QDM-SSD: Quality-Aware Dynamic Masking for Separation-Based Speaker Diarization

Shu-Tong Niu, Jun Du, *Senior Member, IEEE*, Lei Sun, Yu Hu, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We improve iterative separation-based speaker diarization (ISSD) with quality-aware dynamic masking (QDM). We call the proposed framework QDM-SSD. Compared with ISSD, QDM-SSD enhances the simulated data used for model adaptation through QDM to alleviate the influence of errors in speaker priors. In addition to data quality purification, QDM-SSD also makes the adaptation data sparse by automatically adjusting speaker overlap ratios according to data quality. Furthermore, using a sliding window over the adaptation data, clean regions in speech segments can be better localized. Experiments on the two-speaker conversational telephone speech (CTS) corpus show that the proposed QDM-SSD framework can reduce the diarization error rate (DER) by 18.56% relatively compared with ISSD. Moreover, QDM-SSD is shown to generalize to other two-speaker non-conversation telephone speech data sets where ISSD fails to work. Finally, we demonstrate that QDM-SSD can serve as a front-end to improve the performances of back-end automatic speech recognition.**

***Index Terms*—Data quality control, dynamic mask, speaker diarization, speech separation, voice activity detection.**

I. Introduction

SPEAKER diarization is a task to segment an arbitrary audio recording into homogeneous regions according to speaker identities [1], [2]. It is widely used in many applications, including meeting summarization, telephone conversations analysis and speaker based indexing [2], [3], [4]. It can also be placed as a front-end of automatic speech recognition (ASR) in multi-speaker conversation scenarios, such as meetings and home environments [5], [6]. Due to a rising demand in real-world applications, a series of evaluation challenges for speaker diarization have been held [7], [8], [9], [10], [11], [12]. Specifically, DIHARD Challenge [10], [11], [12] aims to promote diarization technology development in realistic and challenging domains.

One mainstream approach is clustering-based speaker diarization (CSD), which generates results by clustering speaker representations [13], [14], [15], [16], [17], [18]. It usually contains multiple independent modules. For speaker representation extraction, powerful speaker embeddings, including i-vector [19], d-vector [20] and x-vector [21], have been explored. For clustering, agglomerative hierarchical clustering (AHC) [13], spectral clustering (SC) [14], mean shift clustering [15], and k-means clustering [16] have been attempted. These clustering-based diarization systems have been proved effective in a variety of domains, and have achieved good rankings in DIHARD-I and DIHARD-II Challenges [17], [18]. In particular, the variational Bayesian hidden Markov model with x-vectors (VBx) [18] has achieved a state-of-the-art performance among CSD systems and won the first place in DIHARD-II Challenge. Nonetheless, conventional CSD methods cannot well handle speaker-overlap regions in conversational speech as conventional clustering algorithms can only assign one specific speaker label to a speech segment. Many approaches have been proposed to handle the overlapping regions in speaker diarization, which can be roughly divided into two categories. The first category employs an external overlap detector to detect the overlapping regions and assigns additional speakers to the detected overlapping segments. For example, the overlap-aware resegmentation method [22] and the overlap-aware spectral clustering method [23]. The second category employs the end-to-end framework to handle the overlapping regions [24], [25], [26], [27], [28]. End-to-end neural speaker diarization (EEND) [24], [25] reformulates speaker diarization as a multi-label classification problem to directly predict activities of all speakers at each time frame. Moreover, audio recordings containing an unknown number of speakers can be handled by some modifications in the EEND framework [25]. Inspired by EEND and personal-VAD [29], target-speaker voice activity detection (TS-VAD) was proposed [26], [27], [28], which can also predict the presence probabilities of all speakers simultaneously. The main difference is that TS-VAD needs embeddings of pre-enrolled target-speakers as additional inputs to avoid speaker permutation problems [24]. Further research was also explored to estimate an unknown number of speakers in TS-VAD [28]. In addition, some recent works utilize auxiliary information from ASR to improve the performance of end-to-end systems [30], [31].

Speech separation is one of the most straightforward techniques to handle the speaker-overlap regions by separating each source speaker from multi-speaker mixtures [32]. Most separation techniques are either time-frequency (T-F) domain

Manuscript received 7 May 2022; revised 21 December 2022; accepted 31 January 2023. Date of publication 13 February 2023; date of current version 24 February 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.)*

Shu-Tong Niu, Jun Du, Lei Sun, and Yu Hu are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: niust@mail.ustc.edu.cn; jundu@ustc.edu.cn; sunlei17@mail.ustc.edu.cn; yuhu@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).

Digital Object Identifier 10.1109/TASLP.2023.3244513

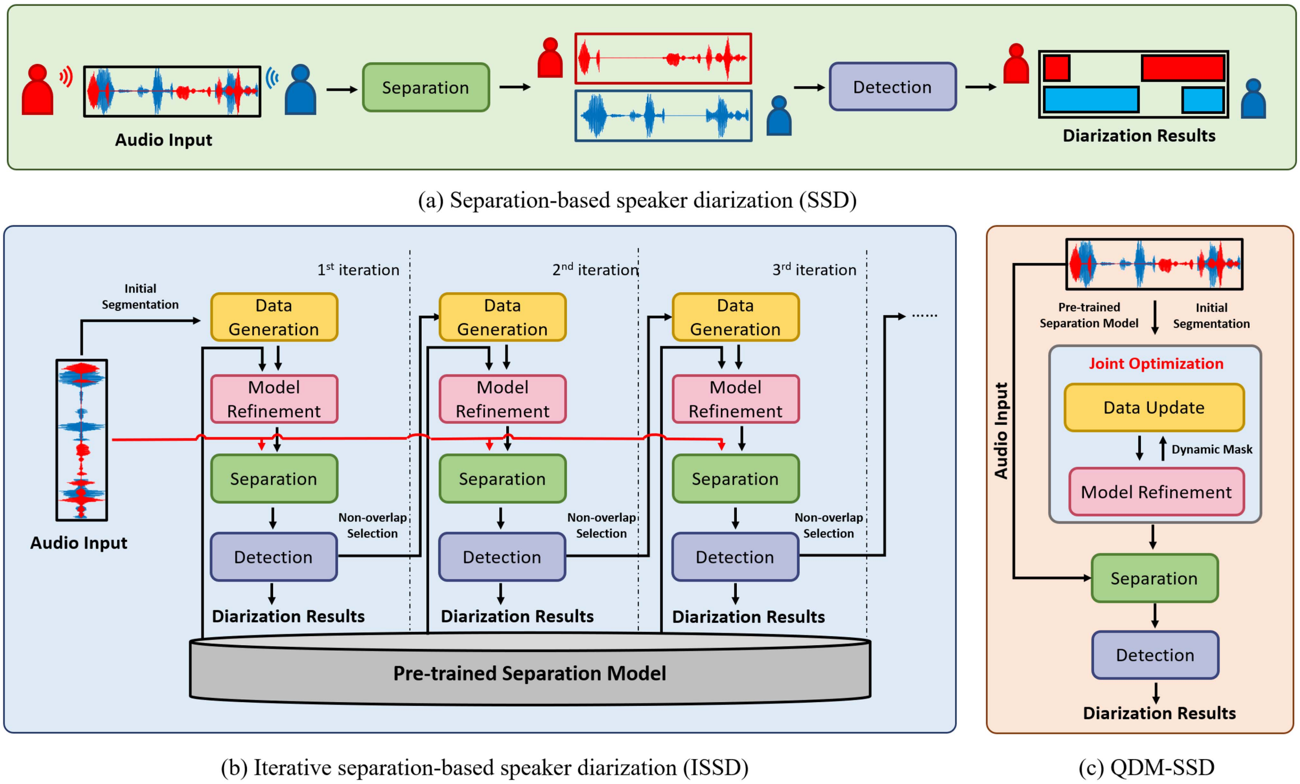
based [33], [34], [35], [36], [37] or time-domain based [38], [39], [40]. The former ones first apply the short-time Fourier transform (STFT) [32] to the input speech mixtures to obtain the corresponding T-F representations (or spectrograms), then use nonlinear regression techniques to directly estimate individual source spectrograms [33] or corresponding time-frequency masks [34], [35], [36], [37]. Numerous model architectures, including feed-forward neural networks [33], recurrent neural networks (RNNs) [34], convolutional neural networks (CNNs) [35], generative adversarial networks (GANs) [36], and self-attention based networks [37], have been used. However, the performances of these techniques remain suboptimal due to issues related to imperfect phase reconstruction [41]. Recently, time-domain based neural networks, mostly utilizing encoder-decoder architectures [42] to directly separate the time-domain mixtures, have achieved exciting results in speech separation tasks [38], [39], [40].

