ANSD-MA-MSE: Adaptive Neural Speaker Diarization Using Memory-Aware Multi-Speaker Embedding

Mao-Kui He, Jun Du, *Senior Member, IEEE*, Qing-Feng Liu, and Chin-Hui Lee, *Life Fellow, IEEE*

***Abstract*—In this paper, we propose a neural speaker diarization (NSD) network architecture consisting of three key components. First, a memory-aware multi-speaker embedding (MA-MSE) mechanism is proposed to facilitate a dynamical refinement of speaker embedding to reduce a potential data mismatch between the speaker embedding extraction and the NSD network. Next, a speaker selection procedure is introduced to handle situations where the detected number of speakers is different from the assumed speaker size in the NSD network. Finally, an adaptive procedure is proposed to improve the required prior information for the nonoverlap speech segments in a given utterance during each iteration. We call our proposed framework adaptive neural speaker diarization with memory-aware multi-speaker embedding (ANSD-MA-MSE). Our method improves diarization performance in realistic operating scenarios, such as adverse acoustic environments, domain mismatches, and a varying, rather than fixed, number of speakers. Having been tested on both the AMI corpus and the DIHARD-III evaluation sets, our proposed approach consistently outperforms other state-of-the-art techniques in diarization error rates, including the results reported by the best single-model system in the DIHARD-III challenge.**

***Index Terms*—Speaker diarization, neural networks, memory-aware speaker embedding, dictionary learning, attention network, adaptive refinement.**

I. Introduction

SPEAKER diarization refers to the task of labeling a given recording with classes corresponding to speaker identity [1], [2]. It is an important front end of speech processing systems that have attracted an ample amount of research attention in recent years. Many speech applications can benefit from good diarization results, including meeting summaries, telephone conversation analysis, transcription of dialog, and so on [2], [3]. Research on speaker diarization is usually conducted on different domains, such as telephone [4], [5], broadcast news [6] and meeting [7]. For real-world scenes with variable speaker numbers, adverse acoustic environments, and a large portion of speech overlap, speaker diarization is still quite challenging.

Conventional clustering-based methods, which include voice activity detection, speech segmentation, speaker feature extraction, and speaker clustering, are widely used in speaker diarization tasks [8], [9]. The whole process can be roughly divided into two main components: speaker representation extraction and clustering. Traditional speaker representation approaches consist of i-vector [10] and neural network-based speaker embedding (e.g., x-vector [11]). Among them, the i-vector and x-vector are segment-level embeddings. For the clustering process, algorithms such as the mean shift [12], agglomerative hierarchical clustering (AHC) [13] and spectral clustering (SC) [5] are commonly adopted. Moreover, different distance measures [14], [15], [16] are also explored to obtain better clustering results. In particular, probabilistic linear discriminant analysis (PLDA) [17] is often used along with the i-vector or x-vector to classify whether two segments are from the same speaker and has shown great effects. Among these methods, Bayesian HMM clustering of x-vector sequences (VBx) diarization system [18] achieved superior performance. VBx uses the x-vector with two-stage clustering: first-stage AHC underclustering and second-stage VB-HMM [19] refining.

Although clustering-based speaker diarization is relatively robust across different domains, these methods cannot deal well with overlap segments because every segment can only be assigned a single label through hard clustering. To address this issue, end-to-end systems have been proposed [20], [21], [22], [23]. In region proposal networks (RPNs) [20], the speech segment are generated by a neural network, and the speaker embedding is calculated accordingly. The diarization predictions are obtained by clustering the extracted embedding. In end-to-end neural speaker diarization (EEND) [21], [22], the task is treated as a multilabel classification problem, which allows the model to deal with overlapping speech and be optimized by directly minimizing diarization errors. A permutation-free objective function was proposed in EEND to address the permutation problem where the output speaker labels are ambiguous

Manuscript received 8 March 2022; revised 9 October 2022 and 1 February 2023; accepted 22 March 2023. Date of publication 6 April 2023; date of current version 24 April 2023. This work was supported by the National Natural Science Foundation of China under Grant 62171427. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alberto Abad. *(Corresponding author: Jun Du.)*

Mao-Kui He, Jun Du, and Qing-Feng Liu are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230026, China (e-mail: hmk1754@mail.ustc.edu.cn; jundu@ustc.edu.cn; qfliu@iflytek.com).

Chin-Hui Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Our code is publicly available at https://github.com/Maokui-He/NSD-MA-MSE.

Digital Object Identifier 10.1109/TASLP.2023.3265199

in the training stage, i.e., not corresponding to any fixed class. Further research was also explored to handle recordings containing unknown numbers of speakers [23]. Recently, target-speaker voice activity detection (TS-VAD) [24] was proposed, which uses speech features along with speaker embedding as input to generate multi-speaker outputs representing speaker presence probabilities for each frame. In [25], a strategy was proposed to handle variable number of speakers in TS-VAD. In [26], they incorporated EEND into TS-VAD and trained the two parts jointly.

The DIHARD challenge [27] was held to facilitate the comparison of different approaches through a unified evaluation dataset, task, and metric to generalize the study of a variety of challenging realistic domains with various recording devices, numbers of speakers, reverberation and background noise. Other challenges, such as NIST SRE series [28], the Fearless Steps series [29], [30], the Iberspeech-RTVE challenge [31], CHiME-6 [32], and VoxSRC 2020 [33], have also included a diarization component. For the DIHARD-I challenge [27], the traditional approaches [34], [35] performed poorly due to the overlapping regions and diversified realistic domains, such as meeting, broadcast, restaurant, clinical, and courtroom. Even the best system with oracle speech activity detection (SAD) information from the team at Johns Hopkins University [34], which explored several key aspects of the state-of-the-art diarization methods, produced a quite high diarization error rate (DER). The team from Brno University of Technology won first place for the subsequent DIHARD-II challenge [36] by utilizing VBx [37]. VBx is robust and achieves satisfactory performance for most of the domains, but it still cannot properly deal with the overlapping segments. Several other teams adjusted the parameters or thresholds during clustering to adapt the system to different domains [38], [39]. In the DIHARD-III challenge [40], most participants [41], [42] utilized neural diarization methods to cope with the overlapping segments and combined them with conventional clustering-based systems to further improve the overall performance. Our USTC-NELSLIP team [43] won the first place of DIHARD-III challenge by combining both separation and improved TS-VAD-based diarization on top of VBx [37].

The original TS-VAD achieved great results on the speaker diarization task of the CHiME-6 challenge, which is the key technology of the champion team [24]. However, its generalizability to realistic unseen or mismatched domains, e.g., DIHARD-III tasks, is still a challenging problem. First, TS-VAD employs a pretrained extractor to obtain speaker embedding (e.g., i-vector) as input. In realistic scenarios, there are no oracle speaker segments for computing speaker embedding and diarized segments are usually unreliable. Second, the number of speakers in TS-VAD is fixed, which restricts its application to domains with a variable number of speakers. Third, the performance of the pretrained TS-VAD model will degrade considerably on mismatched data in different domains.

Accordingly, in this study, to address the above three issues, we propose a novel adaptive neural speaker diarization approach using memory-aware multi-speaker embedding. The proposed method introduces a dedicated memory module to produce a multi-speaker embedding, a set of speaker embeddings for TS-VAD. Whereas most prior work simply extracts i-vectors from roughly estimated (initial) speaker segments, the proposed method retrieves a cleaner and more discriminative multi-speaker embedding from memory via the attention mechanism. The extracted multi-speaker embedding is robust against the quality of initial speaker segments from the clustering-based diarization. The main contributions are in three parts. First, we present a diarization neural network with an auxiliary memory block to extract the multi-speaker embedding. The design of the memory block is inspired by the concept of dictionary learning [44], [45], where the speaker embedding bases (e.g., i-vectors or x-vectors) are extracted from a database consisting of diversified speakers. Although speaker inventory has been used in speaker-attributed automatic speech recognition [46], to the best of our knowledge, this is the first application of dictionary learning to speaker diarization. Then, for one utterance with multiple speakers, with deep extracted features via convolutional neural networks (CNNs) and the speaker mask matrix based on VBx, the new multi-speaker embedding is generated by a weighted sum of all speaker embedding bases via an attention mechanism to learn the weights. In this way, the diarization network can be jointly optimized with the memory block to extract more accurate information for each speaker on the overlapping segments, where the speaker diarization performance is improved compared to the TS-VAD method.

