Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

A Thorough Literature Review on Automatic Speaker Diarization Employing Machine Learning and Deep Learning Methodologies

[1]Sayyada Sara Banu [2]Ratnadeep R. Deshmukh [3*]Jaypalsing N. Kayte

[1, 2] Dept, of CS and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MH), INDIA.

[3] AI Lead, Tech Mahindra Ltd., Hi-Tech City, Hyderabad, Telangana, India

Email: [1]sayyada.sara@gmail.com,[2] rrdeshmukh.csit@bamu.ac.in, [3] jaypalsing@gmail.com

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Speaker Diarization, Neural Speaker Embeddings, Self-Supervised Speech Models, End-to-End Diarization, Speech Representation Learning

Abstract

Automatic Speaker Diarization (ASD) is the process of dividing an audio recording into regions where each speaker is the same and figuring out "who spoke when" with-out knowing who the speakers are ahead of time. It is a necessary part of meeting transcription, conversational analytics, indexing for broadcast media, forensic audio processing, call-center monitoring, and modern systems for human-computer interac-tion. In the past twenty years, diarization research has moved from traditional statis-tical models like Gaussian Mixture Models (GMMs) based on MFCCs and Bayesian Information Criterion (BIC) segmentation to more advanced representation learning methods like i-vectors and Probabilistic Linear Discriminant Analysis (PLDA). Later advances in deep learning led to strong neural embeddings like x-vectors and ECAPA-TDNN, which made it much easier to identify speakers in difficult sound situations. The most current Self-Supervised Learning (SSL) models, such as Wav2Vec 2.0, HuBERT, and WavLM, have set new standards by learning strong speech representations without any labeled input. End-to-End Neural Diarization (EEND), UIS-RNN, and VB-HMM re-segmentation are some of the complementary methods that have improved how well we can handle overlaps and refine time.

This evaluation offers a thorough examination of recent advancements, evaluat-ing the advantages and disadvantages of prominent diarization methodologies, pin-pointing enduring research deficiencies, and delineating prospective avenues for the enhancement of precise, multilingual, and real-time speaker diarization systems.

1. Introduction

Automatic Speaker Diarization is the technique of figuring out "who spoke when" in an audio recording without knowing how many people spoke or who they were. As human-machine interaction, multimedia retrieval, surveillance analytics, meeting transcription, and call-center automation grow increasingly common, strong diarization solutions are now necessary. The goal of diarization is to break up an audio stream into parts that are all the same and then give each part to the right speaker. This assignment may seem easy, but real-world audio recordings include many problems, such as background noise, reverberation, overlapping speech, channel fluctuation, and spontaneous conversational behavior. These complications necessitate advanced modeling of speaker characteristics,

temporal dynamics, and auditory variability, surpassing mere segmentation in diarization.

Early diarization systems used old-fashioned signal processing techniques that relied on statistical feature modeling. Mel-Frequency Cepstral Coefficients (MFCCs), for example, were the major approach to illustrate short-term spectral traits and were used to put speakers into groups based on how similar their voices sounded. Gaussian Mixture Models (GMMs), as discussed in foundational studies by Reynolds and Torres-Carrasquillo [1] and Chen & Gopinath [2], were widely employed for the clustering of MFCC feature vectors into speaker-specific distributions. These models were easy to use and didn't need a lot of computing power, which made them great for early broadcast news and meeting diarization tasks. But GMM-based diarization had many problems when the sound settings changed, and it couldn't handle complex situations like overlapping speech or interactions between multiple speakers in the background [3].

The next big step forward was the creation of total variability modeling using i-vectors. Dehak et al. [4] came up with the i-vector framework, which combined speaker and channel variability into a low-dimensional representation. This made it possible to make small speaker embeddings that might be used in large-scale diarization projects. Later, Garcia-Romero and Espy-Wilson [5] showed that PLDA (Probabilistic Linear Discriminant Analysis) scoring may be used to compare speakers in the i-vector domain. Sell and Garcia-Romero [6] used i-vectors and Agglomerative Hierarchical Clustering (AHC) together to improve diarization in both broadcast and conversational situations.

Even while i-vector-based systems did a better job at diarization than MFCC–GMM models, they still had problems in noisy and reverberant environments. They also had trouble processing speech that overlapped, which is becoming more common in natural conversations. Deep learning was a big change. Neural architectures, especially Time Delay Neural Networks (TDNN), made it possible to obtain very discriminative speaker embeddings called x-vectors. Snyder et al. [7] showed that x-vectors were much better than i-vectors because they could learn non-linear speaker traits over time, and getting more new ideas led to architectures, for example, ECAPA-TDNN, which Desplanques et al. [8] developed. This architecture combines channel attention techniques and squeeze-and-excitation modules [27] to make it easier to tell different speakers apart. These deep embeddings became the most popular method in diarization pipelines, lowering the Diarization Error Rate (DER) by a lot on several benchmark datasets.

Deep learning models still need a lot of labeled data and powerful computers, even though substantial gains have been made. To solve this, the speech community moved toward Self-Supervised Learning (SSL), a way for models to acquire useful representations from raw audio without any human input. Wav2Vec 2.0 [9] is the best SSL model because it can generate strong voice embeddings from masked prediction objectives. SSL models like HuBERT [10] and WavLM [11] have made this direction even better, with WavLM being especially useful for environments with many speakers and a lot of noise. These models are quite resilient to changes in the channel and background noise, and they can even do state-of-the-art diarization in difficult situations like those encountered in DIHARD datasets [23].

