Attention-Based Encoder-Decoder End-to-End Neural Diarization With Embedding Enhancer

Zhengyang Chen, *Graduate Student Member, IEEE*, Bing Han, *Graduate Student Member, IEEE*, Shuai Wang, *Member, IEEE*, and Yanmin Qian, *Senior Member, IEEE*

***Abstract*—Deep neural network-based systems have significantly improved the performance of speaker diarization tasks. However, end-to-end neural diarization (EEND) systems often struggle to generalize to scenarios with an unseen number of speakers, while target speaker voice activity detection (TS-VAD) systems tend to be overly complex. In this paper, we propose a simple attention-based encoder-decoder network for end-to-end neural diarization (AED-EEND). In our training process, we introduce a teacher-forcing strategy to address the speaker permutation problem, leading to faster model convergence. For evaluation, we propose an iterative decoding method that outputs diarization results for each speaker sequentially. Additionally, we propose an Enhancer module to enhance the frame-level speaker embeddings, enabling the model to handle scenarios with an unseen number of speakers. We also explore replacing the transformer encoder with a Conformer architecture, which better models local information. Furthermore, we discovered that commonly used simulation datasets for speaker diarization have a much higher overlap ratio compared to real data. We found that using simulated training data that is more consistent with real data can achieve an improvement in consistency. Extensive experimental validation demonstrates the effectiveness of our proposed methodologies. Our best system achieved a new state-of-the-art diarization error rate (DER) performance on all the CALLHOME (10.08%), DIHARD II (24.64%), and AMI (13.00%) evaluation benchmarks when overlap is considered and no oracle voice activity detection (VAD) is used. Beyond speaker diarization, our AED-EEND system also shows remarkable competitiveness as a speech type detection model.**

***Index Terms*—Neural speaker diarization, attention-based encoder-decoder, CALLHOME, AMI, DIHARD, iterative decoding.**

I. Introduction

SPEAKER diarization is a challenging task in speech processing, aiming to determine "who spoke when" in scenarios with multiple speakers. It serves as a fundamental pre-processing step in various speech-related tasks. For instance, it enables the detection of distinct speaking segments for each individual present in a recording, allowing subsequent speaker recognition models to identify the absolute speaker identity [1]. In meeting scenarios, obtaining the speaking segments for each participant is essential to leverage automatic speech recognition (ASR) systems for generating transcripts for individual speakers [2]. Moreover, speaker diarization and speech separation tasks share similarities, leading researchers to explore methods that leverage one task to improve the other [3], [4], [5].

Conventional speaker diarization systems typically involve multiple stages [6], [7]. Firstly, a voice activity detection (VAD) system is used to filter the non-speech region and the left part is segmented using a specified window length and hop length. Then, a pre-trained speaker embedding extractor is utilized to extract speaker embeddings for each segment. Subsequently, a scoring backend, such as cosine scoring or PLDA [8], [9], is applied to compute similarity scores between pairs of segments. Following this, a clustering algorithm [10], [11], [12] is employed to assign a unique speaker label to each segment. Optionally, compensation algorithms like Variational-Bayesian refinement [10], [13], [14], [15] may be employed to refine the clustering results. However, due to the limitations of clustering algorithms, each segment can only be assigned to a single class. This limitation hinders traditional methods from effectively handling scenarios with speaker overlap.

To address the issue of traditional methods being unable to handle speaker overlap, researchers have proposed end-to-end neural diarization (EEND) methods. In [16], [17], the authors treated speaker diarization as a frame-wise multi-class classification problem and employed the permutation invariant (PIT) loss [18] to optimize the entire system in an end-to-end manner. However, the number of classes in [16], [17] is fixed and determined by the output head dimension, limiting their ability to handle scenarios with a flexible number of speakers. To overcome this limitation, Fujita et al. [19] and Takashima et al. [20] proposed a chain-rule paradigm, enabling the sequential output of diarization results for each speaker. This approach allows for flexibility in handling scenarios with varying numbers of speakers. Additionally, Horiguchi et al. [21] introduced the EEND-EDA system, which utilizes an LSTM encoder-decoder network to model attractors for each speaker. Furthermore, researchers also proposed two-stage hybrid systems [22], [23] to address the challenge of handling a flexible number of speakers. These systems first output diarization results for short segments

Manuscript received 8 August 2023; revised 27 December 2023; accepted 4 February 2024. Date of publication 16 February 2024; date of current version 1 March 2024. This work was supported in part by China NSFC Projects under Grant 62122050 and Grant 62071288 and in part by the Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Yanmin Qian.)*

Zhengyang Chen, Bing Han, and Yanmin Qian are with the Auditory Cognition and Computational Acoustics Lab, Department of Computer Science and Engineering and the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhengyang.chen@sjtu.edu.cn; hanbing97@sjtu.edu.cn; yanminqian@sjtu.edu.cn).

Shuai Wang is with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: wangshuai@cuhk.edu.cn).

Digital Object Identifier 10.1109/TASLP.2024.3366756

2329-9290 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

with a limited number of speakers using EEND, and then employ a clustering algorithm to solve the inter-segment speaker permutation problem.

Recent research has highlighted the potential benefits of incorporating speaker-specific prior information to enhance system performance and enable the output of speaker-related results. For instance, in speech separation systems [24], [25], [26], automatic speech recognition tasks [24], and active speaker detection tasks [27], researchers have successfully integrated the target speaker's speech or embedding to obtain the desired outputs, such as the target speaker's separated speech, transcript, or on-screen person speaking frames. Similarly, in the context of speaker diarization, researchers have explored the integration of target speaker prior information, referred to as target speaker voice activity detection (TS-VAD) [28], [29]. To utilize the TS-VAD system for generating results for all speakers, an additional diarization system is often employed to identify the single-speaker speaking segments for each individual. Subsequently, a pre-trained speaker embedding extractor is used to obtain speaker embeddings. Despite the complexity of TS-VAD systems, their excellent performance in competitions [29], [30] has motivated researchers to investigate various TS-VAD approaches [27], [31], [32], [33].

As mentioned previously, while the EEND-EDA approach can handle scenarios with a variable number of speakers, the authors in [34] acknowledge that the output speaker number of EEND-EDA is empirically constrained by the maximum number of speakers observed during pre-training. Additionally, the TS-VAD system, with its numerous sub-systems, introduces excessive complexity to the overall system. In our previous work [35], we presented a simple attention-based encoder-decoder end-to-end[1] neural diarization system (AED-EEND). In this approach, we replace the LSTM encoder-decoder architecture in EEND-EDA with a transformer decoder and achieve better performance. Furthermore, we introduce a teacher-forcing training strategy that leverages speaker-specific prior information. This strategy effectively mitigates the speaker permutation problem and facilitates faster convergence of the system. Additionally, we propose a heuristic decoding method to iteratively obtain diarization results for each speaker.

Due to space limitation, our analysis of the AED-EEND system in our previous article was not sufficient. In this paper, we will give a more comprehensive analysis of it and evaluate the system on more diarization benchmarks to verify its effectiveness. In our previous work [35], we noticed that our AED-EEND system suffers the same problem as EEND-EDA, which has poor performance when the speaker number of the evaluation set is different from the pre-training simulation set. In this paper, we proposed to add a new *Enhancer* module to our AED-EEND system, which can help the model generalize to the unseen number-of-speaker scenario. During evaluation on datasets with significantly long durations, we observed that the decoding method with clustering operations, as described in [35], exhibited slow execution. In this paper, we introduce improvements to the decoding method to overcome this challenge. Additionally, we discovered a discrepancy between the commonly used simulation dataset in [16], [17], [19], [21] and real data, particularly in terms of the overlap ratio. To address this mismatch, we introduce a more realistic data simulation approach following the guidelines outlined in [36], [37]. In contrast to previous experiments [36], [37] that focused on limited scenarios and employed relatively weaker systems, Notably, our proposed approach achieves state-of-the-art performance across multiple diarization evaluation benchmarks. Furthermore, we explore the replacement of the transformer encoder in AED-EEND with the Conformer encoder, leading to further enhancements. Although AED-EEND is a diarization system, we discovered that we can also utilize it as a standalone speech type detection model to identify non-speech, single-speaker speech, and overlapping speech regions within the audio. The main contributions of this paper can be summarized as follows:

1) We extend our previous conference paper [35] by providing an expanded and in-depth analysis, as well as refining the previous flawed decoding method with clustering operations.
2) We propose a novel Enhancer module to help the frame-level speaker embedding incorporate useful information from the speaker attractor. The experiments show that the Enhancer module can help the system generalize to the unseen number-of-speaker scenario.
3) We explore the impact of simulation data configuration on our system performance and validate it on multiple datasets.
4) We compare the system performance when leveraging the transformer encoder or Conformer encoder, and validate it on multiple datasets.
5) We conduct a thorough investigation of the systems on the relevant datasets. Compared with other diarization systems, we achieve the new state-of-the-art performance on all the CALLHOME, AMI, and DIHARD II evaluation benchmarks when no oracle voice activity information is used.
6) We also evaluate the proposed AED-EEND system as a standalone speech type detection model. Our evaluation reveals that the proposed AED-EEND system exhibits notable competitiveness in detecting non-speech, single-speaker speech, and overlapping speech regions.

II. Novel Attention-Based Encoder-Decoder End-to-End Neural Diarization

In this section, we first review the related EEND systems. Then, we introduced our proposed system. For the introduction of our system, we first describe our designed attention-based encoder-decoder end-to-end neural diarization (AED-EEND) system, highlighting its key components and architecture. Then, we propose the Enhancer module to improve the frame-level speaker embedding representation. And next, we show how we train our system with the teacher-forcing strategy. Finally, we will describe our proposed iterative decoding strategy to output the diarization results for each speaker sequentially.