Second, a new strategy is adopted to address the case of variable speaker numbers in realistic applications. As our diarization network has a fixed number of speakers setting, it can only process utterances with the same number of speakers in one session. The primary idea was proposed in TS-VAD for an unknown number of speakers [25]. If the detected speaker number was smaller than the fixed speaker number, dummy speaker embeddings are selected from the training set to fill the inputs. Instead, only fixed number of speaker embeddings with the longest non-overlapping speaking durations are selected. We extend this strategy to this new framework to handle any number of speakers.

Finally, an adaptive optimization method is designed to adapt our diarization network to different domains and improve the model's generalizability to unseen domains. Based on the analysis of diversified domain characteristics among different utterances in the DIHARD-III challenge, it is found that session-wise adaptation leads to better performance than domain-wise adaptation. Unlike updating the diarization results by re-estimating their speaker embeddings for better results in [24], [47], the main idea is adapting the pretrained model to improve the purity of detected nonoverlapping segments. We detected overlapping segments with the NSD-MA-MSE model, where the speaker labels are extracted from the clustering-based diarization results. Then, we discard the detected overlapping segments and utilize the remaining non-overlapping segments to simulate multiple speaker conversation data for the session-level adaptation. The original idea was proposed in [48] which focused on the entire system description of the competition. The adaptive strategy was only a small part and was not described in detail. In this paper, we described the entire adaptive process for the improved model (NSD-MA-MSE) and explored the effect of the amount of data

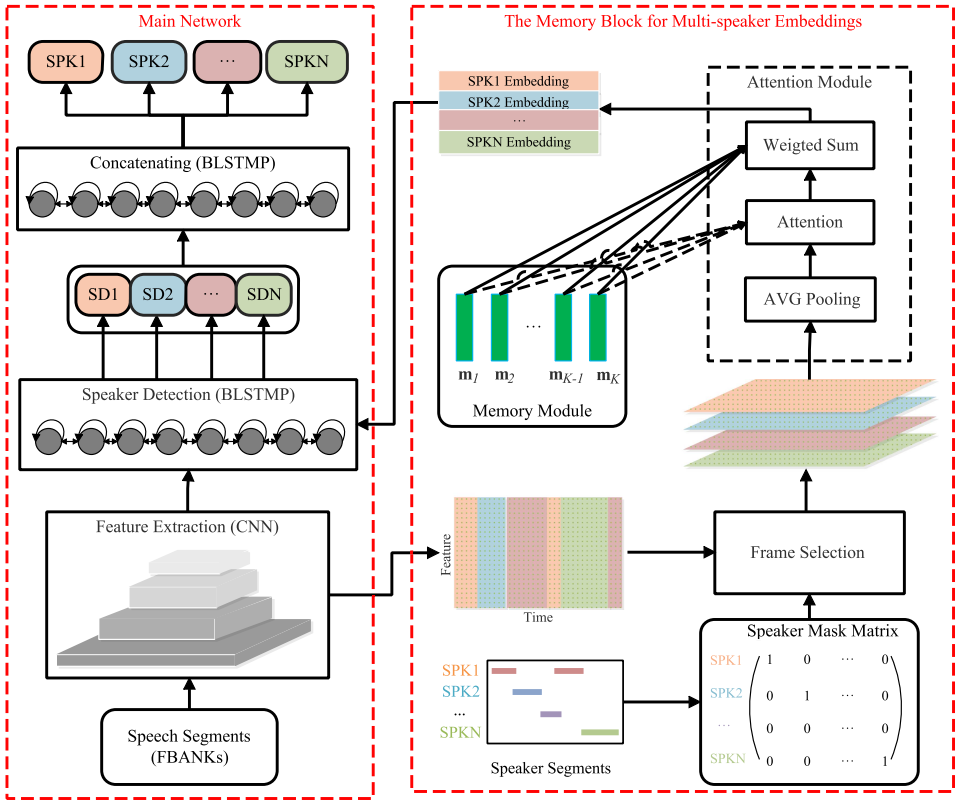


Fig. 1. The architecture of neural speaker diarization network using memory-aware multi-speaker embedding.

and the number of iterations on the results. Our experiments on the DIHARD-III challenge show the effectiveness of the proposed approach, and better performance is achieved than the single best system from the DIHARD-III challenge champion team. We call our proposed framework adaptive neural speaker diarization with memory-aware multi-speaker embedding (ANSD-MA-MSE).

In summary, our main contributions are as follows: First, we construct a neural speaker diarization network using memory-aware multi-speaker embedding. Second, the strategy for processing unknown numbers of speakers is shown to be effective. Finally, the adaptive step improves the diarization performance on both the match and mismatch domains. The remainder of the paper is organized in sections. Section II introduces neural speaker diarization using memory-aware multi-speaker embedding and the strategy for processing unknown numbers of speakers is included in Section II-E. In Section III, the adaptive model strategy is detailed. Experimental results and analyses are presented in Section IV. Finally, we conclude the paper in Section V.

II. Neural Speaker Diarization Using Memory-Aware Multi-Speaker Embedding

A neural speaker diarization approach using memory-aware multi-speaker embedding (NSD-MA-MSE) is proposed to handle the speaker overlap problem in clustering-based methods and the permutation problems in EEND. The architecture of NSD-MA-MSE consists of two parts, namely, the main network and the memory block, as shown in Fig. 1. Inspired by TS-VAD [24], the main network takes conventional speech features (e.g., log-Mel filter-banks) as input to extract frame-level features by CNNs and then combines the features with multi-speaker embedding from the memory block to predict per-frame speech activities for all the speakers simultaneously. To improve discrimination of speaker embeddings and help the diarization network easier to make decisions, a learnable multi-speaker embedding network called memory block is constructed. The memory block adopts the deep extracted features via CNNs from the main network and the speaker mask matrix generated from clustering-based diarization results as inputs. We will elaborate these two parts in the following subsections.

A. Main Network

The input of the main network is a set of acoustic observations denoted by the matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, where $\mathbf{x}_t \in \mathbb{R}^{D'}$ is the D'-dimensional log-Mel filter-bank feature vector (FBANKs) of the t-th frame and T is the frame number of the current utterance. Then, 4 convolutional layers are used to extract a set of deep features denoted by the matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T]$, where $\mathbf{f}_t \in \mathbb{R}^D$ is the D-dimensional feature vector of the t-th frame. The frame-level deep features serve as inputs of both the main network and memory block. Next, we concatenate the deep features with a set of speaker embeddings $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N]$ by copying for all T frames, where $\mathbf{e}_n \in \mathbb{R}^L$ is the L-dimensional vector for the n-th speaker embedding and N is the number of speakers of the main network output,

which are generated from the memory block. Then, a speaker detection (SD) component comprising 2-layer bidirectional long short-term memory with projection (BLSTMP) [49] extracts deeper frame-level speaker-related features from each speaker concatenation. Finally, we concatenate N speakers' SD outputs and pass them to a 1-layer BLSTMP to produce N two-class outputs corresponding to the speech and silence probabilities for each of the N speakers, namely, $\hat{\mathbf{Y}} = (\hat{y}_{nt}) \in \mathbb{R}^{N\times T}$, where $\hat{y}_{nt} \in [0, 1]$ denotes the probability that speaker n is active in frame t of the recording.