Another line of research is looking into end-to-end diarization, which wants to become rid of the clustering stage altogether. Fujita et al. [13, 14] presented End-to-End Neural Diarization (EEND), which directly forecasts speaker activity for many speakers with permutation-free targets. One of the major problems with classical and most deep embedding models is that they don't work well with overlapping speech. EEND is very good at this. Another option, UIS-RNN (Unbounded Interleaved State RNN), which Zhang et al. [12] came up with, uses recurrent neural sequence modeling instead of clustering. Variational Bayesian Hidden Markov Models (VB-HMM) [6, 25] are also often utilized as a re-segmentation phase to increase the quality of diarization across hybrid systems and fine-tune time boundaries.

Benchmark datasets like the AMI Meeting Corpus [17], VoxCeleb1 and VoxCeleb2 [18, 19], and DIHARD challenges [16, 23] have been very important in guiding research on diarization. These datasets include a wide range of acoustic situations, from clear speech to noisy conversations in the real world. They also provide a common testing ground for comparing different diarization methods. At the same time, technologies like Kaldi [20] and SpeechBrain [15] have made it easier for researchers and businesses to use complex diarization designs.

Diarization is still a challenging problem, even though it has made a lot of progress. Overlapping speech is a significant constraint for both classical and neural systems. Domain mismatch—differences across the contexts used for training and testing—can greatly hurt performance, as shown by research on cross-domain speaker recognition [21]. Also, diarization systems for low-resource languages,

such as several Indian languages like Marathi, are still not well-studied because there aren't enough annotated datasets. Real-time deployment is particularly challenging since many SSL models are too expensive to run on edge devices. Lastly, the absence of standardized benchmarks across several disciplines hinders equitable comparison and restricts widespread implementation in practical systems.

Figure 1 shows the whole process of the MFCC + GMM-based speaker diarization system, from how a raw audio stream is turned into time-stamped speaker labels. The procedure starts with the audio input, which could come from meetings, phone calls, broadcast media, or conversations. This raw signal goes through preprocessing first. Voice Activity Detection (VAD) removes quiet or non-speech or voiceless areas, and noise filtering is done to make the feature extraction that comes next better.

The next step is to extract MFCC features. This includes blocking frames, calculating the FFT, applying the Mel filterbank, and finding 13 MFCC coefficients and their delta (Δ) and delta-delta (ΔΔ) derivatives. These attributes do a good job of capturing the characteristics of speakers' vocal tracts. After extracting features, the system does the first step of segmentation, which divides the audio into short, uniform segments of one to two seconds each. Each segment is treated as an individual acoustic unit.

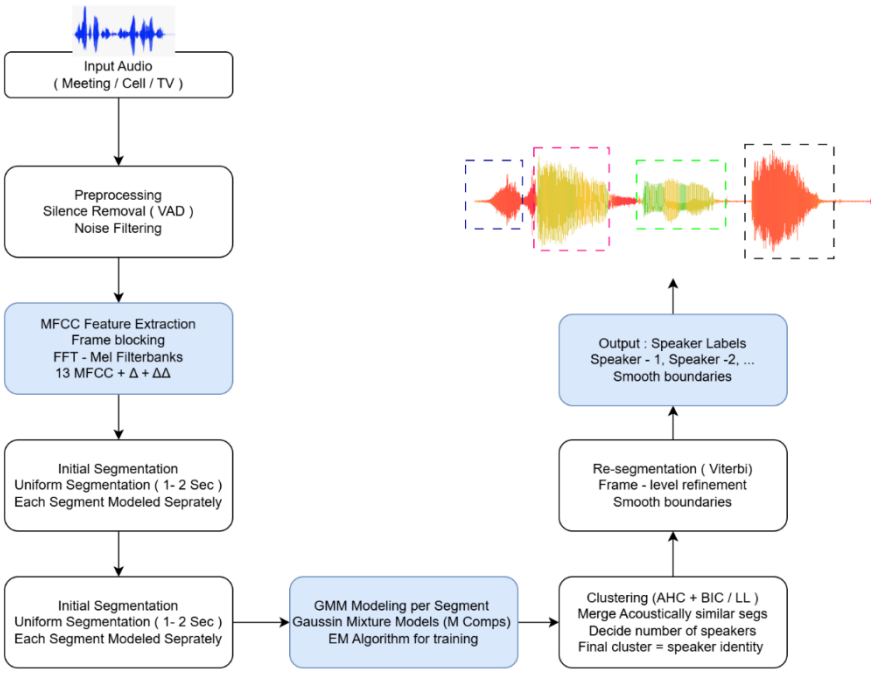

Figure 1. The steps in the MFCC + GMM-Based Speaker Diarization System

The Expectation-Maximization (EM) method guide Gaussian Mixture Models (GMMs) how to find the statistical distribution of MFCC feature vectors for each of these parts. The output GMMs are grouped together using Agglomerative Hierarchical Clustering (AHC) and Bayesian Information Criterion (BIC) or log-likelihood (LL) scoring. This means that segments that sound similar are put together. This step counts the number of speakers and groups the segments that belong to the same person. Finally, the Viterbi algorithm is used to break the data up into smaller pieces again. This makes the frame-level boundaries more accurate and the transitions between speakers smoother. The final output has names for the speakers, such as Speaker-1, Speaker-2, and so on, in the right order of time. The waveform portions on the right, which are color-coded, demonstrate this.

