INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Unsupervised Adaptation with Quality-Aware Masking to Improve Target-Speaker Voice Activity Detection for Speaker Diarization

Shutong Niu[1], *Jun Du*[1,*], *Maokui He*[1], *Chin-Hui Lee*[2], *Baoxiang Li*[3], *Jiakui Li*[3]

[1]University of Science and Technology of China, Hefei, China
[2]Georgia Institute of Technology, Atlanta, USA
[3]SenseTime Research

niust@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose an unsupervised adaptation approach to improve target-speaker voice activity detection (TS-VAD) in speaker diarization (SD) based on quality-aware masking (QM) in order to reduce potential errors in the generated pseudo-labels. Furthermore, the QM-TS-VAD adapted model can be used as a teacher model to fine-tune a student SD model through knowledge distillation (KD) to further mitigate the over-fitting issue. Evaluated on the eight different domains in the DIHARD-III evaluation corpus, our experimental results show that the proposed QM-TS-VAD approach effectively enhances SD performances, and the introduced KD method can further reduce errors in seven of the eight domains. Finally, the proposed framework outperforms the unsupervised adaptation approach in the top-ranked system submitted to the DIHARD-III Challenge.

Index Terms: speaker diarization, unsupervised adaptation, quality-aware masking, teacher-student knowledge distillation, target-speaker voice activity detection

1. Introduction

Speaker diarization is a task to segment audio recordings according to speaker identity, namely "who spoke when" [1, 2]. It can assist various downstream tasks such as speech separation and automatic speech recognition (ASR) [3]. To facilitate the application of diarization under realistic conditions, the DIHARD Challenge was held [4, 5, 6], whose datasets are drawn from different domains in real-world scenarios.

Traditional speaker diarization methods [7, 8] usually comprise several independent components. First, they use a voice activity detection (VAD) model to detect the speech regions. Then, they partition the detected speech regions into many short segments to extract the speaker embeddings (e.g., i-vector [9], x-vector [10]). Finally, they group the extracted embeddings using clustering algorithms [11, 12] and assign the speakers accordingly. These clustering-based speaker diarization (CSD) methods are generally quite robust. For example, the Bayesian HMM clustering of x-vector sequences (VBx) [13] method has achieved first place in the DIHARD-II Challenge [5]. However, one main drawback of CSD methods is that they cannot well handle the overlapping regions where multiple speakers speak simultaneously due to the inherent constraint that each cluster can only be assigned to one speaker.

To handle the overlapping regions in diarization tasks, many approaches have been explored. One category of methods [14, 15] employs the overlap detector to assign the additional speakers in the detected overlapping segments. The other one uses the end-to-end framework to handle the overlapping problem. End-to-end neural speaker diarization (EEND) [16, 17] employs the multi-label classification framework to handle the diarization problem. Recurrent Selective Attention Network (RSAN) [18] recursively extracts the speakers through the time-frequency residual masks. Target-speaker voice activity detection (TS-VAD) [19] takes acoustic features and speaker embeddings as inputs to estimate the activity of speakers.

Although end-to-end systems have shown promising results, there is still one main issue that needs to be solved when applying them under realistic conditions: the real world contains a wide variety of domains, which vary greatly from speaker number, overlap ratios, environmental noises, and so on [6]. To handle the domain variability, end-to-end methods (e.g., EEND and TS-VAD) require labels for all speaker activities at each frame on different domains, which are hard to collect under real conditions [20]. Therefore, unsupervised adaptation approaches attract more and more research attention. For instance, both the first- and second-place systems [21, 22] in DIHARD-III Challenge [6] utilized the unsupervised adaptation techniques. Moreover, [20] proposed an iterative pseudo-label method using unlabeled data for EEND. [23] also proposed a continual training scheme for self-supervision domain adaptation. These methods typically utilize the pseudo-labels to adapt the pre-trained model. However, the generated pseudo labels often contain errors which will mislead the adapted model. To handle this problem, [24] proposed a quality-aware dynamic masking method for two-speaker separation-based speaker diarization (SSD), which employs the separation model to judge the quality of segments and purifies adaptation data through a dynamic mask. However, the methods in [24] are limited to the two-speaker conditions as the separation techniques have not been well established under realistic multi-speaker scenarios.

