

Enhanced Speaker Embeddings With Adaptive Feature Aggregation For Robust Speaker Diarization

Vijay Hothi^{1*}, Dr. N B Gohil², Dr. Mehul K Vala³, Dr. Devang G Jani⁴, Dr. Dhaval R Bhojani⁵, Prof. Rushi J. Trivedi⁶

¹Research Scholar, Gujarat Technological University, Gujarat

²Associate Professor, Government Engineering College, Rajkot

³Assistant Professor, Shantilal Shah Engineering College, Bhavnagar, Gujarat

⁴Associate Professor, Government Engineering College, Bhavnagar, Gujarat

⁵Assistant Professor, Government Engineering College, Rajkot, Gujarat

⁶Assistant Professor, Government Engineering College, Rajkot, Gujarat

*Corresponding Author: Vijay Hothi

Abstract

Speaker diarization, is the process of identifying "who spoke when" in recordings of multiple speakers, is central to applications like meeting transcription, conversational AI, and speech analytics. Yet real-world diarization is hindered by overlapping speech, background noise, and channel variability that corrupt the performance of standard embedding-based systems. To tackle these challenges, they introduce an Enhanced Speaker Embedding approach based on an Adaptive Feature Aggregation Time-Delay Neural Network (AFA-TDNN) for speaker diarization that is robust. The method innovatively adds dynamic feature weighting, multi-scale temporal aggregation, and residual feature fusion to acquire discriminative embeddings. A Batch-Adaptive Data Augmentation (BADA) strategy, which includes spectral masking, noise injection, temporal perturbation, and reverberation simulation, is utilized to improve robustness in varying acoustic conditions. The embeddings are then aggregated with spectral clustering to obtain speaker segmentation with high accuracy even from noisy and overlapped conditions. Experimental evaluation on the AMI Meeting Corpus shows significant gain, with a Diarization Error Rate (DER) of 2.23% and a Jaccard Error Rate (JER) of 5.9%, which beats current baselines. The overall framework guarantees strong, noise-tolerant, and highly accurate speaker diarization and is applicable for real-world multi-speaker applications.

Keywords: Speaker Diarization, AFA-TDNN, BADA, DER, AMI Meeting Corpus.

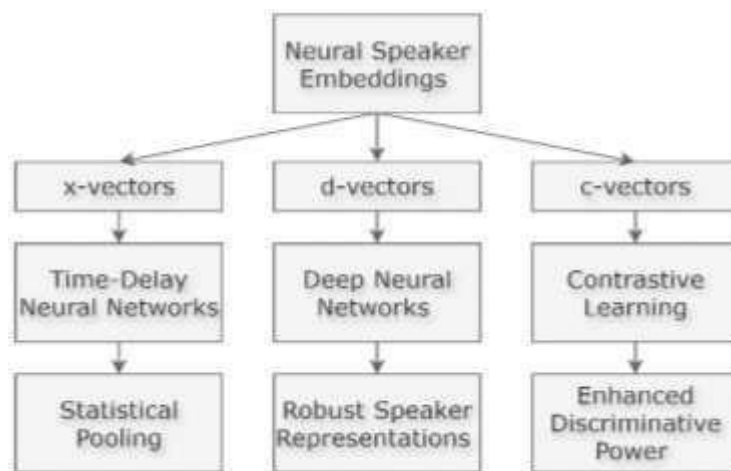
1. Introduction

Speaker diarization, or the allocation of an audio stream into speaker-homogeneous segments based on speaker identity, has become an important task in meeting transcription, broadcast media analysis, call centre monitoring, and conversational AI systems [1, 2]. As opposed to "Automatic Speech Recognition (ASR)", which revolves around speech-to-text conversion, speaker diarization tries to solve the issue of knowing "who spoke when" in multi-speaker settings. Though tremendous progress has been made over the last few years, diarization continues to be problematic owing to variability in speech properties, simultaneous conversations, background noise, and channel distortions [3]. To compensate for these challenges, recent studies have placed stress on the need for speaker embeddings that are robust and coupled with sophisticated aggregation methods capable of extracting discriminative speaker-specific information while being insensitive to environmental and conversational variations [4, 5].

Speaker diarization is especially difficult in the case of overlapping speech, where two or more speakers simultaneously utter words. In contrast, additional background noise adds even more complication to the detection of speaker changes, which tends to mislead

“Deep Learning (DL)” models and lead to incorrect diarization boundaries [6]. In order to counter these problems, various techniques have been tried, including multi-channel signal processing [7], [8], attention mechanisms [9], [10], and end-to-end neural systems [11], [12], which all try to encode speaker traits and temporal dynamics in a better way. In reality, “Speaker Diarization (SD)” carries out segmentation based on the speaker’s identity on an audio recording. Fig. 1, the conventional method generally undergoes a multistage process, incorporating speaker segmentation, local embedding extraction, and clustering.

Figure 1: Recent advancements in speaker diarization



1.1 Limitations of Traditional Speaker Diarization Approaches

Traditional diarization pipelines usually comprise four stages: speech activity detection, feature extraction, speaker embedding generation, and clustering. Though all these stages together form the diarization result, every step is a potential source of error [13]. For example, standard feature extraction strategies are based on “Mel-Frequency Cepstral Coefficients (MFCCs)”, which, though good for simple speech analysis, tend to miss subtle speaker-specific information in noise or overlap speech scenarios. Likewise, earlier representations like i-vectors have issues of performance drop in cases of domain mismatch, rendering them inappropriate for large-scale, real-world tasks [14, 15].

Recent developments in neural speaker embeddings have pushed the state-of-the-art significantly. x-vectors, obtained from “Time-Delay Neural Networks (TDNNs)” with statistical pooling, yield fixed-dimensional vectors from varying-length utterances [16]. d-vectors, obtained using deep neural networks, derive strong speaker-specific features for improving separation over varying conditions [17]. c-vectors use contrastive learning to improve discriminative power across speaker identities [18]. All these embedding methods have together boosted speaker verification and diarization performance, with improved generalization and flexibility.

Nonetheless, the clustering process of diarization remains challenging. As in unsupervised approaches like “Agglomerative Hierarchical Clustering (AHC)” or k-means, embedding quality and similarity metrics are extremely sensitive [19]. Any slight inconsistency in embeddings can introduce drastic diarization errors, especially under multi-speaker settings with short or overlapped utterances [20]. Moreover, static pooling approaches equally treat all frames without any consideration of speaker

dynamics variability, conversational turns, and prosodic features. This restriction does not allow diarization models to cope with real-world situations like noisy telephonic calls or multi-party meetings [21].

1.2 Enhanced Speaker Embeddings with Adaptive Feature Aggregation

To address these issues, stronger speaker embeddings with adaptive feature aggregation have been proposed as promising solutions. In contrast to static techniques, adaptive methods allocate different weights to speech frames dynamically based on their speaker-identity relevance [22]. Techniques based on attention mechanisms, self-supervised learning, and graph-based aggregation frameworks become progressively popular for emphasizing the most speaker-informative areas and masking noise and irrelevant artifacts. This renders embeddings more resilient to overlapping speech and environmental fluctuations [23].

Latest architectures, like the ECAPA-TDNN model [24], have further improved speaker verification by adding channel attention and feature aggregation mechanisms and exhibiting better robustness in verification tasks. Their deployment to speaker diarization, particularly in noisy or distant-talking cases, has yet to be fully explored. Spurred by this shortfall, this paper presents the “Adaptive Feature Aggregation Time-Delay Neural Network (AFA-TDNN)”, an architecture especially designed for speaker diarization with robustness [25].

The AFA-TDNN uses dynamic feature weighting, multi-scale temporal fusion, and a sophisticated augmentation strategy in the creation of speaker embeddings that are both discriminative and robust [26]. In contrast to the previous models, AFA-TDNN utilizes a batch-adaptive augmentation scheme, which injects various perturbed instances of the input in the form of a single training batch in a dynamic manner. This improves robustness to acoustic perturbations, thereby making the model more efficient in the processing of noisy, overlapped, and natural conversational speech [27]. Through the refinement of speaker-specific representations and dynamically combining features, the proposed system improves the state of the art in terms of diarization accuracy and dependability.