[1] In the context of our method, end-to-end refers to the capability of optimizing each module in an end-to-end manner.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

A. End-to-End Neural Diarization System

Unlike traditional stage-wised diarization systems [6], [7], the EEND system takes acoustic features as input and directly outputs the diarization results. In [16], Fujita et al. reformulated the diarization task as a frame-wise multi-class classification problem and optimized the system using the permutation-free objectives in an end-to-end manner. This approach was further evolved by the replacement of the BLSTM encoder with a Transformer encoder, enhancing performance significantly [17]. Additionally, Liu et al. [38] successfully experimented with the Conformer to improve the Transformer encoder, yielding even better results. Exploring different paradigms, Huang et al. [39] approached diarization as a one-dimensional object detection task, applying the renowned region proposal network [40] from object detection to speaker diarization. While the EEND system adeptly addresses overlapping speech, its fixed neural network architecture limits the output speaker count. Horiguchi et al. [21], [34] tackled this limitation using an LSTM-based attractor decoder, and Fujita et al. [19] proposed a chain rule method to transcend the speaker number constraints in EEND systems.

Although the EEND architecture is relatively mature, it still demands significant enhancements for improved performance. Researchers have augmented its capabilities by integrating it with other kinds of systems [22], [32], [41], [42], [43], adding extra loss constraints [44], [45], [46], [47], [48], adding extra useful information [49], [50], and designing more efficient data simulation methods [36], [37], [51]. Besides, the output speaker number of EEND system is always constrained by the training data [34] and the authors in [52] proposed the global and local attractor strategy to alleviate this problem. Additionally, developing effective online EEND systems [53], [54] remains a promising research direction.

In this paper, we focus on designing a new paradigm to improve the EEND system's ability in the variable number-of-speaker scenario.

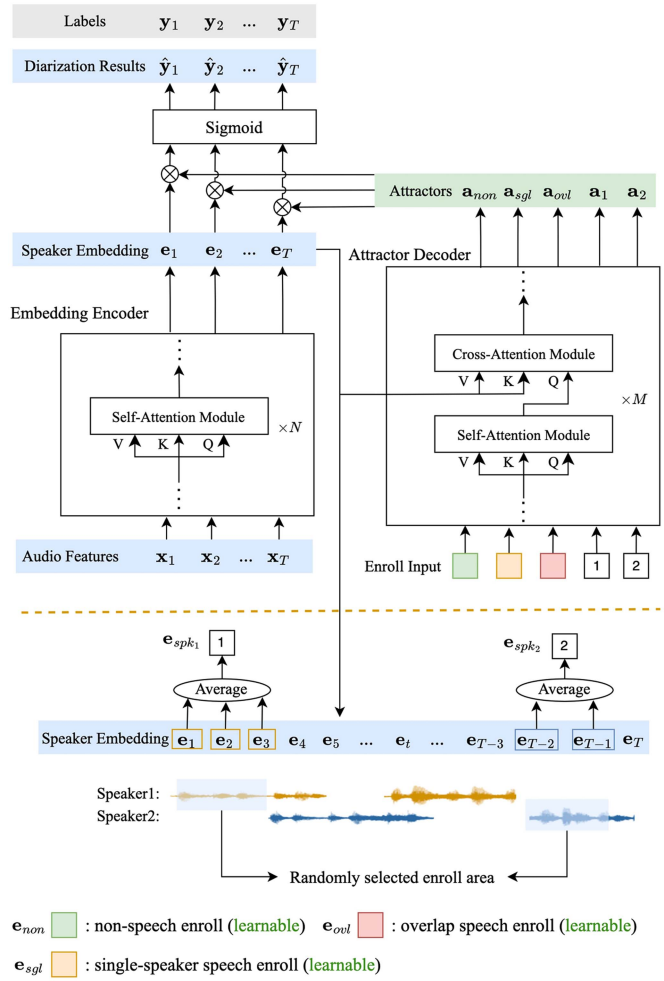

Fig. 1. AED-EEND system architecture when the speaker number is two. The part above the orange dotted line is our main system architecture, which is introduced in Section II-B. The part below the line shows our strategy to get the enrollment embedding in the training process, which is introduced in II-D.

B. Attention-Based Encoder-Decoder Neural Diarzation

In this section, we introduce our proposed AED-EEND system. We just follow the architecture of the original transformer encoder-decoder network proposed in [55]. Following the approach in [34], we made a modification by excluding the use of positional embeddings, and we found that this change hardly affected the system's performance. As depicted in the upper part of Fig. 1, the encoder in our AED-EEND system takes the audio feature sequence, denoted as $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbb{R}^{F\times T}$, as input and produces the frame-level speaker embedding sequence denoted as $E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T] \in \mathbb{R}^{D\times T}$. On the other hand, the decoder, referred to as the attractor decoder, follows the naming convention used in [21], [34]. The attractor decoder takes a sequence of enrollment embeddings, denoted as $E_{\text{enroll}} = [\mathbf{e}_{\text{non}}, \mathbf{e}_{\text{sgl}}, \mathbf{e}_{\text{ovl}}, \mathbf{e}_{\text{spk}_1}, \ldots, \mathbf{e}_{\text{spk}_S}] \in \mathbb{R}^{D\times(S+3)}$, as input. Here, $\mathbf{e}_{\text{non}}$, $\mathbf{e}_{\text{sgl}}$ and $\mathbf{e}_{\text{ovl}}$ represent the enrollment embeddings for non-speech, single-speaker speech, and overlapping speech, respectively. Furthermore, $\mathbf{e}_{\text{spk}_i}$ corresponds to the enrollment embedding for the i-th speaker in the recording. The attractor decoder generates the corresponding attractor, denoted as $A = [\mathbf{a}_{\text{non}}, \mathbf{a}_{\text{sgl}}, \mathbf{a}_{\text{ovl}}, \mathbf{a}_{\text{spk}_1}, \ldots, \mathbf{a}_{\text{spk}_S}] \in \mathbb{R}^{D\times(S+3)}$, for each enrollment input.

Our decoder architecture in the AED-EEND system differs from the LSTM-based attractor decoder presented in [21], as it incorporates both a self-attention module and a cross-attention module. The self-attention module enables the enrollment embeddings to interact with each other, resulting in more distinct attractors being generated. Meanwhile, the cross-attention module allows the enrollment embeddings to attend to all frame-level speaker embeddings, ensuring that the output attractors capture more relevant information from the frame-level speaker embeddings. In contrast to the EEND-EDA system described in [21], which only outputs attractors for existing speakers in the recording, our proposed AED-EEND system also produces attractors for three different speech activities: non-speech, single-speaker speech, and overlapping speech. Originally, this design choice was motivated by the decoding algorithm outlined in Section II-E, which necessitates predicting the speech of a single person. Moreover, we believe that modeling a broader

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

range of attractors contributes to improved system performance. Additionally, the results presented in Section V demonstrate that our AED-EEND system can independently function as a speech type detection system.

With the extracted frame-level speaker embedding sequence E and attractor sequence A, we can calculate the posterior probability that each speaker embedding belongs to the specific attractor based on a simple matrix multiplication operation:

$$\hat{Y} = \sigma\left(A^\top E\right) \in (0,1)^{(S+3)\times T} \tag{1}$$

where the $\sigma(\cdot)$ symbol corresponds to the element-wise sigmoid function and $^\top$ is the matrix transpose operation. Here, we denote the posterior probabilities for all the speech activities and speakers at the tth frame as: $\hat{\mathbf{y}}_t = [\hat{y}_t^{\text{non}}, \hat{y}_t^{\text{sgl}}, \hat{y}_t^{\text{ovl}}, \hat{y}_t^1, \ldots, \hat{y}_t^S]^\top \in (0,1)^{(S+3)}$ and we denote the correponding groud-truth label at t^{th} frame as $\mathbf{y}_t = [y_t^{\text{non}}, y_t^{\text{sgl}}, y_t^{\text{ovl}}, y_t^1, \ldots, y_t^S]^\top \in \{0,1\}^{(S+3)}$. A label value of 1 indicates the presence of speech activity or a speaker in the tth frame, while 0 indicates the absence. Then, we calculate the loss for each utterance by averaging the binary cross-entropy between posterior probability and ground-truth label across all the attractors and frames:

$$\mathcal{L} = \frac{1}{T(S+3)} \sum_{t=1}^{T} \sum_{s\in\mathbb{S}} \left[-y_t^s \log \hat{y}_t^s - (1-y_t^s)\log\left(1-\hat{y}_t^s\right)\right] \tag{2}$$

where $\mathbb{S} = \{\text{non}, \text{sgl}, \text{ovl}, 1, .., S\}$.

