ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746354

IMPROVING SEPARATION-BASED SPEAKER DIARIZATION VIA ITERATIVE MODEL REFINEMENT AND SPEAKER EMBEDDING BASED POST-PROCESSING

Shu-Tong Niu[1], *Jun Du*[1,*], *Lei Sun*[2], *Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]iFlytek Research, Hefei, Anhui, P. R. China
[3]Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

In this paper, we propose an iterative separation-based speaker diarization (ISSD) approach to cope with the realistic data conditions. In the proposed ISSD, we iteratively generate adaptation data according to speaker priors and fine-tune the separation model, which leads to a gradual performance improvement. To further reduce some unavoidable speaker detection errors due to some undesirable prior errors using simple ISSD, we utilize speaker embedding information and propose two post-processing techniques, namely, speaker filtering and speaker recovery. We evaluate the diarization performance on the two-speaker conversational telephone speech (CTS) data set from DIHARD-III Challenge. When compared to state-of-the-art clustering-based speaker diarization (CSD) system, the proposed ISSD approach combined with the two post-processing schemes yields a 47.72 % and 46.97 % relative diarization error rate reduction on the development and evaluation sets, respectively. ISSD is also one key contributing factor to the best-performing system in DIHARD-III Challenge.

Index Terms— Speaker diarization, speech separation, iteration, post-processing, DIHARD-III Challenge

1. INTRODUCTION

Speaker diarization is a task to segment speech into speaker-specific regions [1]. It is an essential component for many applications, such as conference summarization, speech transcription and dominant speaker detection [2, 3]. It also serves as a front-end of automatic speech recognition (ASR) to generate speaker-attributed transcripts [4] for mixed speech.

Traditional clustering-based speaker diarization (CSD) systems [5, 6] can be partitioned into multiple modules, including voice activity detection (VAD), speaker feature extraction and speaker clustering. First, they use VAD to detect the speech segments. Then, they extract segment-based speaker embeddings, such as i-vector [7], d-vector [8] and x-vector [9]. Finally, speaker clustering is utilized to assign the speakers along with timing. In particular, [6] proposes the variational Bayesian hidden Markov model with x-vectors (VBx), which has achieved a state-of-the-art performance among clustering-based systems. However, they can not handle the speaker-overlap regions due to hard cluster assignments. Recently, end-to-end diarization techniques have attracted research attentions due to their potentials to deal with such mixed-speech regions [10, 11]. End-to-end neural speaker diarization (EEND) [10] treats the diarization task as a multi-label classification problem, which can directly minimize diarization errors. Inspired by personal-VAD [12], target-speaker voice activity detection (TS-VAD) [11] is then proposed and uses speaker embeddings as additional inputs to directly predict the activity of each speaker at each time frame.

As for speech separation, it tries to separate each source speaker from multi-speaker speech mixtures [13]. The main separation network models are either time-frequency (T-F) domain based [14, 15] or time-domain based [16, 17]. The former performs separation on the T-F units by applying a short time Fourier transform (STFT) to speech. Nonetheless, the imperfect reconstruction of the phase limits its performance. Recently, time-domain based networks show their potentials [16, 17] in separating speech mixtures directly on time domain through encoder-decoder architectures. However, these separation algorithms are mostly evaluated on simulated data (e.g., the wsj0-2mix data set). When faced with realistic conditions, the performance of these algorithms could often become unstable.

Our earlier separation-based speaker diarization (SSD) system obtains results by detecting speaker presence in separated streams [18]. There are some advantages in the previously proposed SSD systems in handling speaker overlap segments, and the separated streams thus obtained can be directly used for back-end processing, such as ASR. To handle the possible instability in separation performances, [18] proposes a separation guided speaker diarization (SGSD) approach utilizing a complementarity of speech separation and speaker clustering. Nonetheless, the selection strategies in SGSD often depend on the characteristics of the data set. In this paper, we utilize the semi-supervised training algorithms [19, 20] and propose a more general method, namely, iterative separation-based speaker diarization (ISSD). Moreover, we present some post-processing techniques to further improve the ISSD performance inspired by [21]. The key contributions of this paper are three-fold: (i) we alleviate the instability problem of the SSD performance by incorporating an iteration mechanism; (ii) we demonstrate the effectiveness of our proposed ISSD system by analyzing the spectrograms and the corresponding labels of the separated speaker-specific speech streams; and (iii) according to diarization errors in the ISSD system, we propose two post-processing techniques which can filter out the speech of irrelevant speakers and fill the missing segments, respectively. Our experimental results are gathered on the CTS data set from DIHARD-III Challenge, and the proposed ISSD system has achieved better performance than CSD and SGSD systems. Moreover, we also find the proposed post-processing methods contribute to further performance improvements.