In this paper, author contributed the following.

- 1 AFA-TDNN architecture with dynamic feature weighting and multi-scale temporal aggregation for improved speaker discriminability.
- 2 Batch-adaptive augmentation pipeline combining spectral augmentation, dynamic noise injection.
- 3 Extensive evaluation on the AMI meeting corpus, demonstrating improvements, with DERs as low as 2.23% on the evaluation set.

The rest of the paper is organized as follows:

Section 2 describes related work done in that domain, Section 3 describes the AFA-TDNN diarization system, Section 4 details the experimental setup, Section 5 presents results and comparisons, and Section 6 concludes with future directions.

2. Related work

Speaker diarization (SD) continues to pose significant challenges due to the complexity of accurately attributing speech segments to individual speakers in real-world scenarios. Overlapping speech, acoustic variability, and diverse speaker characteristics make this problem particularly difficult, as noted by Sell et al. (2018) [28]. To advance solutions, evaluation campaigns such as DIHARD Ryant et al., (2020), [29] DISPLACE (Baghel et al., 2023), [30] VoxSRCC(Brown et al., 2022), [31] and the NIST Speaker Recognition Evaluation have established benchmarks that fuel innovation. Within these contexts,

neural speaker embeddings have played a central role. For instance, x-vectors Lei et al., (2014) [32] utilize TDNNs and statistical pooling to generate fixed-dimensional representations, while d-vectors Li et al., (2015) [33] leverage deep neural networks for robust speaker-specific features. Similarly, c-vectors Sun et al., (2021) [34] apply contrastive learning to enhance discriminability. Collectively, these embedding techniques marked a shift away from traditional i-vector approaches, providing the basis for modern diarization frameworks.

Building upon these foundations, researchers have explored multiple avenues to improve robust diarization. Carletta et al. (2005) [35], for example, evaluated diarization performance using the widely recognized AMI meeting dataset, which includes multiple audio stream types and conversational dynamics. Their results demonstrated the value of large-scale, naturalistic corpora for testing diarization systems and benchmarking progress. Beyond dataset-driven evaluations, methodological advancements have also driven improvements. Kwon et al. (2021) [36] showed that embedding adaptation through dimensionality reduction, attention-based aggregation, and non-speech clustering reduced diarization error rates by 25.07%. Similarly, Li et al. (2024) [37] enhanced end-to-end neural diarization (EEND) by applying it both locally and globally without explicit embeddings, yielding DER reductions of 13% and 10% on the Call home and RT03-CTS datasets. In another direction, Yaqian et al. (2024) [38] proposed AAMC-SoftMax, which balances intra-class compactness with inter-class separability, surpassing x-vector-based ECAPA-TDNN models by 43.9% and 39.7% on near- and far-field AMI datasets. These works, though diverse in focus, collectively highlight the importance of datasets, embedding adaptation, and training objectives in advancing SD performance.

More recent work has focused on addressing real-world complexities such as domain mismatch, noisy conditions, and long-form conversations. He et al. (2023) [39] introduced the ANSD-MA-MSE framework, which dynamically refines embeddings through a memory-aware multi-speaker mechanism, thereby enhancing robustness under mismatched acoustic conditions. Complementarily, cord et al. (2023) [40] developed a teacher-student embedding system at frame rate with block-wise processing; significantly reducing performance drops when diarizing conversations involving a large number of speakers. In parallel, Alvarez et al. (2025) [41] explored feature fusion by integrating ECAPA-TDNN embeddings with Geneva Minimalistic Acoustic Parameter Sets (GeMAPS) in the EEND-EDA model, achieving a 29% DER reduction on the CallHome corpus. Together, these studies illustrate a clear progression: while earlier contributions centred on strengthening embedding representations, recent research increasingly emphasizes adaptive architectures, multimodal feature fusion, and scalable solutions capable of generalizing across diverse and challenging diarization environments. Despite these advancements, existing studies still struggle to achieve consistent performance in scenarios with highly overlapping speech, background noise, and distant-talking conditions. Moreover, while models such as ECAPA-TDNN and EEND have improved robustness, their application to adaptive feature aggregation for speaker diarization remains underexplored. Thus, there is a pressing need for architectures that can dynamically weight features and enhance speaker embeddings to ensure reliable diarization across diverse and challenging acoustic environments.

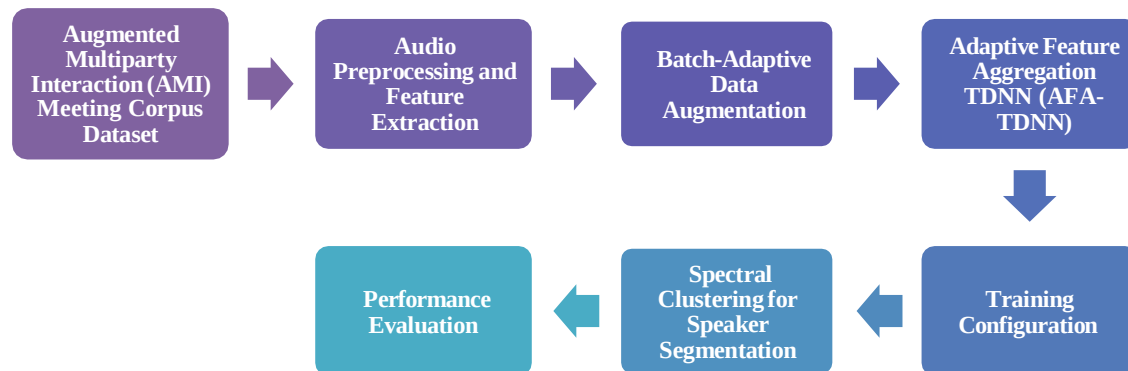
Speaker diarization, the task of segmenting and tagging audio recordings with speaker information, still challenges researchers and practitioners in achieving good accuracy due to variation in acoustic contexts, overlap, background noise and speaker variability. Traditional and common speaker embeddings, e.g. x-vectors and d-vectors capture insufficient speaker-discriminative characteristics even in known noisy and informal contexts, leading to bad diarization performance. Furthermore, the feature aggregation

methods often employed in combination with x-vectors and d-vectors are characteristic and do not accommodate potentially dynamic acoustic contexts, making them less robust and generalizable. Hence, they require an improved method utilizing variations of speaker embedding and incorporate adaptive methods of feature aggregation to account for multiple speakers features to potentially enhance speaker disambiguation power and reliability for speaker separation tasks involving noise, overlap speech and environmental constraints.

3. Research Methodology

The proposed research methodology presents an advanced speaker diarization framework integrating “Adaptive Feature Aggregation Time-Delay Neural Network (AFA-TDNN)” with Spectral Clustering for robust speaker segmentation. The process begins with audio preprocessing and feature extraction from the AMI Meeting Corpus, where multi-channel recordings are enhanced using BeamformIt and transformed into 80-dimensional log-Mel filter bank features. In an effort to enhance performance against increased generalization, BADA adopts Batch Adaptive Data Augmentation, including the introduction of spectral masking, noise injection, temporal perturbation, and reverberation simulation. The AFA-TDNN extracts discriminative speaker embeddings with noise resiliency, while spectral clustering is used to separate speakers effectively in noisy and overlapping environments.

Figure 2: Framework of Research Methodology



3.1 Dataset Collection

All experiments in the current study use the “Augmented Multiparty Interaction (AMI)” Meeting Corpus, a popular benchmark dataset for speech recognition and speaker diarization tasks. The AMI corpus contains simulated and real multi-party meeting recordings captured under control and natural acoustic conditions. The dataset contains multiple microphone types, including headset microphones, close-talking microphones, lapel microphones, and microphone arrays, all with varied audio quality suitable for test-resilient model evaluation.