Speech separation techniques can be naturally adopted in speaker diarization task. In [43], streams estimated by separation models are fed into embedding extraction modules to help the CSD systems handle overlapping speech. Some other studies jointly perform source separation, speaker counting and diarization through multi-task frameworks [44], [45], which implicitly generate diarization results through speech separation. In addition, continuous speech separation (CSS) has attracted more and more research attention recently [37], [46], using data sets similar to those used for diarization tasks. Since the separation models are often trained with simulated data, there is a mismatch between training and testing data when dealing with realistic recordings due to the unavailability of clean sources, which causes the separation performances to be unstable [47]. To alleviate this problem, our team has proposed different approaches to separation-based speaker diarization (SSD) [47], [48], [49]. In [47], [48], we investigated some strategies leveraging upon complementary properties of speech separation and speaker clustering, and automatically chose the clustering-based results when separation results were outrageous. In [49], we introduced an iterative separation-based speaker diarization (ISSD) approach which alternately updated a set of adaptation data and fine-tuned the separation model, leading to a significant performance improvement. Compared with the end-to-end systems, ISSD can also handle overlapping speech, and the separated streams can be directly used for back-end processing. By incorporating ISSD, our diarization system ranked first among all submitted systems in DIHARD-III Challenge [50].

Although ISSD achieved top diarization performances when compared to CSD systems [50], [51], there are still two main unresolved issues in ISSD. First of all, in generating the adaptation data, information obtained with CSD or from the previous iteration is utilized. However, these speaker priors often contain errors which contaminate the training data during model adaptation. This will lead to an over-fitting problem as the separation model is fine-tuned with only a small amount of adaptation data. In fact, from the perspective of unsupervised adaptation, it is common to see errors in priors (also known as pseudo labels) [52], [53], [54], [55], which is harmful to adaptation, thereby limiting the performance of the adapted model. Second, the gap between the separation and diarization tasks introduces an upper bound on the ISSD performance. For all SSD-based techniques, the diarization performance heavily relies on the quality of the separated streams. Nonetheless, different from the diarization data which contain both overlapped and overlap-free regions, the separation model in ISSD is trained and fine-tuned only on fully overlapping speech mixtures like in most mainstream speech separation studies [33], [34], [35], [36], [37], [38], [39], [40]. This will bias the separation model towards assuming that the test recordings are also fully overlapped. Correspondingly, the false alarm errors in ISSD will increase because many regions are misclassified as overlapped speech. Although in previous works [49], we have proposed some post-processing techniques to mitigate over-detection of overlapped regions in ISSD, the task mismatch problem has not been fully solved and limits the performance of both speaker diarization and its application on back-end ASR.

Here, we propose separation-based speaker diarization with quality-aware dynamic mask (QDM-SSD) to address the above two issues, which can jointly perform data update and model refinement in each iteration through quality-aware dynamic masking (QDM). In QDM-SSD, we first adopt the separation ability of the adapted model to judge the quality of speech segments used for adaptation data simulation. Then, we generate QDMs whose active lengths are variable according to the quality of the corresponding segments. Through applying the QDMs on the original speech segments, most poor-quality parts are masked. This can effectively reduce the influence of the errors in speaker priors during model adaptation, which alleviates the first problem in ISSD. At the same time, utilizing the speech segments masked by QDMs to simulate the adaptation data can make the data sparse and the styles of the adaptation data similar to those of diarization data, which effectively alleviates the second problem in ISSD. Furthermore, to better control the quality of the adaptation data, we propose a start-point localization (SL) technique to capture the clean regions in speech segments through a sliding window. We verify the effectiveness of the proposed techniques with a realistic two-speaker conversational telephone speech (CTS) [15] data set and other two-speaker non-conversation telephone speech (NCTS) data sets (in which ISSD fails to work) from DIHARD-III Challenge [12]. The experimental results show that, in different scenarios, QDM-SSD decreases both speaker misclassification and false alarm errors when compared to ISSD. Moreover, the SL method can help QDM-SSD achieve further improvements. Finally, we employ QDM-SSD as the front-end for ASR tasks and demonstrate that QDM-SSD can make the back-end ASR system better handle overlapped speech.

The remainder of this paper is organized as follows. In Section II, we give an overview of our previous SSD works. In Section III, we elaborate on the proposed QDM-SSD framework. In Section IV, we present experimental results with detailed analyses. In Section V, we discuss the scalability of the proposed method under multi-speaker scenarios. Finally, we draw our conclusions in Section VI.



(a) Separation-based speaker diarization (SSD)

(b) Iterative separation-based speaker diarization (ISSD)

(c) QDM-SSD

Fig. 1. Overall framework comparison for different SSD methods.

II. Prior Work

A. Separation-Based Speaker Diarization

Our previously-proposed separation-based speaker diarization (SSD) [47], [48] framework is shown in Fig. 1(a), which mainly contains two modules: separation and detection. Given an audio mixture $y(t)$ which contains two speakers:

$$y(t) = x_1(t) + x_2(t) \tag{1}$$

where t is the sample index in the time domain. $x_i(t)$ denotes the i-th source signal. We first perform a separation process:

$$f_\theta(y(t)) = \{\hat{x}_1(t), \hat{x}_2(t)\} \tag{2}$$

where $f_\theta(\cdot)$ denotes a separation model with parameter set θ, and $\hat{x}_1(t)$ and $\hat{x}_2(t)$ are two separated streams. Here, we train a widely used time-domain model, Conv-TasNet [38], using the data set simulated from Librispeech [56]. For the two outputs of $f_\theta(\cdot)$, we form an overall loss using a permutation invariant training (PIT) [57] objective:

$$E = -\frac{1}{2}\sum_{n=1}^{2} l(\hat{x}_n(t), x_\phi(t)) \tag{3}$$

where $\hat{x}_n(t)$ is the n-th separated stream, $x_\phi(t)$ is reference speech with a permutation index ϕ that minimizes E. For l, we use a commonly adopted metric for separation, namely scale-invariant source-to-noise ratio (Si-SNR):

$$l(\hat{x}_n(t), x_\phi(t)) = 10\log_{10}\frac{||\hat{x}'_n(t)||^2}{||\hat{x}_n(t) - \hat{x}'_n(t)||^2} \tag{4}$$

where $\hat{x}'_n(t) = \frac{\langle \hat{x}_n(t), x_\phi(t)\rangle x_\phi(t)}{||x_\phi(t)||^2}$ means the projection of separated stream $\hat{x}_n(t)$ onto reference speech $x_\phi(t)$, with $\langle \hat{x}_n(t), x_\phi(t)\rangle$ being the dot product of $\hat{x}_n(t)$ and $x_\phi(t)$.

In the detection part in Fig. 1(a), we use voice activity detection (VAD) [58] to detect the speaker presence in $\hat{x}_1(t)$ and $\hat{x}_2(t)$, obtaining a variable number of M_n segments as:

$$\mathbf{X}_n = \{\hat{\mathbf{x}}_{n,j} | j = 1, 2, \ldots, M_n\} \tag{5}$$

where $\hat{\mathbf{x}}_{n,j}$ represents the j-th segment in the n-th stream. M_n is the number of the detected segments. By combining the time labels and channel properties of the speech segments in $\mathbf{X}_1$ and $\mathbf{X}_2$, we can generate the corresponding speaker distributions, including information from the overlapping regions as shown in the ‘Diarization Results’ of Fig. 1(a).

B. Iterative Separation-Based Speaker Diarization

In Fig. 1(b), we illustrate the ISSD framework [49], [51]. A key in ISSD improvement is the use of an iterative label-learn-relabel process in model refinement. As shown in Fig. 1(b), we first perform an initial segmentation using the speaker priors from CSD. Then, we adopt the obtained speech segments to generate adaptation data and conduct a light fine-tuning on the separation model. Finally, we alternately generate the adaptation data using speaker priors from the previous iteration and refine the separation model. The above processes are reflected in the two new components in Fig. 1(b) added to SSD in Fig. 1(a), namely ‘Data Generation’ and ‘Model Refinement’ to be described in the following.

1) Data Generation: We first obtain the diarization prior $\mathbf{y}_i \in \{0,1\}^{1\times T}$ for speaker i in $y(t)$, whose elements are:

$$y_{i,t} = \begin{cases} 0 & \text{Speaker } i \text{ is inactive at } t, 1 \leq t \leq \mathrm{T} \\ 1 & \text{Speaker } i \text{ is active at } t, 1 \leq t \leq T \end{cases} \tag{6}$$

Then we perform the non-overlap selection on $\mathbf{y}_i$:

$$\mathbf{y}'_i = \mathbf{y}_i \odot \mathbf{m} \tag{7}$$

where $\mathbf{m}$ is a time mask with elements $m_t = 1$ if only one speaker is present at t, and $m_t = 0$ otherwise, and $\odot$ is an element-by-element multiplication. We first obtain the time indexes of the parts in $\mathbf{y}'_i$ with a value of 1, and then we can use the obtained time indexes to capture the single-speaker speech segments in $y(t)$ for speaker i as:

$$\mathbf{S}_i = \{\mathbf{s}_{i,j} | j = 1, 2, \ldots, N_i\} \tag{8}$$

where $\mathbf{s}_{i,j}$ represents the j-th segment for the i-th speaker, and N_i is the total number of segments. We simulate paired mixtures by randomly selecting and mixing two L-second-long segments from $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively.