2. Traditional Methods for Speaker Diarization

2.1 Clustering Based on MFCC and GMM

Mel Frequency Cepstral Coefficients (MFCCs) have been the primary acoustic features for initial speaker diarization systems due to their efficacy in capturing transient spectral characteristics of speech and These traits were often combined with Gaussian Mixture Models (GMMs), that show how the acoustic frames are distributed statistically and group them into clusters that are similar to each other. The combination of MFCCs and GMMs was the most important part of traditional diarization research. Reynolds and Torres-Carrasquillo [1]

demonstrated one of the early comprehensive applications of MFCC + GMM for meeting diarization, suggesting that statistical clustering of MFCC features can effectively differentiate speakers in moderately controlled environments. As per Chen and Gopinath [2] help a comprehensive explanation of Gaussian Mixture Models (GMMs) as a probabilistic framework suitable for speech modeling, significantly easing the application of GMM-based clustering in speech and diarization projects. Anguera et al. [3] adapted this methodology for multi-microphone meeting environments, utilizing spatial cues to improve diarization quality in multi-channel recordings.

$$p(y_i) = \sum_{g=1}^{G} T_g f_g(y_i \mid \emptyset_g),$$

Where f_g is a probability density function with parameter $\emptyset_g, T_g$ is the corresponding mixture probabitlity where
$\sum_{g=1}^{G} T_g = 1$. Then its most basic form, model-based clustering sees each part of the mixture model as a cluster, figures out the model parameters, and puts each observation into the cluster that best matches its most likely mixture components.
Early research used MFCC + GMM-based diarization a lot, but it didn't work well in all acoustic situations. In clean and controlled environments, these models usually worked well, but they weren't as accurate when faced with real-world problems like background noise, reverberation, and the natural flow of discourse. Furthermore, GMM-based clustering lacked the necessary discriminative strength to distinguish speakers with analogous voice characteristics and was inherently unable to simulate overlapping speech, which frequently occurs in natural dialogues. Because of this, MFCC + GMM systems were a simple and efficient starting point for diarization, but their flaws led to the creation of stronger modeling methods in the years that followed.

2.2 Bayesian Information Criterion (BIC) Segmentation

The Bayesian Information Criterion (BIC) became one of the first and most extensively used ways to determine speaker changes in diarization systems and comparing the likelihoods of single vs. dual Gaussian models, that BIC presents a systematic technique to decide whether two nearby audio segments are better characterized as coming from the same speaker or two independent speakers. This approach works without needing speaker labels ahead of time, which makes it perfect for unsupervised diarization pipelines. Early diarization systems used BIC as a key aspect of their segmentation since it was a mathematically sound way to identify speaker boundaries, especially in recordings of meetings and news broadcasts when the sound settings were quite steady. Research conducted by Reynolds and Torres-Carrasquillo [1] and Chen & Gopinath [2] demonstrated that statistical model comparison methods, such as BIC, may efficiently distinguish homogenous speech segments prior to their transmission to subsequent clustering modules.But BIC-based segmentation had many problems when diarization shifted to more varied and realistic audio settings. BIC depends a lot on the idea that statistical qualities stay the same segments. This is often false in noisy, echoey, or very dynamic conversation recordings done. The method struggles when speaker changes are tiny or when acoustic variances between speakers are small and making it less able to tell the difference between them. As stronger feature representations happen i-vectors [4], x-vectors [7], and self-supervised embeddings like Wav2Vec2.0 [9] and WavLM [11] grew more common, BIC became less useful. This is because newer systems work better with learned embeddings that hold more speaker-specific information. As a result, advanced neural and embedding-based algorithms have mostly taken the place of BIC. These methods are more resilient, scalable, and accurate in a wider range of acoustic situations.

2.3 Diarization Based on i-vectors and PLDA

The i-vector framework was a big step forward for speaker diarization since it made it possible to make small, fixed-length representations of speech segments that may be any length. Classical MFCC + GMM systems used statistical clustering to represent speakers, whereas i-vectors incorporated variability in both speakers and channels in a single low-dimensional space and Dehak et al. [4] were the first to use this total variability modeling method, which put both speaker-specific and environmental elements into one latent space. This enabled diarization systems to function on substantially compressed feature vectors while maintaining essential speaker attributes. Because of this, i-vector technologies made it much easier to accurately diarize televised news, phone calls, and meeting records.Probabilistic Linear Discriminant Analysis (PLDA) is an important part of i-vector diarization. It looks at how similar i-vectors are by modeling variability within and between speakers. Garcia-Romero and Espy-Wilson [5] showed that length-normalized i-vectors used with PLDA scoring worked well and consistently

in a wide range of acoustic settings. Sell and Garcia-Romero [6] further combined i-vectors with Agglomerative Hierarchical Clustering (AHC) to establish a robust baseline for diarization that is used in many systems that have been tested by NIST and DIHARD. I-vectors provided markedly superior speaker discrimination compared to BIC or GMM-based clustering, even with constrained speech data.But even while i-vector/PLDA-based systems did better than older statistical methods, they still had many problems when it came to real-world diarization tasks. One big problem was that they were sensitive to domain mismatch. For example, if we are try i-vectors trained on clear or telephone voice generally didn't work well on far-field or loud microphone recordings. Also, i-vectors weren't very effective at dealing with speech that overlaps, which is prevalent in natural discussions and meetings. The technique depended on fixed statistical assumptions and couldn't learn non-linear speaker features, which is different from newer neural embedding models like x-vectors [7] or ECAPA-TDNN [8]. Even with these problems, i-vector/PLDA pipelines were a key step between traditional acoustic modeling and modern deep learning methods. They also set the stage for embedding-based diarization architectures.