2. PRIOR WORK

As shown in [18], our earlier separation-based speaker diarization (SSD) framework consists of two parts: separation and detection.

*corresponding author

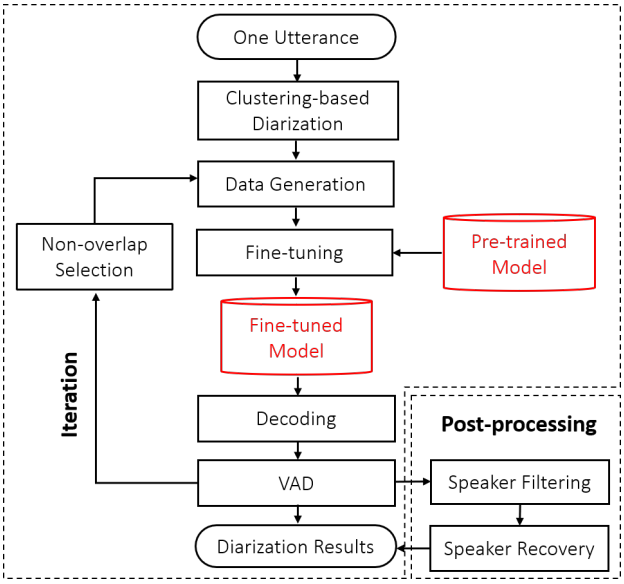


Fig. 1: Overall framework of the proposed approaches.

Given an utterance $y(t)$ with a two-speaker mixture, we first segregate the two voices of different speakers:

$$f_{\text{sep}}(y(t)) = \{\hat{x}_1(t), \hat{x}_2(t)\} \tag{1}$$

where $f_{\text{sep}}(\cdot)$ is the separation model, $\hat{x}_1(t)$ and $\hat{x}_2(t)$ are two separated streams. In [18], we employ Conv-TasNet [16] as the separation model $f_{\text{sep}}(\cdot)$ which is trained on dataset simulated from Librispeech [22]. We use a permutation invariant training (PIT) [23] based learning objective to update the model parameters:

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

where $\hat{x}_n(t)$ means n-th separated stream, $x_\phi(t)$ denotes reference speech with the permutation ϕ that minimizes L. l is calculated by the scale-invariant source-to-noise ratio (Si-SNR):

$$l(\hat{x}(t), x(t)) = 10\log_{10}\frac{||x_{\text{tg}}(t)||^2}{||\hat{x}(t) - x_{\text{tg}}(t)||^2} \tag{3}$$

where $x_{\text{tg}}(t) = \frac{<\hat{x}(t), x(t)>x(t)}{||x(t)||^2}$. $\hat{x}(t)$ and $x(t)$ are the estimates and targets respectively. In detection part, we use VAD model to detect the speaker presence in separated streams, and obtain the diarization results through combining the detection results along the time axis.

3. THE PROPOSED APPROACH

Fig. 1 illustrates the framework of the proposed methods, which mainly contains two parts: iterative separation-based speaker diarization (ISSD) and speaker embedding based post-processing. In the ISSD part, we alternately update the adaptation data and fine-tune the separation model. In post-processing part, the speaker filtering can filter out the speech of irrelevant speakers, and the speaker recovery can recover the speakers in the missing segments. We will elaborate on these methods in the following subsections.