C. Frame-Level Speaker Embedding Enhancer

In the cross-attention module of the attractor decoder within our proposed AED-EEND system, we utilize the enrollment input or intermediate outputs as the query, while the frame-level embeddings from the encoder serve as the key and value. This architectural design enables the model to extract valuable information from the frame-level embeddings, leading to the generation of more informative attractors. In a similar vein, we explore the possibility of designing a complementary structure that allows the frame-level embeddings to gather more useful information from the attractors, thereby improving the quality of the frame-level embeddings. To this end, we introduce the Embedding Enhancer (EE) module, which is illustrated in Fig. 2. The EE module takes the frame-level speaker embeddings and all the attractors as input. In contrast to the attractor decoder shown in Fig. 1, the EE module treats the speaker embeddings as the query in the cross-attention module, while the attractors are used as the key and value. This arrangement facilitates the flow of information from the attractors to the frame-level speaker embeddings. Surprisingly, results in Tables VII and XI indicate that this module can help the system generalize better to evaluation sets with unseen numbers of speakers. Here, we denote the enhanced embedding obtained from the EE module as $\bar{E} = [\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2, \ldots, \bar{\mathbf{e}}_T] \in \mathbb{R}^{D\times T}$. By applying the similar operation described in (1), we derive the updated posterior probabilities as $\bar{\mathbf{y}}_t = [\bar{y}_t^{\text{non}}, \bar{y}_t^{\text{sgl}}, \bar{y}_t^{\text{ovl}}, \bar{y}_t^1, \ldots, \bar{y}_t^S]^\top \in (0,1)^{(S+3)}$. The loss function for the enhanced embedding can be defined as

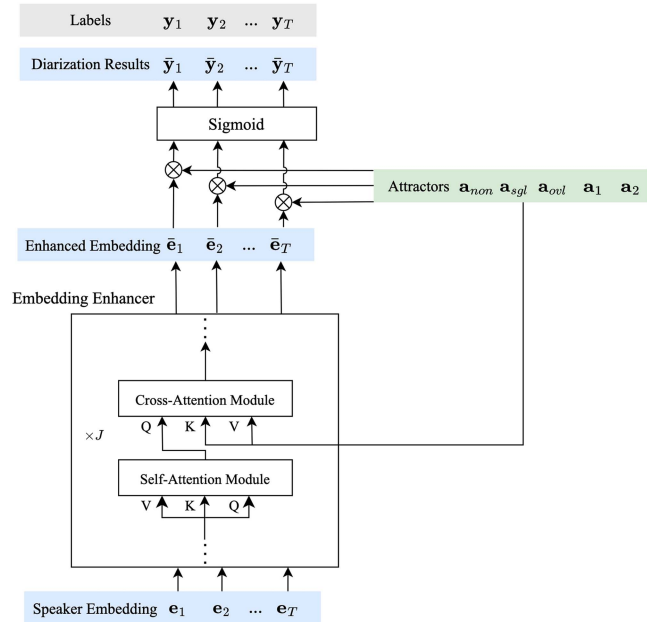

Fig. 2. Embedding Enhancer module when the speaker number is two.

follows:

$$\mathcal{L}_{EE} = \frac{1}{T(S+3)} \sum_{t=1}^{T} \sum_{s\in\mathbb{S}} \left[-y_t^s \log \bar{y}_t^s - (1-y_t^s)\log\left(1-\bar{y}_t^s\right)\right] \tag{3}$$

When we add the EE module to our AED-EEND system and get the AED-EEND-EE system, we optimize the whole system with the summation of loss $\mathcal{L}$ and loss $\mathcal{L}_{EE}$:

$$\mathcal{L}_{total} = \mathcal{L} + \mathcal{L}_{EE} \tag{4}$$

D. Model Optimization With Teacher Forcing Strategy

As described in Section II-B, our attractor decoder requires enrollment embeddings for each kind of speech activity and all the speakers existing in the utterance as input. This concept of enrollment embedding closely resembles that of the TS-VAD system [28]. In the TS-VAD system, the enrollment embedding is obtained from a pre-trained speaker embedding extractor. In Fig. 1, the bottom part illustrates that we derive the speaker enrollment embedding directly from the frame-level speaker embedding sequence by averaging a subset of embeddings associated with the specific speaker. Moreover, unlike the speakers' identities, which vary across utterances, the three types of speech activities exist in all utterances. Hence, we directly set the enrollment embeddings for the three distinct types of speech activity as learnable vectors.

In the training of automatic speech recognition (ASR) systems, the teacher forcing strategy [56] is commonly employed. This strategy involves feeding the system with the ground-truth word (token) to predict the subsequent word (token). By adopting this approach, model training becomes more stable, and convergence is achieved at a faster rate. Inspired by a similar concept, we adopt a comparable strategy in our training phase. We extract the single-speaker speaking region for each individual

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

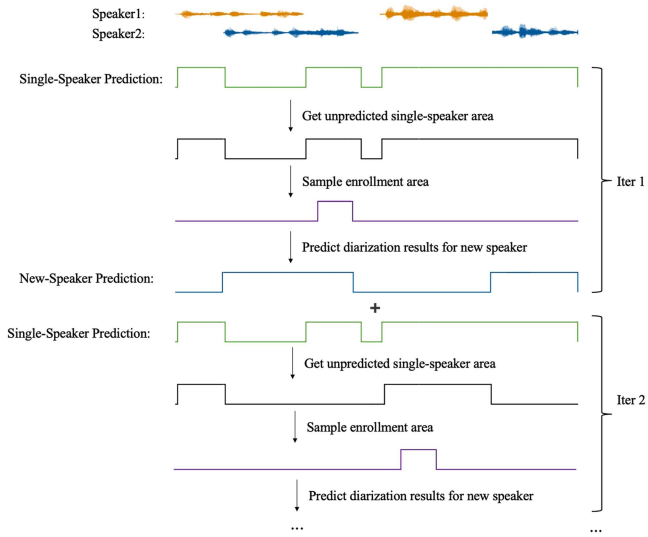

Fig. 3. Overall pipeline of the iterative decoding method.

from the ground-truth label and randomly sample a contiguous region with a predefined enrollment length (EL), denoted as L_{Enroll}, as the enrollment area. Subsequently, we compute the enrollment embedding by averaging the speaker embeddings within the enrollment area.

E. Iterative Decoding Algorithm for Inference

In this section, we present our proposed iterative decoding method for the AED-EEND system. As discussed in Section II-D, we utilize the ground-truth label to find the enrollment area for each speaker in the training process, but such a strategy can not be used in the evaluation process. In the TS-VAD method [28], an extra diarization system is used to get the single-speaker speaking area for each person. We can certainly use the same method as TS-VAD to obtain the enrollment area, but for the simplicity of the system, we propose an iterative decoding method based on some heuristic strategies to find the single-speaker speaking area for each speaker.

The proposed iterative decoding method follows a general process, as illustrated in Fig. 3, which encompasses three key operations:

1) Get the unpredicted single-speaker speaking area.
2) Sample a continuous area from the unpredicted single-speaker speaking area, while imposing constraints to ensure that only one speaker is present within the chosen area.
3) Use the sampled continuous area as the new enrollment area to predict the diarization results for a new speaker.
4) Repeat steps one to three until the results of all speakers have been decoded.

The detailed implementation of our iterative decoding method is outlined in Algorithm 1. Initially, we input the enrollment embeddings $[\mathbf{e}_{\text{non}}, \mathbf{e}_{\text{sgl}}, \mathbf{e}_{\text{ovl}}]$ into our AED-EEND system to obtain the index list I_{sgl} representing frames with single-speaker activity. Subsequently, we derive the index list I containing the frames that exist in I_{sgl} but not in I_{spk}, where I_{spk} encompasses

Algorithm 1: Proposed Iterative Decoding Algorithm.

Data: $I = [t_1, t_2, ...]$: frame indexes list for embedding in E
$C = [I_1, I_2, ..., I_K]$: K time-continuous segments from I, and the indexes in I_k are sorted in order
Enroll Length (EL): L_{Enroll}
Stop Decoding Length (SDL): L_{stop}

```
// get the active frame indexes list for three kinds of speeches
I_non, I_sgl, I_ovl = AED-EEND(e_non, e_sgl, e_ovl)
// active frame indexes list for speakers; enroll embedding set
I_spk = []; E = {}
while True do
    // get the un-predicted single-speaker area
    I = I_sgl − I_sgl ∩ I_spk
    // get segments with length ≥ L_Enroll
    C' = [I'_1, I'_2, ..., I'_K'] = filter_segs(C, L_Enroll)
    // get the segment with the longest length
    I_longest = get_longest_seg(C')
    L_Enroll_tmp = min(length(I_longest), L_Enroll)
    if length(C') == 0) then
        C'.add(I_longest)
    if L_stop > length(I_longest) then
        break; // stop decoding
    case Init-Decode do
        I_enroll = [t_1, t_2, ..., t_{L_Enroll_tmp}] ∈ I'_1
    case Rand-Decode do
        // randomly select a segment from C'
        I'_k = random_select_seg(C')
        // randomly select a consecutive sub-segment with length
           L_Enroll_tmp
        I_enroll = random_select_sub-seg(I'_k, L_Enroll_tmp)
    case SC-Decode do
        // The index in I is clustered based on corresponding embedding
        [I^cls_1, I^cls_2, ...] = spectral_cluster(I)
        I^cls_k = get_longest_seg([I^cls_1, I^cls_2, ...])
        I_enroll = random_select_sub-seg(I^cls_k, L_Enroll_tmp)
    case SC-Decode-Local do
        [I^cls_1, I^cls_2, ...] = spectral_cluster(I_longest)
        I^cls_k = get_longest_seg([I^cls_1, I^cls_2, ...])
        I_enroll = random_select_sub-seg(I^cls_k, L_Enroll_tmp)
    // average the embeddings with indexes in I_enroll
    e = average(E_{I_enroll})
    E.add(e)
    I_spk = (AED-EEND(E))
```

Output: I_{spk}

the frame indices predicted for the speakers in previous iterations. The un-predicted single-speaker area is thus captured by I. Next, our aim is to find a continuous area of enrollment length (EL) L_{Enroll} from I to serve as the new enrollment area, with the objective of ensuring that this area contains only one speaker. To accomplish this, we have devised four heuristic strategies:

- **Init-Decode**: In this strategy, we sample the initial L_{Enroll} length area from the first continuous segment in I as the new enrollment area.
- **Random-Decode**: Here, we randomly select a continuous segment, denoted as I'_k, from I. From I'_k, we further randomly select a continuous area with a length of L_{Enroll} as the new enrollment area
- **SC-Decode**: This strategy involves performing spectral clustering [11], [12] on all the embeddings in I to obtain the embedding index cluster list I^{cls}_k with the longest length. From I^{cls}_k, we randomly sample a segment with a length of L_{Enroll} as the new enrollment area.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SIMULATION DATASET CONFIGURATION. OVL CORRESPONDS TO THE OVERLAP RATIO

Dataset	Split	#Spk	#Mixtures	β	Ovl (%)	Duration (hrs)
SM-1spk	Train	1	100,000	2	0.0	2159
	Test	1	500	2	0.0	10.76
SM-2spk	Train	2	100,000	2	34.1	2484
	Test	2	500	2	34.4	12.33
SM-3spk	Train	3	100,000	5	34.2	4226
	Test	3	500	5	34.7	20.92
SM-4spk	Train	4	100,000	9	31.5	6647
	Test	4	500	9	32.0	33.03
SM-5spk	Train	5	100,000	13	30.3	9202
	Test	5	500	13	30.7	45.20
SC-1spk	Train	1	36,989	-	0.0	2159
	Test	1	223	-	0.0	12.45
SC-2spk	Train	2	24,343	-	8.18	2481
	Test	2	118	-	8.13	12.41
SC-3spk	Train	3	29,297	-	11.5	4226
	Test	3	86	-	11.1	12.44
SC-4spk	Train	4	35,640	-	13.4	6647
	Test	4	66	-	12.9	12.43
SC-5spk	Train	5	40,249	-	14.5	9202
	Test	5	55	-	14.8	12.50

- **SC-Decode-Local**: We observed that spectral clustering can be time-consuming for long sequences. In this strategy, we limit the spectral clustering process to the longest continuous segment within I. Subsequently, we apply the same procedure as SC-Decode to identify the new enrollment area. It is worth noting that, because this method only performs clustering locally, it is much faster than the SC-Decode.

