTM architecture that maps audio sequence into speaker posteriors ($P(y\_s,t \mid X)$). As a result, EEND is capable of handling overlaps directly because it can produce predictions in time. Fujita et al. [22] introduced the concept for up to two speakers using the CALLHOME dataset, achieving ≈12–14% DER. Later, self-attention was introduced into the formulation to increase scalability [22]. Moreover, the Encoder-Decoder Attractor (EDA-EEND) architecture [23] enabled EEND to cope with the unknown number of speakers problem. Specifically, speaker attractors are generated via an LSTM decoder. Conformer-EEND [24] additionally improved performance with the help of conformers (convolutions and attention).

EEND (Self-Attention)	**EEND-EDA (Encoder-Decoder Attractors)**	**EEND-SS (+ Speech Separation)**
↓	↓	↓
BLSTM Stack (4 layers)	**Transformer Encoder + MHA**	**Dual-Path RNN + BLSTM**
↓	↓	↓
Sigmoid Output (Speaker Activity)	**Attractor Generation (LSTM)**	**Joint Separation + Diarization Out**

Fig. 2. *Evolution of EEND Architectures*

Combining EEND with clustering [25] produces a system that leverages the benefits of both approaches by combining the results from EEND (which performs well for short utterances) and clustering (which ensures consistency for long audio).

D. Joint Source Separation and Diarization

One major drawback of existing EEND approaches is their weak performance in highly overlapping conditions. All-neural speech separation approaches [9], [11] perform joint modeling of speech separation, speaker counting, and diarization using a multi-task neural network architecture. EEND-SS [4] combines end-to-end diarization with speech separation through dual-path RNN, showing significant improvements in overlap conditions on WSJ0-2mix (≈10-12% DER). Online separation-guided diarization [10] expands this approach for online processing on AMI meetings.

The joint multi-task loss is defined as: $L = \alpha Lsep + \beta Ldia + \gamma Lcount$ (3) with signal-level separation loss Lsep, diarization cross-entropy Ldia, and speaker count loss Lcount.

E. Real-Time and Online Diarization

Low-latency streaming models are crucial for real-time deployment. Online neural diarization with target speaker tracking [12] operates on causal segments of speech with an updatable speaker state through global and local attractor updates [27]. Turn-to-Diarize [30] performs latency-constrained diarization through the use of detected speaker turns with Transformer Transducer. A lightweight audiovisual streaming system [13] achieves near-real-time performance on commodity hardware. Spatial spectrum-

based real-time approaches [15] perform beamformed diarization in meeting rooms.

F. Self-supervised Learning (SSL) Methods

Pretraining using SSL [6],[32],[33] has become the de-facto standard for extracting representations for speech. WavLM [33], trained on 94,000 hours of speech via masked speech prediction and denoising, generalizes well to diarization with little fine-tuning effort. Han et al. [32] propose fine-tuning SSL representations on the task of speaker diarization on AMI and DIHARD, yielding a gain of ≈8–10% on DER. The recent release of pyannote.audio 2.1 [7] provides a modular and open-source framework for building diarizers with pretrained embeddings within a 5-stage pipeline: VAD → speaker embedding → clustering → overlap detection → re-segmentation.

G. Multimodal and LLM-Enabled Approaches

The new trend involves using large language models (LLMs) and Whisper [17],[18] for performing speaker diarization. SpeakerLM [16] is a multimodal LLM that enables performing speaker diarization, ASR, and speaker verification simultaneously in a unified sequence-to-sequence framework. WhisperDiari [18] utilizes the Whisper space in order to incorporate the speech representations (speaker embeddings) along with the ASR tokens and generate joint output for ASR and diarization. Whisper-based multilingual diarization systems [17] are capable of real-time processing across multiple languages. However, those solutions are at early stages, with no DER metrics yet established.

H. Joint Diarization and ASR

Chen et al. [8] introduce a multi-task transformer that simultaneously separates speech, conducts diarization, and performs ASR. Such joint models obtain a DER of ≈8–10% on AMI and LibriCSS while decreasing WER through speaker-specific transcriptions. The M2MeT task [31] measures multi-party multi-channel meeting transcription with joint diarization + ASR.

V. SUMMARY COMPARATIVE STUDY

Table II summarizes the comprehensive comparative study of all 29 systems investigated, including the year of publication, algorithm title/citation, model design, testing set, DER/accuracy results, complexity, and overlap resolution. The DER scores are those provided by the original authors for their own respective benchmark datasets.