Each of the audio recordings in the corpus is prefixed with manually annotated word-level transcripts and time-aligned speaker labels, allowing precise training and testing of diarization models. The meetings include four non-native speakers of English discussing real and scenario-based meetings, bringing in natural variability in speech patterns, accents, and overlap. The richness of the multi-channel recordings, together with rich annotations, makes the AMI corpus a perfect option for the development and testing of

advanced speaker embedding methods. The dataset is released under a Creative Commons License [42].

3.2 Audio Preprocessing and Feature Extraction

In the framework to be proposed, pre-processing of audio and feature extraction are key to obtaining strong speaker representations for Enhanced Speaker Embeddings with Adaptive Feature Aggregation. The system consists of two primary steps: multi-channel audio enhancement and acoustic feature calculation.

• Audio Preprocessing

To improve raw multi-channel recordings quality, researcher uses the BeamformIt toolkit [43], which uses delay-and-sum beam forming to minimize background noise and enhance “Signal-To-Noise Ratio (SNR)”. The beam forming procedure does the following: It takes signals from M microphones and aligns them according to estimated time delays τ_m and sums them to get a cleaner signal. Mathematically, the enhanced audio signal $y(t)$ is computed as:

$$y(t) = \frac{1}{M} \sum_{m=1}^M x_m(t + \tau_m)$$

where $x_m(t) \rightarrow$ input signal from the m^{th} microphone, $\tau_m \rightarrow$ estimated delay for the m^{th} channel, $M \rightarrow$ total number of microphones. This approach effectively suppresses interfering sounds and enhances speaker-relevant acoustic information, which is crucial for robust speaker diarization.

• Feature Extraction

After beamforming, they extract 80-dimensional log-Mel “Filter Bank (FBANK)” energies from each enhanced audio segment. The Mel filter bank energies are computed by passing the “Short-Time Fourier Transform (STFT)” magnitude spectrum $|X(f)|^2$ through a set of triangular Mel-scaled filters $H_m(f)$. The Mel-energy for the m^{th} filter is computed as:

$$E_m = \sum_f |X(f)|^2 \cdot H_m(f)$$

Subsequently, the logarithmic compression is applied to obtain log-Mel features:

$$F_m = \log E_m$$

Finally, to ensure distributional stability and reduce channel-specific variability, mean and variance normalization is applied to each segment:

where $F_m \rightarrow$ original log-Mel feature, $\mu \rightarrow$ segment-wise mean and $\sigma \rightarrow$ segment-wise standard deviation. This two-stage pre-processing pipeline—BeamformIt-based multi-channel enhancement followed by log-Mel FBANK feature extraction and normalization—ensures that the input to the proposed AFA-TDNN architecture is noise-suppressed, channel-invariant, and speaker-discriminative, enabling more robust and accurate diarization performance.

3.3 Batch-Adaptive Data Augmentation (BADA)

To enhance the robustness and generalization capability of the proposed AFA-TDNN model across diverse and challenging acoustic environments, to implement a BADA strategy during training. This technique introduces controlled perturbations to the input audio data, allowing the model to learn speaker-invariant embeddings that remain stable under noise, channel variations, and reverberation effects.

For each training utterance X , they generate six perturbed versions $\{X1, X2, X3, X4, X5, X6\}$ using multiple augmentation techniques, effectively expanding the diversity of the training batch. The augmented features are concatenated with the original features within the same mini batch, increasing the effective batch size from B to $B'=6B$.

- **Spectral Augmentation (Spec Augment)**

To simulate channel distortions and missing frequency information, random frequency masking is applied on the log-Mel FBANK features.

$F \in \mathbb{R}^{T \times M}$, where T is the number of time frames and M is the number of Mel bands:

$$\tilde{F}(t, m) = \begin{cases} 0 & \text{if } m \in [f_0, f_0 + F_w] \\ F(t, m), & \text{otherwise} \end{cases}$$

Where f_0 = starting frequency band (randomly selected) and F_w = width of the masked frequency region. This forces the network to rely on contextual information rather than fixed frequency patterns, improving robustness to microphone channel mismatches.

- **Dynamic Noise Injection**

To improve noise robustness, additive noise N is sampled from the MUSAN dataset [44] and mixed with the original waveform X at a randomly selected SNR in the range 0–12 dB:

$$X_{\text{noisy}} = X + \alpha N$$

Where:

$$\alpha = \sqrt{\frac{P_x}{P_N \cdot 10^{\text{SNR}/10}}}$$

- P_x = power of clean speech
- P_N = power of noise signal
- α = noise scaling factor

This ensures the model learns to distinguish speaker characteristics even under varying noise levels.

- **Temporal Perturbation**

To simulate variations in speaker tempo and conversational dynamics, researchers apply speed perturbation and time-domain shifts during the data augmentation process. In speed perturbation, the audio speed is modified by a scaling factor $s \in \{0.9, 1.0, 1.1\}$, which either slows down, maintains, or speeds up the original signal. Mathematically, the transformation is represented as:

$$X_{\text{speed}}(t) = X(st)$$

where $X(t)$ is the original speech signal, s is the speed scaling factor, and $X_{\text{speed}}(t)$ is the perturbed signal. Additionally, time shifting is applied by introducing a random temporal offset Δt , defined as:

$$X_{\text{shift}}(t) = X(t + \Delta t)$$

Where Δt is uniformly sampled within a predefined range. These techniques effectively mimic natural variations in speaking rate and speech alignment, helping the model become more robust and invariant to speaker tempo fluctuations and asynchronous speaker turns.

- **Reverberation Simulation**

Real-world conversations often occur in environments with varying room acoustics. To simulate reverberation, the clean speech signal X is convolved with a “Room Impulse Response (RIR)” filter $h(t)$ from the RIR dataset [45]:

$$X_{\text{reverb}}(t) = X(t) * h(t)$$

Where $*$ denotes convolution. This exposes the model to different acoustic environments and improves domain adaptability.

- **Batch Concatenation and Consistency Loss**

After generating six augmented versions for each utterance, the effective batch size increases:

$$B' = B \times (1 + N_{\text{aug}})$$

Where:

- B = base batch size (e.g., 32)
- N_{aug} = number of augmentations per utterance (here, 6)
- $B' = 32 \times (1 + 6) = 192$

To enforce speaker embedding consistency across all augmentations, introduce a “Speaker Consistency Loss (L_{SC})”:

$$L_{\text{SC}} = \frac{1}{N_{\text{aug}}} \sum_{i=1}^{N_{\text{aug}}} \|E_i - E_0\|^2$$

Where E_0 = embedding of original utterance and E_i = embedding of the i^{th} augmented version. This ensures that embedding’s generated from the same speaker remain close in feature space, regardless of distortions.

- **Final Augmentation Strategy**

The complete BADA process can be summarized as:

$$X' = \{X, X_{\text{spec}}, X_{\text{noisy}}, X_{\text{speed}}, X_{\text{shift}}, X_{\text{reverb}}\}$$

Where X' represents the augmented training set containing seven variations per utterance (one original + six augmented).

This adaptive augmentation strategy significantly improves generalization, noise robustness, and speaker invariance, allowing the proposed AFA-TDNN model to perform effectively in real-world multi-speaker environments.

3.4 Adaptive Feature Aggregation TDNN (AFA-TDNN)

The proposed AFA-TDNN is an enhancement of the traditional x-vector TDNN framework [16], designed to extract robust and discriminative speaker embeddings. Unlike the standard x-vector architecture, AFA-TDNN introduces adaptive feature aggregation mechanisms that dynamically emphasize speaker-specific characteristics while suppressing irrelevant noise and channel distortions. The architecture comprises five key components, described below with relevant mathematical formulations.