2) Model Refinement: Starting with the separation model in SSD as the pre-trained model, we adapt it to $y(t)$ and fine-tune the pre-trained model utilizing the simulated adaptation data via (3) and (4) as in SSD.

The above two processes gradually adapt the separation model according to $y(t)$, which can effectively improve the stability of the separation model and refine the diarization results. When compared with the state-of-the-art CSD system, the ISSD framework with proper post-processing methods can yield approximately 47% relative diarization error rate reduction on the realistic CTS data set [49]. By incorporating ISSD method, our overall system ranked the first place among all submitted systems in the DIHARD-III Challenge [50].

III. Separation-Based Speaker Diarization With Quality-Aware Dynamic Mask

Fig. 1(c) shows the overall framework of the proposed separation-based speaker diarization with quality-aware dynamic mask (QDM-SSD). In the proposed QDM-SSD, 'QDM' means the quality-aware dynamic mask, which is a vector composed of 0/1 elements. The active regions (the values are equal to 1) of QDM are determined by the quality of the corresponding speech segments, and that's why we call it 'quality-aware'. In this paper, the quality of speech segments can be defined as follows: (1) the good-quality speech segments are the speech segments that only contain one speaker; (2) the poor-quality speech segments are the speech segments that include more than one speaker, which may contain a large proportion of interfering speech. Similar to the ISSD framework, QDM-SSD also requires the initial speaker priors to adapt the pre-trained separation model. However, QDM-SSD does not directly use the speech segments obtained from the speaker priors, which may contain some inevitable interfering speech. Instead, QDM-SSD first applies QDMs on these speech segments to make them pure and sparse, and then uses them in the model adaptation process, as illustrated in the 'Data Update' module in Fig. 1(c). At the same time, the adapted model fine-tuned on the purified and sparse data can help generate more accurate QDMs, resulting in better-quality adaptation data. Therefore, from the perspective of system optimization, the main advantage of QDM-SSD over ISSD is that the former can jointly perform the data update and model refinement through the QDM, which can also effectively mitigate the two problems mentioned in Section I. On the one hand, QDM can block out most misleading speech segments in adaptation data generation, leading to an increase in data purity, which can alleviate the impact of errors in speaker priors. On the other hand, the active lengths of QDMs for different speech segments are variable according to the quality of the corresponding segments, which enables the overlap ratio of adaptation data to be set automatically. It makes the adaptation data similar to the test data of the diarization task, and thus bridges the gap between speech separation and speaker diarization. In addition, unlike ISSD which needs to regenerate data in each iteration, the proposed QDM-SSD updates adaptation data through online mixing, which effectively simplifies the process.

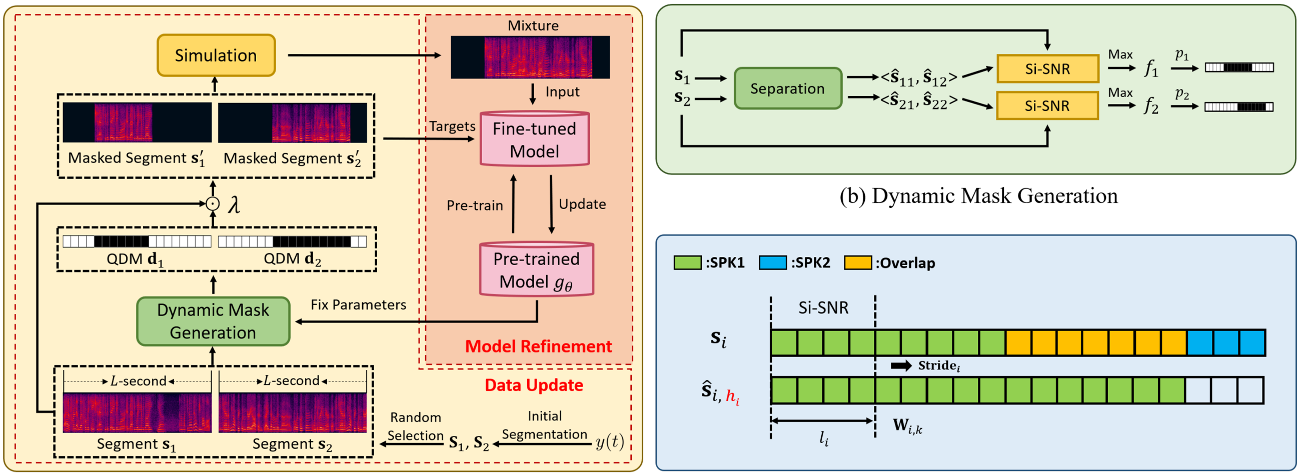
A. Joint Optimization

Fig. 2(a) details joint optimization, the core module of QDM-SSD shown in Fig. 1(c), consisting of two interacting parts, namely, data update and model refinement. In the data update part, the initial speaker priors are utilized to perform the segmentation on the two-speaker input audio signal $y(t)$, obtaining the two sets, $\mathbf{S}_1$ and $\mathbf{S}_2$, which contain the corresponding hypothetical single-speaker speech segments like in (8). To generate the simulated mixture, two L-second-long segments $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^{1\times\lfloor L\times f_s\rfloor}$ are randomly selected from each of $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, as shown at the bottom of Fig. 2(a), where f_s is the sampling frequency. $\lfloor\cdot\rfloor$ denotes the floor function. Then, we fix the parameters of the pre-trained separation model $g_\theta(\cdot)$ to judge the quality of different speech segments, which is essential for generating the corresponding quality-aware dynamic masks. Through the element-wise multiplication between the original speech segments and the generated QDMs, we can obtain the sparse speech segments used for adaptation data generation:

$$\mathbf{s}'_i = \mathbf{s}_i \odot \mathbf{d}_i \tag{9}$$

where $i = 1, 2$ denotes the hypothetical speaker index for the corresponding segment. $\mathbf{d}_i \in \{0,1\}^{1\times\lfloor L\times f_s\rfloor}$ represents the quality-aware dynamic mask for segment $\mathbf{s}_i$, which contains both active part (i.e., the values are equal to 1) and inactive part (i.e., the values are equal to 0). $\mathbf{s}'_i$ is the corresponding masked (or sparse) segment as illustrated in Fig. 2(a). Furthermore, at the beginning of the iterative phase,[1] we cannot fully trust the separation ability of the model. Our previous SSD results [47] have demonstrated that the simple pre-trained separation model is unstable on realistic mismatched conditions. So in order to reflect our confidence in the separation abilities of the pre-trained models in different iterations, we employ a parameter $\lambda \in [0, 1]$

[1] To be consistent with ISSD, the 'iteration' in QDM-SSD is defined by the adaptation data generation process. That is, if the conditions to generate adaptation data change, we define this as a new iteration.



(a) Joint Optimization Framework

(b) Dynamic Mask Generation

(c) Chunk-wise Start-point Localization

Fig. 2. An illustration of the main components of the proposed QDM-SSD.

to control the probability of applying QDMs to the original speech segments as shown in Fig. 2(a):

$$\lambda = \min\left(\alpha \times (N_I - 1), 1\right) \tag{10}$$

where N_I denotes the number of iterations. $\alpha > 0$ is a pre-defined coefficient which guarantees that λ increases with the iteration number, i.e., as the number of iterations increases, our confidence in the separation ability of the adapted model also increases. Then, we can simulate the sparse mixture utilizing the masked segments $\mathbf{s}'_1$ and $\mathbf{s}'_2$. In model refinement, we employ the generated sparse mixture as the input, and $\mathbf{s}'_1$ and $\mathbf{s}'_2$ as the targets to adapt the separation model, as shown in Fig. 2(a). In this process, we still adopt (3) and (4) to optimize the model parameters. Ultimately, the adapted model will be used as the pre-trained model in the next iteration.

As shown in Fig. 2(a), data update needs the pre-trained model g_θ to judge speech quality to generate the corresponding QDMs (the details will be introduced in Section III-B). The model refinement needs data update to generate data for model adaptation. The more the adapted model can generate accurate dynamic masks to simulate higher quality adaptation data, the better the simulated data can further improve the adapted model. Therefore, in general, the two modules of the joint optimization framework dynamically form a closed loop to improve the diarization performance. Our experimental results also confirm the effectiveness of QDM-SSD for alleviating the two problems mentioned in Section I. It is noted that mask generation is crucial for joint optimization. We will elaborate on the QDM generation process next.