While the Init-Decode and Random-Decode strategies do not explicitly impose constraints to ensure the selection of areas with only one speaker, the experiments conducted in Section IV-A1 demonstrate that using a very short enrollment area can yield remarkably good performance. The utilization of a short enrollment area significantly increases the likelihood of capturing an area containing only one speaker. Furthermore, we have developed a criterion to determine when to terminate the iterative decoding process. Specifically, when the length of the longest continuous segment within I is below a pre-defined stop decoding length (SDL) of L_{stop}, the decoding process is terminated.

III. EXPERIMENTAL SETUP

A. Data Corpus

1) Dataset With Simulation Recordings: To have a fair comparison with the EEND-EDA system, we just follow [34] to simulate 5 different sub-sets of simulation data. The detailed information regarding the simulation data can be found in the upper part of Table I.

The utterances in each subset have a fixed number of speakers, and the number of speakers varies across different subsets, ranging from 1 to 5. Although this simulation setup has been widely used in previous works [16], [17], [19], [21], [34], it should be noted that there is a significant mismatch in the overlap ratio between the simulated data and the real recordings, as indicated in Table II. We believe that this configuration may not be optimal.

TABLE II
DATASETS OF REAL RECORDINGS. THE THREE NUMBERS IN THE DURATION COLUMN CORRESPOND TO THE MINIMUM DURATION/MAXIMUM DURATION/AVERAGE DURATION. OVL CORRESPONDS TO THE OVERLAP RATIO

Dataset	Split	#Spk	#Utt	Ovl (%)	Duration (mins)
CALLHOME-2spk [57]	Part 1	2	155	14.0	0.862/2.211/1.234
	Part 2	2	148	13.1	0.875/2.233/1.202
CALLHOME-3spk [57]	Part 1	3	61	19.6	0.947/6.352/2.071
	Part 2	3	74	17.0	0.774/8.210/2.418
CALLHOME [57]	Part 1	$2-7$	249	17.0	0.862/10.11/2.097
	Part 2	$2-6$	250	16.7	0.774/10.01/2.053
AMI headset mix [58]	Train	$3-5$	136	13.4	7.965/90.25/35.59
	Dev	4	18	14.1	15.73/49.50/32.22
	Test	$3-4$	16	14.6	13.98/49.53/33.98
DIHARD II [59]	Dev	$1-10$	192	9.8	0.447/11.62/7.442
	Test	$1-9$	194	8.9	0.631/13.50/6.957

In recent studies [36], [37], the authors proposed a methodology that leverages the statistics from real recordings to guide the synthesis of simulated data, resulting in improved performance on real datasets. Additionally, the authors in [36] referred to the simulation data in [21], [34] as simulated mixtures (SM), while they named the simulation data in their own paper as simulated conversion (SC). In our paper, we adopt these terminologies. Following the approach outlined in [36], we extracted statistics from Part1 of the CALLHOME dataset (as shown in Table II) and simulated an equal amount of SC data to match the SM data. The statistics presented in Table I illustrate that the SC data exhibits a significantly smaller overlap ratio compared to the SM data. While the effectiveness of SC data has been validated in previous studies [36], [37], it is important to note that the experiments in [36] were primarily conducted using only the 2-speaker CALLHOME data and DIHARD III CTS data, and the baselines employed in [37] were relatively weaker. In our experiments, we aim to evaluate the effectiveness of SC data in more diverse scenarios and employ a strong system.

2) Dataset With Real Recordings: In our experiments, we evaluate our systems on three different datasets with real recordings, the CALLHOME dataset [57], AMI dataset with mixed headset [58] and the DIHARD II [59] dataset.[2] The CALLHOME dataset is divided into two parts according to the kaldi recipe[3]. Part 1 is used for model adaptation, while Part 2 is used for evaluation. The AMI dataset consists of three parts: Train, Dev, and Test. We perform model adaptation on the Train part and evaluate the system on both the Dev and Test parts. For the DIHARD II dataset, we conduct model adaptation on the Dev part and evaluate on the Test part. The CALLHOME dataset comprises 8 k telephone-channel recordings, the AMI dataset consists of 16 k meeting recordings, and the DIHARD II dataset contains 16 k recordings from a diverse range of sources. To simplify the training setup, we downsampled the AMI and DIHARD II datasets to 8 k in our experiments.

B. Model Configuration

For all the audio data in our experiments, we extract 345-dimensional acoustic features following the

[2]Unfortunately, we do not report results for the DIHARD III dataset as it is not freely available, and we only have access to the DIHARD II dataset.

[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v2

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

TABLE III
MODEL CONFIGURATION

Model	Param #	Layer# / Feed-Forward Dimension		
		Encoder	Decoder	Enhancer
AED-EEND	11.6M	4/2048	4/2048	0/-
AED-EEND (conformer)	10.4 M	4/1024	4/1024	0/-
AED-EEND-EE	11.6M	4/2048	4/2048	4/2048
AED-EEND-EE (small)	6.4M	4/1024	4/512	4/512
AED-EEND-EE (conformer)	10.4M	4/1024	4/1024	4/1024
EEND-EDA	6.4M	4/2048	-	-

methodology described in [21], [34]. Consequently, the input dimension for all our systems is set to 345. In the attention module, we configure the number of heads to 4 and the attention unit number to 256. Additional system configurations can be found in Table III. It should be noted that, when we use the Embedding Enhancer (EE) module, we share parameters between the second to fourth layers of the decoder and Enhancer, which equals that the decoder and Enhancer only have two layers containing parameters. With this setup, our AED-EEND-EE system has the same number of parameters as the AED-EEND system. Additionally, we experimented with replacing the transformer encoder with the Conformer [60] module. The Conformer module uses a convolution kernel size of 31, and we adjusted the feed-forward dimension of the encoder and decoder to achieve a similar parameter count as the AED-EEND system. Although the AED-EEND system shares the same encoder as the EEND-EDA system, the attention-based decoder in the AED-EEND system has more parameters compared to the LSTM-based decoder in the EEND-EDA system. To ensure a fair parameter comparison, we designed a smaller system called AED-EEND-EE (small), which matches the parameter count of the EEND-EDA system.

C. Training Setup

In our experiment, the training process is divided into two stages: pre-training with the simulated dataset and adaptation with the real dataset. The model trained solely on the simulation data is evaluated on the simulation evaluation set, while the adapted model is evaluated on the real dataset. We randomly sample the enrollment length introduced in Section II-D from 1 s to 3 s. We also randomly drop the enrollment embeddings for all the speakers existing in each utterance with a probability of 0.5 in the training process, which ensures the system's robustness when encountering scenarios where not all speakers' enrollment embeddings are available during the iterative decoding process.

The acoustic features mentioned in Section III-B, with a dimensionality of 345, are extracted using a window size of 25 ms and a hop length of 10 ms. During the training process, we downsample the acoustic feature sequence by a factor of ten, resulting in a frame resolution of 0.1 s. For the pre-training stage, all utterances are divided into segments of 50 s in length, from which we randomly sample 64 segments to construct the training batch. As for the adaptation stage, following the approach in [34], we use 50 s segments for the CALLHOME dataset and 200 s segments for the AMI and DIHARD II datasets. In the adaptation stage, we set the batch size to 32. During the pre-training stage, we utilize the Noam optimizer [55] and configure the training process with 100 epochs and 200,000 warmup steps. For the adaptation stage, we employ the Adam optimizer with a learning rate of 0.00001.

In our experiment, the evaluation of the system can be categorized into two scenarios: the fixed number-of-speaker scenario and the flexible number-of-speaker scenario. In the fixed number-of-speaker scenario, the pre-training data, adaptation data, and evaluation data have the same number-of-speaker. In the flexible number-of-speaker scenario, we perform pre-training by combining the simulation data from Table I for a range of speaker numbers, either from 1 speaker to 4 speakers or from 1 speaker to 5 speakers, depending on the specific experiment. Subsequently, we adapt the pre-trained model using the specific real adaptation set.

D. Evaluation Setup

To ensure a fair comparison with the EEND-EDA system [34], we adopted most of their evaluation configurations. For the simulation and CALLHOME dataset, we downsampled the acoustic feature sequence by 10 times and evaluated the DER (Diarization Error Rate) with a collar tolerance of 0.25 s. In their study, the authors downsampled the feature sequence by 5 times for the AMI and DIHARD II evaluations, without using collar tolerance. However, in our experiments, we found that the AMI dataset is not highly sensitive to the downsampling rate and contains significantly longer recordings. As a result, we used a downsampling rate of 10 times for the AMI evaluation and 5 times for the DIHARD II evaluation. No collar is used for AMI and DIHARD II evaluation. We use 0.5 as the threshold to get the decision for diarization results. Additionally, it is worth noting that our experiments did not employ any oracle speech segments. This deliberate choice allowed us to assess the performance of our single system independently and evaluate its capabilities in the diarization task without relying on additional information.