- **Dynamic Feature Weighting**

To capture speaker-relevant information under diverse acoustic conditions, AFA-TDNN employs a modified “Squeeze-and-Excitation (SE)” block. For an input feature map $\mathbf{F} \in \mathbb{R}^{T \times M}$, where T represents time frames and M represents feature dimensions, global context statistics are first computed using temporal average pooling:

$$\mathbf{z} = \frac{1}{T} \sum_{t=1}^T \mathbf{F}(t)$$

The statistics vector $\mathbf{z}$ is passed through a two-layer fully connected gating mechanism with nonlinear activations:

$$\mathbf{s} = \sigma(W_2 \cdot \delta(W_1 \cdot \mathbf{z}))$$

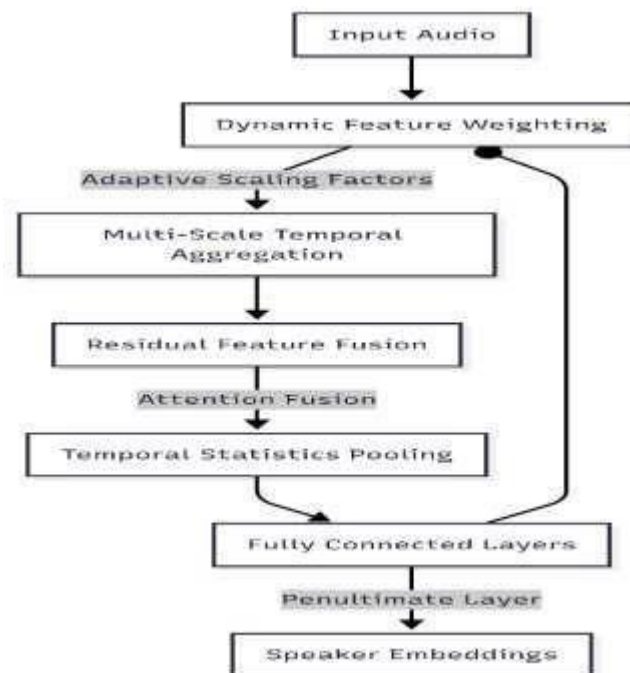
where W_1 and W_2 are learnable weight matrices, $\delta(\cdot)$ is the ReLU activation, and $\sigma(\cdot)$ is the sigmoid function.

The resulting scaling vectors adaptively modulate the original feature representation:

$$\tilde{F} = s \odot F$$

where $\odot$ denotes element-wise multiplication. This allows the network to highlight speaker-relevant frequency bands while suppressing irrelevant noise and overlapping speech components.

Figure 3: Block diagram of the AFA-TDNN.



- **Multi-Scale Temporal Aggregation**

Speech signals contain both short-term phonetic patterns and long-term prosodic structures. To effectively capture these variations, AFA-TDNN introduces parallel dilated convolutional layers with dilation rates $d \in \{1, 2, 4\}$. For an input sequence F , the output of a convolutional layer with dilation d is:

$$H_d(t) = \sum_{k=1}^K W_d(k) \cdot F(t - d \cdot k)$$

where K is the kernel size and $W_d(k)$ are learnable convolutional weights. The outputs from all dilation branches are concatenated:

$$H = \text{Concat}(H_1, H_2, H_4)$$

This multi-resolution feature representation enables the model to learn fine-grained phonetic cues while retaining long-term contextual dependencies, improving diarization performance for short-turn and overlapping speaker scenarios.

- **Residual Feature Fusion**

To effectively integrate the multi-scale temporal features, AFA-TDNN applies attention-guided residual fusion. The fused representation F_{fused} is computed as:

$$F_{\text{fused}} = H + \alpha \cdot A(H)$$

Where $A(H)$ is the attention-weighted feature representation, and α is a learnable scaling parameter.

This design stabilizes gradient flow during training while combining local phonetic and global contextual information. As a result, the embeddings become more robust in low-SNR and noisy environments.

- **Temporal Statistics Pooling**

Since speech recordings have variable lengths, AFA-TDNN employs a statistics pooling layer to generate fixed-length embedding from variable-length sequences. Given frame-level features $F(t)$, the mean μ and variance σ^2 are computed over the entire segment:

$$\mu = \frac{1}{T} \sum_{t=1}^T F(t), \quad \sigma^2 = \frac{1}{T} \sum_{t=1}^T (F(t) - \mu)^2$$

The pooled feature vector P is then formed by concatenating μ and σ :

$$P = [\mu; \sigma]$$

This transformation enables the network to map variable-length utterances into fixed-dimensional embeddings, making them suitable for clustering-based speaker diarization.

- **Fully Connected Layers and Embedding Extraction**

The pooled features PPP are passed through two “Fully Connected (FC)” layers to project them into a compact embedding space:

$$E = f(W_{fc2} \cdot \delta(W_{fc1} \cdot P + b_1) + b_2)$$

where W_{fc1} , W_{fc2} are learnable weight matrices, b_1 , b_2 are bias terms, $\delta(\cdot)$ is a ReLU activation, E represents the speaker embedding vector extracted from the penultimate layer. These embeddings E are then used for spectral clustering in the diarization pipeline.

3.5 Training Configuration

The AFA-TDNN model, as proposed, is trained on 3-second random crops of speaker utterances from the AMI Meeting Corpus. For acoustic representation, 80-dimensional log-Mel filter bank energies are employed as input features. The model is optimized with the Adam optimizer for a “Cyclical Learning Rate (CLR)” strategy with a triangular policy in order to prevent local minima and enhance convergence stability. Training is conducted for 12 epochs using a starting batch size of 32 that is effectively increased to 192 segments per iteration by the use of BADA. While the model maintains the fundamental hyperparameters of the baseline x-vector TDNN architecture, architectural changes are made to include multi-scale temporal aggregation and dynamic feature weighting to allow the extraction of stronger, noise-resistant, and discriminative speaker embeddings.

3.6 Spectral Clustering for Speaker Segmentation

Speaker diarization aims to determine “who spoke when” in multi-speaker conversations, which is particularly challenging in real-world scenarios involving overlapping speech, environmental noise, and variable speaking styles. In the proposed framework, they adopt “Spectral Clustering (SC)” over conventional “Agglomerative Hierarchical Clustering (AHC)” because SC better exploits the global structure of speaker embedding rather than relying on purely local similarity measures. By leveraging the graph-theoretic properties of embedding, SC achieves more accurate speaker separation even under noisy and overlapping conditions.

Spectral clustering treats the diarization problem as a graph partitioning task. Let us consider N speech segments, each represented by an embedding vector $e_i \in \mathbb{R}^d$ obtained from the AFA-TDNN model. A similarity graph $G = (V, E)$ is then constructed, where $V = \{v_1, v_2, v_N\}$ represents the nodes corresponding to the speech segments, and E denotes the weighted edges representing the similarity between them. The goal is to partition this graph into K distinct clusters such that embedding within the same cluster exhibit high intra-cluster similarity, while embeddings across different clusters demonstrate low inter-cluster similarity. Spectral clustering achieves this by leveraging the eigenvalues and eigenvectors of the graph Laplacian, enabling the discovery of natural divisions in the embedding space for accurate and robust speaker segmentation. The complete spectral clustering pipeline consists of the following steps:

- **Affinity Matrix Construction**

Given a set of extracted speaker embeddings $\{e_1, e_2, \dots, e_N\}$, they compute the cosine similarity between all embedding pairs to construct an affinity matrix $A \in \mathbb{R}^{N \times N}$:

$$A_{ij} = \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|}$$

Here, $A_{ij} \in [-1, 1]$ measures the similarity between embedding e_i and e_j . To ensure robustness, low-similarity edges below a threshold δ are pruned:

$$A_{ij} = \begin{cases} A_{ij}, & \text{if } A_{ij} \geq \delta \\ 0, & \text{otherwise} \end{cases}$$

This step reduces noise influence and prevents spurious correlations.

- **Graph Laplacian Formation**

From the affinity matrix A , they construct the degree matrix D , defined as:

$$D_{ii} = \sum_{j=1}^N A_{ij}$$

Using A and D , they calculate the unnormalized graph Laplacian L :

$$L = D - A$$

Instead, the normalized Laplacian L_{sym} is defined as:

$$L_{\text{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

This normalization helps stabilize eigen decomposition, particularly when speaker embeddings have uneven similarities.