B. Dynamic Mask Generation

Fig. 2(b) shows the details of dynamic mask generation, which essentially utilizes the separation ability of the pre-trained separation model $g_\theta(\cdot)$ to estimate the quality of input segments and then generates the dynamic masks accordingly. For the two input segments, $\mathbf{s}_1$ and $\mathbf{s}_2$, we first perform separation utilizing the pre-trained model, which can be expressed as:

$$\hat{\mathbf{S}} = g_\theta\left(\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix}\right) = \begin{bmatrix}\hat{\mathbf{s}}_{11} & \hat{\mathbf{s}}_{12}\\ \hat{\mathbf{s}}_{21} & \hat{\mathbf{s}}_{22}\end{bmatrix} \tag{11}$$

where $\hat{\mathbf{s}}_{i1}$ and $\hat{\mathbf{s}}_{i2}$ are two separated streams of $\mathbf{s}_i$ with $i = 1, 2$ representing the hypothetical speaker index. Then, we calculate the relative Si-SNR between the original speech segments and the two corresponding separated streams:

$$v_{i,j} = f_{\text{Si-SNR}}\left(\hat{\mathbf{s}}_{i,j}, \mathbf{s}_i\right) \tag{12}$$

where $v_{i,j}$ denotes the score of the j-th separated stream from the i-th speech segment. $f_{\text{Si-SNR}}(\cdot)$ is the Si-SNR function as indicated in (4). After calculating the relative Si-SNRs for all separated streams in (11), we can obtain a score matrix, whose elements correspond to those in $\hat{\mathbf{S}}$ of (11):

$$\mathbf{A} = \begin{bmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{bmatrix} \tag{13}$$

Next, we employ the larger of the scores obtained in the two separated streams from the i-th speech segment as the final score for the i-th speech segment:

$$f_i = \max\left(v_{i1}, v_{i2}\right) \tag{14}$$

where f_i denotes the final score. We can also obtain the index of the separated stream with a higher score for the i-th segment, which is represented as h_i. In the proposed QDM-SSD framework, we assume that a fully-adapted separation model can assign most speech segments from the same speaker to the same channel in the separation results. The experimental results in Section IV also prove the rationality of this assumption. Therefore, for the good-quality speech segments which contain only one speaker, one of the corresponding separated streams should be assigned to almost all speech, which is very similar to the original input segment, and the other separated stream should contain almost no speech, with a very low relative Si-SNR score. So overall, the maximum values of the two scores obtained in the separated streams, i.e., f_i in (14) for the single-speaker speech segments, are usually very high. For the poor-quality speech segments that contain more than one speaker (also include the overlapping regions), the adapted model can segregate the input into two streams according to the speaker identities, resulting in two separated streams being quite different from the original input segment. This is likely to result in low relative Si-SNRs (as defined in (12)) for all separated streams. Correspondingly, the final scores for these poor-quality segments are also quite low. In summary, the final scores of the speech segments defined in (14) can effectively reflect the quality of the corresponding speech segments: for the high-quality speech segments with only one speaker, the corresponding scores will be very high, and conversely, for the low-quality speech segments containing more than one speaker, the corresponding scores will be very low. To facilitate the generation of dynamic masks, we scale the final scores to $[0, 1]$ as follows:

$$p_i = \begin{cases} 0 & f_i \leq \tau_1 \\ \max\left(\frac{1}{1+\exp\left(-\beta\times\left(f_i - \frac{\tau_1+\tau_2}{2}\right)\right)}, p_{\min}\right) & \tau_1 < f_i < \tau_2 \\ 1 & f_i \geq \tau_2 \end{cases} \tag{15}$$

where p_i can be considered as the probability of active speech presence when generating the QDM. As can be seen, the main component (namely, the sub-function in the second row) of (15) is a deformation of the Sigmoid function. τ_1 and τ_2 are two predefined interval endpoints, whose negative mean value $-\frac{\tau_1+\tau_2}{2}$ is also employed as an offset added to the variable f_i, controlling the horizontal translation relative to the original Sigmoid function. $\beta > 0$ represents the scaling factor which controls the horizontal stretch in the formula deformation. Therefore, in general, we scale the range of the scores to $[0, 1]$ through a monotone increasing function in (15). Then, according to the obtained p_i, we can calculate the length of the active part for the corresponding dynamic mask:

$$l_i = \lfloor p_i \times L \rfloor \tag{16}$$

where L is the length of the selected segments. Note that there are some QDMs whose active lengths are zero ($p_i = 0$). For these QDMs, the corresponding segments usually contain a large proportion of interfering speech, and the masked signals (i.e., the target speech) are zeros, so we simply discard these segments to avoid the undefined values in Si-SNR calculation. It is also noted that we set the minimum active length of the generated dynamic masks to $\frac{L}{10}$, i.e., the value of $p_{\min}$ in (15) is 0.1. Our experimental results indicate that this constraint can avoid generating too many fragmentary segments and attain a better ability to detect overlapping speech. In essence, it is a tradeoff between under-detection and over-detection for overlapping speech. After obtaining l_i, the elements of the dynamic mask can be calculated as:

$$d_{i,t} = \begin{cases} 1 & \lfloor q_i \rfloor \leq \frac{t}{f_s} \leq \lfloor q_i \rfloor + l_i \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where t denotes the index of the sampling point in the speech segment with the sampling frequency f_s. q_i represents the starting point for the active part (i.e., the values are equal to 1) in the dynamic mask. In dynamic mask generation, it's important to locate the starting points for the active parts accurately. The reason is that if we can not find good starting points for the active parts in the dynamic masks, the poor-quality segments may be selected when simulating the adaptation data, which will introduce misleading information in the model adaptation. To find the starting point q_i which can accurately locate good-quality segments, we propose a chunk-wise start-point localization (SL) method as illustrated in Fig. 2(c), which can localize the start-point q_i, and benefit in selecting the good-quality parts in the original segments when applying the QDMs. As can be seen, we employ a sliding window whose length is the same as l_i in (16), and we move the sliding window forward by $\lfloor \frac{L}{100} \rfloor$ (namely, $\text{stride}_i = \lfloor \frac{L}{100} \rfloor$) for each time step. Thus the total number K of windows in different positions for the i-th speech segment can be calculated as $\lfloor \frac{L - l_i}{\text{stride}_i} \rfloor$. Then, we obtain the window-level relative Si-SNR:

$$w_{i,k} = f_{\text{Si-SNR}}\left(\mathbf{W}_{i,k}(\hat{\mathbf{s}}_{i,h_i}),\ \mathbf{W}_{i,k}(\mathbf{s}_i)\right) \tag{18}$$

where k is the position index ranging from 1 to K. In the i-th segment, $\mathbf{W}_{i,k}$ denotes the sliding window at the k-th position, and h_i is the index of the separated stream with a higher score. Therefore, the score $w_{i,k}$ is calculated as the relative Si-SNR between the higher-score separated stream $\hat{\mathbf{s}}_{i,h_i}$ and the original segment $\mathbf{s}_i$ within the window $\mathbf{W}_{i,k}$. Now, we can select the starting points corresponding to the relatively high $w_{i,k}$, obtaining a start-point candidate set:

$$C_i = \left\{ k \times \text{stride}_i \mid w_{i,k} \geq \frac{\tau_1 + \tau_2}{2} \right\} \tag{19}$$

where we assume that the parts with a window-level score of $\frac{\tau_1+\tau_2}{2}$ or greater can be regarded as good-quality sub-segments. Through (19), we can maintain enough good-quality parts and discard most poor-quality parts to enhance the performance of the adapted model.

IV. Experiments and Result Analyses

We focus on the two-speaker conversations, where both speech separation and speaker diarization techniques have been well established. The training set of the pre-trained separation model was simulated by the Librispeech corpus [56]. We generated about 250 hours of fully overlapped mixtures by randomly mixing two speech segments from different speakers. Furthermore, we adopted the development subset (about 3 hours) from CALLHOME American English Speech (LDC97S42) as our development set.

For diarization, to verify the effectiveness and generalization of our techniques, realistic recordings from three domains in the DIHARD-III corpus [12] were used as the evaluation data, including 'Conversational Telephone Speech' (or CTS), 'Map Task' (or MT) and 'Sociolinguistic Lab Recordings' (or SLR). The CTS domain consists of 61 ten-minute conversations between two native English speakers (about 10 hours), drawing from the unreleased Phase II data of the Fisher English collection [59]. The overall overlap ratio is about 12%, which is quite large in common two-speaker scenarios. The MT domain consists of 23 recordings (about 2.5 hours) where pairs of speakers engage in a map task. All recordings are drawn from the DCIEM Map Task Corpus (LDC96S38). Statistically, the overlap ratio of this domain is about 3%. The SLR domain includes 16 sociolinguistic interviews (about 2.5 hours) recorded in a controlled environment, and the overlap ratio is about 5%. All the recordings of this domain are taken from the LDC Mixer 6 collection (LDC2013S03).