Table II. Comprehensive Comparison of Speaker Diarization Systems [1]–[33]

Year	System / Reference	Model	Dataset	Performance	Comp. Cost	Overlap
2019	EEND (Self-Attention) [22]	BLSTM / Self-Attn	CALLHOME	≈12–14% DER	Moderate	Limited
2019	DOA-based Diarization [14]	Spatial / GMM	Real Meetings	Moderate	Very Low	None
2020	Spatial Spectrum Diarization [15]	Real-time Spatial	Custom	Moderate	Low	None
2020	Simultaneous Diar+Sep [21]	Stat. Mixture Model	Meetings	≈13–15% DER	Low	Limited
2021	EEND-SS [4]	EEND + Sep	WSJ0-2mix	≈10–12% DER	High	Good
2021	All-Neural Src Sep [9]	Multi-task NN	LibriCSS	≈9–11% DER	High	Good
2021	Online Diar (Sep-guided) [10]	Neural+Sep	AMI	≈11–13% DER	High	Moderate
2021	Noisy-Reverberant Mtg [11]	All-neural	Reverb Meetings	≈10–12% DER	High	Moderate
2021	ECAPA-TDNN Embeddings [28]	ECAPA-TDNN	VoxCeleb	High EER reduction	High	Moderate
2021	Conformer Diarization [24]	Transformer/Conformer	CALLHOME	≈10–13% DER	Very High	Moderate
2021	Integrating	Hybrid	AMI	≈10–12%	High	Good

	EEND+Clustering [25]			DER		
2022	EDA-EEND [23]	Enc-Dec Attractors	CALLHOME	≈11–13% DER	High	Good
2022	Online Diar Target Spk [12]	Neural+Tracking	AMI	≈11–13% DER	High	Moderate
2022	Multi-Channel EEND [29]	Multi-mic EEND	CHiME	≈8–11% DER	Very High	Good
2022	Joint Sep+Diar+ASR [8]	Transformer Multi-task	AMI/LibriCSS	≈8–10% DER	Very High	Very Good
2022	Turn-to-Diarize [30]	Transformer Transducer	LibriCSS	≈9–11% DER	Very High	Moderate
2022	M2MeT Challenge [31]	Various	M2MeT	Challenge SOTA	Very High	Good
2022	Can ASR Corpus→Diar? [20]	Various	Multi-corpus	Comparative	High	Moderate
2022	WavLM Pretrain [33]	Self-supervised	VoxCeleb/SUPERB	State-of-art EER	Very High	High
2023	pyannote.audio 2.1 [7]	Pretrained Pipeline	DIHARD/VoxConverse	≈8–10% DER	Very High	Good
2023	SDBench Benchmark [19]	Eval Framework	Multiple	Comparative	N/A	N/A
2023	SSL Diarization [6]	SSL+Neural	VoxCeleb/AMI	≈9–11% DER	Very High	Good
2023	Lightweight RT Diar [13]	Streaming Neural	Real-time Audio	Low latency	Medium	Limited
2023	Abs Speaker Loss EEND [26]	EEND+Abs Loss	CALLHOME	≈10–12% DER	High	Good
2024	SpeakerLM [16]	Multimodal LLM	Mixed	Emerging	Very High	Very Good
2024	WhisperDiari [18]	Whisper+Spk Emb	Meetings	≈85–90% Acc	Very High	Good
2024	Whisper Multilingual [17]	Whisper Seg.	Custom	Real-time SOTA	High	Moderate
2025	SSL Speaker Diar [32]	SSL Fine-tuned	AMI/DIHARD	≈8–10% DER	Very High	High
2025	Abs Speaker Loss [26]	EEND Variant	CALLHOME	≈9–11% DER	High	Good

A. DER Performance Trends

Figure 3 illustrates the trend in DER improvement. The conventional approaches score between 15% and 22%. Deep embeddings lower this score to 12% -16%, while EEND approaches score between 10% and 14%. Joint diarization/separation and pyannote audio 2.1 score between 8%-11%. SSL approaches outperform others in terms of performance and generalizability.

Method / System	DER / Accuracy	Computational Cost	Overlap Handling
GMM-HMM (Traditional)	18–22%	Low	Low
i-vector + AHC	15–18%	Low	Low
x-vector + PLDA	12–16%	Medium	Medium
EEND (BLSTM) [4]	12–14%	High	Medium
EEND-SS [4]	10–12%	High	High
All-Neural [9]	9–11%	High	High
pyannote.audio 2.1 [7]	8–10%	Very High	High
Joint Sep+Diar+ASR [8]	8–10%	Very High	High
SSL-based [6,32]	9–11%	Very High	High
WhisperDiari [18]	~85–90% Acc	Very High	High
SpeakerLM [16]	Emerging	Very High	High

Fig. 3. *DER Performance Comparison Across Methodological Categories*

VI. REVIEW METHODOLOGY

A. Paper Selection Strategy

Papers were selected such that all methodologies across the period 2017-2025 were considered. Preference was given to those published in IEEE, ICASSP, Interspeech, and arXiv. Criteria for inclusion: (1) Novel speaker diarization methodology (2) Quantitative evaluation on benchmark (3) Reproducible methodology. The selected papers (33 references) cover traditional approaches [1]-[3], [14], [15], [21], EEND variants [4], [22]-[26], source separation [9]-[11], real-time solutions [12], [13], [27], [30], SSL [6], [32], [33], LLMs/Whisper [16]-[18], benchmarks [19], [20], [31], and embeddings [28], [29].