- **Eigen Decomposition & Speaker Count Estimation**

Spectral clustering practises eigenvalue decomposition of the Laplacian matrix:

$$L V_k = \lambda_k V_k$$

The eigengap heuristic is applied to estimate the number of speakers K :

$$K = \arg \max_k (\lambda_{k+1} - \lambda_k)$$

where $\lambda_k \rightarrow k^{\text{th}}$ smallest eigenvalue, $K \rightarrow$ optimal number of clusters (speakers). A large eigengap implies clear separation between speaker clusters.

- **Spectral Embedding Construction**

They take the top K 'eigenvectors $\{v_1, v_2, \dots, v_K\}$ corresponding to the smallest eigenvalues and form a low-dimensional representation':

$$V = [v_1, v_2, \dots, v_K] \in \mathbb{R}^{N \times K}$$

Each row of V represents a spectral embedding of a speech segment in a transformed space where clusters are more separable.

- **K-Means Clustering**

Finally, relate k -means clustering on the spectral embeddings V to assign speaker labels:

$$C = \arg \min_C \sum_{i=1}^N ||V_i - \mu_{c_i}||^2$$

where $V_i \rightarrow$ spectral embedding of the i^{th} segment, $\mu_{c_i} \rightarrow$ centroid of the cluster assigned to V_i and $C \rightarrow$ final speaker cluster assignments.

Spectral clustering significantly improves diarization accuracy by exploiting global relationships in embeddings, making it particularly effective for multi-party meetings and overlapping speech scenarios. By combining AFA-TDNN embeddings with spectral clustering, the proposed system ensures robust, noise-resilient, and accurate speaker diarization. The method leverages the geometric properties of embeddings and graph spectral theory to outperform conventional clustering techniques, achieving superior separation of speakers in real-world meeting environments.

3.7 Performance Evaluation

The proposed method is calculated using “Diarization Error Rate (DER)”, which primarily explanations for missed detections, false alarms, and speaker confusion and computed as:

$$DER = \frac{MD + FA + SC}{TRD} \times 100\%$$

where “MD is Missed Detection Time”, “FA is False Alarm Time”, “SC is Speaker Confusion Time” and “TRD is Total Reference Duration”. A lower value of DER indicates better performance. The evaluations are done under oracle and estimated speaker conditions. Different “Jaccard Error Rate (JER)” is also examined to provide segment-wise evaluation of diarization performance.

The JER is an assessment tool in speaker diarization that is used to extend the overlap among reference speaker segments (truth) and hypothesized speaker segments (predicted output). The JER assesses the “Intersection-Over-Union (IoU)” of the reference and system output to provide a more balanced measure of accuracy with regards to segmentation accuracy.

$$JER = 1 - \frac{|R \cap H|}{|R \cup H|}$$

Where $R \rightarrow$ Set of reference speaker segments, $H \rightarrow$ Set of hypothesized speaker segments, $|R \cap H| \rightarrow$ Total length of duration of overlapping correctly predicted speech segments, and $|R \cup H| \rightarrow$ the total length of duration covered by either reference or hypothesis. This formula is considered the proportion of mismatched regions to the total length of duration combined, and the greater the JER, the better the diarization performance.

DER and JER describe the accuracy of diarization performance, while Cohen's Kappa coefficient (κ) indicates the agreement that are expected to be associated with the speakers and refer to the speaker labels in conjunction to determine whether the agreement is by chance. It is defined as:

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

Where P_0 is Observed agreement (accuracy), P_e is Expected agreement by chance.

4. Result and Analysis

This section shows the performance of the suggested AFA-TDNN model based on several performance measures, such as DER, JER, and Kappa coefficient. This analysis also compares various augmentation strategies and compares the suggested model with state-of-the-art models to estimate its performance and reliability.

4.1 Impact of Batch-Adaptive Data Augmentation (BADA)

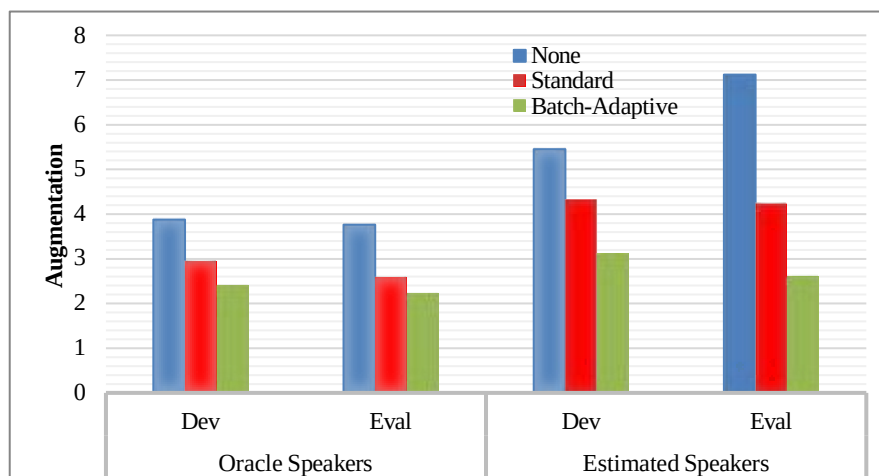
For examine the performance of the suggested BADA method over no augmentation and classical augmentation methods. Table 1 provides the DER under various augmentation conditions.

Table1: DERs (%) with Different Augmentation Strategies on AMI Data

Augmentation	Oracle Speakers		Estimated Speakers	
	Dev	Eval	Dev	Eval
None	3.88	3.76	5.45	7.12
Standard	2.95	2.58	4.32	4.23
Batch-Adaptive	2.41	2.23	3.12	2.61

Figure 4 presents the comparison of DER using three different augmentation strategies: No Augmentation, Standard Augmentation, and the plannedBADA on the “AMI meeting corpus”. The results are reported for both oracle speaker counts and estimated speaker counts on the development (Dev) and evaluation (Eval) sets. Without augmentation, DER remains relatively high, with 3.88% (Dev) and 3.76% (Eval) under oracle conditions and 5.45% (Dev) and 7.12% (Eval) under estimated speakers. Applying standard augmentation improves performance, reducing DER to 2.95% (Dev) and 2.58% (Eval) for oracle speakers. However, the proposed Batch-Adaptive Augmentation accomplishes the finest efficiency, with the lowest DER of 2.41% (Dev) and 2.23% (Eval) for oracle speakers, and 3.12% (Dev) and 2.61% (Eval) for estimated speakers. These results demonstrate that BADA significantly enhances robustness, improving diarization accuracy across diverse acoustic conditions.

Figure 4: DER comparison of different Augmentation techniques



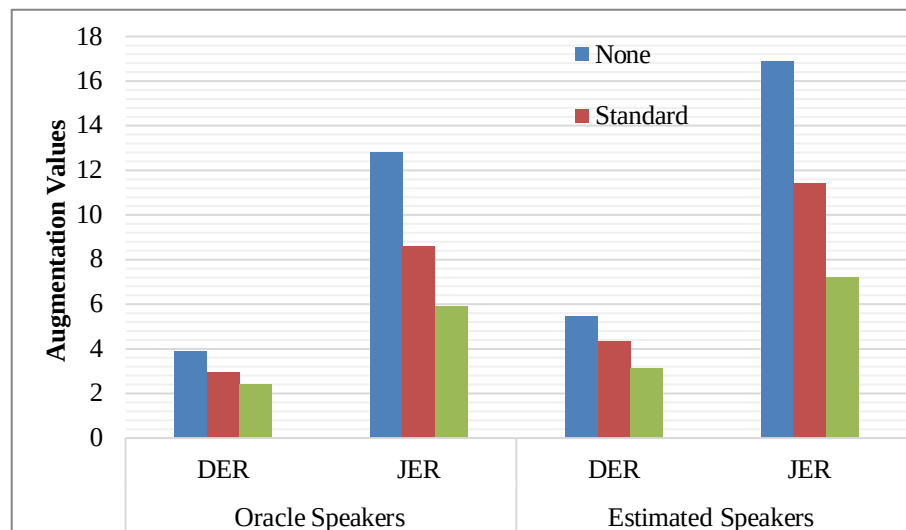
They further analyze the contribution of BADA compared to no augmentation and standard augmentation techniques. Table 2 presents both DER and JER results across different augmentation strategies.