We also adopted QDM-SSD as front-end processing for speech recognition on the HUB5 English evaluation data (LDC2002S09 + LDC2002T43) [60], which is a widely used ASR data set with approximately 11-hour conversational telephone speech, drawing from two sources: (i) 20 telephone conversations from Switchboard (SWB) and (ii) 20 telephone conversations from CALLHOME (CH). The overlap ratios in SWB and CH are 4.5% and 8.6%, respectively.

A conventional clustering-based speaker diarization (CSD) system, namely VBx [18], was taken as our baseline, and the obtained results were adopted as the speaker priors for ISSD and QDM-SSD to start the first iteration. Speech used in our experiments was sampled at $f_s = 8000$. Conv-TasNet [38] was employed as our separation model. The Asteroid toolkit [61] was used in our separation model pre-training and fine-tuning with a learning rate of 0.001. Adam [62] was adopted as the optimizer. Using speech segments of 3-second long, the pre-training process in QDM-SSD was the same as that in ISSD [49], where we ran 75 epochs. In the iterative phase, we simulated about 4-hour of adaptation data with an online style in each QDM-SSD iteration, and each simulated mixture is one-second-long (namely, $L = 1$), which is consistent with ISSD [49]. In each ISSD or QDM-SSD iteration, we fine-tuned the pre-trained model for one epoch, and we also updated the pre-trained model utilizing the adapted model obtained from the last iteration. We used the WebRTC VAD[2] to detect speech segments in each stream, which is the default configuration. In all experiments, we used the oracle VAD segment boundaries to refine the detected results, which is allowed in the first track of DIHARD-III Challenge [12]. Specifically, we used the oracle VAD to mask the silent segments and filled the neighborhood speaker in the undetected segments. In addition, to show the superiority of the QDM-SSD, we also employed a more conservative DNN-VAD as in [51] to detect the separated results in ISSD, which can help reduce false alarm errors. The constant α in (10) was set to 0.5. The β in (15) was set to 0.3. The constants, τ_1 and τ_2 in (15), were set to 10 and 30, respectively. All hyper-parameters were tuned with the development set mentioned earlier. The diarization error rate (DER) [63], consisting of miss (MI), false alarm (FA), and confusion (CF) errors, was adopted to evaluate all algorithms.

For back-end speech recognition, we simply used the standard recipe available in Kaldi.[3] We adopted a hybrid system, which includes a bidirectional LSTM (BLSTM) based acoustic model (AM) trained with aligned senones obtained from a set of tri-phone HMMs. The LSTM model has 3 hidden layers with 1024 units in each direction. In the decoding phase, a tri-gram language model (LM) was used. Following the Kaldi recipe, both the AM and the LM were trained on the Switchboard-1 Release 2 (LDC97S62) data set, which includes 260-hour of speech data with corresponding transcriptions. To ensure a fair comparison, we fixed the back-end ASR framework and employed different SSD methods in the front-end processing to generate transcripts. Therefore, the qualities of the corresponding transcripts can reflect the effectiveness of different SSD methods. For the ASR evaluation metric, we adopted the concatenated minimum-permutation WER (cpWER), which was used in CHiME-6 Challenge [64].

TABLE I
DER (%) Comparisons on the CTS Domain in DIHARD-III Challenge

Sys. / N_I	ISSD	ISSD (DNN-VAD)	Oracle ISSD	QDM-SSD	QDM-SSD + SL
Prior	16.22	16.22	-	16.22	16.22
1	11.51	10.05	9.61	11.51	11.51
2	10.35	9.51	8.79	9.02	8.70
3	10.03	9.36	8.67	8.73	8.45
4	**9.86**	9.23	8.60	8.67	8.24
5	9.99	**9.19**	**8.47**	**8.47**	**8.03**

N_I denotes the number of iterations.

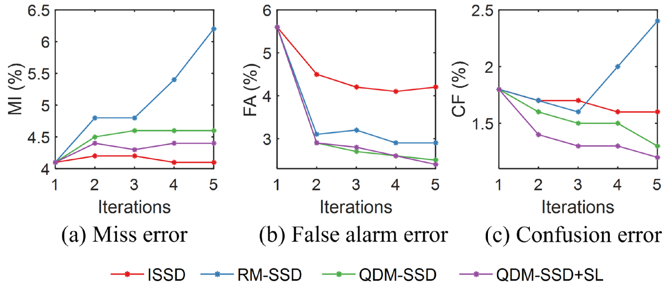
A. Diarization Performance and Result Analyses on CTS

We first used CTS to evaluate speaker diarization as in [14], [15], [24], [25], [65]. To further validate the generalization ability later, we also tested on two-speaker non-conversation telephone speech (NCTS) in the MT and SLR domains.

Table I presents an overall DER (%) comparisons among different SSD methods for five iterations on the DIHARD-III CTS corpus. 'ISSD (DNN-VAD)' means employing a more conservative DNN-VAD for ISSD system. 'Oracle ISSD' means employing the ground-truth labels as the speaker priors to start the iteration. 'SL' denotes the start-point localization techniques illustrated in Section III-B. To show the effectiveness of the SL method, we also present the results of QDM-SSD without the SL method which uses random starting points for the active parts in the dynamic masks. The QDM-SSD without and with SL method are denoted as 'QDM-SSD' and 'QDM-SSD + SL' in our experiments, respectively. As can be observed, The ISSD results converge in five iterations, achieving up to 39.12% relative DER reduction compared with the baseline. Then, with a more conservative DNN-VAD to reduce the FAs of ISSD, the smaller DERs can be obtained. Next, the oracle speaker priors can significantly improve the ISSD performance as shown in the third column, achieving up to a 47.78% relative DER reduction compared with baseline. This can be regarded as the upper bound of the single-system ISSD performance. For the QDM-SSD framework with random starting points, we achieve better DER compared with the original ISSD and DNN-VAD-based ISSD, which is very close to the upper bound of the oracle ISSD performance. Finally, with SL, QDM-SSD can attain about an 18.56% relative DER reduction when compared with original ISSD, and

[2] [Online]. Available: https://github.com/wiseman/py-webrtcvad

[3] [Online]. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/swbd/s5c



(a) Miss error (b) False alarm error (c) Confusion error

Fig. 3. Detailed error (%) comparisons in 1–5 iterations.

outperform oracle ISSD as shown in the bottom row in Table I. To clearly analyze how QDM improves SSD, we compare detailed errors among different SSD techniques, starting with the same speaker priors, as shown in Fig. 3. To illustrate the validity of the QDM, we also show SSD results with random-length masks (random length in $(0, L]$), called RM-SSD. In Fig. 3(a), we can see that ISSD has the smallest miss errors. This is because in ISSD model adaptation, all simulated data are fully overlapping, which biases towards no MIs. Although QDM-SSD produces more MIs compared with ISSD due to data sparsification, it can still maintain a good overlap-detection performance as shown in the green curve of Fig. 3(a). The purple curve shows that the proposed SL techniques can slightly improve the overlap-detection performance. However, RM-SSD gets more MIs as the iteration number increases due to the uncontrolled segment qualities in adaptation data generation. Fig. 3(b) compares the false alarm errors. It can be seen that ISSD gets the most FAs when compared with all the dynamic mask based SSD methods due to the task mismatch problem. The FAs of QDM-SSD with or without the SL methods are still smaller than those of RM-SSD. The confusion error is a quite important indicator which can directly reflect the purity of the adaptation data. Fig. 3(c) shows a CF comparison among different systems. As can be seen, apart from the RM-SSD which cannot control the qualities of speech segments, all others can maintain the error decreasing tendency in the iterative process. Furthermore, compared with the ISSD, the proposed QDM-SSD can further purify the adaptation data, leading to smaller CFs. This can also reflect the validity of data update in joint optimization. The SL method can further improve the purity of the adaptation data in QDM-SSD, achieving the smallest CFs among all systems. Specially, from Fig. 3 we can see that, although making the adaptation data sparse can help the SSD methods reduce the FAs, simply setting the overlap ratio randomly (as most diarization data simulation algorithms do [24], [26]) when adapting the model to realistic recordings will not bring much improvement to SSDs. It even degrades their performances (as shown in ‘RM-SSD’) due to the uncontrolled segment qualities in the mixing processes. The proposed QDM-SSD techniques can effectively alleviate this problem and ensure the quality of the simulated data when making them sparse, as illustrated in Fig. 3.