IV. EVALUATION ON SPEAKER DIARIZATION TASK

A. Fixed Number of Speakers Scenario

1) Decoding Methods Comparison: In Section II-E, we introduced four decoding methods that are utilized for generating the diarization results iteratively. Additionally, we defined two crucial hyperparameters, namely, the enrollment length (EL) and stop decoding length (SDL). These hyperparameters play a significant role in assisting the algorithm to identify the enrollment area and determine the appropriate time to stop the iteration process. In this section, we will provide a comprehensive comparison of the different decoding methods and explore the impact of varying hyperparameter values on the results. Table IV presents the corresponding results obtained from the SM simulation dataset and CALLHOME dataset when the number of speakers is 2.

The upper part of Table IV investigates the influence of different enrollment length values and the oracle speaker number is used. When utilizing a short enrollment length with the SM dataset, we observe that all of our proposed decoding methods

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
DER (%) RESULTS COMPARISON FOR DIFFERENT DECODING METHODS WITH VARYING ENROLL LENGTHS AND STOP DECODING LENGTHS. ALL THE RESULTS ARE BASED ON OUR 2-SPK SYSTEMS. EL: ENROLL LENGTH. SDL: STOP DECODING LENGTH. GT-DECODE STANDS FOR GROUND TRUTH DECODE, WHICH REFERS TO THE TEACHER-FORCING STRATEGY EMPLOYED DURING THE TRAINING PROCESS. RESULTS MARKED WITH A GRAY BACKGROUND INDICATE VERY POOR PERFORMANCE. IN THE RESULTS IN THE UPPER PART OF THE TABLE, WE ASSUME THAT THE NUMBER OF SPEAKERS IS KNOWN, WHICH IS 2. IN THE RESULTS OF THE LOWER PART OF THE TABLE, THE NUMBER OF SPEAKERS IS DETERMINED THROUGH SDL

Decoding Method	DER (%) on 2-spk SM with Different EL							DER (%) on 2-spk CALLHOME with Different EL						
	0.1s	0.5s	1s	2s	3s	5s	10s	0.1s	0.5s	1s	2s	3s	5s	10s
GT-Decode	3.10	3.22	3.22	2.95	3.05	2.87	2.88	8.39	8.03	7.58	7.56	7.18	7.27	7.19
Init-Decode	5.18	3.14	3.14	6.34	23.8	67.1	97.5	40.78	10.9	7.90	8.71	15.1	35.1	86.1
Rand-Decode	3.36	3.13	3.18	6.73	24.1	67.2	97.5	11.36	8.32	8.38	9.39	16.0	35.5	86.1
SC-Decode	3.13	3.14	3.12	3.17	3.16	3.17	8.27	8.61	7.75	7.86	8.04	7.68	7.81	8.78
SC-Decode-Local	3.47	3.48	3.50	6.94	24.63	67.36	97.50	8.39	7.81	7.76	8.93	15.95	35.48	86.15
	DER (%) on 2-spk SM with Different SDL							DER (%) on 2-spk CALLHOME with Different SDL						
	0.5s	0.8s	1s	1.5s	2s	2.5s	-	0.5s	0.8s	1s	1.5s	2s	2.5s	-
Init-Decode	4.92	3.55	3.40	3.56	4.88	10.03	-	19.4	10.9	10.7	10.8	10.6	12.2	-
Rand-Decode	4.06	3.68	3.50	3.51	5.53	10.85	-	13.2	9.35	8.58	8.20	8.58	10.4	-
SC-Decode	3.98	3.55	3.25	3.53	6.59	13.08	-	13.0	8.28	8.34	8.47	9.36	11.5	-
SC-Decode-Local	4.35	3.71	3.57	3.61	6.75	13.80	-	11.05	7.98	7.87	8.24	9.13	11.49	-

demonstrate reasonable performance, with some approaches even approaching the performance of the GT-Decode method. Because there is a great deal of randomness in Init-Decode and Rand-Decode, the performance degrades seriously when enrollment length becomes longer, which may cause the enrollment part to contain more than one speaker. Besides, the SC-Decode-Local method also demonstrates poorer performance with longer enrollment lengths. Similar patterns are observed when evaluating different decoding methods on the CALLHOME dataset. Additionally, the Init-Decode method performs poorly when an extremely short enrollment area (0.1 s) is used for the CALLHOME dataset, possibly due to inaccurate boundary predictions. Considering the robustness of a short enrollment area, we adopt a default enrollment length setup of 0.5 s for subsequent experiments.

In the subsequent analysis, we depart from the assumption of knowing the oracle number of speakers and instead utilizing a pre-defined stop decoding length (SDL) to determine the number of speakers. The corresponding results are presented in the lower part of Table IV. Our findings indicate that employing excessively long or short SDL values yields less desirable results. Remarkably, the most robust outcomes were achieved when utilizing a SDL of 1 s. This observation aligns with intuitive expectations, as an overly long SDL may overlook certain speakers, while an excessively short SDL can lead the system to falsely predict additional speakers. Moving forward, we will adopt a default SDL of 1 s in our subsequent experiments when the speaker number is unknown. Additionally, for scenarios with a fixed number of speakers, we will use the SC-Decode strategy as the default decoding method. To make decoding faster, we will employ the SC-Decode-Local method for scenarios with a flexible number of speakers.

2) Comparison Between Different Systems: In this section, we compare our proposed system with others when the speaker number is fixed and known in advance. The results in Table V show that our system achieves the best results on all conditions. Besides, equipped with our proposed Enhancer module, most of the evaluation sets are improved.

TABLE V
DER (%) RESULTS ON THE FIXED NUMBER OF SPEAKERS CONDITION

Method	SM data DER(%)		CALLHOME DER(%)	
	2-spk	3-spk	2-spk	3-spk
x-vector clustering [21]	28.77	31.78	11.53	19.01
BLSTM-EEND [16]	12.28	-	26.03	-
SA-EEND [17], [21]	4.56	8.69	9.54	14.00
SC-EEND [19]	-	-	8.86	-
EEND-EDA [21]	2.69	8.38	8.07	13.92
EEND-EDA †	4.20	10.51	8.32	17.07
TS-VAD [31]	-	-	9.51	-
MTFAD [31]	-	-	7.82	-
AED-EEND	3.14	5.16	**7.75**	12.87
AED-EEND-EE	**2.46**	**4.26**	8.18	**12.21**

†: our implementation

TABLE VI
DER (%) RESULTS ON THE SM SIMUATION DATASET WITH FLEXIBLE NUMBER OF SPEAKERS. SYSTEMS ARE TRAINED ON 1-4 SPEAKERS SM DATASET

Method	spk#			
	1	2	3	4
x-vector clustering [21]				
Estimated #spk	37.42	7.74	11.46	22.45
Oracle #spk	1.67	28.77	31.78	35.76
SC-EEND [19]				
Estimated #spk	0.76	4.31	8.31	12.50
EEND-EDA [21]				
Estimated #spk	0.39	4.33	8.94	13.76
Oracle #spk	0.16	4.26	8.63	13.31
AED-EEND				
Estimated #spk	0.07	2.72	5.56	10.07
Oracle #spk	0.07	2.93	6.00	9.72
AED-EEND-EE				
Estimated #spk	**0.07**	**2.45**	**4.71**	7.04
Oracle #spk	0.09	2.73	5.83	**6.65**

B. Flexible Number of Speakers Scenario

In this section, all the systems will be evaluated on the dataset that the speaker number is flexible.

1) Results on Simulation Dataset: Firstly, we assess the performance of our system on the SM evaluation set and present the results in Table VI, considering scenarios where the speaker number is estimated or the oracle speaker number is known.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

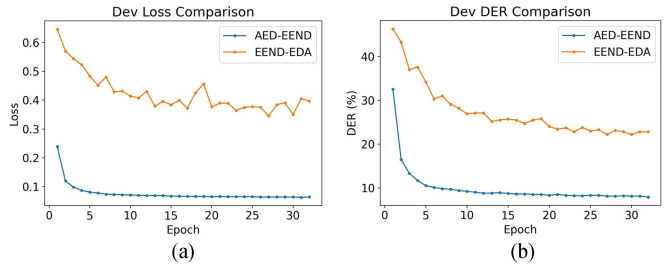

Fig. 4. Convergence speed comparison between EEND-EDA and our proposed AED-EEND in the first 30 epoches. Both systems are trained on the SM dataset with 1-5 speakers.

TABLE VII
DER (%) RESULTS ON THE CALLHOME DATASET WITH FLEXIBLE NUMBER OF SPEAKERS. SYSTEMS ARE PRE-TRAINED ON 1-4 SPEAKERS SM DATASET

Method	spk#					
	2	3	4	5	6	all
x-vector clustering [21]						
Estimated #spk	15.45	18.01	22.68	31.40	34.27	19.43
Oracle #spk	8.93	19.01	24.48	32.14	34.95	18.98
EEND-EDA [21]						
Estimated #spk	8.50	13.24	21.46	33.16	40.29	15.29
Oracle #spk	8.35	13.20	21.71	33.00	41.07	15.43
AED-EEND						
Estimated #spk	**6.18**	**11.51**	18.44	30.79	39.90	13.25
Oracle #spk	6.35	11.54	19.08	29.58	34.59	13.16
AED-EEND-EE						
Estimated #spk	6.93	11.92	**17.12**	**28.22**	31.97	**12.91**
Oracle #spk	6.97	12.04	17.41	28.23	**31.78**	13.02

Notably, our systems exhibit the highest performance among all evaluated systems, and the introduction of our proposed Enhancer module notably enhances results for the 3-speaker and 4-speaker evaluation sets. Furthermore, our systems demonstrate consistent performance regardless of whether the speaker number is known or estimated, indicating the effectiveness of our proposed decoding strategy in accurately predicting the number of speakers.