Table 2: Effect of Augmentation on DER (%) and JER (%)

Augmentation	Oracle Speakers		Estimated Speakers	
	DER	JER	DER	JER
None	3.88	12.8	5.45	16.9
Standard	2.95	8.6	4.32	11.4
Batch-Adaptive	2.41	5.9	3.12	7.2

Figure 5 illustrates the impact of BADA compared to No Augmentation and Standard Augmentation on both DER and JER using the AMI meeting corpus. Under oracle speaker counts, the DER without augmentation is 3.88% with a JER of 12.8%. Applying standard augmentation improves the results, reducing DER to 2.95% and JER to 8.6%. The proposed BADA strategy achieves the best performance, further lowering DER to 2.41% and JER to 5.9%. Similarly, under estimated speaker counts, the DER drops from 5.45% (None) to 4.32% (Standard) and finally to 3.12% (BADA), while the JER decreases from 16.9% (None) to 11.4% (Standard) and 7.2% (BADA). These results clearly demonstrate that BADA significantly enhances diarization accuracy by improving both DER and JER across all testing conditions.

Figure 5: Effect of different augmentation techniques on DER and JER performance



4.2 Performance Evaluation

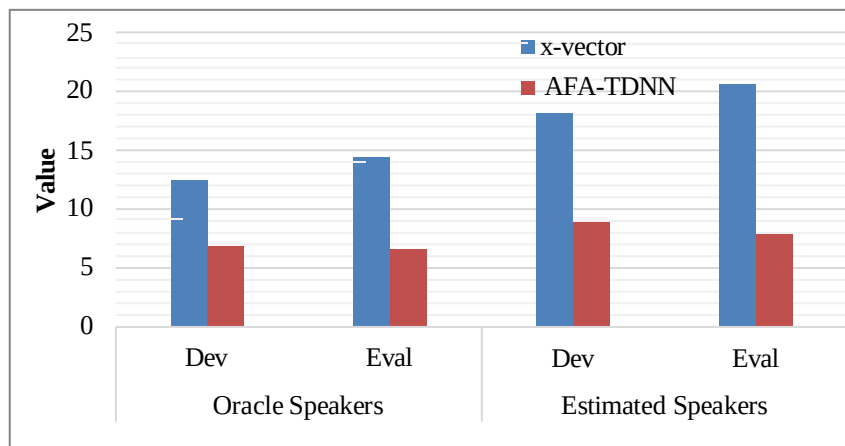
• Jaccard Error Rate (JER) Analysis

Lower JER values indicate better diarization performance. Table 2 compares the JER performance of x-vector and AFA-TDNN embeddings.

Table 3: JERs (%) on AMI dataset

Embedding	Oracle Speakers		Estimated Speakers	
	Dev	Eval	Dev	Eval
x-vector	12.42	14.36	18.15	20.62
AFA-TDNN	6.81	6.54	8.92	7.84

Figure 6 illustrates the JER comparison between the traditional x-vector TDNN model and the proposed AFA-TDNN embeddings using the AMI meeting corpus with a spectral clustering backend. Under oracle speaker counts, the x-vector model records a JER of 12.42% on the Dev set and 14.36% on the Eval set, whereas the proposed AFA-TDNN significantly improves performance, achieving much lower JER values of 6.81% (Dev) and 6.54% (Eval). Similarly, under estimated speaker counts, the x-vector model shows a higher JER of 18.15% (Dev) and 20.62% (Eval), while AFA-TDNN maintains superior performance with 8.92% (Dev) and 7.84% (Eval). These results demonstrate that AFA-TDNN provides a substantial improvement in segmentation accuracy, achieving nearly 50% lower JER compared to the baseline x-vector approach.

Figure 6: JER comparison of x-vector and proposed AFA-TDNN model

- **Diarization Error Rate (DER) Analysis**

Lower DER indicates better diarization performance.

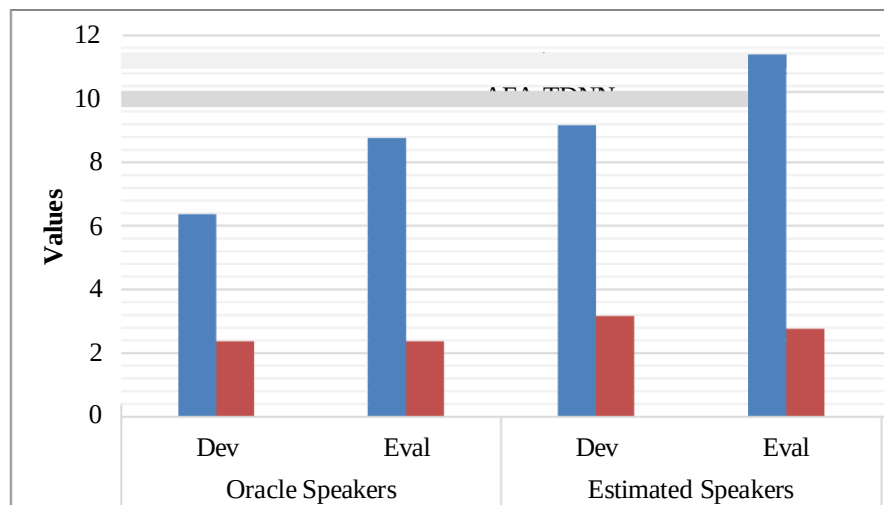
Table 4: DERs (%) on AMI dataset

Embedding	Oracle Speakers		Estimated Speakers	
	Dev	Eval	Dev	Eval
x-vector	6.14	8.57	9.21	11.04
AFA-TDNN	2.41	2.23	3.12	2.61

Figure 7 presents the DER comparison between the traditional x-vector TDNN model and the proposed AFA-TDNN embeddings using the AMI meeting corpus with a spectral

clustering backend. Under oracle speaker counts, the x-vector model achieves a DER of 6.14% on the Dev set and 8.57% on the Eval set. In contrast, the proposed AFA-TDNN significantly reduces the DER to 2.41% (Dev) and 2.23% (Eval). Similarly, in the estimated speaker count scenario, the DER for the x-vector model rises to 9.21% (Dev) and 11.04% (Eval), whereas the AFA-TDNN achieves much lower values of 3.12% (Dev) and 2.61% (Eval). These results highlight that the proposed AFA-TDNN achieves over a 60% reduction in DER compared to x-vectors, demonstrating superior diarization accuracy and robustness in both oracle and estimated speaker conditions.

Figure 7: DER comparison of x-vector and proposed AFA-TDNN model



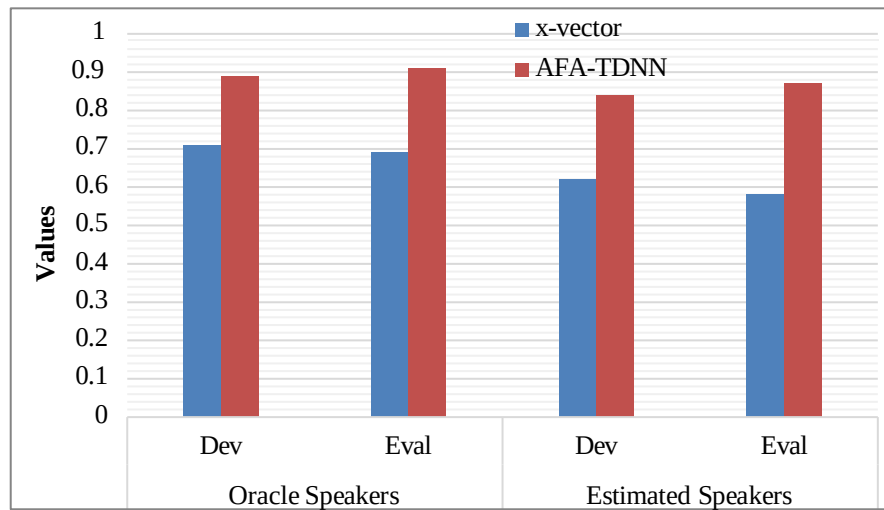
- **Kappa Coefficient Analysis**

Higher the value of Kappa Coefficient the better is performance of model.