3.1. Iterative separation-based speaker diarization (ISSD)

ISSD aims to adapt the separation model to realistic data by leveraging upon the information from prior diarization results. First, we use the clustering-based speaker diarization (CSD) results as priors to generate the adaptation data. Then, we fine-tune the pre-trained model to adapt it to the test utterances. Ultimately, we use the results generated by the adapted model to update adaptation data for the next iteration. The diarization results can be refined iteratively through this process. The main components of the proposed ISSD framework are summarized as follows.

Data update - The data updating part contains non-overlap selection and data generation processes in Fig. 1. We first obtain the diarization prior $\mathbf{y}_i \in \{0,1\}^{1\times T}$ which corresponds to speaker i in test utterance $y(t)$. Here, the elements of $\mathbf{y}_i$ are defined as:

$$y_{i,t} = \begin{cases} 0 & \text{Speaker } i \text{ is inactive at } t \\ 1 & \text{Speaker } i \text{ is active at } t \end{cases} \tag{4}$$

Then we eliminate the overlap segments in $\mathbf{y}_i$:

$$\mathbf{y}_i' = \mathbf{y}_i \cdot \mathbf{m} \tag{5}$$

where $\mathbf{m} \in \{0,1\}^{1\times T}$ is the time mask with elements $m_t = 1$ if only one speaker is present at time t in prior results, and $m_t = 0$ otherwise. We cut the test utterance $y(t)$ according to $\mathbf{y}_i'$ and obtain the set including all speech segments belonging to speaker i:

$$\boldsymbol{S}_i = \{s_{i,j} | j = 1, 2, \ldots, N\} \tag{6}$$

where $s_{i,j}$ is a speech segment of speaker i. N is the number of segments, which varies among different speakers. We simulate paired speech mixtures by mixing two segments randomly selected from $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ respectively.

Model adaptation - We use the trained separation model in [18] as the pre-trained model. To adapt it to test utterances, we fine-tune the pre-trained model utilizing the simulated adaptation data as in Fig. 1 via Eqs. (2) and (3) as in SSD [18].

Result generation - As shown in the bottom middle part of Fig. 1, we use the adapted model to separate test utterance and obtain two separated streams $\hat{x}_1(t)$ and $\hat{x}_2(t)$, and then use VAD to get a variable number of M segments belonging to the n-th stream as:

$$\hat{\boldsymbol{S}}_n = \{\hat{s}_{n,j} | j = 1, 2, \ldots, M\}. \tag{7}$$

By combining the time labels of the VAD segments in $\hat{\boldsymbol{S}}_1$ and $\hat{\boldsymbol{S}}_2$, we can get the corresponding diarization results, which will be used to update the adaptation data in the next iteration. In this study, we focus on the two-speaker case. It's also worth noting that there are no overlap segments in the CSD results, so non-overlap selection in the left module in Fig. 1 is not performed in the first iteration.

3.2. Post-processing

Speaker filtering - Although the iterative process can significantly improve the separation performance in ISSD system, there might be still some residual errors in separation results. Fig. 2 shows an example of residual errors, in which the speech segments of one speaker are assigned to the two streams. To handle this problem, we propose a speaker filtering method. In speaker filtering, we use the mean value of the embeddings extracted from the segments in $\boldsymbol{S}_i$ as the reference embedding of speaker i:

$$\boldsymbol{e}_i = \frac{1}{N}\sum_{j=1}^{N} f_{\text{emb}}(s_{i,j}) \tag{8}$$

where $f_{\text{emb}}(\cdot)$ denotes the speaker embedding extraction model. Then we use the cosine similarity between $\boldsymbol{e}_i$ and the speaker embedding $f_{\text{emb}}(\hat{s}_{n,j})$ as the score of segment $\hat{s}_{n,j}$ on speaker i:

$$\text{score}_{n,j,i} = \cos(\boldsymbol{e}_i, f_{\text{emb}}(\hat{s}_{n,j})) \tag{9}$$