Table 1. Strengths and Weaknesses of Classical and Statistical Diarization Methods

Approach	Pros	Cons
Clustering with MFCC and GMM	• Low computational cost: MFCC + GMM systems don't need a lot of resources and can be used for early diarization and processing in real time. • Works well in clean, regulated spaces: Works well when there isn't much background noise or echo.Bad performance in loud or echoey situations: Noise messes with spectral characteristics, which makes GMM clusters overlap.	• Bad performance in loud or echoey situations: Noise messes with spectral characteristics, which makes GMM clusters overlap. • Can't handle overlapping speech: Only one speaker can be active at a time. • Limited discriminative power: It can't pick up on complicated speaker traits and does worse than newer deep learning models.
Segmentation using the Bayesian Information Criterion (BIC)	• Unsupervised segmentation: No requirement for speaker labels; it works by comparing statistical models. Based on math: Uses likelihood and penalized complexity to find boundaries. • Works best in clear, stationary recordings like early broadcast news when the sound is regulated.Very sensitive to noise and echoes: Statistical assumptions do not hold in varying acoustic environments.	• BIC has a limited ability to tell speakers apart because it can't tell the difference between people with comparable acoustic profiles. • Not good at handling overlapping speech: It only works with one active speaker at a time, which makes it bad for real conversations.
i-vector + PLDA-Based Diarization	• Compact and informative representation: It uses low-dimensional vectors to capture important speaker traits [4]. • Better speaker discrimination: PLDA rating makes it easier to tell speakers apart [5]. • Strong base for a lot of diarization tasks: Widely used in NIST and DIHARD evaluations because it works well [6]. • Better than classical GMM systems because they can handle more data and are less sensitive to changes in sound.	• Sensitive to domain mismatch: Performance diminishes when the training and testing environments are quite different. • Bad at dealing with overlapping speech: It assumes that only one person is speaking at a time and can't represent several speakers overlapping. • Limited ability to represent non-linear relationships: Doesn't do as well as neural embeddings in finding complicated speaker patterns. • Not as good as recent deep learning systems: ECAPA-TDNN, x-vectors, and SSL models like Wav2Vec2.0 and WavLM [7–11] did better.

3. Models for Deep Learning Embedding

Deep learning changed speaker diarization in a big way by letting models learn highly discriminative speaker embeddings that work far better than previous statistical methods for example i-vectors, and deep neural networks train non-linear and hierarchical representations directly from raw or feature-derived voice data, capturing complicated speaker-specific properties and This approach is different from i-vector systems, which use linear total variability modeling. The advent of x-vectors, introduced by Snyder et al. [7], was the initial significant advancement in this domain. Time-Delay Neural Networks (TDNNs) used to extract these embeddings. They combine long-term temporal context and create fixed-length representations that function well for both speaker verification and diarization system. Desplanques et al. [8] built on this work by creating ECAPA-TDNN, a more complex architecture that includes like channel attention, residual connections, and squeeze-and-excitation modules [27] that is make it even easier to tell speakers apart and These new ideas greatly lowered the Diarization Error Rate (DER) for extremely difficult datasets, including VoxCeleb, AMI, and DIHARD.

The deep learning paradigm also made it possible to create hybrid diarization pipelines that combine neural embeddings with more advanced clustering methods with spectral clustering and Variational Bayesian HMM (VB-HMM) re-segmentation. This makes the temporal boundaries more accurate and stable. Deep neural embeddings outperform traditional methods GMM clustering [1–3] or BIC segmentation in terms of generalization in noisy, reverberant, and multi-speaker settings. This change from using statistical models to using data-driven representation learning has made deep learning embeddings a key part of modern diarization systems. They provide the basis for new developments, for example, self-supervised learning and end-to-end diarization architectures.

3.1 x-vectors (Deep Embeddings Based on TDNN)

The emergence of x-vectors was a big step to forward for diarization systems since they used deep neural networks to get very distinct speaker embeddings and as per Snyder et al. [7] proposed x-vectors, which are created using a Time-Delay Neural Network (TDNN) architecture that simulates long-term temporal context with analyzing frame-level auditory information over lengthy periods of time. This helps the network learn strong speaker traits that go beyond short-term spectrum qualities. After going through a number of TDNN layers and statistical pooling combines frame-level activations into a fixed-dimensional representation. This makes it easier to group and compare speakers and X-vectors learn complicated non-linear transformations directly from speaker-labeled data, which makes them far better at separating speakers in embedding space than i-vectors [4–6], which use linear generative modeling. When used with spectral clustering or PLDA scoring, x-vectors make a big difference in the Diarization Error Rate (DER), especially in noisy and conversational contexts. These improvements made x-vectors the basic deep learning method for modern diarization pipelines and set the groundwork for more improvements to the architecture.