B. Memory-Aware Multi-Speaker Embedding (MA-MSE)

The memory block employs $\mathbf{F}$ from the main network and the speaker mask matrix $\mathbf{S} = (s_{nt}) \in \mathbb{R}^{N\times T}$ as inputs, where $s_{nt} = 0/1$ denotes that the speaker n is not appearing or appears in frame t of the recording, respectively. $\mathbf{S}$ is from the ground truth when training and the auxiliary speaker diarization system (VBx in our experiments) when testing. Note that VBx did not provide overlap information for $\mathbf{S}$. First, the frame selection module is designed to select the frame-level features of each active speaker, as the silence frames are usually removed in traditional speaker embedding extraction. The corresponding output fed to the attention module is a tensor with three-dimensional arrays $\mathbf{F}^{\mathrm{s}} = (\mathbf{f}^{\mathrm{s}}_{tn}) \in \mathbb{R}^{D\times T\times N}$, where $\mathbf{f}^{\mathrm{s}}_{tn}$ is the D-dimensional feature vector of the t-th frame for the n-th speaker, which can be calculated as:

$$\mathbf{f}^{\mathrm{s}}_{tn} = s_{nt}\mathbf{f}_t \tag{1}$$

Then, the memory module, as one key component in NSD-MA-MSE, is designed to provide another input for the subsequent attention module. The memory here refers to a set of speaker embedding bases inspired by the concept of dictionary learning [44], [45], which can be used to predict a more discriminative embedding than i-vectors and x-vectors for a new speaker. In this study, the design of the memory module is flexible. It may contain different types of memories, e.g., i-vector or x-vector. The speaker embedding basis vectors in a memory are easily distinguished from each other by their corresponding speaker. To build a memory, we should train a speaker recognition model in advance. For example, a conventional i-vector system can be built first based on the setup for a universal background model with a Gaussian mixture model (GMM-UBM) [10]. Then, speaker embedding i-vectors are extracted through the pretrained speaker recognition model. Finally, a clustering algorithm such as K-means [51] is adopted to control the number of basis vectors in a memory, and the cluster centroids are taken as a memory. In addition, we can construct a new memory using x-vectors. Extraction of the x-vector uses the discriminative model via deep neural networks in an end-to-end manner, which is quite different from i-vector extraction based on the generative model via GMM-UBM. Accordingly, these two types of speaker embedding might be complementary, which motivates us to combine them in the memory module to further improve the performance of multi-speaker representations. In our experiments, we show its effectiveness for different scenes.

Based on $\mathbf{M} = \{\mathbf{m}_k \in \mathbb{R}^L | k = 1, 2, \ldots, K\}$ with K speaker embedding basis vectors from the memory module and $\mathbf{F}^{\mathrm{s}}$ from the frame selection module, the attention module is equipped to select the speaker embedding bases of the memory that are most similar to the current speech segment from each memory. Speaker embedding is usually computed from a speech segment with whole frames to improve robustness. The information gathering part is designed to obtain speaker information about the whole speech segment to improve the accuracy of the attention module. However, if we use a complex model such as a neural network for information gathering, we may face the overfitting problem. Therefore, the average pooling layer is employed to output a matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N]$ and

$$\mathbf{p}_n = \frac{1}{T_n}\sum_{t=1}^{T}\mathbf{f}^{\mathrm{s}}_{tn}, \quad T_n = \sum_{t=1}^{T} s_{nt} \tag{2}$$

where T_n is the speech duration of speaker n. $\mathbf{p}_n$ is a D-dimensional vector representing the information gathering from the frame selection module for speaker n.

The combination block calculates the similarity between memory and the current speech segment and combines them into a vector named the aggregated speaker vector. The additive attention mechanism is adopted to learn the similarity scores between $\mathbf{p}_n$ and each vector in memory, as described in the following formula:

$$c_{nk} = \mathbf{v}^{\top}\tanh(\mathbf{W}\mathbf{p}_n + \mathbf{U}\mathbf{m}_k) \tag{3}$$

where c_{nk} is the attention value that scores the similarity between $\mathbf{p}_n$ and $\mathbf{m}_k$. The matrices $\boldsymbol{W} \in \mathbb{R}^{D^{\mathrm{a}}\times D}$ and $\mathbf{U} \in \mathbb{R}^{D^{\mathrm{a}}\times L}$ and the vector $\mathbf{v} \in \mathbb{R}^{D^{\mathrm{a}}}$ are the parameters of the attention model for memory. D^{a} denotes the attention dimension. The attention values are normalized through the logistic sigmoid operation instead of the softmax operation to avoid the sparsity problem [52]:

$$a_{nk} = \frac{1}{1 + \exp(-c_{nk})} \tag{4}$$

where the attention weights a_{nk} are used to compute a weighted sum of the vectors in each memory as follows:

$$\mathbf{e}_n = \sum_{k=1}^{K} a_{nk}\mathbf{m}_k \tag{5}$$

where $\mathbf{e}_n$ is the aggregated speaker vector for speaker n. $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N]$ denotes the multi-speaker embedding of the current utterance; this is the memory block output.

C. Model Optimization

In this section, we discuss the optimization for the parameter set of the main network $\boldsymbol{\Lambda}_1$ and the parameter set of the memory block $\boldsymbol{\Lambda}_2$. $\boldsymbol{\Lambda}_1$ can be further divided into two subsets $\boldsymbol{\Lambda}^{\mathrm{F}}_1$ and $\boldsymbol{\Lambda}^{\mathrm{C}}_1$. $\boldsymbol{\Lambda}^{\mathrm{F}}_1$ denotes the parameter set of the feature extraction part in the main network, while $\boldsymbol{\Lambda}^{\mathrm{C}}_1$ denotes the parameter set of the remaining part in the main network after combining the multi-speaker embedding from the memory block.

With the input acoustic observation set $\mathbf{X}$ and speaker mask matrix $\mathbf{S}$, the whole model with the main network and the

auxiliary memory block predicts $\hat{\mathbf{Y}}$ speech/silence probabilities for each of the N speakers with $\hat{\mathbf{Y}} = (\hat{y}_{nt}) \in \mathbb{R}^{N\times T}$. We adopt the binary cross-entropy loss of multiple speakers as the learning objective to jointly optimize the parameter set in both the main network and memory block $\mathbf{\Lambda} = (\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$:

$$L_1(\mathbf{\Lambda}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}[y_{nt}\log(\hat{y}_{nt}) + (1-y_{nt})\log(1-\hat{y}_{nt})] \tag{6}$$

where $\hat{y}_{nt}$ and y_{nt} are the predicted output and label of the n-th speaker at the t-th frame, respectively.

In addition, an auxiliary loss is introduced to assist the convergence of the optimization of the memory block by minimizing the mean square error between the memory-aware multi-speaker embedding and the conventional speaker embedding:

$$L_2(\mathbf{\Lambda}_1^{\mathrm{F}}, \mathbf{\Lambda}_2) = \sum_{n=1}^{N}||\mathbf{e}_n - \mathbf{e}_n^*||^2 \tag{7}$$

$\mathbf{e}_n$ is the n-th speaker embedding vector of the memory block output, and $\mathbf{e}_n^*$ corresponds to the conventional speaker embedding vector (i-vector or x-vector) of the n-th speaker computed with its speech segments. Please note that this auxiliary loss updates not only the parameters in the memory block ($\mathbf{\Lambda}_2$) but also the parameters in the feature extraction part of the main network ($\mathbf{\Lambda}_1^{\mathrm{F}}$), as shown in Fig. 1.

The final loss of the proposed NSD-MA-MSE model is:

$$L(\mathbf{\Lambda}) = L_1(\mathbf{\Lambda}) + \lambda L_2(\mathbf{\Lambda}_1^{\mathrm{F}}, \mathbf{\Lambda}_2) \tag{8}$$

where λ is a weight to control the effect on the network from the auxiliary loss.

D. Model Inference

In the inference stage, first, the estimated speaker mask matrix $\hat{\mathbf{S}}$ is generated by a clustering-based diarization method from the testing session utterance. Then, with input features $\mathbf{X}$ and $\hat{\mathbf{S}}$, the NSD-MA-MSE model outputs speaker presence probabilities $\hat{\mathbf{Y}}$. Next, we perform thresholding on $\hat{\mathbf{Y}}$ to obtain hard speaker labels $\hat{\mathbf{L}} = (\hat{l}_{nt}) \in \{0,1\}^{N\times T}$ indicating speech ($\hat{l}_{nt} = 1$) or silence ($\hat{l}_{nt} = 0$) for speaker n in frame t. Finally, we perform postprocessing with reliable SAD labels $\mathbf{V} = (v_t) \in \{0,1\}^T$ (from reference SAD labels or a good SAD system), which is quite important for achieving good diarization performance. We perform postprocessing both with and without reference SAD results. In this step, the speech segments in the network outputs that are silent segments in the SAD labels are marked as silent. The silent segments in the network outputs, which are speech segments in the SAD labels, are marked as the active speaker with the longest speech duration in the neighborhood speech segments of SAD.