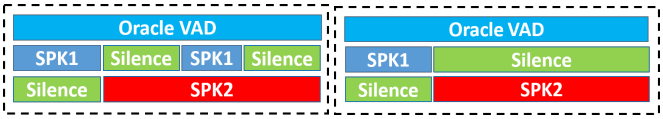
where $\hat{s}_{n,j} \in \hat{\boldsymbol{S}}_n$ is a VAD segment in stream n, which is defined in Eq. (7). Then we get the corresponding speaker of the segment $\hat{s}_{n,j}$ by finding the index i which maximizes $\text{score}_{n,j,i}$, namely:

$$c_{n,j} = \arg\max_i(\text{score}_{n,j,i}) \tag{10}$$

Through Eq. (10), we can get the set whose elements denote the speaker indices corresponding to all segments in $\hat{\boldsymbol{S}}_n$, namely $\boldsymbol{C}_n = \{c_{n,j}|j = 1, 2, \ldots, M\}$. We use the most elements of $\boldsymbol{C}_n$ as the corresponding speaker of stream n:

$$b_n = \arg\max_i |\{i \in \boldsymbol{C}_n\}| \tag{11}$$

where $|\cdot|$ means finding the cardinality of a set. To filter out residual speech in stream n, we calculate $\cos(\boldsymbol{e}_{b_n}, f_{\text{emb}}(\hat{s}_{n,j}))$, the cosine similarity between embedding $f_{\text{emb}}(\hat{s}_{n,j})$ and the reference embedding of speaker b_n, and then compare it with a threshold in order to exclude segment $\hat{s}_{n,j}$ with less cosine similarity in the n-th stream. Through this process, we can filter out the speech regions that are most likely to belong to irrelevant speakers.



(a) *ISSD result* (b) *Oracle result*

Fig. 2: An example of residual errors in ISSD system.

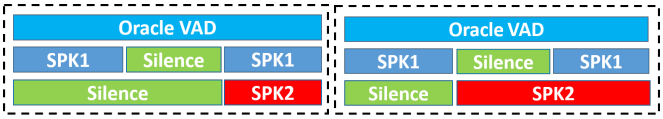
Speaker recovery - Another problem in the ISSD system is illustrated in Fig. 3. When we use the oracle VAD information in the diarization task, some segments that contain speech in oracle VAD are labeled as non-speech in the ISSD results. To handle these missing segments, we propose a recovery mechanism. For each utterance, we assume the set of missing segments is defined as:

$$\boldsymbol{M} = \{m_j|j = 1, 2, \ldots, K\} \tag{12}$$

where K is the number of missing segments in the current utterance. Since the duration of missing segments is usually very short, we assume that there is only one speaker in m_j, and we get the index of this speaker by maximizing the cosine similarity between the embedding of m_j and reference embedding like Eqs. (9) and (10):

$$d_j = \arg\max_i(\cos(\boldsymbol{e}_i, f_{\text{emb}}(m_j))). \tag{13}$$

We add speaker d_j to the time corresponding to segment m_j in the ISSD results. Through this process, we can label the missing segments with relatively accurate speaker identities.



(a) *ISSD result* (b) *Oracle result*

Fig. 3: An example of missing errors in ISSD system.

4. EXPERIMENTS

The Librispeech corpus [22] was adopted to simulate the training set of the pre-trained separation model. We simulated about 250-hour training data by randomly mixing two utterances from different speakers. The Voxceleb1 & 2 data sets [24] were used to train the speaker embedding extraction model. The CTS dataset from DIHARD-III Challenge [25] was adopted to evaluate the proposed methods, which contains development set and evaluation set. There are 61 ten-minute utterances in each set, and each utterance contains two speakers. The overlap ratio is quite large in both sets (11.9% in the development set and 10.5% in the evaluation set), which can reflect the model's ability to handle the overlap segments.