Table 5: Kappa Scores on AMI Data

Embedding	Oracle Speakers		Estimated Speakers	
	Dev	Eval	Dev	Eval
x-vector	0.71	0.69	0.62	0.58
AFA-TDNN	0.89	0.91	0.84	0.87

Figure 8 illustrates the comparison of Cohen's Kappa coefficient between the x-vector TDNN model and the proposed AFA-TDNN embeddings using the AMI meeting corpus. The Kappa coefficient measures the agreement between predicted speaker labels and ground-truth annotations, where higher values indicate better diarization performance. Under oracle speaker counts, the x-vector model achieves Kappa scores of 0.71 on the Dev set and 0.69 on the Eval set, whereas the proposed AFA-TDNN achieves significantly higher scores of 0.89 (Dev) and 0.91 (Eval). Similarly, under estimated speaker counts, the x-vector model obtains lower Kappa values of 0.62 (Dev) and 0.58 (Eval), while AFA-TDNN demonstrates superior performance with 0.84 (Dev) and 0.87 (Eval). These results indicate that AFA-TDNN provides stronger agreement with reference annotations, ensuring higher reliability and robustness in multi-speaker diarization tasks.

Figure 8: Kappa coefficient comparison of x-vector and proposed AFA-TDNN model

4.3 Comparison with State-of-the-Art Models

To assess the efficacy of the suggested AFA-TDNN, juxtapose its performance with two prevalent baseline models: the conventional x-vector TDNN embeddings and the ECAPA-TDNN embeddings. All studies are performed using the beamformed AMI meeting corpus dataset utilizing SC as the backend for speaker segmentation. Two cases are included in the evaluation: (i) Oracle Speaker Counts, in which the real speaker count is known, and (ii) Estimated Speaker Counts, in which the speaker count is dynamically deduced using the eigengap-based clustering algorithm. The diarization quality is evaluated using two main metrics: DER and JER. Lower values signify greater performance. Table 1 reports DER and JER values under both oracle speaker counts and estimated speaker counts.

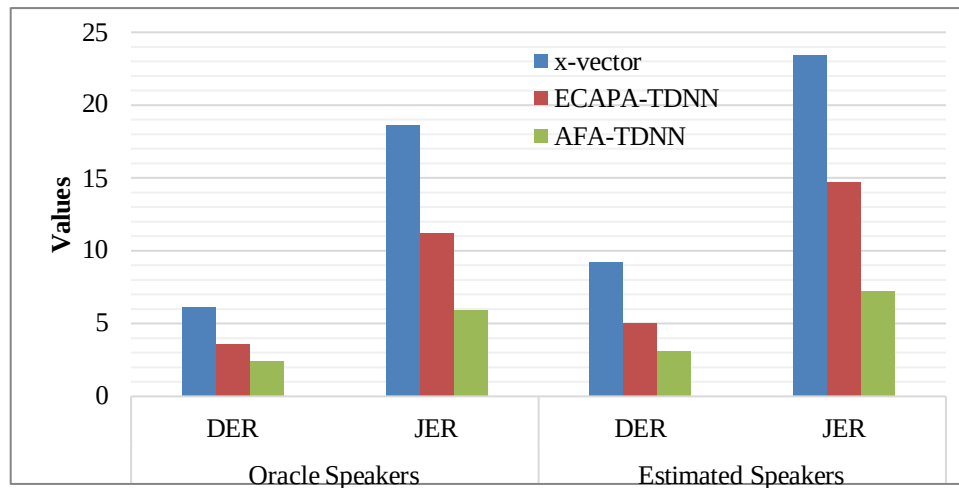
Table 6: DER (%) and JER (%) Comparison of AMI Dataset

Embedding	Oracle Speakers		Estimated Speakers	
	DER	JER	DER	JER
x-vector	6.14	18.6	9.21	23.4
ECAPA-TDNN	3.62	11.2	5.04	14.7
AFA-TDNN	2.41	5.9	3.12	7.2

Figure 9 compares the performance of x-vector TDNN, ECAPA-TDNN, and the proposed AFA-TDNN embeddings on the AMI meeting corpus using a spectral clustering backend. The evaluation considers both DER and JER under two scenarios: oracle speaker counts and estimated speaker counts. Under oracle conditions, the x-vector model achieves a DER of 6.14% and JER of 18.6%, while ECAPA-TDNN improves performance with a DER of 3.62% and JER of 11.2%. The proposed AFA-TDNN achieves the best results, reducing DER to 2.41% and JER to 5.9%. Similarly, for estimated speakers, AFA-TDNN maintains superior performance with a DER of 3.12% and JER of 7.2%, compared to ECAPA-TDNN (5.04%, 14.7%) and x-vector (9.21%, 23.4%). These results confirm that AFA-TDNN

significantly outperforms existing methods, providing highly robust and accurate speaker diarization.

Figure 9: DER and JER comparison of x-vector, ECAPA-TDNN, and proposed AFA-TDNN models.



5. Conclusion

In this study, they introduced an enhanced speaker diarization system using AFA-TDNN along with spectral clustering for robust and accurate speaker separation in multi-speaker conditions. Contrary to the conventional x-vector and ECAPA-TDNN methods, the new system incorporates dynamic feature weighting, multi-scale temporal aggregation, by fusing residual features to train speaker embeddings that are both highly discriminative and noise-resistant. Moreover, a BADA approach was proposed to facilitate model generalization across heterogeneous acoustic environments by mimicking real-world variations such as noise, reverberation, and temporal disturbances. Comparative experimental evaluations on the “AMI Meeting Corpus” show that the method proposed considerably surpasses available baselines with consistently lower DER and JER, in addition to a better segmentation reliability, reflected by higher Kappa coefficients.

To conclude, the AFA-TDNN model overcomes important challenges like overlapping speech, variability of channel, and background noise, allowing for better diarization performance. The results set AFA-TDNN as a robust, scalable, and noise-resistant solution to be applied on real-world tasks like meeting transcription, conversational analytics, and intelligent voice-based systems. Future research will be directed towards scaling the approach to fully end-to-end diarization pipelines, multimodal fusion, and real-time deployment for large-scale conversational AI systems.

References

- 1 O'Shaughnessy, Douglas. "Speaker Diarization: A Review of Objectives and Methods." *Applied Sciences* 15, no. 4 (2025): 2002.
- 2 Wooters, Chuck, and Marijn Huijbregts. "The ICSI RT07s speaker diarization system." In *International Evaluation Workshop on Rich Transcription*, pp. 509-519. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- 3 Mao, Huanru Henry, Shuyang Li, Julian McAuley, and Garrison Cottrell. "Speech recognition and multi-speaker diarization of long conversations." *arXiv preprint arXiv:2005.08072* (2020).