To visually observe the effectiveness of the proposed approaches, we use different SSD methods to process a ten-second-long speech mixture example in the DIHARD-III CTS corpus and compare the corresponding separation results in Fig. 4. The spectrogram of the mixture and the oracle segment boundaries for each speaker (Stream 1 in red and Stream 2 in blue) are shown in the top two rows. Spectrograms and their corresponding detection results of the two separated streams obtained by each method are shown in the four rows in the left and right columns, respectively. The gray bars indicate the false alarm regions. As can be seen, simple SSD can cause considerable over-detection of overlapping speech in both separated streams, leading to the large false alarm errors (DER = 40.66%, FA = 37.9%) due to the mismatch between the data used in the training and testing phases. The ISSD [49] can directly utilize the speech segments from realistic recordings to adapt the model, alleviating the mismatch to some extent (DER = 21.99%, FA = 18.4%). We can also see that the Stream 1 of ISSD results almost exactly matches the corresponding ground-truth label. In the Stream 2 of ISSD results, although the energy of residual speech from the unrelated speakers is significantly weakened when compared with SSD (as shown in the blue rectangles), there are still some over-detection errors of speaker presence. Compared with SSD and ISSD, QDM-SSD can effectively suppress residual speech from unrelated speakers (DER = 9.84%, FA = 3.7%) as can be seen in the corresponding separated streams in Fig. 4. The clear improvement of QDM-SSD over ISSD can be seen from the Stream 2, in which most FA errors have been eliminated, resulting from data sparsification through QDM. Moreover, for the Stream 1, QDM-SSD+SL can improve over QDM-SSD by further suppressing residual speech (DER = 8.03%, FA = 1.6%) as shown in red rectangles. Similarly, QDM-SSD+SL generates the best separated results in Stream 2. In addition, benefiting from data purification, the confusion errors in QDM-SSD and QDM-SSD+SL (CF = 0.2% in both) are also smaller than those of SSD (CF = 0.6%) and ISSD (CF = 0.7%).

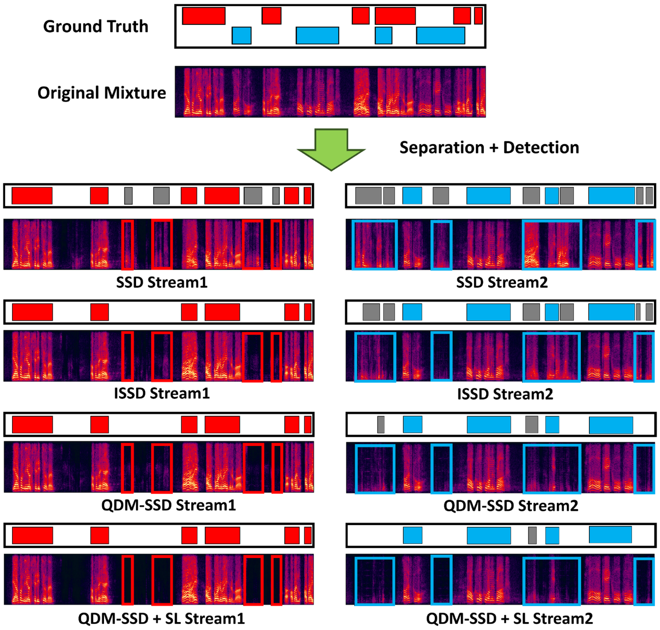
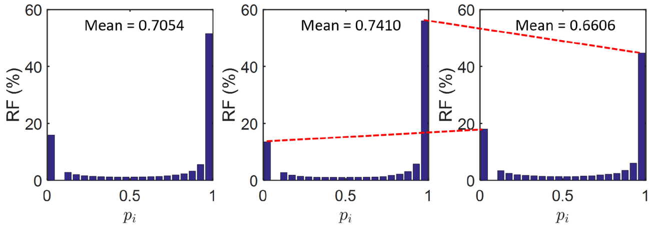


Fig. 4. The spectrograms and the corresponding detection results of the separated streams from different SSD methods. The regions which are falsely separated in the simple SSD framework are marked with red and blue rectangles.



(a) Overall results (b) Ten-best results (c) Ten-worst results

Fig. 5. Relative frequency (RF) histograms of p_i for different speaker priors from CSD baseline: (a) p_i from all priors, (b) p_i from the ten-best priors, and (c) p_i from the ten-worst priors.

TABLE II
DETAILED MI, FA, CF AND DERS (%) OF DIFFERENT SYSTEMS ON THE CTS DOMAIN FROM DIHARD-III CHALLENGE

Systems	MI	FA	CF	DER
VBx [50]	12.0	0.0	4.2	16.22
EEND [66]	-	-	-	9.29
TS-VAD	5.8	0.9	2.5	9.18
ISSD	4.1	4.1	1.6	9.86
QDM-SSD	4.6	2.5	1.3	8.47
QDM-SSD + SL	4.4	2.4	1.2	8.03
ITS-VAD (TS-VAD priors)	5.6	1.4	2.0	9.00
ITS-VAD (ISSD priors) [50]	4.6	2.0	1.2	7.76
ITS-VAD (QDM-SSD + SL priors)	4.3	1.7	1.2	**7.27**

As mentioned earlier, the active length of QDM (i.e., l_i in (16) or p_i in (15)) is related to the corresponding segment quality and the quality of speaker priors. To validate this, we show in Fig. 5 the distributions of p_i for speaker priors with different quality. In our experiments, we use VBx to obtain the speaker priors. We first present the distribution of p_i from all CSD priors in Fig. 5(a). Then we choose the ten-best and ten-worst CSD results as speaker priors and show their corresponding distributions in Fig. 5(b) and 5(c), respectively. We can see from Fig. 5(a) that more than half of the speech segments are entirely used to simulate adaptation data (i.e., $p_i = 1$) statistically, or we generate more than 25% of fully overlapping mixtures in the simulated data, which ensures the ability of detecting overlapping speech with the adapted model. At the same time, we also generate sufficient sparse speech segments whose lengths are approximately uniform-distributed in $[0.1\,L, L)$ supported by the quality of the corresponding speech segments. This improves the performance of the adapted model on speech regions with only one active speaker. Moreover, from Fig. 5(b) and 5(c), we can see that the distribution of p_i is relevant to the quality of speaker priors. For better speaker priors which can generate good-quality speech segments, the probability of $p_i = 1$ increases and the probability of $p_i = 0$ decreases when compared with the worse speaker priors, as marked by two red dotted lines. The overall mean value of p_i also shows the same trend, which equals 0.7410 for the ten-best and 0.6606 for the ten-worst CSD priors. Finally, as shown in Table II, we compare the proposed QDM-SSD with the advanced end-to-end speaker diarization systems which have attracted widespread attention recently. We directly list reported results of VBx [50] and EEND systems [66] in DIHARD-III Challenge, and we also show the results of TS-VAD in our DIHARD-III [50] system. We then list results obtained with our SSD-based systems. Note that the VBx and EEND methods do not know the number of speakers during inference, which makes it more difficult for these two methods to achieve good results compared with SSD methods. Moreover, these methods may use different training sets, so our goal is not to strictly compare the DERs, but to see if the proposed QDM-SSD can achieve similar performance to other advanced end-to-end systems. In Table II, we divide all results into three blocks. The top three-row block includes the existing results. The middle three-row block contains the results of different SSD systems. ITS-VAD in the bottom three-row block denotes the iterative TS-VAD algorithm proposed in our DIHARD-III system [50], which uses the iteration mechanism similar to ISSD and also requires prior information. From Table II, we can see that QDM-SSD with or without SL method is on par with EEND or TS-VAD systems. Comparing the TS-VAD and ISSD results, we can observe that ISSD yields more FA errors due to the task mismatch between speech separation and speaker diarization as discussed earlier. QDM can help ISSD alleviate this issue and suppresses residual speech from unrelated speakers, though there are still more false alarm errors in QDM-SSD (or + SL) than in the end-to-end systems. Nonetheless, the MIs and CFs in different SSD systems tend to be smaller than those in TS-VAD. The above analysis shows that there is a strong complementarity between SSD and end-to-end techniques. Therefore, we take the results of different SSDs as the priors for ITS-VAD to obtain better results, as we have done in DIHARD-III Challenge [50]. As we can observe in the bottom block, for ITS-VAD, using ISSD results as priors can obtain better results than using the results of pre-trained TS-VAD as priors, although TS-VAD outperforms ISSD. Moreover, using the results of the proposed 'QDM-SSD + SL' as priors can help ITS-VAD achieve further improvement (DER = 7.27%), yielding a better performance than that of our best-performing system using ISSD results as priors (DER = 7.76%) on the CTS domain from DIHARD-III Challenge.