E. Strategy to Handle Unknown Number of Speakers

We can see that the above model optimization and inference procedure is applicable to the NSD-MA-MSE model with a fixed number of output speakers N which is usually chosen to be larger than the number of speakers for most sessions in the dataset. However, in realistic scenarios, the number of interactive speakers N^* might be variable and unknown in both the training and testing stages. We need to devise a strategy to handle the cases where $N^* \neq N$.

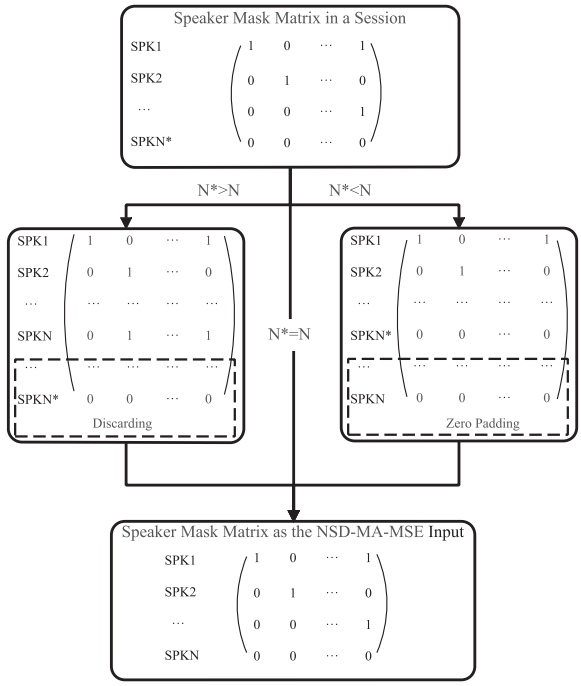


Fig. 2. Strategy for handling an unknown number of speakers. For the speaker mask matrix, the discarding operation (left branch) is conducted when $N^* > N$, while the zero padding operation (right branch) is performed when $N^* < N$.

If the detected speaker number in one session is smaller than the fixed speaker number, we randomly assign the output nodes of the diarization network to the speakers embedded in the current session, and the dummy speakers selected from the training set. Dummy speakers are not speaking, so they are labeled as silent in the training stage and discarded after model inference. If the detected speaker number is larger, we randomly select a fixed number of speakers to correspond to the output nodes when training and select a fixed number of speakers with the longest nonoverlapping speaking durations when testing. The remaining $N^* - N$ speakers with the shortest durations will be discarded directly because those speakers have little influence on the diarization results. In the training stage, N^* for each utterance is easy to obtain as the reference information. In the testing stage, as a clustering-based diarization system is adopted to generate the initial speaker mask matrix for NSD-MA-MSE, the estimation of N^* is a byproduct without incurring any computational overhead. If this estimate N^* is equal to N, then no further effort is needed. Otherwise, we deal with two cases using the above strategy, as illustrated in Fig. 2.

III. Adaptive NSD-MA-MSE Refinement

To solve the problem of mismatch between training and testing data caused by realistic diversified domains, in this section, we propose an adaptive neural speaker diarization using

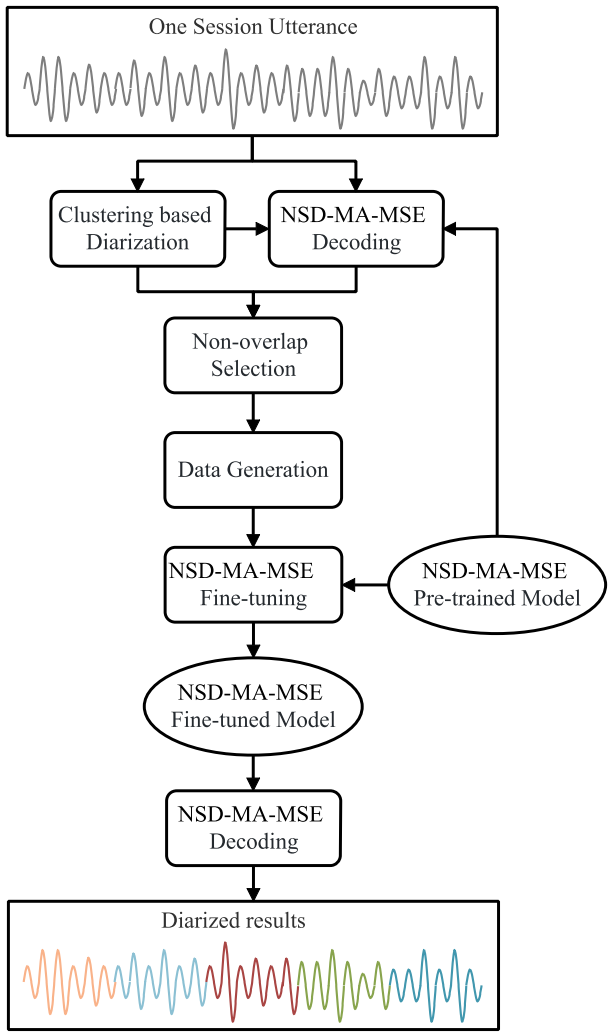


Fig. 3. Flowchart of adaptive neural speaker diarization using memory-aware multi-speaker embedding (ANSD-MA-MSE).

memory-aware multi-speaker embedding (ANSD-MA-MSE) to improve NSD-MA-MSE by model adaptation.

The overall ANSD-MA-MSE procedure is illustrated in Fig. 3. First, to obtain the speaker diarization prior information, we decode each testing session utterance with the pretrained NSD-MA-MSE model based on the speaker mask matrix from the clustering-based diarization results. In the adaptation stage, based on the first-pass decoding results, we discard the overlapping segments detected by NSD-MA-MSE and use the remaining nonoverlapping segments to simulate multi-speaker dialog data for the subsequent NSD-MA-MSE model fine-tuning. Next, using the fine-tuned model, the second-pass decoding is performed with the speaker masks matrix from the nonoverlapping segments to generate the refined diarization results.

The whole procedure includes the following 6 steps:

1) *Initialization:* The speaker diarization results are initialized with both overlapping and nonoverlapping segments by decoding each testing session utterance with the pretrained NSD-MA-MSE model based on the speaker mask matrix from the clustering-based diarization results (VBx). If only one speaker is detected, this entire process is skipped.
2) *Nonoverlap Selection:* Nonoverlap speech data are selected based on the speech overlap detection results from the pretrained or fine-tuned NSD-MA-MSE model.
3) *Data Generation:* Multiple speaker dialogs are simulated using the selected nonoverlapping speech data in 2) for model adaptation.
4) *NSD-MA-MSE Model Fine-tuning:* The NSD-MA-MSE pretrained model is fine-tuned with the simulated adaptation data in 3).
5) *NSD-MA-MSE Model Decoding:* Each testing session utterance is decoded using a fine-tuned NSD-NSD-MA-MSE model with the speaker mask matrix generated in 2).
6) *Adaptive Processing:* **2)** is followed for model adaptation with refined speaker diarization prior information decoded by the fine-tuned NSD-MA-MSE model.

In the following subsections, several key parts of ANSD-MA-MSE are elaborated.

A. Nonoverlap Segment Selection

The accuracy of nonoverlap selection plays an important role in the diarization performance after model adaptation. The clustering-based diarization methods are usually unable to provide overlap detection results. Meanwhile, we found that for the nonoverlapping segments, the clustering-based diarization sometimes performs better than the pretrained NSD-MA-MSE results in mismatched conditions. Accordingly, in the first-pass decoding, we employ the pretrained NSD-MA-MSE model to obtain speaker overlapping segments and filter out these segments from cluster-based results to obtain the nonoverlapping segments for the subsequent data generation. However, due to the mismatch between training and testing, the overlap detection in the first-pass decoding might not be accurate in adverse environments, which can lead to the impurity of the obtained nonoverlapping segments. In the following multipass decoding scheme, the fine-tuned NSD-MA-MSE model is adopted to improve the purity of the detected nonoverlapping segments. Moreover, one main difference from the first-pass decoding is that only the refined NSD-MA-MSE model is used to generate the diarization results without the intervention of clustering-based diarization.