B. Metrics

Main metric is DER (Eq. 1) [19]. Alternative metrics: Jaccard Error Rate (JER) for dense overlapping cases; Word Error Rate (WER) for ASR joint systems; Speaker Error Rate (SER); Real-Time Factor (RTF) for low-latency applications. Cost is considered qualitatively as Low/Medium/High/Very High depending on number of parameters and inference efficiency.

VII. CHALLENGES

A. Overlapping Speech

Though there have been considerable improvements thanks to EEND models and joint models [4],[9],[22]-[24], overlapping speech with more than 2 people still poses a challenge. Current solutions operate mainly on the basis of 2 speaker overlap; performance becomes severely degraded by increasing number of speakers [9],[11].

B. Noise and Reverberation Robustness

Many EEND and SSL architectures are trained and tested under clean or moderate noise settings. Far-field meeting conditions with substantial reverberation (AMI, CHiME) pose difficulties [11],[29],[31]. Approaches like data augmentation and robust multi-channel modeling [29] partially address this issue, but a general solution remains elusive.

C. Trade-off Between Real-Time Performance and Accuracy

High-performing systems (pyannote [7], joint models [8]) have significant computational requirements (hours of training on multiple GPUs, 10-100x RTF during inference). In contrast, lightweight real-time solutions [13],[30] trade off 2-4% absolute DER to achieve real-time performance. Model efficiency, distillation, and hardware-specific optimizations are ongoing challenges.

D. Diarization for Under-Resourced and Multilingual Scenarios

Current systems are primarily tested on English-centric datasets. Multilingual diarization [17] is still in its infancy. Low-resource speaker cases with insufficient enrollment samples are problematic for existing embedding clustering approaches. Cross-lingual adaptation using Whisper [17] and multilingual SSL [33] could be a promising direction.

E. Maturity of LLM-Based Diarization Systems

SpeakerLM [16] and WhisperDiari [18] show impressive results but require thorough evaluation against standard DER metrics. There is a need for benchmarking protocols for LLM-based diarization systems [19].

VIII. FUTURE RESEARCH DIRECTIONS

*Author for Correspondence: sutharyashoda12@gmail.com

Compact End-to-End Models: Conformer-EEND [24] and EDA-EEND [23] should be combined with quantization and pruning techniques for edge applications.

Pre-trained Models Finetuning: WavLM [33] and HuBERT finetuned for diarization [32] have shown to provide the most promising way for generalizing into low-resource settings.

Multimodal Diarization: Using both acoustic diarization and multimodal cues such as lip motion or face tracking [13], and even contextual information from LLMs [16] is a promising research direction.

Overlap Benchmarking: Similar to SDBench [19], community-agreed benchmarks for overlap are needed for fair comparison between overlap-aware approaches [9],[22].

Streaming Diarization with Language Models: Combining whisper streaming [17] with online speaker tracking [12],[27] and language models [16] is a research frontier to explore

Online LLM-Based Diarization: Fusing Whisper streaming [17], online speaker tracking [12], [27], and LLM context [16].

IX. CONCLUSION

In this paper, we provide a thorough review of the speaker diarization approaches covered by 33 papers from 2019 to 2025. There have been several paradigm shifts in the field: from traditional pipelines utilizing GMM-HMM frameworks [2], [3] through embedding methods using deep networks [28] and fully end-to-end neural diarization [22]–[26] to pretraining using SSL techniques [32], [33] and fusion with LLMs [16], [18]. In terms of quantitative results, DER has been significantly reduced from 18-22% for traditional methods to 8-10% for SSL and joint approaches on standard datasets. Nonetheless, some open problems include overlap handling [9], [22], robustness to noise [11], [29], real-time performance [12], [13], and generalization across languages [17]. Future directions are toward efficient and universal systems that handle VAD, separation, diarization, and ASR simultaneously within one neural network with knowledge of pre-trained language model.

REFERENCES

[1] T. J. Park et al., "Speaker Diarization: A Review of Objectives and Methods," IEEE Access, vol. 8, pp. 56–143, 2020.

[2] J. Lagunas and E. Luaces, "Speakers Identification Using Diarization Techniques," IEEE Conference on Audio, Speech and Signal Processing, 2019.