- 4 Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. "Speaker diarization: A review of recent research." *IEEE Transactions on audio, speech, and language processing* 20, no. 2 (2012): 356-370.
- 5 Stan, Adriana. "Residual information in deep speaker embedding architectures." *Mathematics* 10, no. 21 (2022): 3927.
- 6 Irum, Amna, and Ahmad Salman. "Speaker verification using deep neural networks: A." *International Journal of Machine Learning and Computing* 9, no. 1 (2019).
- 7 S. Horiguchi, Y. Takashima, P. García, S. Watanabe, and Y. Kawaguchi, "Multi-Channel End-To-End Neural Diarization with Distributed Microphones," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7332–7336. doi: 10.1109/ICASSP43922.2022.9746749.
- 8 N. Zheng et al., "Multi-Channel Speaker Diarization Using Spatial Features for Meetings," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7337–7341. doi: 10.1109/ICASSP43922.2022.9747343.
- 9 Z. Chen, B. Han, S. Wang, and Y. Qian, "Attention-based encoder-decoder end-to-end neural diarization with embedding enhancer," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1636–1649, 2024.
- 10 X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE TASLP*, 2012.
- 11 M. Rybicka, J. Villalba, T. Thebaud, N. Dehak, and K. Kowalczyk, "End-to-End Neural Speaker Diarization With Non-Autoregressive Attractors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3960–3973, 2024, doi: 10.1109/TASLP.2024.3439993.
- 12 F. Landini, "From Modular to End-to-End Speaker Diarization," *ArXivPrepr. ArXiv240708752*, 2024.
- 13 Serafini, Luca, Samuele Cornell, Giovanni Morrone, Enrico Zovato, Alessio Brutti, and Stefano Squartini. "An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings." *Computer speech & language* 82 (2023): 101534.
- 14 Ancilin, J., and A. Milton. "Improved speech emotion recognition with Mel frequency magnitude coefficient." *Applied Acoustics* 179 (2021): 108046.
- 15 Pudasaini, Anil, Muna Al-Hawawreh, Mohamed Reda Bouadjene, Hakim Hacid, and Sunil Aryal. "A comprehensive study of audio profiling: Methods, applications, challenges, and future directions." *Neurocomputing* (2025): 130334.
- 16 Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware deep neural network," in *ICASSP 2014*, 2014.
- 17 L. Li, D. Wang, Z. Zhang, and T. F. Zheng, "Deep Speaker Vectors for Semi Text-independent Speaker Verification," *ArXivPrepr. ArXiv150506427*, 2015.
- 18 G. Sun, C. Zhang, and P. Woodland, "Combination of Deep Speaker Embeddings for Diarisation," *ArXivPrepr. ArXiv201012025*, 2021.
- 19 Kinoshita, Keisuke, Marc Delcroix, and Naohiro Tawara. "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7198-7202. IEEE, 2021.
- 20 Kinoshita, Keisuke, Marc Delcroix, and Naohiro Tawara. "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech." *arXiv preprint arXiv:2105.09040* (2021).
- 21 Anidjar, Or Haim, Yannick Esteve, Chen Hajaj, Amit Dvir, and Itshak Lapidot. "Speech and multilingual natural language framework for speaker change detection and diarization." *Expert Systems with Applications* 213 (2023): 119238.

- 22 Chauhan, Neha, Tsuyoshi Isshiki, and Dongju Li. "Enhancing speaker recognition models with noise-resilient feature optimization strategies." In *Acoustics*, vol. 6, no. 2, pp. 439-469. MDPI, 2024.
- 23 Kwon, Youngki, Jee-weon Jung, Hee-SooHeo, You Jin Kim, Bong-Jin Lee, and Joon Son Chung. "Adapting speaker embeddings for speaker diarisation." *arXiv preprint arXiv:2104.02879* (2021).
- 24 B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834. doi: 10.21437/Interspeech.2020-2650.
- 25 Gao, Zhifu, Yan Song, Ian McLoughlin, Pengcheng Li, Yiheng Jiang, and Li-Rong Dai. "Improving Aggregation and Loss Function for Better Embedding Learning in End-to-End Speaker Verification System." In *Interspeech*, pp. 361-365. 2019.
- 26 Tu, Youzhi. "Deep speaker embedding for robust speaker verification." (2022).
- 27 Johannessen, Runar Ask. "Aggregation of speaker embeddings for speaker diarization." Master's thesis, NTNU, 2020.
- 28 G. Sell, D. Garcia-Romero, D. Snyder, A. McCree, D. Povey, and S. Khudanpur, "Diarization is Hard: Some Experiences and Lessons Learned," in *Proceedings of Interspeech*, 2018, pp. 2808–2812. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1576.pdf
- 29 Ryant, Neville, Prachi Singh, VenkatKrishnamohan, RajatVarma, Kenneth Church, Christopher Cieri, Jun Du, SriramGanapathy, and Mark Liberman. "The third DIHARD diarization challenge." *arXiv preprint arXiv:2012.01477* (2020).
- 30 Baghel, Shikha, ShreyasRamoji, Prachi Singh, Somil Jain, Pratik Roy Chowdhuri, KaustubhKulkarni, SwapnilPadhi, DeepuVijayasanen, and SriramGanapathy. "Displace challenge: Diarization of speaker and language in conversational environments." *arXiv preprint arXiv:2303.00830* (2023).
- 31 Brown, Andrew, Jaesung Huh, Joon Son Chung, ArshaNagrani, Daniel Garcia-Romero, and Andrew Zisserman. "Voxsrc 2021: The third voxceleb speaker recognition challenge." *arXiv preprint arXiv:2201.04583* (2022).
- 32 Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *ICASSP 2014*, 2014.
- 33 Li, Lantian, Dong Wang, Zhiyong Zhang, and Thomas Fang Zheng. "Deep speaker vectors for semi text-independent speaker verification." *arXiv preprint arXiv:1505.06427* (2015).
- 34 Sun, Guangzhi, Chao Zhang, and Philip C. Woodland. "Combination of deep speaker embeddings for diarisation." *Neural Networks* 141 (2021): 372-384.
- 35 Carletta, Jean, Simone Ashby, SebastienBourban, Mike Flynn, Mael Guillemot, Thomas Hain, JaroslavKadlec et al. "The AMI meeting corpus: A pre-announcement." In *International workshop on machine learning for multimodal interaction*, pp. 28-39. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.
- 36 Kwon, Youngki, Jee-weon Jung, Hee-SooHeo, You Jin Kim, Bong-Jin Lee, and Joon Son Chung. "Adapting speaker embeddings for speaker diarisation." *arXiv preprint arXiv:2104.02879* (2021).
- 37 Li, Xiang, VivekGovindan, RohitPaturi, and SundararajanSrinivasan. "Speakers unembedded: Embedding-free approach to long-form neural diarization." *arXiv preprint arXiv:2406.18679* (2024).
- 38 Li, Yaqian, Xiaolong Zhang, Haibin Li, and Wengming Zhang. "Enhancing Speaker Diarization with Deep Neural Network Embeddings and Spectral Clustering." Available at SSRN 4761198.

- 39 He, Mao-Kui, Jun Du, Qing-Feng Liu, and Chin-Hui Lee. "Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023): 1561-1573.
- 40 Cord-Landwehr, Tobias, Christoph Boeddeker, Cătălin Zorilă, Rama Doddipatla, and Reinhold Haeb-Umbach. "Frame-wise and overlap-robust speaker embeddings for meeting diarization." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.
- 41 Alvarez-Trejos, Juan Ignacio, Alicia Lozano-Diez, and Daniel Ramos. "Feature Integration Strategies for Neural Speaker Diarization in Conversational Telephone Speech." *Applied Sciences* 15, no. 9 (2025): 4842.
- 42 AMI Project, "The AMI Meeting Corpus." University of Edinburgh, IDIAP Research Institute, University of Sheffield, TNO, 2005. [Online]. Available: <https://groups.inf.ed.ac.uk/ami/corpus/>
- 43 X. Anguera, C. Wooters, and J. Hernando, "BeamformIt: An Open-source Toolkit for Robust Multi-microphone Speech Processing," in *Proceedings of Interspeech*, 2006. [Online]. Available: https://www.iscaspeech.org/archive/interspeech_2006/i06_2601.html
- 44 D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus." 2015. [Online]. Available: <https://arxiv.org/abs/1510.08484>
- 45 T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition," in *Proceedings of ICASSP*, 2017, pp. 5220–5224. doi: 10.1109/ICASSP.2017.7953075.