B. Data Generation

After the nonoverlapping selection for each testing session utterance, we collect a set of truncated utterances without silence segments for each of the detected N^* speakers. We simulate each session utterance of the single-microphone output with N'-speaker dialog ($2 \leq N' \leq N^*$) in the waveform domain as:

$$s^{\mathrm{m}}(l) = \Sigma_{u=1}^{U} s_u^{\mathrm{r}}(l - \tau_u) \tag{9}$$

where $\{s_u^{\mathrm{r}}(l), 1 \leq l \leq T_u\}$ is the reverberant signal of the u-th utterance:

$$s_u^{\mathrm{r}}(l) = s_u(l) * h_{n'}(l). \tag{10}$$

$s_u(l)$ is the l-th sample of randomly selected nonoverlapping utterances with the speaker ID n'. $h_{n'}(l)$ is the synthetic room impulse response (RIR) for speaker n' using the image method [53], where the room size (length, width, height) ranges from (5 m, 5 m, 2.5 m) to (12 m, 12 m, 4.5 m) while the distance between speaker and microphone ranges from (0.5 m, 0.5 m, 0.1 m) to (4 m, 4 m, 1 m). U is the utterance number for the simulated session with the N'-speaker, and at most 10 utterances are randomly selected for each speaker per session. The parameter τ_u is used to control the overlap time ratio of multiple speakers randomly picked from 0 to 40% and silence time between two adjacent sentences from 0 to 2 s. We repeatedly use the truncated utterances 50 times. In total, we generate session utterances for approximately 4 hours as the adaptation data for a 10-minute recording. We also adopt this method to simulate data for NSD-MA-MSE pretraining while the utterances are from LibriSpeech [54].

C. Model Fine-Tuning and Decoding With NSD-MA-MSE

With the simulated adaptation data, the pretrained NSD-MA-MSE model is fine-tuned with the loss function in (8), where λ is the same as that in the pretraining stage. It is worth noting that the model parameters are optimized with only one epoch due to the limited adaptation data. Then, decoding using the procedure in Section II-D is conducted with the fine-tuned model to generate the diarization results for adaptive processing.

IV. Experimental Setup

A. Datasets

We evaluate the proposed method with two datasets, namely, the AMI corpus [55] and the DIHARD-III challenge corpus.

The AMI meeting corpus consists of 100 hours of meeting recordings. Each session includes 3-5 speakers. They use a range of signals that are synchronized to a common timeline and collected by close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. The meetings were recorded in English using three rooms with different acoustic properties and included mostly nonnative speakers. We evaluate our model following [18] to create training, development, and test sets where only words are considered as speech. We evaluate both Mix-Headset and beamformed audio on AMI.

The goal of the DIHARD-III challenge is to automatically detect and label all speaker segments for each recording. The development (DEV) and evaluation (EVAL) sets include sample selections of 5-10 minutes drawn from 11 domains exhibiting wide variations in recording equipment, recording environment, ambient noise, number of speakers (from 1 to 10), and speaker demographics. These domains range in difficulty from the trivial such as reading audiobooks recorded under clean conditions by a single speaker to the extremely challenging such as conversations between up to 6 diners recorded by a binaural microphone in restaurants with varying room acoustics and noise levels. They define two partitions of the evaluation data: 1) core evaluation set " a balanced evaluation set in which the total duration of each domain is approximately equal. 2) full evaluation set " a larger evaluation set that uses all available selections for each domain and which is, thus, unbalanced with some domains having more audio than others; it is a proper superset of the core evaluation set. The core sets of DEV and EVAL are 23.94 and 22.73 hours, respectively. Full sets of DEV and EVAL have 34.15 and 33.01 hours, respectively.

The accuracy of speaker diarization systems in this paper is measured by the diarization error rate (DER) [28]. No forgiveness collar is applied to the reference segments prior to scoring, and overlapping speech is also evaluated for both AMI and DIHARD-III.

B. Speech Activity Detection

To explore the effect of speech activity detection (SAD), we perform time delay neural network (TDNN) SAD [56] on both AMI and DIHARD-III. We train two SAD models for AMI with Mix-Heatset and beamformed data. The SAD model on DIHARD-III is trained with DIHARD-III DEV full sets.

C. VBx-Based Diarization

To better illustrate the effectiveness of our proposed method, we use a state-of-the-art clustering-based method, namely, Bayesian hidden Markov model (BHMM) x-vector diarization [18] (i.e., VBx), to compute the speaker mask matrix for decoding NSD-MA-MSE models and model adaptation. The x-vectors are first extracted with the deep neural network architecture based on ResNet101 [57] for each speech segment divided by SAD. Then, they are clustered using AHC with a similarity metric based on probabilistic linear discriminant analysis (PLDA) log-likelihood ratio scores [58], followed by VB-HMM-based clustering to create the diarization results. For the AMI dataset, we provide the results of our system evaluated on beamformed mic-array audio as well as on Mix-Headset audio. For the DIHARD-III dataset, we directly use the VBx parameters tuned on the DIHARD-II dataset [36].

D. TS-VAD Based Diarization

In the following experiments, several TS-VAD models with different types of speaker embedding are designed. First, for the original TS-VAD using the i-vector as the input embedding, we follow the method in [25], which is proposed to handle an unknown number of speakers and used in our DIHARD-III challenge system. Second, for the TS-VAD using the x-vector as the input embedding, we adopt the method in [59], where the ResNet block for extracting the x-vector is also optimized during the training stage. Finally, we create a new structure that combines both the i-vector and x-vector as the input embedding, and the parameters of the x-vector part are fine-tuned when training the TS-VAD, as in [25], [59]. All those models are optimized with

the same training dataset as in the NSD-MA-MSE model for the AMI corpus. No additional iterations were implemented for the speaker embedding purification since there wasn't a significant improvement.

E. NSD-MA-MSE Based Diarization

We investigate two types of speaker embedding as the memory module. The first one is 100-dim i-vectors following the Kaldi process for the DIHARD-I challenge [27]. The VoxCeleb1 and VoxCeleb2 datasets are used for building the i-vectors. We compute each speaker's i-vector by averaging all of the i-vectors from his/her utterances and clustering them into K classes. Each class center is taken as the memory module input, named 'NSD-MA-MSE(i-vector)'. The second one is the 256-dim x-vectors described in [18], which are extracted with deep neural network architecture based on ResNet101 [57], [60]. Like i-vectors, the memory module consists of K clustered centers of all speakers' averaged x-vectors extracted from the VoxCeleb1 and VoxCeleb2 datasets, named 'NSD-MA-MSE(x-vector)'. Meanwhile, to take advantage of both i-vectors and x-vectors, we also concatenate these two types of memory-aware multi-speaker embedding in the original NSD-MA-MSE models, named 'NSD-MA-MSE(i-vector+x-vector)'.

Most of our experiments are performed on the AMI corpus to determine configuration parameters where both training and testing data are provided. First, a series of experiments over different settings of speaker embedding and K are designed with the loss function in (8), where $\lambda = 0.1$. Then, another set of experiments is conducted for different settings of speaker embedding and λ, where K is set to 128. All the models are trained using the AMI training set, and the number of output speakers N is 4, which is the maximum number of speakers in most sessions of the AMI dataset.

Based on the experiments on the AMI corpus, the best configuration parameters are applied directly to the DIHARD-III dataset, and the number of output speakers N is set to 8, which covers most sessions in DIHARD-III. Since the training data for DIHARD-III are not provided, we use real recording datasets, including Switchboard-1 Release 2 LDC97S62 [61], AMI training set, Voxconverse DEV set and simulated multi-speaker dialogs using LibriSpeech, which are mentioned in Section III-B. We use Adam with a learning rate of 0.0001 to optimize the entire model on 4 RTX 3090 and get the best model after 2 epochs.

In the testing/inference stage, the VBx results in Section IV-C are used to generate the speaker mask matrix. Following the procedure in Section II-D, we decode both Mix-Headset and Beamformed data on AMI and the original data on DIHARD-III. No additional iterations were implemented for the speaker mask matrix purification since there wasn't a significant improvement.