B. Diarization Performance on NCTS

To verify the generalization ability of the proposed QDM-SSD, we also perform the evaluation on the two-speaker non-conversation telephone speech (NCTS) domains. In particular, the results in Table I show that the DERs of different SSD methods mostly converge after the third iteration. Therefore, we only run three iterations in subsequent experiments. The detailed DERs among different systems on the map task (MT) and sociolinguistic lab recordings (SLR) domains are listed in Table III. For the MT domain, we can observe that our earlier ISSD method (DER = 6.17%) underperforms the baseline system (DER = 4.97%), indicating that the ISSD doesn't work in this domain. A reason is that in a task with a low overlap ratio such as MT (overlap ratio is about 3%), although simple ISSD can reduce MIs by processing overlapping speech, it will increase FAs and degrade the DER performance. In essence, this is still caused by

TABLE III
DETAILED MI, FA, CF AND DER (%) COMPARISONS AMONG DIFFERENT SYSTEMS ON THE MAP TASK (MT) AND SOCIOLINGUISTIC LAB RECORDINGS (SLR) DOMAINS FROM DIHARD-III CHALLENGE

Map Task (MT)					
Systems	N_I	MI	FA	CF	DER
VBx [50]	-	2.9	0.0	2.0	4.97
ISSD	1	0.6	4.3	1.3	6.17
QDM-SSD	2	0.8	1.8	0.5	3.05
	3	0.7	1.7	0.6	**2.95**
QDM-SSD + SL	2	0.7	1.7	0.5	**2.88**
	3	0.7	1.7	0.5	2.90
Sociolinguistic Lab Recordings (SLR)					
Systems	N_I	MI	FA	CF	DER
VBx [50]	-	4.8	0.0	3.0	7.79
ISSD	1	1.7	10.4	1.4	13.49
QDM-SSD	2	1.8	5.9	1.4	9.11
	3	1.8	3.9	1.4	**7.22**
QDM-SSD + SL	2	1.8	4.0	1.3	7.13
	3	1.9	2.9	1.3	**6.05**

N_I denotes the number of iterations. We use the speech enhancement method in [67] to produce enhanced speech for all systems on the SLR domain.

the task mismatch between separation and diarization. Since the results of the first ISSD iteration are worse than the initial CSD priors, we only perform one iteration. The proposed QDM-SSD can effectively alleviate this problem, achieving a 40.6% relative DER reduction when compared with the baseline on the MT domain. The SL method can still bring a slight improvement, yielding a 42.1% relative DER reduction when compared with the baseline.

From Table III, we also observe a similar effect in the SLR domain, which is more challenging due to interfering noises captured in the interviews. To address this issue, we adopt speech enhancement method [67] in our DIHARD-III system to produce enhanced speech for all systems here. However, ISSD still produces much worse results than the baseline due to a considerable amount of residual noises. In this case, it is more important to control the quality of speech segments used for adaptation data simulation. Therefore, we can see that SL method can bring significant improvements to QDM-SSD, achieving 22.3% and 16.2% relative DER reductions compared with the baseline and the original QDM-SSD, respectively.

TABLE IV
WERS (%) COMPARISON AMONG DIFFERENT FROND-END PROCESSES ON THE HUB5 ENGLISH EVALUATION SPEECH, INCLUDING SWITCHBOARD (SWB) AND CALLHOME (CH) AMERICAN ENGLISH SPEECH SUBSETS

WER (%) for ASR on HUB5 English evaluation speech						
Input	N_C	Segmentation		SWB	CH	Overall
		VAD	Oracle			
Mixture	1	✓		31.8	40.0	35.9
Mixture	1		✓	26.3	28.8	27.9
ISSD	2	✓		19.4	26.9	23.2
QDM-SSD	2	✓		17.8	26.0	22.0
QDM-SSD + SL	2	✓		16.0	25.8	**20.9**
Oracle Channel	2	✓		14.9	23.7	19.3
Oracle Channel	2		✓	11.7	21.3	**16.5**

N_C denotes the number of channels.

C. ASR Performance on SWB-HUB5 and CH-HUB5

One advantage of the SSD methods over CSD and end-to-end systems is that the separated streams thus obtained can be directly used for back-end processing, such as ASR. Therefore, we adopt the proposed SSD-based methods as front-end processing for speech recognition on HUB5 English evaluation speech. Table IV lists WER comparisons among different systems. The original HUB5 evaluation set includes two subsets, namely Switchboard (SWB) and CALLHOME (CH) American English Speech,[4] and each recording contains two channels corresponding to the two parties in the telephone conversations. As is typical [14], [16], [68], [69], they are merged into a single channel to evaluate the diarization performance, which is denoted as 'Mixture' in Table IV. We also employed the original two-channel recordings as inputs for ASR system to obtain an upper bound on ASR performance, namely the 'Oracle Channel' in Table IV. To verify the effectiveness of applying our methods in back-end ASR task, we use different SSD methods to process the single-channel mixtures and obtain the two separated streams as the inputs for ASR. The number of iterations is set to 3 for all different SSD methods. We still use WebRTC VAD to attain the speech segments (denoted as 'VAD'), which is consistent with the previous experiments. For the 'Mixture' and 'Oracle Channel,' we also present the ASR results obtained with the oracle boundary information for each speaker (denoted as 'Oracle' in the 'Segmentation' column), which is actually equivalent to giving the ground-truth labels of diarization. For a fair comparison, all systems use the same back-end framework to generate ASR transcription results. From Table IV, we can observe that even with the oracle boundary information, the ASR system that takes the single-channel mixtures as inputs still generates fairly poor results (WER = 27.9%). The reason is that the single-channel mixtures contain some overlapping speech regions, which significantly reduces the quality of the ASR results. However, overlapping segments are common in conversation speech recorded in realistic conditions [5], [64], [69]. To address this issue, we use different SSD methods to process the single-channel mixtures, which can handle overlapping speech, thus improving the ASR performance. As shown in Table IV, the results of back-end ASR are consistent with our previous results on front-end diarization, i.e., the performance of the proposed QDM-SSD is better than that of the original ISSD. Applying the SL method can help QDM-SSD achieve a further improvement. It is worth noting that under the same segmentation conditions

[4] Note that the recordings in CH of the HUB5 evaluation set are different from those in CALLHOME American English Speech (LDC97S42) where we used a subset to serve as our development set.

TABLE V
DETAILED DER (%) COMPARISONS AMONG QDM-SSD + SL METHOD WITH DIFFERENT PARAMETER VALUES ON THE DEVELOPMENT AND EVALUATION SETS

Development set			
α / DER	β / DER	$(\tau_1,\ \tau_2)$ / DER	$p_{\min}$ / DER
0.1 / 10.28	0.1 / 9.34	(5, 25) / 9.45	0.0 / 9.33
0.3 / 9.35	**0.3 / 9.04**	**(10, 30) / 9.04**	**0.1 / 9.04**
0.5 / 9.04	0.5 / 9.32	(15, 35) / 9.19	0.2 / 9.30
0.7 / 9.11	0.7 / 10.17	(20, 40) / 9.43	0.3 / 9.28
Evaluation set (CTS)			
α / DER	β / DER	$(\tau_1,\ \tau_2)$ / DER	$p_{\min}$ / DER
0.1 / 8.91	0.1 / 8.34	(5, 25) / 8.49	0.0 / 8.47
0.3 / 8.21	**0.3 / 8.03**	**(10, 30) / 8.03**	**0.1 / 8.03**
0.5 / 8.03	0.5 / 8.38	(15, 35) / 8.17	0.2 / 8.39
0.7 / 8.06	0.7 / 8.75	(20, 40) / 8.29	0.3 / 8.37

The evaluation set is from the conversational telephone speech (CTS) domain in the DIHARD-III corpus.

using WebRTC VAD, the ASR system using speech processed by 'QDM-SSD + SL' as inputs (WER = 20.9%) can achieve similar performance to that obtained by the ASR system using oracle channel speech as inputs (WER = 19.3%). Furthermore, we find that for the same recordings, using WebRTC VAD will remove more speech segments with low energy, which is harmful to the back-end ASR system. This is the reason for the performance degradation caused by using WebRTC VAD when compared with using oracle boundary information when adopting 'Oracle Channel' as inputs.

D. Robustness Analysis

Since the proposed method contains a lot of parameters, we show how the proposed method is sensitive/robust to those parameters (including α, β, τ_1, τ_2, and $p_{\min}$) in Table V. We present the results of QDM-SSD + SL with different parameter values on the development set and evaluation set (CTS). The bold fonts in this table indicate the parameter values we have chosen in the proposed method. From this table, we can observe that the optimal parameter values in the development set can still achieve the best performance in the evaluation set, and the deviations from these values can lead to performance degradation, which illustrates the robustness of the proposed method to the parameter values. Moreover, the values of α and β have a relatively large impact on the performance. The main reason is that these two parameters influence the proportion of simulated data using dynamic masks, which will affect the results to some extent.