We used the VBx system [6] as the baseline and adopted the VBx results to start the first ISSD iteration. We employed the Conv-TasNet [16] as the separation model. In pre-training, we ran 75 epochs on 3-second long segments. In the iterative phase, we generated about 4-hour adaptation data according to the speaker priors for each utterance. We fine-tuned the pre-trained model for only 3 epochs to prevent overfitting. The WebRTC VAD[1] was used to obtain the VAD segments of separated streams. To further improve the diarization performance, we employed DOVER-Lap (DL) [26] to combine different systems. In post-processing, the well-known ResNet-34 architecture [27] was used as speaker embedding extractor. To alleviate the impact of segment duration on embedding extraction, we split the segments into multiple one-second sub-segments and applied post-processing on these data. The threshold in speaker filtering was tuned on development set and fixed to be 0.3. We evaluated the proposed algorithms with diarization error rate (DER) [28] which consists of miss (MI), false alarm (FA), and confusion (CF) errors. The tolerance collar was set to zero and the oracle VAD boundary information was used in all experiments.

4.1. Evaluation of the ISSD Framework

Table 1 lists a detailed DER performance comparison among different systems. N_I and N_E denote the number of iterations and the number of epochs, respectively. We applied the Dover-Lap (marked "DL" in Table 1) to combine systems of the baseline and three epochs in the first and second ISSD iterations. There are four result blocks in Table 1. The SSD and SGSD systems in the second and third blocks adopt the ISSD pre-trained model as the separation model, which is the same as in [18]. Several observations could be made here. First, when compared with CSD, SSD gets more false alarms (FAs) and confusion errors (CFs). This is caused by an unstable separation performance of the pre-trained model, which is our motivation for adopting the iterative strategy in ISSD. It also indicates that CSD is a good choice for providing priors in the first ISSD iteration, which is sufficiently stable and accurate for generating adaptation data. Second, ISSD achieves the best performance among the four systems. On the one hand, when compared with the CSD and SGSD systems, ISSD can handle more overlap segments, observing smaller MIs. On the other hand, when compared with the SSD system, ISSD conducts a more robust separation, which can be seen in smaller FAs and CFs. Moreover, we add the CSD system as a voter in the fusion algorithm to mitigate the over-detection of overlap regions (i.e., the false alarms) in the ISSD system, which is proven effective in the results of 'DL'. Finally, by comparing the CSD results and the DL results in the first ISSD iteration, we can see that the latter have smaller MIs and CFs (e.g., 10.5% and 3.7% versus 4.1% and 1.8% for "Eval"). Smaller MIs can help discard more

[1]https://github.com/wiseman/py-webrtcvad

Table 1: Detailed DER (%) comparison among different systems on the CTS development and evaluation sets from DIHARD-III Challenge. DER consists of Miss (MI), False alarm (FA), and Confusion errors (CF). DL indicates Dover-Lap. N_I and N_E denote the number of iterations and the number of epochs, respectively.

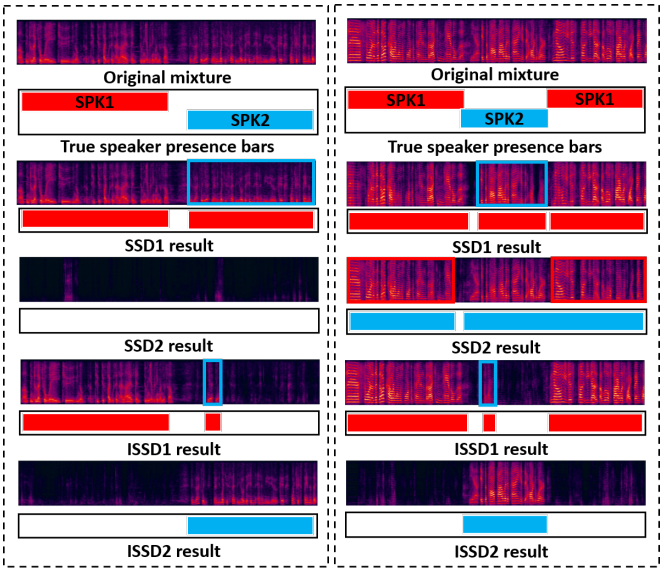
System	N_I	N_E	Dev				Eval			
			MI	FA	CF	DER	MI	FA	CF	DER
CSD	\	\	12.0	0.0	4.2	16.22	10.5	0.0	3.7	14.20
SSD	\	\	3.6	11.0	6.2	20.84	3.3	8.7	6.9	18.93
SGSD	\	\	7.6	2.6	2.7	12.95	6.4	2.6	2.2	11.24
ISSD	1	1	4.1	5.6	1.8	11.51	3.6	6.0	1.5	10.97
		2	4.3	5.0	2.0	11.36	3.5	4.6	1.5	9.67
		3	4.4	4.9	2.1	11.31	3.6	4.5	1.7	9.80
		DL	4.9	2.6	2.1	**9.56**	4.1	2.6	1.8	**8.48**
	2	1	4.2	4.2	1.6	9.99	3.5	4.2	1.4	9.09
		2	4.2	3.9	1.6	9.69	3.5	3.7	1.4	8.60
		3	4.2	3.8	1.5	9.61	3.5	3.6	1.4	8.54
		DL	4.6	2.6	1.6	**8.83**	3.8	2.5	1.4	**7.77**