F. ANSD-MA-MSE Based Diarization

We conduct the experiments for the ANSD-MA-MSE method on the DIHARD-III dataset due to the high mismatch between the training set and testing sets in different domains. As mentioned in Section III-A, we first select nonoverlap speech data with the VBx and NSD-MA-MSE results. Then, approximately 4-hour data are generated as in Section III-B for each session. Next, the pretrained NSD-MA-MSE model for DIHARD-III is fine-tuned with those data on the session level. Finally, each session is decoded on the fine-tuned model with the nonoverlapping speaker mask matrix generated from the VBx and NSD-MA-MSE results, and the same inference step for the NSD-MA-MSE model is adopted. This process can be adaptively performed to refine the diarization results.

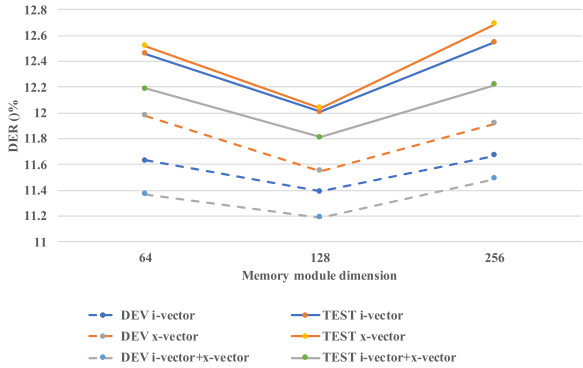


Fig. 4. DER comparison of NSD-MA-MSE methods with different speaker embedding and memory module dimensions on the AMI Mix-Headset DEV and TEST set.

V. Results and Analysis

In this section, we first show the results of the proposed NSD-MA-MSE method on both the AMI and DIHARD-III datasets. Then, we demonstrate the effectiveness of the proposed ANSD-MA-MSE method on the mismatched DIHARD-III dataset. Further analysis is also given to illustrate the advantages of the proposed method over other methods. Finally, our proposed method achieves the best diarization performance out of the top single systems of DIHARD-III final submissions.

A. NSD-MA-MSE Results on AMI

As shown in Fig. 4, the first experiment is designed for different settings of speaker embedding (i-vector, x-vector, and i-vector+x-vector) and memory module dimension ($K = 64, 128, 256$) in NSD-MA-MSE. The DER results on the AMI DEV set decoded with the Mix-Headset channel show that the memory module with 128 cluster centers performs consistently better across all three types of speaker embedding, making a good tradeoff between the speaker diversity and model generalizability. The 'NSD-MA-MSE(i-vector+x-vector)' system yields the best DER among different speaker embeddings, demonstrating the complementarity of speaker embedding based on the generative model (i-vector) and discriminative model (x-vector).

The second experiment in Fig. 5 examines the effect of weights ($\lambda = 0, 0.1, 0.5, 1.0$) for the loss function in (8) over different types of speaker embedding in the memory module of NSD-MA-MSE. $\lambda = 0$ indicates the NSD-MA-MSE model optimized using (6) without the auxiliary loss in (7). Based on the DER results of the Mix-Headset channel on the AMI DEV

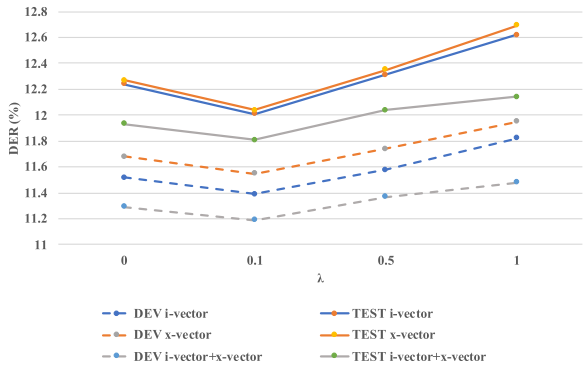


Fig. 5. DER comparison of NSD-MA-MSE methods with different speaker embedding and weights for auxiliary loss on the AMI Mix-Headset DEV and TEST set.

TABLE I
DER COMPARISON OF DIFFERENT DIARIZATION METHODS FOR BOTH MIX-HEADSET AND BEAMFORMED DATASETS OF AMI WITH **REFERENCE** SAD

Method	Speaker embedding	Mix-Headset		Beamformed	
		DEV	TEST	DEV	TEST
VBx	x-vector	16.48	17.86	18.31	19.67
TS-VAD	i-vector	14.12	14.32	16.84	16.55
TS-VAD	x-vector	14.45	14.68	17.21	16.89
TS-VAD	i-vector+x-vector	14.25	14.43	17.06	16.77
NSD-MA-MSE	i-vector	11.39	12.01	13.94	13.94
NSD-MA-MSE	x-vector	11.55	12.04	13.76	13.73
NSD-MA-MSE	i-vector+x-vector	**11.19**	**11.81**	**13.74**	**13.69**

TABLE II
DER COMPARISON OF DIFFERENT DIARIZATION METHODS FOR BOTH MIX-HEADSET AND BEAMFORMED DATASETS OF AMI WITH TDNN SAD

Method	Speaker embedding	Mix-Headset		Beamformed	
		DEV	TEST	DEV	TEST
VBx	x-vector	23.14	23.93	24.92	26.20
TS-VAD	i-vector	19.63	18.81	21.61	21.86
TS-VAD	x-vector	19.96	19.14	22.05	22.17
TS-VAD	i-vector+x-vector	19.83	19.02	21.84	22.03
NSD-MA-MSE	i-vector	16.89	17.18	18.89	19.54
NSD-MA-MSE	x-vector	16.98	17.10	19.27	19.63
NSD-MA-MSE	i-vector+x-vector	**16.71**	**16.95**	**18.74**	**19.53**

set, the NSD-MA-MSE model with the help of auxiliary loss ($\lambda = 0.1$) achieves consistent and remarkable improvements in the diarization performance across all three types of speaker embedding. Similar to the first experiment, the 'NSD-MA-MSE(i-vector+x-vector)' system performs the best.

With the best configuration of $K = 128$ and $\lambda = 0.1$, Table I illustrates the DER comparison of different diarization methods on AMI DEV and TEST sets for both Mix-Headset channel and Beamformed data with reference SAD. Table II shows the results with TDNN VAD. First, the neural speaker diarization methods of both NSD-MA-MSE and TS-VAD outperform VBx across all settings. Second, combining the i-vector and x-vector in TS-VAD does not always yield better DERs. Finally, for all types of speaker embedding, the DERs of our NSD-MA-MSE method is correspondingly lower than those of the TS-VAD method.

TABLE III
DER COMPARISON OF DIARIZATION METHODS ON DIHARD-III DEV AND EVAL FULL SET WITH BOTH REFERENCE AND TDNN SAD

SAD	Set	VBx	TS-VAD	NSD-MA-MSE
Reference	DEV	15.52	14.26	12.00
	EVAL	14.96	13.97	11.73
TDNN	DEV	18.86	17.85	14.95
	EVAL	21.30	20.12	17.32

TABLE IV
THE SPEAKER EMBEDDING (I-VECTOR) DISTANCE COMPARISON BETWEEN TS-VAD AND NSD-MA-MSE

Speaker embedding	**Type-I distance**	**Type-II distance**
i-vectors in TS-VAD	0.30	0.17
i-vectors in NSD-MA-MSE	0.37	0.11

We also compared initializing $\mathbf{S}$ with the ground truth and VBx in the testing stage and found that there was little performance difference, which means that the model is robust to the speaker mask matrix $\mathbf{S}$.

B. NSD-MA-MSE Results on DIHARD-III

In previous experiments, both model optimization and inference were conducted on the AMI corpus, where the data distributions between the training and testing sets were relatively matched. By switching to DIHARD-III, we have only the limited development and evaluation sets from 11 domains without the specific training data. Therefore, with the neural diarization models trained on real and simulated data as described in Section IV-E, there might be a high mismatch between training and testing for DIHARD-III. Here we performed experiments on DIHARD-III with the model 'TS-VAD with i-vector' for TS-VAD and 'NSD-MA-MSE with (i-vector+x-vector)' for NSD-MA-MSE results.