3.2 ECAPA-TDNN (Enhanced Channel Attention, Propagation, and Aggregation)

The ECAPA-TDNN architecture developed by Desplanques et al. [8], built on the success of x-vectors and made speaker embedding extraction even more complex by adding advanced architectural modules that improve representational power. ECAPA-TDNN combines channel attention, ResNet-style skip connections, and squeeze-and-excitation (SE) methods [27]. This lets the network dynamically focus on the feature maps and frequency channels that are most important, and its improved embedding extractor uses multi-layer aggregation and propagation blocks to combine information from different time resolutions in a way that works correct and these new features make it easier for ECAPA-TDNN to pick up on small differences between speakers This makes it especially useful for difficult diarization situations, like recordings with a lot of noise, conversations with more than one speaker, and surroundings that transcend domains. Empirical evaluations of the VoxCeleb1 and VoxCeleb2 datasets by [18, 19] demonstrate that ECAPA-TDNN significantly outperforms prior TDNN-based x-vector systems, achieving state-of-the-art results in speaker recognition and diarization tasks and ECAPA-TDNN has become one of the most popular architectures in recent diarization frameworks since it is better at distinguishing between different types of data and is more stable.

3.3 A look at the differences between i-vector, x-vector, and ECAPA-TDNN embeddings

The transition from i-vectors to x-vectors and finally to ECAPA-TDNN embeddings signifies a significant advancement in speaker diarization techniques. I-vectors provide concise low-dimensional representations obtained from total

variability modeling [4]; nonetheless, they are constrained by their linear assumptions and exhibit inadequate robustness in environments characterized by noise, channel variability, or overlapping speech. Deep learning techniques like x-vectors get around these problems by learning non-linear speaker representations directly from enormous amounts of labeled data. This made it much easier to tell speakers apart and lowered the DER [7]. Nonetheless, x-vectors encountered difficulties in navigating intricate acoustic environments.

ECAPA-TDNN improved the quality of embeddings even further by adding things like channel attention and squeeze-and-excitation blocks [8, 27]. This made it work better for both speaker verification and diarization. Even while i-vectors are quicker to train and don't take up much processing power, x-vectors and ECAPA-TDNN are far better at generalizing across domains and are more robust in real-world situations. When used with more complex clustering methods like spectral clustering or VB-HMM re-segmentation, contemporary neural embeddings do far better than older statistical methods, getting top scores on the AMI, CALLHOME, and DIHARD benchmarks [16, 23].

3.4 Wav2Vec 2.0

Baevski et al. [9] unveiled Wav2Vec 2.0, an innovative SSL architecture that acquires contextualized speech representations using contrastive learning and masked prediction. Wav2Vec 2.0 works directly on raw waveforms instead of mel-based inputs like older feature extractors. This lets it learn both high-level speech patterns and fine-grained acoustic characteristics. When used in diarization pipelines, especially with Variational Bayesian Hidden Markov Model (VB-HMM) re-segmentation, Wav2Vec 2.0 achieves very high diarization accuracy, even better than standard deep embeddings like x-vectors and ECAPA-TDNN. Its excellent performance on DIHARD challenge datasets shows that it can handle severe noise, reverberation, and spontaneous conversations. This is why Wav2Vec 2.0 is one of the most important SSL models for diarization.

3.5 HuBERT

Hsu et al. [10] came up with HuBERT (Hidden-Unit BERT), which takes SSL for speech representation to the next level by combining hidden unit clustering with masked prediction. HuBERT doesn't just learn from raw audio; it first makes fake labels by grouping acoustic units together over and over again. These clusters are the targets for a BERT-style masked prediction job. This helps the model acquire speech qualities that are more stable and have more content. HuBERT embeddings have proven to be quite strong when there is noise and more than one speaker, which is prevalent in natural conversations and recordings of meetings. It is very useful for downstream diarization tasks since it may collect phonetic, speaker-specific, and temporal information. This is especially true when speech overlaps or when the domain doesn't match.

3.6 WavLM

Chen et al. [11] developed WavLM, an advanced SSL model meticulously tailored for voice separation, diarization, and challenging multi-talker settings. WavLM, on the other hand, has multi-task pretraining with clear goals for speaker masking, denoising, and predicting multiple speakers. This lets it model complicated sound interactions and deal with difficult recordings with more than one speaker at a time. WavLM has reached the highest level of performance on several diarization benchmarks and is now one of the top diarization models in the world. It is a popular choice for high-performance diarization systems since it can handle a lot of noise and knows how to deal with overlapping sounds.

Table 2. Advantages of SSL Models

Advantages	Description
Better performance when there is a lot of noise	SSL models learn strong representations that stay the same even when there is background noise, echo, and changes in real-world sound.
More compact and strong embeddings	SSL embeddings hold more detailed information about each speaker than i-vectors or x-vectors, which makes it easier to group and separate them.
No requirement for data that has been categorized for training	SSL models learn from a lot of speech that isn't tagged, which cuts down on the need for expensive and time-consuming annotation work.

Good generalization across different areas	Models trained with SSL work well with telephony, broadcast, far-field microphones, and conversational datasets.