As introduced in Section III-C, we directly trained our AED-EEND system on the variable number-of-speaker simulation dataset, whereas the EEND-EDA [34] system is first trained on the 2-speaker simulation data set and then further trained on the simulation set with more speakers. This additional step was necessary for the EEND-EDA system due to its utilization of permutation invariant (PIT) loss, which poses challenges during training. The pre-training on the 2-speaker simulation dataset enabled the EEND-EDA model to better adapt to datasets with a flexible number of speakers. In our case, our system does not encounter speaker permutation issues, which facilitates easier convergence during training. We validate this assumption in Fig. 4, where we compare the training progress of our AED-EEND system and the EEND-EDA system on the SM dataset with 1-5 speakers. The plotted loss and diarization error rate (DER) statistics for the development set clearly demonstrate that our AED-EEND system achieves significantly faster convergence compared to the EEND-EDA system. This observation implies that we can employ simpler training strategies to achieve superior diarization performance.

2) Results on CALLHOME Dataset: We proceeded to evaluate our systems using the CALLHOME real dataset, and the results are presented in Table VII. Interestingly, both the EEND-EDA system and our proposed AED-EEND system exhibit worse performance compared to the traditional x-vector clustering system in the 6-speaker condition. In [21], the authors suggest that this discrepancy arises from the pre-training stage, where the models only encounter data containing four or fewer speakers. Surprisingly, our proposed AED-EEND-EE system demonstrates a noteworthy capability to generalize to unseen speaker number scenarios, surpassing the performance of the x-vector clustering system in the 5-speaker and 6-speaker conditions. This is because our proposed EE module facilitates the flow of information from segment-level attractors to frame-level speaker embeddings, which enhances the speaker identity discriminative power of frame-level speaker embeddings. For the systems without the EE module, the speaker identity discriminative power of frame-level speaker embeddings is constrained by the speaker number present in the training set. However, compared to the results in Table VI, the overall performance gain from the Enhancer module becomes smaller and limited. We believe that our proposed Enhancer module does have strong modeling capabilities, but due to the mismatch between simulation data and real data, the model cannot generalize. The relevant experiments in the next section will also verify this point.

3) Enhance the Results on CALLHOME Dataset With Optimized Setup: To further enhance the performance on the CALLHOME dataset, we made modifications to the experimental setup used in Table VII and present the updated results in Table VIII. Firstly, we followed the approach outlined in [34] by incorporating a 5-speaker simulation dataset into our simulation training process. This change resulted in notable improvements for our proposed AED-EEND system in the evaluation sets with multiple speakers. Moreover, the results from Table VII demonstrated that our AED-EEND-EE system already obtains the ability to generalize to situations with an unknown number of speakers. Therefore, the performance improvement after adding the 5-speaker simulation data was insignificant. Additionally, our smaller version, AED-EEND-EE (small), introduced in Section III-B, which has a similar parameter count to the EEND-EDA system, also outperformed the EEND-EDA system. Moreover, AED-EEND-EE (small) exhibited comparable performance to the AED-EEND-EE system, indicating that the dimension of the transformer has minimal impact on system performance.

Statistics in Tables I and II highlight the significant difference in overlap ratio between the SM simulation data and the real data. To mitigate the mismatch between the simulation dataset and the real dataset, we replaced the SM dataset with the SC dataset, as described in Section III-A1. This substitution led to further improvements in the performance of the AED-EEND system. Additionally, we explored the replacement of the transformer encoder with the Conformer, which yielded additional enhancements. Surprisingly, our AED-EEND-EE system, trained on the SC data, even outperformed the two-stage systems listed at the bottom of Table VIII, achieving a new state-of-the-art performance on the CALLHOME evaluation set. Furthermore, we observed that the Enhancer module provided

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
CALLHOME DER (%) RESULTS WITH OPTIMIZED SETUP

Method	Simulation Data		spk#					
	Spk #	Datatype	2	3	4	5	6	all
EEND-EDA [21]	$k \in \{1, ..., 4\}$	SM	8.50	13.24	21.46	33.16	40.29	15.29
SC-EEND [19]	$k \in \{1, ..., 4\}$	SM	9.57	14.00	21.14	31.07	37.06	15.75
MTFAD [21]	$k \in \{1, ..., 4\}$	-	-	-	-	-	-	14.31
AED-EEND	$k \in \{1, ..., 4\}$	SM	6.18	11.51	18.44	30.79	39.90	13.25
AED-EEND-EE	$k \in \{1, ..., 4\}$	SM	6.93	11.92	17.12	28.22	31.97	12.91
EEND-EDA [34]	$k \in \{1, ..., 5\}$	SM	8.09	12.20	15.32	27.36	29.21	12.88
AED-EEND	$k \in \{1, ..., 5\}$	SM	6.26	11.54	17.28	29.05	30.61	12.56
AED-EEND-EE	$k \in \{1, ..., 5\}$	SM	6.42	11.47	17.20	27.15	29.36	12.43
AED-EEND-EE (small)	$k \in \{1, ..., 5\}$	SM	6.07	11.94	17.78	28.52	24.58	12.44
AED-EEND	$k \in \{1, ..., 5\}$	SC	5.84	11.02	15.32	27.68	27.19	11.61
AED-EEND + Conformer	$k \in \{1, ..., 5\}$	SC	**5.58**	10.49	13.06	26.47	24.06	10.66
AED-EEND-EE	$k \in \{1, ..., 5\}$	SC	5.69	9.81	12.44	23.35	21.72	**10.08**
AED-EEND-EE (Init-Decode)	$k \in \{1, ..., 5\}$	SC	5.83	**9.66**	13.73	24.46	25.74	10.57
AED-EEND-EE (Random-Decode)	$k \in \{1, ..., 5\}$	SC	5.66	9.68	14.62	24.35	22.44	10.55
AED-EEND-EE + Conformer	$k \in \{1, ..., 5\}$	SC	5.61	9.78	13.56	26.50	**20.04**	10.35
VBx [61] †	-	-	9.44	13.89	16.05	13.87	24.73	13.28
EEND-post [42] *	-	-	9.87	13.11	16.52	28.65	27.83	14.06
EEND-vector clust. [22] *	-	-	7.94	11.93	16.38	21.21	23.10	12.49
EEND-GLA-Large [23] *	-	-	7.11	11.88	14.37	25.95	21.95	11.84
EEND-vector clust. + WavLM [62] *	-	-	6.46	10.69	**11.84**	**12.89**	20.70	10.35
EEND-OLA + SOAP [63] *	-	-	5.73	10.31	11.96	23.89	20.39	10.14

†: Oracle speech segments are used
*: Two-stage systems

more significant improvements when trained on the SC data compared to the SM data. This phenomenon is consistent with the results in Tables VI and VII, where the Enhancer module exhibited significant improvements on the simulation evaluation set but yielded only marginal improvements on the real dataset. While we acknowledge the exceptional modeling capabilities of the Enhancer module, the substantial mismatch between the SM dataset and the real dataset limits its generalizability to the real dataset. Besides, we provide the results for the AED-EEND-EE (SC) system using the decoding methods Init-Decode and Random-Decode, where no explicit constraint is added to ensure there is one speaker in the enrollment area. Supervisingly, based on these two decoding methods, we also achieved a pretty good performance on CALLHOME evaluation set.

4) Speaker Counting Evaluation on CALLHOME: Next, we assess the accuracy of our system in predicting the number of speakers and present the speaker counting confusion matrix in Table IX. The results reveal that our systems exhibit lower performance in the speaker counting task compared to EEND-EDA, indicating that the performance of speaker counting and speaker diarization may vary across different system types. Interestingly, we observe a positive correlation between the performance of diarization and speaker counting in our AED-EEND system. Perhaps in the future, we can enhance our AED-EEND system with the speaker counting ability form EEND-EDA to achieve better performance.

5) Results on DIHARD II Dataset: In this section, we evaluate our system on the DIHARD II dataset, which offers a more challenging evaluation with a larger number of speakers and diverse conversation scenarios. The results are presented in Table XI. In the table, both our AED-EEND and

TABLE IX
SPEAKER COUNTING CONFUSION MATRIX EVALUATED ON CALLHOME DATASET. ALL THE SYSTEMS ARE PRE-TRAINED ON THE SIMULATION DATASET WITH 1-5 SPEAKERS

(a) EEND-EDA [34] (Accuracy=84.4 %)

Pred. #Speakers \ Ref. #Speakers	1	2	3	4	5	6
1	**0**	1	0	0	0	0
2	0	**142**	7	1	0	0
3	0	5	**54**	4	0	0
4	0	0	13	**14**	4	1
5	0	0	0	1	**1**	2
6	0	0	0	0	0	**0**

(b) AED-EEND (SM) (Accuracy=73.6 %)

Pred. #Speakers \ Ref. #Speakers	1	2	3	4	5	6
1	**0**	5	0	0	0	0
2	0	**139**	31	4	0	0
3	0	4	**42**	14	3	2
4	0	0	1	**2**	2	0
5	0	0	0	0	**0**	0
6	0	0	0	0	0	**1**

(c) AED-EEND-EE (SM) (Accuracy=75.6 %)

Pred. #Speakers \ Ref. #Speakers	1	2	3	4	5	6
1	**0**	5	0	0	0	0
2	0	**141**	25	4	0	0
3	0	2	**48**	16	2	2
4	0	0	1	**0**	2	0
5	0	0	0	0	**0**	1
7	0	0	0	0	1	0

(d) AED-EEND-EE (SC) (Accuracy=77.6 %)