In Table III, we observe that the proposed NSD-MA-MSE method still yields better performance than the TS-VAD method for most domains. To further illustrate the superiority of NSD-MA-MSE over TS-VAD, we compare the static speaker embedding from a separate extractor in TS-VAD and the dynamic memory-aware speaker embedding from a jointly optimized diarization network in NSD-MA-MSE. Two types of distance measures are defined based on i-vectors for analysis. Type-I distance is used to examine the discrimination between different speakers, which is calculated using the Euclidean distance of two speaker i-vectors averaged across all two-speaker combinations among N speakers in one session. Type-II distance is used to study the effect of the initial VBx-based diarization results on the multi-speaker embedding extraction, which is computed using the Euclidean distance of two i-vectors extracted based on the oracle and VBx-based speaker mask matrix for each speaker in one session.

Table IV shows the speaker embedding (i-vector) distance comparison between TS-VAD and NSD-MA-MSE. We observe that the Type-I distance of NSD-MA-MSE is larger than that

TABLE V
DER OF ANSD-MA-MSE ON DIHARD-III DEV AND EVAL FULL SET WITH REFERENCE AND TDNN SAD

SAD	Set	NSD-MA-MSE	ANSD-MA-MSE
Reference	DEV	12.00	11.56
	EVAL	11.73	11.12
TDNN	DEV	14.95	14.49
	EVAL	17.32	16.76

TABLE VI
DER COMPARISON OF DIARIZATION METHODS ACROSS 11 DOMAINS ON DIHARD-III EVAL FULL SET WITH REFERENCE SAD

Domain	VBx	TS-VAD	NSD-MA-MSE	ANSD-MA-MSE
AUDIOBOOK	0.01	0.01	0.01	0.01
BROADCAST	4.25	4.48	4.29	4.18
CLINICAL	9.19	14.82	9.96	7.78
COURTROOM	3.07	3.90	2.85	2.86
CTS	13.87	5.98	5.78	5.69
MAPTASK	3.32	4.66	2.64	1.63
MEETING	31.33	29.21	28.41	26.20
RESTAURANT	38.93	51.28	37.80	37.65
SOCIO FIELD	8.04	9.41	7.45	6.68
SOCIO LAB	5.89	6.78	4.98	3.39
WEBVIDEO	36.75	37.43	36.45	35.75
ALL	14.96	13.97	11.73	11.12

of TS-VAD, implying that a set of more discriminative multi-speaker embeddings can be generated via the joint optimization of speaker embedding extraction and diarization. Furthermore, the smaller Type-II distance of NSD-MA-MSE than TS-VAD indicates that the proposed memory-aware multi-speaker embedding is less affected by the initial speaker diarization results.

C. ANSD-MA-MSE Results on DIHARD-III

Table V lists the overall performance of the ANSD-MA-MSE on DIHARD-III development (DEV) and evaluation (EVAL) full sets. The ANSD-MA-MSE achieves substantial improvements over NSD-MA-MSE. With the model adaptation, ANSD-MA-MSE reduces the DER of NSD-MA-MSE from 12.00% to 11.56% on the development set and from 11.73% to 11.12% on the evaluation set, where reference SAD is provided.

Table VI displays the DER comparison of VBx, TS-VAD, NSD-MA-MSE, and ANSD-MA-MSE across 11 domains on the DIHARD-III EVAL full set. Both TS-VAD and NSD-MA-MSE outperform VBx on well-matched domains such as CTS and MEETING corresponding to Switchboard and AMI in the training set. For other domains, the high mismatch between training and testing leads to the performance degradation of TS-VAD, while NSD-MA-MSE still improves the diarization results more than VBx. ANSD-MA-MSE consistently performs better than NSD-MA-MSE, indicating the effectiveness of modal adaptation. For matched domains (CTS and MEETING), additional gains can be observed from NSD-MA-MSE to ANSD-MA-MSE. For most of the other mismatched domains, ANSD-MA-MSE achieves considerable DER reductions over NSD-MA-MSE. Even for the WEBVIDEO and RESTAURANT domains, ANSD-MA-MSE outperforms VBx, which includes a large number of overlapping speech and noisy environments.

TABLE VII
OVERALL COMPARISON OF PROPOSED ANSD-MA-MSE METHOD WITH TOP SINGLE SYSTEMS ON DIHARD-III EVAL SET WITH REFERENCE SAD

	FULL	CORE
Hitachi-JHU	12.74	15.34
BUT	16.54	15.50
USC-SAIL	18.19	19.76
USTC-NELSLIP	12.41	14.86
ANSD-MA-MSE	11.12	14.04

Fig. 6 is an example of two-speaker diarization results aligned with the input spectrogram for the comparison among VBx, NSD-MA-MSE, and ANSD-MA-MSE. The top image corresponds to the input spectrogram. There are 4 bar graphs representing oracle labels (Reference), clustering-based diarization (VBx), and the proposed NSD-MA-MSE and ANSD-MA-MSE. Four colors are used to denote the nonspeech segment (white), the first speaker segment (red), the second speaker segment (blue), and the two-speaker overlap segment (green). In the first black dotted box from the left, ANSD-MA-MSE detects overlap segments more accurately than NSD-MA-MSE (SPK2 to Overlap). In the second black dotted box from the left, ANSD-MA-MSE corrects the speaker errors (SPK1 to Overlap) of NSD-MA-MSE from VBx. In the third black dotted box from the left, ANSD-MA-MSE reduces the speaker errors on single speaker segments more than VBx. In the last black dotted box from the left, both ANSD-MA-MSE and NSD-MA-MSE reduce the errors of VBx (overlap to SPK2). These results indicate the superiority of ANSD-MA-MSE over NSD-MA-MSE and VBx for both overlapping and nonoverlapping segments.

D. Overall Comparison on DIHARD-III

Finally, we make an overall comparison of our proposed ANSD-MA-MSE method with the top single-model systems in the DIHARD-III challenge, as shown in Table VII. Here is a simple brief description of these systems.

- *Hitachi-JHU [42]:* EEND is used as postprocessing to update the diarization results of the TDNN-based x-vector system.
- *BUT [41]:* VBx is based on applying BHMM to TDNN x-vectors, while the PLDA model is replaced by a heavy-tailed PLDA (HTPLDA).
- *USC-SAIL [62]:* Based on multiple embedding extractors, the domain adaptive speaker diarization system employs two different approaches using hard decisions and soft decisions.
- *USTC-NELSLP [43]:* This is our champion system that combines various front-end techniques to solve the diarization problem, including speech separation, TS-VAD, and iterative system optimization.

The proposed ANSD-MA-MSE system achieves the best results among all systems for both FULL and CORE settings on DIHARD-III Track 1 with oracle SAD, which outperforms our USTC-NELSLIP champion system and other top systems considerably.