overlap segments in the non-overlap selection, and smaller CFs benefit in attaining more accurate separation targets. Hence, compared with the first ISSD iteration, we can generate purer adaptation data in the second ISSD iteration, which leads to better performance as shown in Table 1. Note that the proposed ISSD approach is also one key contributing factor to our top-performing system submitted to DIHARD-III Challenge [29].

To better illustrate the effectiveness of our proposed ISSD system, we selected two speech segments that were falsely separated in the SSD system (i.e., two failure cases in [18]). We present the corresponding ISSD results in Fig. 4. In the first case, as shown in Fig. 4 (a), the SSD system assigns speech segments from different speakers to one data stream, which results in many confusion errors (DER = 42.5%, CF = 36.1%). At the bottom of Fig. 4 (a), we can see that the ISSD system can better distinguish between these two speakers, and generates a more accurate result (DER = 4.62%). In Fig. 4 (b), the speech segments of one speaker are separated into two streams in SSD, resulting in many false alarm errors (DER = 82.4%, FA = 78.1%). The ISSD system can assign these segments to a single speaker more accurately and generate much better results (DER = 9.62%), as shown in the bottom of Fig. 4 (b).

4.2. Evaluation of the Proposed Post-processing Techniques

Table 2 lists the results after post-processing with speaker filtering (or 'SF') and speaker recovery (or 'SR'). We only show the results of the third epoch ($N_I = 2$, $N_E = 3$) due to a space limitation, when applying post-processing to the results after the second ISSD iteration. Dover-Lap is marked 'DL', which is used to combine the three processed epochs and the baseline system. From this table, we can observe that speaker filtering can help the ISSD system achieve better performance. The improvement mainly comes from the reduction of FAs, which means speaker filtering can filter out speech from irrelevant speakers and reduce the false alarm errors. Moreover, it is shown that speaker recovery can further improve the ISSD performance after the speaker filtering process. The performance improvement is due to the reduction of CFs rather than the reduction of MIs. One reason is that we simply assign the neighborhood speaker to missing segments in the original ISSD system, and the speaker recovery process fills in a more accurate speaker in these missing segments. Finally, we performed a DL fusion between the ISSD results processed by SF and SR and the results of the baseline system, yielding the lowest overall DERs (8.48% for "Dev" and 7.53% in "Eval") as shown in the bottom row in Table 2. Moreover, our SF and SR approaches are also suitable for the multi-speaker (more than two speakers) scenarios, not just for the two-speaker case.



(a) *Selected segment 1* (b) *Selected segment 2*

Fig. 4: The spectrograms and diarization labels comparison between SSD system and ISSD system in two selected speech segments. The regions which were falsely separated are marked with rectangles.

Table 2: Detailed DERs (%) of post-processing methods on the CTS domain of development set and evaluation set from DIHARD-III Challenge. N_I and N_E denote the number of iterations and the number of epochs, respectively. DL indicates Dover-Lap. SF and SR indicate speaker filtering and speaker recovery, respectively.