4. End-to-End and Hybrid Approaches

Speaker diarization has changed over time from traditional clustering-based systems to end-to-end neural architectures and hybrid re-segmentation algorithms that directly represent how speakers act. Standard diarization pipelines use a series of steps, such as feature extraction, embedding creation, clustering, and refinement. These steps add up to errors and make the system less effective in complicated acoustic situations. To solve these problems, modern methods try to bring all of these parts together into one neural model or make existing pipelines better by using probabilistic refinement methods. End-to-End Neural Diarization (EEND), UIS-RNN, and Variational Bayesian Hidden Markov Models (VB-HMM) are some of the most important new ideas that have come out. They make diarization much more accurate, especially in real-world conversations and when people are talking at the same time.

4.1 End-to-End Neural Diarization (EEND)

Fujita et al. [13, 14] presented End-to-End Neural Diarization (EEND) as revolutionary method that immediately forecasts speaker activity for many speakers, eliminating the need for a distinct clustering phase. EEND treats diarization as a supervised sequence prediction issue with permutation-free aims, while embedding-based methods approach clustering as an unsupervised post-processing step. This training method eliminates the problem of label ambiguity that often happens when there are multiple speakers and lets the model consistently assign speaker activity labels over time. One of the best things about EEND is that it works well in situations when speech overlaps, which is still a big problem for standard diarization pipelines. EEND greatly lowers the Diarization Error Rate (DER) in conversational and meeting datasets where people often talk at the same time by modeling speaker overlaps and temporal dependencies at the same time. EEND is a big step forward in the search for fully integrated diarization solutions.

4.2 UIS-RNN

Zhang et al. [12] put forward UIS-RNN (Unbounded Interleaved-State Recurrent Neural Network) as a supervised alternative to clustering-based diarization. Instead of using unsupervised clustering, UIS-RNN uses a recurrent neural network that directly simulates changes in speakers and changes over time. EEND predicts activity from more than one speaker at the same time, while UIS-RNN gives each speaker a label one at a time and learns the chance of staying with the same speaker or switching to a new one. This helps UIS-RNN better understand long-range dependencies and how people talk to each other. UIS-RNN employs supervised training, which means it may use labeled diarization datasets and work in varied acoustic circumstances as long as there is enough training data. The model has been successfully used with x-vector embeddings [7] to create hybrid supervised diarization systems that work better than unsupervised clustering methods.

4.3 VB-HMM Re-Segmentation

Variational Bayesian Hidden Markov Model (VB-HMM) re-segmentation is a very useful way to increase the temporal alignment and speaker boundary accuracy of embedding-based diarization pipelines. VB-HMM was first used for speaker detection and diarization evaluation [6]. Since then, it has become a popular post-processing technique in competitive diarization systems, including entries to DIHARD challenges [25]. VB-HMM enhances diarization efficacy by integrating temporal continuity and modeling speaker transition probabilities more proficiently than conventional clustering methods. It smooths out the boundaries between segments that come from embedding and cuts down on fragmentation mistakes, which are typical in pipelines that use AHC or spectral clustering. When used with embeddings from i-vectors [6], x-vectors [7], ECAPA-TDNN [8], or SSL models like Wav2Vec 2.0 [9], VB-HMM always improves diarization accuracy by making both segment duration and speaker assignment more constant.

5. Comparative Performance Summary

The evolution of speaker diarization techniques indicates a distinct tendency towards more resilient and discerning modeling methodologies. Classical systems like MFCC + GMM are easy to use and don't cost much to run, but they don't work well in noisy or overlapping situations, which leads to high Diarization Error Rates (DER). The addition of i-vector + PLDA made things better by making speaker representations smaller and clustering performance more stable. But a big step forward happened when deep learning-based embeddings were used. X-vectors (TDNN) cut DER by a lot by finding complicated patterns that

are unique to each speaker. ECAPA-TDNN took this progress even further by adding channel attention and multi-layer feature aggregation to make it more stable, especially in noisy and echoey environments. The biggest improvements in performance came from Self-Supervised Learning (SSL) models, especially Wav2Vec 2.0 paired with VB-HMM. This model presently has the best diarization performance across a wide range of datasets since it uses unlabeled speech pretraining and overlap-aware refinement. In general, modern SSL-based methods always work better than classical and deep learning-based systems. They show better generalization and robustness in real-world diarization situations.

Table 3. Comparative Efficacy of Principal Diarization Methodologies

Model / Approach	DER (%)	Key Strengths
MFCC + GMM [1]	12.8	Low resource requirement; simple implementation
i-vector + PLDA [6]	9.4	Small representation; stable grouping
x-vector (TDNN), [7]	6.3	Very accurate; good at telling the difference between things
ECAPA-TDNN, [8]	5.8	Noise resilience; cutting-edge deep embedding model
Wav2Vec 2.0 + VB-HMM [9, 11, 25]	4.7	Best overall performance; better handling of generalization and overlap

6. Identified Research Gaps

Even though there have been big improvements in speaker diarization, such as moving from classical MFCC–GMM approaches [1–3] and i-vector/PLDA systems [4–6] to modern deep embeddings [7–8] and self-supervised models [9–11], there are still some problems that need to be solved before diarization systems can be used in the real world. State-of-the-art designs have significantly lowered the Diarization Error Rate (DER), especially on benchmark datasets like AMI, CALLHOME, and DIHARD [16, 23]. However, there are still many important research gaps that need to be filled. To make strong, multilingual, and real-time diarization solutions, especially in places like Indian languages and mobile apps, we need to fill in these gaps. The next subsections list the main gaps that have been found in existing research and the difficulties that our study is trying to solve.