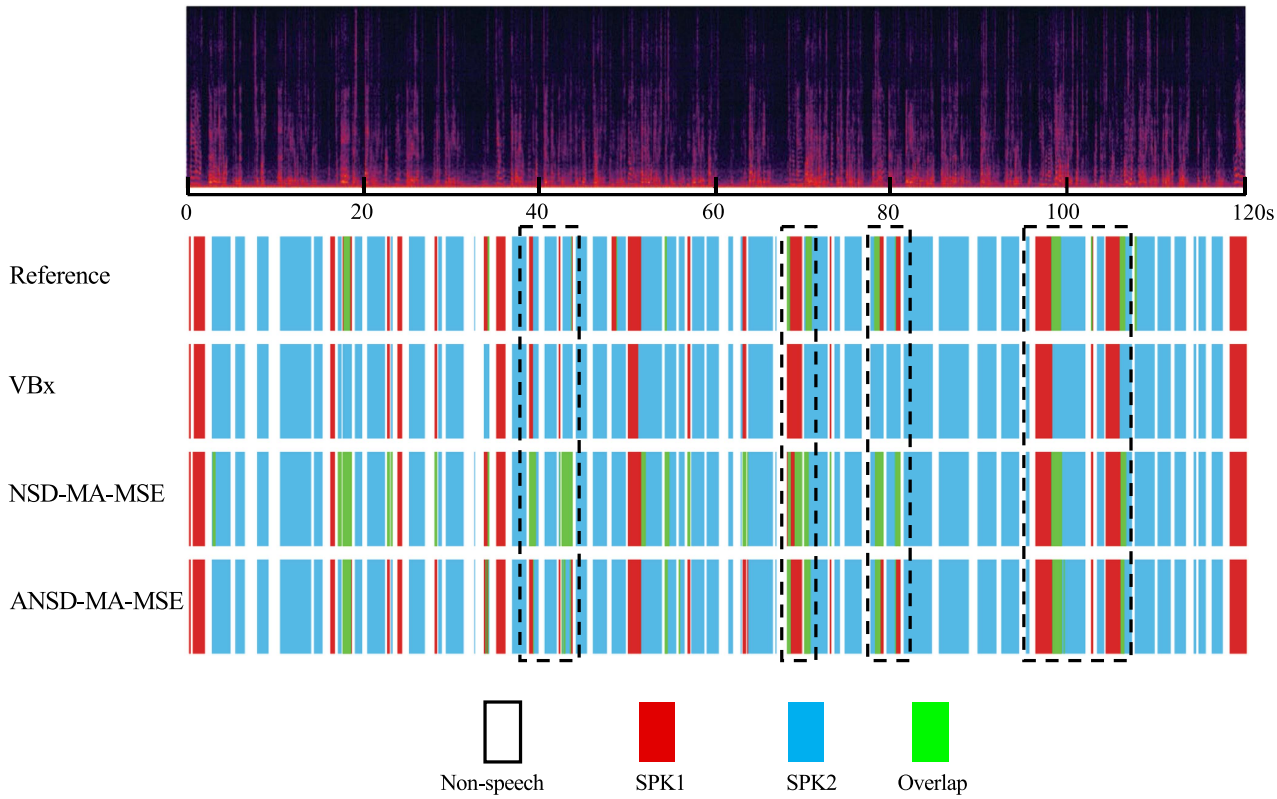


Fig. 6. An example of two-speaker diarization results aligned with the input spectrogram for the comparison among VBx, NSD-MA-MSE, and ANSD-MA-MSE.

VI. Conclusion

In this study, we first propose a novel speaker diarization method named NSD-MA-MSE to handle overlapping segments and to improve the performance over the TS-VAD method under matched conditions. To apply NSD-MA-MSE to mismatched conditions, we introduce a model adaptation strategy named ANSD-MA-MSE that substantially improves the diarization performance for most domains of the DIHARD-III challenge. In the future, we will investigate how to improve the results of those extremely challenging domains (many speakers, large overlap ratio, and adverse environments), such as WEBVIDEO and RESTAURANT, in the DIHARD-III challenge.

References

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101317.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[3] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1382–1393, Sep. 2009.

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4930–4934.

[5] 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.

[6] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Interspeech*, 2013, pp. 1481–1487.

[7] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial neural network features for speaker diarization," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 402–406.

[8] 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.

[9] 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.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.

[12] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

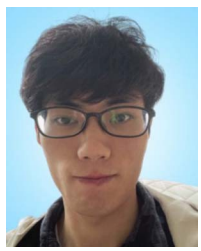
[13] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 1853–1856.

[14] S. Chen et al., "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, vol. 8, pp. 127–132.

[15] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, 2006, vol. 5, pp. 521–524.

[16] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 217–227, Jan. 2014.

[17] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[18] 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.
[19] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Cernockỳ, "Bayesian HMM based x-vector clustering for speaker diarization," in *Proc. Interspeech*, 2019, pp. 346–350.
[20] Z. Huang et al., "Speaker diarization with region proposal network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6514–6518.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," in *Proc. Interspeech*, 2021, pp. 3555–3559.
[26] 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.
[27] N. Ryant et al., "First DIHARD challenge evaluation plan," Tech. Rep., 2018.
[28] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2006, pp. 309–322.
[29] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. Interspeech*, 2018, pp. 2758–2762.
[30] J. H. Hansen et al., "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *Proc. Interspeech*, 2019, pp. 1851–1855.
[31] A. Ortega et al., "Albayzin evaluation: IberSpeech-RTVE 2018 speaker diarization challenge," 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/EvalPlan-SpeakerDiarization-v1.3.pdf
[32] S. Watanabe et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th Int. Workshop Speech Process. Everyday Environ.*, et al., pp. 1–7.
[33] A. Nagrani et al., "Voxsrc 2020: The second voxceleb speaker recognition challenge," 2020, *arXiv:2012.06867*.
[34] 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.
[35] M. Diez et al., "BUT system for DIHARD speech diarization challenge," in *Proc. Interspeech*, 2018, pp. 2798–2802.
[36] N. Ryant et al., "Second DIHARD challenge evaluation plan" Linguistic Data Consortium, Tech. Rep., 2019.
[37] F. Landini et al., "BUT system for the second DIHARD speech diarization challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6529–6533.
[38] M. Sahidullah et al., "The speed submission to DIHARD II: Contributions & lessons learned," in *Proc. Interspeech*, 2019, pp. 999–1002.
[39] Z. Zajíc, M. Kunešová, M. Hrúz, and J. Vaněk, "UWB-NTIS speaker diarization system for the DIHARD II 2019 challenge," in *Proc. Interspeech*, 2019, pp. 993–997.
[40] N. Ryant et al., "The third DIHARD diarization challenge," in *Proc. Interspeech*, 2021, pp. 3570–3574.
[41] F. Landini et al., "BUT system description for the third DIHARD speech diarization challenge," in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.
[42] S. Horiguchi et al., "The Hitachi-JHU DIHARD III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by DOVER-Lap," in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.
[43] Y. Wang et al., "USTC-NELSLIP system description for DIHARD-III challenge," in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.
[44] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.
[45] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[46] N. Kanda et al., "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," in *Proc. Interspeech*, 2020, pp. 36–40.
[47] 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.
[48] Y.-X. Wang, J. Du, M. He, S. Niu, L. Sun, and C.-H. Lee, "Scenario-dependent speaker diarization for DIHARD-III Challenge," in *Proc. Interspeech*, 2021, pp. 3106–3110.
[49] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.
[50] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.
[51] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, 1967, vol. 1, no. 14, pp. 281–297.
[52] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA : Curran Associates, Inc., 2015.
[53] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[54] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[55] I. Mccowan et al., "The AMI meeting corpus," in *Proc. Meas. Behav.*, 2005.
[56] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "3rd DIHARD challenge evaluation plan," 2020, *arXiv:2006.05815*.
[57] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to voxceleb speaker recognition challenge 2019," in *Proc. The VoxCeleb Challange Workshop*, 2019.
[58] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[59] W. Wang, Q. Lin, D. Cai, L. Yang, and M. Li, "The DKU-Duke-Lenovo system description for the third DIHARD speech diarization challenge," in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[61] J. Godfrey and E. Holliman, "Switchboard-1 release 2 LDC97S62," 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC97S62
[62] T. J. Park, R. Peri, A. Jati, and S. Narayanan, "USC-SAIL system for DIHARD III: Domain adaptive diarization system," in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.



Mao-Kui He received the B.S. degree in 2018 from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, where he is currently working toward the Ph.D. degree with the National Engineering Research Center of Speech and Language Information Processing. His research includes speech enhancement and speaker diarization.

Jun Du (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was with iFlytek Research as a Team Leader, working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an Associate Researcher, working on handwriting recognition, and OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and a Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team was the recipient of several champions of the CHiME-4/CHiME-5/CHiME-6 Challenge, SELD Task of 2020 DCASE Challenge, and DIHARD-III Challenge.

Qing-Feng Liu received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 1998 and 2003, respectively. He is the Founder, CEO, and President of iFlytek, the Director of the National Speech and Language Engineering Laboratory of China, a Professor and Doctoral Advisor with USTC, the Director General of the Union of Speech Industry of China, and the Director General of the Union of National University Student Innovation and Entrepreneurship.

Chin-Hui Lee (Life Fellow, IEEE) is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience, ending at Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff, and the Director of the Dialogue Systems Research Department. He has authored or coauthored more than 550 papers and 31 patents. They have been more than 55,000 times with an h-index of 80 on Google Scholar. He was the recipient of numerous awards, including the Bell Labs President's Gold Award in 1998. He was also the recipient of the 2006 Technical Achievement Award from IEEE Signal Processing Society for Exceptional Contributions to the Field of Automatic Speech Recognition. In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal for Scientific Achievement for Pioneering and Seminal Contributions to the Principles and Practices of Automatic Speech and Speaker Recognition.