Method	Dev				Eval			
	MI	FA	CF	DER	MI	FA	CF	DER
ISSD ($N_I = 2$, $N_E = 3$)	4.2	3.8	1.5	9.61	3.5	3.6	1.4	8.54
+ SF	4.4	3.4	1.6	9.40	3.7	3.4	1.4	8.41
+ SR	4.3	3.4	1.3	9.09	3.7	3.4	1.2	8.25
+ DL	4.9	2.2	1.4	**8.48**	4.1	2.2	1.3	**7.53**

5. CONCLUSION

In this paper, we propose the ISSD framework to handle the instability of separation performances in the original SSD systems. Moreover, we propose two post-processing techniques with speaker filtering and speaker recovery to reduce the impacts of residual speech and missing segments in ISSD. Experimental results on the CTS data of DIHARD-III Challenge demonstrate that our proposed ISSD approach can achieve better diarization performances when compared with the CSD and SSD systems. When combined with the two proposed post-processing algorithms, further improvements can also be observed. In the future, we will explore ISSD in multi-speaker scenarios where more than two speakers are mixed.

6. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62171427, and Ximalaya Inc.

7. REFERENCES

[1] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *arXiv preprint arXiv:2101.09624*, 2021.

[2] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. on ASLP*, vol. 17, no. 7, pp. 1382–1393, 2009.

[3] Douglas A Reynolds and P Torres-Carrasquillo, "Approaches and applications of audio diarization," in *ICASSP*, 2005, vol. 5, pp. v–953.

[4] Naoyuki Kanda, Yusuke Fujita, Shota Horiguchi, Rintaro Ikeshita, Kenji Nagamatsu, and Shinji Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP*, 2019, pp. 6630–6634.

[5] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. on ASLP*, vol. 21, no. 10, pp. 2015–2028, 2013.

[6] Mireia Diez, Lukáš Burget, Federico Landini, and Jan Černockỳ, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," *IEEE Trans. on ASLP*, vol. 28, pp. 355–368, 2019.

[7] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on ASLP*, vol. 19, no. 4, pp. 788–798, 2010.

[8] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP*, 2014, pp. 4052–4056.

[9] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[10] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[11] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, et al., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," *arXiv preprint arXiv:2005.07272*, 2020.

[12] Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, and Ignacio Lopez Moreno, "Personal vad: Speaker-conditioned voice activity detection," *arXiv preprint arXiv:1908.04284*, 2019.

[13] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. on ASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[14] Jun Du, Yanhui Tu, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE Trans. on ASLP*, vol. 24, no. 8, pp. 1424–1437, 2016.

[15] Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, and Bo Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *ICASSP*, 2018, pp. 711–715.

[16] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. on ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[18] Shu-Tong Niu, Jun Du, Lei Sun, and Chin-Hui Lee, "Separation guided speaker diarization in realistic mismatched conditions," *arXiv preprint arXiv:2107.02357*, 2021.

[19] Douglas A Reynolds and P Torres-Carrasquillo, "Approaches and applications of audio diarization," in *ICASSP*, 2005, vol. 5, pp. v–953.

[20] Jacob Kahn, Ann Lee, and Awni Hannun, "Self-training for end-to-end speech recognition," in *ICASSP*, 2020, pp. 7084–7088.

[21] Xiong Xiao, Naoyuki Kanda, Zhuo Chen, Tianyan Zhou, Takuya Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Yu Wu, Jian Wu, et al., "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," in *ICASSP*, 2021, pp. 5824–5828.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[23] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[24] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[25] Neville Ryant, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[26] Desh Raj, Leibny Paola Garcia-Perera, Zili Huang, Shinji Watanabe, Daniel Povey, Andreas Stolcke, and Sanjeev Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *SLT*, 2021, pp. 881–888.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[28] "The 2009 (rt-09) rich transcription meeting recognition evaluation plan," *http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf*, 2009.

[29] Yu-Xuan Wang, Jun Du, Maokui He, Shu-Tong Niu, Lei Sun, and Chin-Hui Lee, "Scenario-dependent speaker diarization for dihard-iii challenge," *Proc. Interspeech 2021*, pp. 3106–3110, 2021.