Pred. #Speakers \ Ref. #Speakers	1	2	3	4	5	6
1	**0**	4	0	0	0	0
2	0	**134**	20	2	0	0
3	0	10	**53**	10	0	1
4	0	0	1	**6**	3	1
5	0	0	0	2	**1**	1
7	0	0	0	0	1	0

AED-EEND-EE systems pre-trained on SM data share the same training data setup as the EEND-EDA system, yet both of our systems outperform the EEND-EDA system on the DER metric. In [34], the iterative inference+ strategy is proposed to handle the problem of the number of outputs of EEND-EDA being empirically limited by its training dataset. It is worth noting that our AED-EEND-EE (SM) system outperforms the EEND-EDA (Iterative inference+) system a lot on both DER and JER metrics, which further improves that our AED-EEND-EE system can generalize better to the unseen number-of-speaker scenario.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

TABLE X
SPEECH TYPE DETECTION RESULTS. "OURS" DENOTES OUR AED-EEND-EE + CONFORMER SYSTEM TRAINED ON THE 1-5 SPEAKERS SC SIMULATION DATASET. FA REPRESENTS THE FALSE ALARM RATE (%) OR FALSE POSITIVE RATE (%). MISS REPRESENTS THE MISS ERROR RATE (%) OR FALSE NEGATIVE RATE (%). F_1 DENOTES THE F_1-SCORE (%). FOR COMPARISON WITH OTHER SYSTEMS, WE INVERTED OUR SYSTEM'S PREDICTION FOR NON-SPEECH TO OBTAIN THE PREDICTION FOR SPEECH. THE SILERO-VAD MODEL IS A VOICE ACTIVITY DETECTION MODEL, WHICH CAN ONLY DISTINGUISH SPEECH AND NON-SPEECH. THE PYANNOTE TOOLKIT ONLY PROVIDES THE INTERFACE FOR SPEECH AND OVERLAP AREA PREDICTION, AND WE DERIVE THE PREDICTION FOR SINGLE-SPEAKER SPEECH AREA FROM THESE TWO PREDICTIONS

Dataset	System	Speech			Single-Speaker			Overlap		
		FA	MISS	F_1	FA	MISS	F_1	FA	MISS	F_1
CALLHOME	Silero-Vad [64]	24.40	8.24	94.29	-	-	-	-	-	-
	Pyannote [65], [66]	**17.10**	7.07	95.34	38.38	11.73	87.65	**3.58**	57.59	52.12
	Ours	25.56	**3.15**	**96.92**	**27.00**	**8.67**	**91.08**	3.98	**33.77**	**70.13**
DIHARD II	Silero-Vad	40.50	9.87	88.23	-	-	-	-	-	-
	Pyannote	21.72	6.04	93.23	28.93	8.77	88.93	**1.99**	63.46	44.37
	Ours	**13.87**	**3.91**	**95.64**	**16.56**	**8.61**	**91.68**	3.03	**36.67**	**61.37**
AMI Dev	Silero-Vad	15.92	8.09	93.65	-	-	-	-	-	-
	Pyannote	11.12	5.05	95.91	18.65	8.42	91.30	2.19	35.34	70.97
	Ours	**9.99**	**2.24**	**97.52**	**12.42**	**5.85**	**94.08**	**2.17**	**21.84**	**79.96**
AMI Eval	Silero-Vad	9.14	14.19	91.27	-	-	-	-	-	-
	Pyannote	**8.85**	5.79	95.96	20.37	8.54	91.13	**1.66**	41.68	68.30
	Ours	10.61	**1.98**	**97.73**	**13.61**	**5.00**	**94.43**	1.99	**20.48**	**81.76**

TABLE XI
DIARIZATION RESULTS ON DIHARD II

Method	DER (%)	JER (%)
TS-VAD [28]	39.80	41.79
TS-VAD (Multi-Channel) [28]	37.57	40.51
SA-EEND [34]	32.14	54.32
EEND-EDA [34]	29.57	51.50
EEND-EDA (Iterative inference+) [34]	28.52	49.77
EEND-GLA-Large [23]	28.33	50.62
DIHARD II Track 2 Winner [67]	27.11	49.07
+ EEND Post-Processing [42]	26.88	48.43
AED-EEND (Pre-trained on SM)	27.06	51.72
AED-EEND-EE (Pre-trained on SM)	25.34	47.15
AED-EEND	25.92	49.53
AED-EEND + Conformer *	27.11	49.48
AED-EEND-EE	**24.64**	**47.02**
AED-EEND-EE *	26.13	47.28
AED-EEND-EE + Conformer *	25.12	47.56

*: the downsampling rate is set to 10 because the convolution operation in Conformer is sensitive to the time resolution.

TABLE XII
DIARIZATION RESULTS ON AMI DATASET

Method	Dev		Eval	
	DER (%)	JER (%)	DER (%)	JER (%)
x-vec AHC + VB + OVL [42], [68]	-	-	28.15	41.00
+ EEND Post-Processing [42]	-	-	27.97	40.57
SA-EEND [34]	31.66	39.20	27.70	37.50
RPNSD [39] †	-	-	25.08	32.12
Transcribe-to-Diarize [44]	23.51	-	24.43	-
Multi-Class Spec-Clustering [69] *	-	-	23.60	-
CmpEm + Overlap Detector [70]	-	-	22.93	-
EEND-EDA [34]	21.93	25.86	21.56	29.99
Multi-scale SD [71] *	22.20	-	21.19	-
NSD-MA-MSE [72]	16.71	-	16.95	-
AED-EEND (Pre-trained on SM)	20.74	24.75	19.86	28.48
AED-EEND-EE (Pre-trained on SM)	18.89	23.19	16.91	25.32
AED-EEND	18.94	22.93	18.77	25.89
AED-EEND-EE	15.89	19.57	16.33	23.73
AED-EEND + Conformer	13.86	17.58	13.19	19.01
AED-EEND-EE + Conformer	**13.63**	**17.17**	**13.00**	**18.52**

*: Oracle speech segments are used
†: Oracle speaker number is used

Then, we replace the SM training data with SC training data, similar to the results introduced in IV-B3, and we achieve further improvement. Besides, we also tried to replace the transformer encoder with a Conformer in our DIHARD experiment and we use a different downsampling rate introduced in Section III-D. Remarkably, even with the change in downsampling ratio, the system utilizing the Conformer encoder yields excellent results. In Table XI, we also include the results of other works reported on the DIHARD II dataset, showcasing that our best-performing system achieves state-of-the-art performance on the DIHARD II evaluation benchmark.

6) Results on AMI Dataset: Finally, we evaluate our system on the AMI dataset, and the results are presented in Table XII. Despite the AMI dataset consisting of only 3 or 4 speakers, the average duration of the AMI dev and test sets is 33 minutes, posing a significant challenge for the model's ability to handle long speech recordings. Similar to the results obtained on the CALLHOME and DIHARD II datasets, the Enhancer module and the utilization of the SC data consistently improve the system's performance. Remarkably, the Conformer encoder proves to be highly beneficial for the AMI dataset, further enhancing the system's capabilities and achieving state-of-the-art performance on the AMI evaluation benchmark.

V. EVALUATION ON SPEECH TYPE DETECTION TASK

In this section, we evaluate the performance of our AED-EEND systems as an independent speech type detection model and compare it with the Pyannote [65], [66] and Silero-Vad [64] systems, as displayed in Table X. For Pyannote and Silero-Vad, we utilize their pre-trained models[4],[5] directly. Meanwhile, for our system, in cases where the predictions of the three types of speech activities conflict with each other, we independently assess each prediction. It should be noted that the Pyannote system in [66] is trained on AMI and DIHARD III datasets,

[4]https://huggingface.co/pyannote/segmentation
[5]https://github.com/snakers4/silero-vad/tags

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

while the training data used for Silero-Vad remains unclear. It is observed that our system achieves the highest F1-Score among the three systems for all the speech types. Although the comparisons may not be entirely fair due to the different training data configurations, these results still demonstrate the excellent speech type detection capability of our newly proposed AED-EEND-EE system. This ability significantly contributes to the overall success on the diarization task.

VI. Conclusion

This paper proposed an innovative paradigm for speaker diarization by employing a simple attention-based encoder-decoder network for the task. Within this paradigm, we proposed a teacher-forcing training strategy to simplify the training pipeline and speed up the training process. We also proposed an iterative decoding method to output the diarization results for each speaker sequentially. Moreover, we propose a novel Embedding Enhancer module designed to enhance our system's ability to adapt to scenarios involving an unseen number of speakers. Recognizing a significant disparity between commonly used simulation datasets and real-world data, we advocate for the adoption of a more realistic simulation dataset to elevate the system performance. Additionally, we explore replacing the transformer encoder with the Conformer model, aiming to better capture local information nuances. All these methodologies can be combined to boost the system performance, and it achieves the state-of-the-art across the established diarization benchmarks. Beyond diarization task, we identify our system's potential as a competitive speech-type detection model.

References

[1] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin, "Automatic named identification of speakers using diarization and ASR systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4557–4560.

[2] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu, and S. Watanabe, "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 31–38.

[3] D. Raj et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 897–904.

[4] T. Von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 91–95.

[5] S. Maiti et al., "EEND-SS: Joint end-to-end neural speaker diarization and speech separation for flexible number of speakers," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 480–487.

[6] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, Oct. 2013.

[7] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 413–417.

[8] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification," *Digit. Signal Process.*, vol. 31, pp. 93–101, 2014.

[9] J. Villalba, M. Diez, A. Varona, and E. Lleida, "Handling recordings acquired simultaneously over multiple channels with PLDA," in *Proc. Interspeech*, 2013, pp. 2509–2513.

[10] G. Sell et al., "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[11] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5239–5243.

[12] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM based similarity measurement with spectral clustering for speaker diarization," in *Proc. Interspeech* 2019, 2019, pp. 366–370.