6.1 Overlapping Speech

One of the biggest problems with speaker diarization is that it can't handle speech that overlaps a lot, especially when the overlap is more than 30–40%. Even though end-to-end models like EEND [13, 14] and SSL-based representations like WavLM [11] have made it easier to find overlaps, performance still drops a lot in real-world conversations where more than one person is talking at the same time. Classical pipelines and even x-vector or ECAPA-TDNN systems treat overlap as noise, which means that speech is ignored, segments are assigned incorrectly, and DER is too high. Real-world diarization, especially in meetings, homes, and noisy Indian settings, necessitates models that can achieve multi-speaker separation without relying on extensive annotated datasets.

6.2 Domain Mismatch

Diarization models are very sensitive to domain mismatch, which is the difference between the conditions in which they were trained and the situations in which they are tested in the actual world. For instance, algorithms that were trained on clean or controlled datasets don't work well with far-field microphones, telephone speech, consumer-grade mobile recordings, or echo-rich indoor settings. Research indicates that even high-performing SSL models, including Wav2Vec 2.0 [9] and WavLM [11], demonstrate significant performance declines when faced with inconsistent acoustic settings. This is still a big problem for using diarization in real-world processes, especially when it comes to analyzing massive amounts of audio, using mobile devices, and working in multilingual settings.

6.3 Languages with Few Resources (Including Our Work)

There is a big vacuum in research on how to make diarization systems for languages with less resources, especially Indian languages like Hindi, Marathi, Telugu, and Kannada. Most diarization models are trained on English-centric datasets like VoxCeleb [18, 19], AMI [17], or DIHARD [16, 23]. These datasets don't show how Indian speech sounds, how people talk, or how phonetic diversity works.

Our present study especially addresses this issue by constructing speaker diarization and speaker identification verification algorithms for Marathi speech, utilizing proprietary datasets and diverse acoustic contexts. The lack of extensive multilingual diarization corpora complicates the development of resilient ASR-driven diarization and multi-speaker segmentation technologies.

This research gap underscores the necessity for transfer learning, cross-lingual modeling, and self-supervised pretraining specifically designed for underrepresented languages.

6.4. Limitations on Real-Time Deployment

Another big problem is getting diarization models to work in real time. Wav2Vec 2.0 and WavLM are examples of state-of-the-art SSL models that need strong GPUs for both training and inference. Because of this, they aren't good for edge devices, Android apps, IoT sensors, and embedded systems, where processing power and memory are restricted. Even lightweight deep embeddings like x-vectors [7] may have a hard time working well when latency is very low unless they are heavily optimized. Our research also helps fill this gap by focusing on effective diarization methods for mobile and on-device contexts. This makes it possible to use speaker segmentation and identity authentication in real-world applications.

6.5 There aren't any common benchmarks or standards for evaluation.

A significant deficiency in contemporary diarization studies is the absence of cohesive, domain-inclusive benchmarks. Most evaluation datasets are for certain areas, including conference rooms (AMI), phone conversations (CALLHOME), or very loud environments (DIHARD). Because of this, models that work well in one context typically don't work well in another, which makes it hard to generalize across domains. Also, discrepancies in evaluation methodologies, SAD systems, scoring pipelines, and overlap detection criteria make it hard to compare papers. There is an urgent need for standardized multilingual, multi-domain diarization benchmarks that appropriately reflect real-world deployment situations, including spontaneous Indian languages, hybrid acoustic sources, and mobile device audio.

Table 4. A summary of the gaps in research and what we did to fill them

There is a gap in research	Problem	What We Contributed to Our Work
Speech that overlaps	Bad performance when the overlap is more than 30–40%	Investigating hybrid SSL and clustering models for multi-speaker Marathi speech.
Mismatch in the domain	Degradation in telephony, far-field, and noisy data	Making diarization models with different Marathi datasets, like mobile, meeting, and outdoor data
Languages with few resources	There aren't enough datasets for Marathi and Hindi.	Creating a system for Marathi speakers to keep track of their conversations and confirm their identities
Deployment in real time	SSL models are too big for devices that are built in.	Targeting MFCC/Mel-based diarization on Android devices
Unified tests	There are no standard sets that work in more than one language or area.	Making organized Marathi exam sets for research on diarization

7. Conclusion

Over the past twenty years, speaker diarization has changed a lot. It has gone from classical MFCC-GMM clustering methods [1–3] to statistical modeling with i-vectors and PLDA [4–6], and finally to deep learning-based x-vectors and ECAPA-TDNN embeddings [7–8]. Self-supervised learning models like Wav2Vec 2.0, HuBERT, and WavLM [9–11] have changed the field even more by letting diarization systems build strong speech representations without needing a lot of labeled data. These enhancements have made the Diarization Error Rate (DER) much better, made it more resistant to noise, and made it possible to mimic complicated speakers in real-world settings.

Even with these successes, there are still some big problems to solve. When people talk to each other when there are more than one person talking at once, overlapping speech still causes a lot of diarization errors. Domain mismatch is still a problem that won't go away. Models trained on clean datasets generally don't work well when they are used in telephony, outdoors, or with microphones that are far away. Low-resource languages, such as Marathi and numerous other Indian languages, are inadequately represented in diarization research owing to the limited availability of annotated datasets. Also, it is hard

to use state-of-the-art SSL models in real-time and resource-limited settings, like Android apps, which makes them less useful in practice.