V. Discussion

In this paper, we focus on realistic two-speaker scenarios because the robust blind separation techniques for realistic two-speaker (and multi-speaker) scenarios have not been well established. However, multi-speaker (more than two speakers) scenarios are crucial in the diarization task. Therefore, we discuss the scalability of the proposed methods under multi-speaker scenarios in this section. We think that the following methods are possible ways to extend the proposed methods from two-speaker scenarios to multi-speaker scenarios:

1) Applying the two-speaker separation model to generate the dynamic masks under multi-speaker scenarios. In general, three or more people rarely speak simultaneously. For example, the analysis in [70] shows that over 90% of the overlaps involve only two speakers for most meeting domains, even though the meetings involve more than two speakers. The separation method in [71] also assumes the maximum speaker number of overlaps to be two to handle the multi-speaker scenarios. In this case, the relatively robust two-speaker separation model can be employed to judge the quality of speech under multi-speaker conditions, thus generating dynamic masks.
2) Adjusting the separation model from 'one versus one' to 'one versus rest'. Similar to [44], [45], we can extract the speakers recursively to handle the multi-speaker scenarios. Specifically, we only extract one speaker at each iteration and then feed the rest speech to the next iteration to extract subsequent speakers. In this case, the multi-speaker scenarios can be handled. We can judge the quality of speech segments by the number of iterations when performing the recursive extraction on them, and then generating the dynamic masks accordingly.
3) Adjusting blind separation to speaker-dependent separation (possibly using speaker embedding information). At the same time, instead of only relying on speech separation, the more robust end-to-end method (e.g., EEND) can be used to generate speaker priors, leading to more stable speech separation performance. In this case, speech separation can produce reliable dynamic masks to clean the speaker-specific segments.

VI. Conclusion

In this paper, we propose a QDM-SSD framework which can utilize quality-aware dynamic masks to perform adaptive data purification and sparsification for separation-based speaker diarization. Experimental results demonstrate the effectiveness of the proposed QDM-SSD framework in both speaker diarization and speech recognition tasks. In the future, we intend to extend QDM-SSD to handle realistic issues, such as a variable number of speakers, adverse environments, and an effective utilization of poor-quality speech segments.

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. A. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. v/953–v/956.

[5] Ö. Çetin and E. Shriberg, “Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site,” in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2006, pp. 212–224.

[6] I. Medennikov et al., “The STC system for the CHiME-6 challenge,” in *Proc. CHiME Workshop Speech Process. Everyday Environ.*, 2020, pp. 36–41.

[7] 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.

[8] A. Nagrani et al., “VoxSRC 2020: The second voxceleb speaker recognition challenge,” 2020, *arXiv:2012.06867*.

[9] F. Yu et al., “M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6167–6171.

[10] N. Ryant et al., “First DIHARD challenge evaluation plan,” 2018. [Online]. Available: https://zenodo.org/record/1199638

[11] N. Ryant et al., “Second DIHARD challenge evaluation plan,” Linguistic Data Consortium, Tech. Rep, 2019.

[12] N. Ryant et al., “The third DIHARD diarization challenge,” in *Proc. INTERSPEECH*, 2021, pp. 3570–3574.

[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] 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.

[15] 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.

[16] D. Dimitriadis and P. Fousek, “Developing on-line speaker diarization system,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2739–2743.

[17] G. Sell et al., “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2808–2812.

[18] 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.

[19] 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.

[20] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4052–4056.

[21] 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.

[22] L. Bullock, H. Bredin, and L. P. Garcia-Perera, “Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7114–7118.

[23] 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.

[24] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, and K. Nagamatsu, “End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification,” 2020, *arXiv:2003.02966*.

[25] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 269–273.

[26] I. Medennikov et al., “Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 274–278.

[27] W. Wang, Q. Lin, D. Cai, L. Yang, and M. Li, “The DKU-Duke-Lenovo system description for the third DIHARD speech diarization challenge,” 2021, *arXiv:2102.03649*.

[28] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3555–3559.

[29] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, “Personal VAD: Speaker-conditioned voice activity detection,” in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 433–439.

[30] 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.

[31] 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.

[32] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[33] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, “A regression approach to single-channel speech separation via high-resolution deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1424–1437, Aug. 2016.

[34] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[35] P. Chandna, M. Miron, J. Janer, and E. Gómez, “Monoaural audio source separation using deep convolutional neural networks,” in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 258–266.

[36] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, “CBLDNN-based speaker-independent speech separation via generative adversarial training,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 711–715.

[37] S. Chen et al., “Don’t shoot butterfly with rifles: Multi-channel continuous speech separation with early exit transformer,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6139–6143.

[38] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[39] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2642–2646.

[40] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 21–25.

[41] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, “End-to-end speech separation with unfolded iterative phase reconstruction,” in *Proc. INTERSPEECH*, 2018, pp. 2708–2712.

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

[43] X. Xiao et al., “Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5824–5828.

[44] K. Kinoshita, M. Delcroix, S. Araki, and T. Nakatani, “Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 381–385.

[45] T. v. 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.

[46] Z. Chen et al., “Continuous speech separation: Dataset and analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7284–7288.

[47] S.-T. Niu, J. Du, L. Sun, and C.-H. Lee, “Separation guided speaker diarization in realistic mismatched conditions,” 2021, *arXiv:2107.02357*.

[48] X. Fang et al., “A deep analysis of speech separation guided diarization under realistic conditions,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 667–671.

[49] S.-T. Niu, J. Du, L. Sun, and C.-H. Lee, “Improving separation-based speaker diarization via iterative model refinement and speaker embedding based post-processing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8387–8391.

[50] Y. Wang et al., “USTC-NELSLIP system description for DIHARD-III challenge,” 2021, *arXiv:2103.10661*.

[51] Y.-X. Wang, J. Du, M. He, S.-T. Niu, L. Sun, and C.-H. Lee, “Scenario-dependent speaker diarization for DIHARD-III challenge,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3106–3110.

[52] Q. Wang and T. Breckon, “Unsupervised domain adaptation via structured prediction based selective pseudo-labeling,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6243–6250.

[53] X. Zhang, Y. Ge, Y. Qiao, and H. Li, “Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3436–3445.

[54] Y. Zhang, B. Deng, K. Jia, and L. Zhang, “Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 781–797.

[55] Y. Takashima, Y. Fujita, S. Horiguchi, S. Watanabe, P. García, and K. Nagamatsu, “Semi-supervised training with pseudo-labeling for end-to-end neural diarization,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3096–3100.

[56] 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.

[57] 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.

[58] J. Kola, C. Espy-Wilson, and T. Pruthi, “Voice activity detection,” Merit Bien, pp. 1–6, 2011.

[59] S. M. Strassel, “Linguistic resources for effective, affordable, reusable speech-to-text,” in *Proc. Lang. Resour. Eval. Conf.*, 2004, pp. 65–68.

[60] L. D. Consortium, “2000 Hub5 english evaluation speech LDC2002S09,” [Web Download]. Linguistic Data Consortium, Philadelphia, PA, USA, 2002.

[61] M. Pariente et al., “Asteroid: The pytorch-based audio source separation toolkit for researchers,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2637–2641.

[62] D. P. Kingma and J. Ba, “ADAM: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[63] “The (RT-09) rich transcription meeting recognition evaluation plan,” 2009. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf

[64] S. Watanabe et al., “CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” in *Proc. 6th Int. Workshop Speech Process. Everyday Environ.*, 2020, pp. 1–7.

[65] M. Diez, L. Burget, F. Landini, and J. Černocký, “Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 355–368, 2019.

[66] F. Landini et al., “BUT system description for the third DIHARD speech diarization challenge,” in *Proc. 3rd DIHARD Speech Diarization Challenge Workshop*, 2021.

[67] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, and C.-H. Lee, “Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7099–7103.

[68] 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.

[69] S. R. Chetupalli and S. Ganapathy, “Speaker conditioned acoustic modeling for multi-speaker conversational ASR,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 3834–3838.

[70] Ö. Çetin and E. Shriberg, “Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition,” in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 293–296.

[71] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, “Low-latency speaker-independent continuous speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6980–6984.

Shu-Tong Niu received the B.S. degree from the Department of Electronic Information Engineering, Dalian University of Technology, Dalian, China, in 2019. He is currently working toward the Ph.D. degree with the University of Science and Technology of China, Hefei, China. His main research interests include speech separation 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, 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 won several champions of CHiME-4/CHiME-5/CHiME-6 Challenge, SELD Task of 2020 DCASE Challenge, and DIHARD-III Challenge.

Lei Sun received the B.S. degree from the Northeastern University at Qinhuangdao, Qinhuangdao, China, and the D.E. degree from the University of Science and Technology of China, Hefei, China, in 2020. He is currently with iFlytek Research, Hefei, China. His research interests include speech enhancement, speaker diarization, and robust speech recognition.

Yu Hu received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2000, 2003, and 2009, respectively. In 1999, he became a Research Engineer with iFlytek, Ltd., Anhui, China, as the Co-founder, working on Mandarin speech synthesis and speech prosody analysis. He was the one of researchers who built first few generations of iFlytek Mandarin speech synthesis engines. Since 2004, his research interest has been changed to robust speech recognition, and began to work on the iFlytek Mandarin speech recognition system.

Chin-Hui Lee (Fellow, IEEE) is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience, ending with 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 30 patents, and has been cited more than 50 000 times for his original contributions 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, and SPS’s 2006 Technical Achievement Award 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 in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition. He is a Fellow of ISCA.