[13] G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4794–4798.

[14] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 147–154.

[15] M. Diez, S. Wang, and J. Rohdin, "Bayesian HMM based x-vector clustering for speaker diarization," in *Proc. Interspeech*, 2019, pp. 346–350.

[16] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech* 2019, pp. 4300–4304.

[17] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 296–303.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.

[19] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, J. Shi, and K. Nagamatsu, "Neural speaker diarization with speaker-wise chain rule," 2020, *arXiv:2006.01796*.

[20] Y. Takashima, Y. Fujita, S. Watanabe, S. Horiguchi, P. García, and K. Nagamatsu, "End-to-end speaker diarization conditioned on speech activity and overlap detection," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 849–856.

[21] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Proc. Interspeech* 2020, pp. 269–273.

[22] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Proc. Interspeech* 2021, pp. 3565–3569.

[23] S. Horiguchi, S. Watanabe, P. García, Y. Takashima, and Y. Kawaguchi, "Online neural diarization of unlimited numbers of speakers using global and local attractors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 706–720, 2023.

[24] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5554–5558.

[25] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[26] K. Žmolíková et al., "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, Aug. 2019.

[27] Y. Jiang, R. Tao, Z. Pan, and H. Li, "Target active speaker detection with audio-visual cues," in *Proc. Interspeech*, 2023, pp. 3152–3156.

[28] I. Medennikov et al., "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech* 2020, pp. 274–278.

[29] I. Medennikov et al., "The STC system for the CHiME-6 challenge," in *Proc. Workshop Speech Process. Everyday Environments*, 2020.

[30] W. Wang, X. Qin, M. Cheng, Y. Zhang, K. Wang, and M. Li, "The DKU-SMIIP diarization system for the voxceleb speaker recognition challenge 2022," in *Proc. Voxsrc Workshop*, 2022.

[31] C.-Y. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Multi-target extractor and detector for unknown-number speaker diarization," *IEEE Signal Process. Lett.*, vol. 30, pp. 638–642, 2023.

[32] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[33] M. Cheng, W. Wang, Y. Zhang, X. Qin, and M. Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

[34] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1493–1507, 2022.

[35] Z. Chen, B. Han, S. Wang, and Y. Qian, "Attention-based encoder-decoder network for end-to-end neural speaker diarization with target speaker attractor," in *Proc. Interspeech*, 2023, pp. 3552–3556.

[36] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From simulated mixtures to simulated conversations as training data for end-to-end neural diarization," in *Proc. Interspeech* 2022, pp. 5095–5099.

[37] F. Landini, M. Diez, A. Lozano-Diez, and L. Burget, "Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[38] Y. C. Liu, E. Han, C. Lee, and A. Stolcke, "End-to-end neural diarization: From transformer to conformer," in *Proc. Interspeech* 2021, pp. 3081–3085.

[39] Z. Huang, M. Delcroix, L. P. Garcia, S. Watanabe, D. Raj, and S. Khudanpur, "Joint speaker diarization and speech recognition based on region proposal networks," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101316.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[41] W. Wang and M. Li, "Incorporating end-to-end framework into target-speaker voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8362–8366.

[42] S. Horiguchi, P. Garcia, Y. Fujita, S. Watanabe, and K. Nagamatsu, "End-to-end speaker diarization as post-processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7188–7192.

[43] C. Zhang, J. Shi, C. Weng, M. Yu, and D. Yu, "Towards end-to-end speaker diarization with generalized neural speaker clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8372–8376.

[44] N. Kanda et al., "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8082–8086.

[45] S. Maiti, H. Erdogan, K. Wilson, S. Wisdom, S. Watanabe, and J. R. Hershey, "End-to-end diarization for variable number of speakers with local-global networks and discriminative speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7183–7187.

[46] Y. Yu, D. Park, and H. K. Kim, "Auxiliary loss of transformer with residual connection for end-to-end speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8377–8381.

[47] S. Horiguchi, Y. Takashima, S. Watanabe, and P. Garcia, "Mutual learning of single-and multi-channel end-to-end neural diarization," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 620–625.

[48] Y.-R. Jeoung, J.-Y. Yang, J.-H. Choi, and J.-H. Chang, "Improving transformer-based end-to-end speaker diarization by assigning auxiliary losses to attention heads," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[49] S. Horiguchi, Y. Takashima, P. Garcia, S. Watanabe, and Y. Kawaguchi, "Multi-channel end-to-end neural diarization with distributed microphones," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7332–7336.

[50] A. Khare, E. Han, Y. Yang, and A. Stolcke, "ASR-aware end-to-end neural diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8092–8096.

[51] N. Yamashita, S. Horiguchi, and T. Homma, "Improving the naturalness of simulated conversations for end-to-end neural diarization," in *Proc. Speaker Lang. Recognit. Workshop*, Beijing, China, 2022, pp. 133–140.

[52] S. Horiguchi, S. Watanabe, P. García, Y. Xue, Y. Takashima, and Y. Kawaguchi, "Towards neural diarization for unlimited numbers of speakers using global and local attractors," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 98–105.

[53] E. Han, C. Lee, and A. Stolcke, "BW-EDA-EEND: Streaming end-to-end neural speaker diarization for a variable number of speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7193–7197.

[54] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, P. García, and K. Nagamatsu, "Online end-to-end neural diarization with speaker-tracing buffer," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 841–848.

[55] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[56] A. Woodward, C. Bonnín, I. Masuda, D. Varas, E. Bou-Balust, and J. C. Riveiro, "Confidence measures in encoder-decoder models for speech recognition," in *Proc. Interspeech*, 2020, pp. 611–615.

[57] "2000 NIST speaker recognition evaluation," 2000. Accessed: Aug. 6, 2023. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2001S97

[58] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Mach. Learn. Multimodal Interaction: 2nd Int. Workshop*, Edinburgh, U.K., Springer, 2006, pp. 28–39.

[59] N. Ryant et al., "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Proc. Interspeech* 2019, pp. 978–982.

[60] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech* 2020, pp. 5036–5040.

[61] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Comput. Speech Lang.*, vol. 71, 2022, Art. no. 101254.

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

[63] J. Wang, Z. Du, and S. Zhang, "Told: A novel two-stage overlap-aware framework for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[64] S. Team, "Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," 2021. Accessed: Aug. 6, 2023. [Online]. Available: https://github.com/snakers4/silero-vad

[65] H. Bredin et al., "Pyannote.audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, Spain, 2020, pp. 7124–7128.

[66] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, Brno, Czech Republic, 2021, pp. 3111–3115.

[67] F. Landini et al., "But system description for dihard speech diarization challenge 2019," 2019, *arXiv:1910.08847*.

[68] L. G. Perera et al., "Speaker detection in the wild: Lessons learned from JSALT 2019," in *Proc. Speaker Lang. Recognit. Workshop*, 2020, pp. 415–422.

[69] D. Raj, Z. Huang, and S. Khudanpur, "Multi-class spectral clustering with overlaps for speaker diarization," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 582–589.

[70] Z. Li and J. Whitehill, "Compositional embedding models for speaker identification and diarization with simultaneous speech from 2 speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7163–7167.

[71] T. J. Park, N. R. Koluguri, J. Balam, and B. Ginsburg, "Multi-scale speaker diarization with dynamic scale weighting," in *Proc. Interspeech* 2022, pp. 5080–5084.

[72] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee, "ANSD-MA-MSE: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1561–1573, 2023.

Zhengyang Chen (Graduate Student Member, IEEE) received the B.Eng. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His research interests mainly include speaker recognition and speaker diarization.

Bing Han (Graduate Student Member, IEEE) received the B.Eng. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently working toward the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His research mainly focuses on speaker recognition.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.

Shuai Wang (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2020. He is currently a Research Scientist with the Shenzhen Research Institute of Big Data, Chinese University of Hong Kong, Shenzhen (SRIBD, CUHK-SZ), Shenzhen, China. Prior to that, he was with Tencent, Shenzhen, China, as a Senior Application Scientist, leading the Group working on speaker recognition, voice conversion and speech synthesis. He has authored or coauthored more than 30 papers on the topic of speaker modeling. He initiated the popular "WeSpeaker" project, utilized by numerous research groups across academia and industry. He was the recipient of IEEE Ganesh N. Ramaswamy Memorial Award (ICASSP2018). He was also the main contributor to the winning systems of VoxSRC 2019 and DIHARD 2019. He is also a Regular Reviewer for conferences and journals including ICASSP, INTERSPEECH, ASRU, TASLP, and CSL. He is the Member of ISCA, and SPS.

Yanmin Qian (Senior Member, IEEE) received the B.S. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2012. Since 2013, he has been with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently a Full Professor. From 2015 to 2016, he was an Associate Research with the Speech Group, Cambridge University Engineering Department, Cambridge, U.K. He has authored or coauthored more than 200 papers in peer-reviewed journals and conferences on speech and language processing, including T-ASLP, Speech Communication, ICASSP, INTERSPEECH, and ASRU. He has also applied for more than 80 Chinese and American patents. His research interests include automatic speech recognition and translation, speaker and language recognition, speech separation and enhancement, music generation and understanding, speech emotion perception, multimodal information processing, natural language understanding, deep learning, and multi-media signal processing. He was the recipient of the 5 championships of international challenges, several top academic awards in China, including Chang Jiang Scholars Program of the Ministry of Education, Excellent Youth Fund of the National Natural Science Foundation of China, and the First Prize of Wu Wenjun Artificial Intelligence Science and Technology Award (First Completion), and several awards from international research committee, including the Best Paper Award in Speech Communication and Best Paper Award from IEEE ASRU in 2019. He is the Member of IEEE Signal Processing Society Speech and Language Technical Committee.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 12,2026 at 21:19:46 UTC from IEEE Xplore. Restrictions apply.