In general, this assessment shows how quickly things are changing in the field of diarization and how many problems still need to be solved. It emphasizes the necessity of creating diarization solutions that are not only precise but also computationally efficient, multilingual, and adaptable to real-world contexts. The findings from this study establish a robust basis for the progression of diarization technologies and the direction of future research endeavors, including the contributions of our current initiatives in Marathi speaker diarization and streamlined on-device diarization systems.

Future Work

Future research in speaker diarization should concentrate on enhancing robustness, multilingual support, and real-time deployment capabilities. Even while deep learning and self-supervised models have greatly cut down on diarization errors, problems like overlapping speech, domain mismatch, and the lack of datasets for low-resource languages are still not solved. Our current work on Marathi speaker diarization helps fill in gaps in language resources and create fast, on-device diarization systems that operate well in real-world Indian settings. Future directions also stress the need for lightweight model design for embedded apps, multilingual benchmarks that function together, and better interface with ASR and conversational AI frameworks.

Important Future Work Directions

- Better handling of overlap with hybrid EEND + SSL models.
- Adapting to telephony, far-field, and loud real-world sound circumstances.
- Multilingual SSL and dataset development for low-resource languages like Marathi and Hindi.
- Real-time on-device diarization using model compression, pruning, and lightweight neural architectures.
- Unified multilingual benchmarking to make it possible to compare things across languages and fields.
- Integration with ASR/NLP systems for meeting analytics, assigning roles, and conversational intelligence.

References

[1] D. A. Reynolds and P. A. Torres-Carrasquillo, “Approaches and applications of audio diarization,” ICASSP 2005 Proceedings, vol. 5, pp. 953–956, IEEE, 2005.

[2] S. S. Chen and P. S. Gopinath, “Gaussian mixture models for speech processing,” IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, pp. 23–41, 2000.

[3] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2011–2022, 2007.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[5] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” Interspeech 2011, pp. 249–252.

[6] G. Sell and D. Garcia-Romero, “Speaker diarization with PLDA-based clustering,” ICASSP 2014, pp. 3692–3696, IEEE.

[7] Sujan Hiregundagal Gopal Rao. (2023). A Review of Cybersecurity Threats in Automotive Semiconductor Control Units. International Journal of Intelligent Systems and Applications in Engineering, 12(1), 927–932.

[8] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” Interspeech 2020, pp. 3830–3834.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” NeurIPS 2020, pp. 12449–12460.

[10] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

[11] S. Chen, C. Wang, Z. Chen, et al., “WavLM: A unified pre-trained model for speech processing tasks,” NeurIPS 2022, pp. 1–14.

[12] Y. Zhang, J. Geiger, J. Azcarreta, S. Khudanpur, and D. Povey, “Fully supervised speaker diarization,” ICASSP 2019, pp. 6301–6305.

[13] Y. Fujita et al., “End-to-end neural speaker diarization with permutation-free objectives,” IEEE SLT Workshop, pp. 190–197, 2018.

[14] Y. Fujita, S. Watanabe, et al., “End-to-end speaker diarization for an unknown number

of speakers with encoder–decoder attractors," Interspeech 2020, pp. 876–880.
[15] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," arXiv preprint arXiv:2106.04624, 2021.
[16] K. Ko, M. Zelenak, et al., "The 2018 DIHARD Speech Diarization Challenge," Interspeech 2018, pp. 2798–2802.
[17] J. Carletta, "Unleashing the potential of the AMI meeting corpus," Machine Learning for Multimodal Interaction, vol. 3869, pp. 1–10, Springer, 2005.
[18] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," Interspeech 2017, pp. 2616–2620.
[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," Interspeech 2018, pp. 1086–1090.
[20] D. Povey et al., "The Kaldi speech recognition toolkit," IEEE ASRU Workshop, 2011.
[21] J. Villalba et al., "State-of-the-art speaker recognition with neural embeddings: A review," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1692–1715, 2022.
[22] S. Watanabe et al., "CHiME-6 challenge: Automatic speech recognition, diarization, and multi-channel processing," Interspeech 2020, pp. 1–5.
[23] N. Ryant, et al., "The Third DIHARD Diarization Challenge," Interspeech 2021, pp. 1–5.
[24] J. Thienpondt, B. Desplanques, and K. Demuynck, "Robust speaker recognition with domain-adversarial training," IEEE/ACM TASLP, vol. 32, pp. 120–130, 2024.
[25] S. Horiguchi et al., "The Hitachi/JHU DIHARD III system: Improvements to VB-HMM diarization and multi-stream ASR," arXiv preprint arXiv:2102.01363, 2021.
[26] A. Bredin and G. Chollet, "Segmental speech diarization," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1180–1190, 2007.
[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," CVPR 2018, pp. 7132–7141.
[28] J. Xu et al., "Self-supervised learning for speaker diarization: A review," arXiv preprint arXiv:2301.00087, 2023.
[29] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," Interspeech 2018, pp. 2252–2256.
[30] R. Yin et al., "Neural speech segmentation and diarization: A review," IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 84–96, 2022.