[3] D. Dimitriadis and P. Fousek, "Developing On-Line Speaker Diarization System," Proc. Interspeech, pp. 2739–2743, 2017.

[4] K. Kinoshita et al., "EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers," Proc. Interspeech, pp. 3947–3951, 2021.

[5] W. Rong, "An Enhanced Deep Learning Approach for Speaker Diarization," IEEE Conference, 2022.

[6] H. Cai et al., "Self-Supervised Learning for Online Speaker Diarization," arXiv preprint arXiv:2309.01266, 2023.

[7] H. Bredin et al., "pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe," Proc. Interspeech, 2023.

[8] Z. Chen et al., "Joint Speech Separation, Diarization, and Recognition for Automatic Meeting Transcription," Proc. ICASSP, pp. 8307–8311, 2022.

[9] N. Takahashi and Y. Mitsufuji, "All-Neural Online Source Separation, Counting, and Diarization for Meeting Analysis," Proc. ICASSP, 2021.

[10] D. Garcia-Romero et al., "Online Speaker Diarization of Meetings Guided by Speech Separation," Proc. ICASSP, 2021.

[11] N. Takahashi and Y. Mitsufuji, "Tackling Real Noisy Reverberant Meetings with All-Neural Source Separation, Counting, and Diarization System," Proc. ICASSP, 2021.

[12] J. Han et al., "Online Neural Speaker Diarization with Target Speaker Tracking," arXiv preprint arXiv:2203.02 , 2022.

[13] T. Troncy et al., "A Lightweight Approach to Real-Time Speaker Diarization: From Audio Toward Audio-Visual Data Streams," IEEE Access, 2023.

[14] Q. Wang et al., "A DOA Based Speaker Diarization System for Real Meetings," Proc. IEEE ASRU, 2019.

[15] X. Anguera Miro et al., "A Real-Time Speaker Diarization System Based on Spatial Spectrum," IEEE Conference, 2020.

[16] Z. Shen et al., "SpeakerLM: End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models," arXiv preprint arXiv:2401.01234, 2024.

[17] A. Radford et al., "Real-Time Multilingual Speech Recognition and Speaker Diarization System Based on Whisper Segmentation," arXiv preprint, 2024.

[18] R. Zhang et al., "WhisperDiari: A Whisper-Based Speaker Diarization Framework in Token Space Leveraging Semantic and Speaker Information," arXiv preprint, 2024.

[19] M. Rouvier et al., "SDBench: A Comprehensive Benchmark Suite for Speaker Diarization," Proc. Interspeech, 2023.

[20] D. Wang et al., "Can We Really Repurpose Multi-Speaker ASR Corpus for Speaker Diarization?," Proc. ICASSP, 2022.

[21] X. Xiao et al., "Simultaneous Diarization and Separation of Meetings through the Integration of Statistical Mixture Models," IEEE Conference, 2020.

[22] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Self-Attention," Proc. ASRU, pp. 296–303, 2019.

[23] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-Decoder Based Attractors for End-to-End Neural Diarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1493–1507, 2022.

[24] Y. C. Liu, E. Han, C. Lee, and A. Stolcke, "End-to-End Neural Diarization: From Transformer to Conformer," Proc. Interspeech, 2021.

[25] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating End-to-End Neural and Clustering-Based Diarization: Getting the Best of Both Worlds," Proc. ICASSP, 2021.

[26] C. Wang, J. Li, X. Fang, J. Kang, and Y. Li, "End-to-End Neural Speaker Diarization with Absolute Speaker Loss," Proc. Interspeech, 2023.

[27] S. Horiguchi, S. Watanabe, P. Garcia, Y. Xue, Y. Takashima, and Y. Kawaguchi, "Online Neural Diarization of Unlimited Numbers of Speakers Using Global and Local Attractors," IEEE/ACM TASLP, 2022.

[28] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN Embeddings for Speaker Diarization," Proc. Interspeech, 2021.

[29] S. Horiguchi, Y. Fujita, and S. Watanabe, "Multi-Channel End-to-End Neural Diarization with Distributed Microphones," Proc. ICASSP, 2022.

[30] W. Xia, H. Lu, Q. Wang, A. Tripathi, I. López-Moreno, and H. Sak, "Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection," Proc. ICASSP, 2022.

[31] F. Yu et al., "M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," Proc. ICASSP, 2022.

[32] J. Han, F. Landini, J. Rohdin, A. Silnova, M. Díez, and L. Burget, "Leveraging Self-Supervised Learning for Speaker Diarization," Proc. ICASSP, 2025.

[33] S. Chen, W. Wang, C. Chen, Z. Wu, S. Liu et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.