In this paper, we extend the quality-aware masking (QM) method from two-speaker to multi-speaker conditions based on TS-VAD, which we call quality-aware masking TS-VAD (QM-TS-VAD). Moreover, we employ the knowledge distillation (KD) method [25] on QM-TS-VAD to further alleviate the over-fitting problem. Experiments on eight domains in the DIHARD-III corpus show that the proposed QM-TS-VAD can effectively alleviate the interference of errors in pseudo-labels, and the introduced KD method can help the adapted SD model achieve further improvement. The key contributions of this paper are three-fold: (1) we extend the quality-aware masking method from two-speaker to multi-speaker conditions based on the TS-VAD model and propose the QM-TS-VAD framework; (2) we introduce the KD method to further alleviate the over-fitting problem caused by errors in pseudo-labels; and (3) by incorporating these two methods, our final system achieves better results than adaptation method in the champion system [21] of the DIHARD-III Challenge.

* corresponding author

10.21437/Interspeech.2023-1502

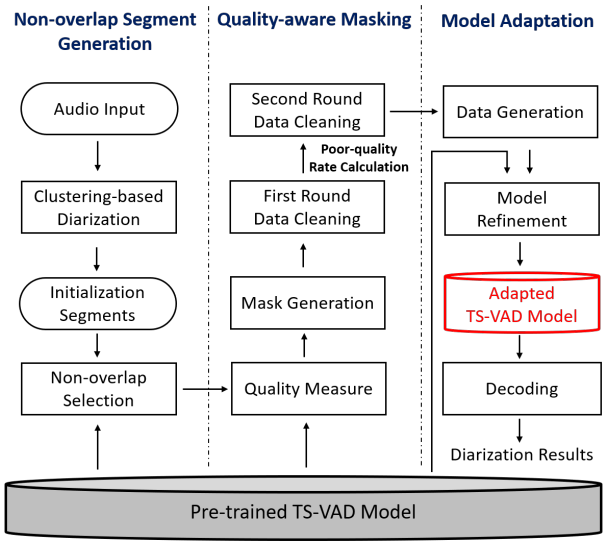

Figure 1: *Overall framework of the QM-TS-VAD.*

2. Prior Work

In [21], Wang et al. proposed an unsupervised adaptation framework called iterative TS-VAD (ITS-VAD). Given an audio mixture, they use the pre-trained TS-VAD model to remove the overlapping regions in CSD results to obtain the non-overlap speaker prior $\mathbf{y}_i \in \{0,1\}^{1\times T}$ for speaker i as:

$$y_{i,t} = \begin{cases} 1 & \text{Only speaker } i \text{ is active at } t,\ 1 \le t \le T \\ 0 & \text{Otherwise},\ 1 \le t \le T \end{cases} \tag{1}$$

where T is the utterance length. The hypothetical single speaker segments for speaker i can be obtained according to $\mathbf{y}_i$:

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

where $\mathbf{s}_{i,j}$ is the j-th non-overlap segment for i-th speaker. N_i is the total number of non-overlap segments for speaker i. These segments will be used to simulate the adaptation data as in [21]. In model adaptation, the binary cross-entropy loss is used:

$$L_1 = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\left[c_{i,t}\log\left(c'_{i,t}\right) + \left(1-c_{i,t}\right)\log\left(1-c'_{i,t}\right)\right] \tag{3}$$

where M is the speaker number in TS-VAD outputs. $c_{i,t} \in \{0,1\}$ is the pseudo label of speaker i at frame t. $c'_{i,t}$ is the corresponding posterior probability estimated by adapted model. Moreover, the adapted model can be used to generate speaker priors in the next iteration. The results in [21] show that the ITS-VAD can effectively improve the generalization ability of the model, which is also one key technique in their champion system of DIHARD-III Challenge [6]. However, the segments in $\mathbf{S}_i$ usually contain misleading information such as the interfering speech from other speakers due to the errors in speaker priors, which limits the performance of the adapted model.

3. The Proposed Framework

The overall framework of the proposed QM-TS-VAD is illustrated in Fig. 1. As can be observed, the whole framework comprises three key components: non-overlap segment generation, quality-aware masking, and model adaptation. In the non-overlap segment generation process, we utilize the speaker priors (pseudo-labels) from the CSD method and the pre-trained TS-VAD model to obtain the hypothetical non-overlap segments. In the quality-aware masking process, we purify the obtained segments through the quality-aware masks. Finally, we use the purified segments to simulate the adaptation data and refine the model. Moreover, in order to further mitigate the over-fitting problem, we also introduce the knowledge distillation technique. Note that the non-overlap segment generation and model adaptation are the same as in ITS-VAD [21]. Therefore, we mainly introduce the quality-aware masking and KD processes in the subsequent subsections.

3.1. Quality-aware masking

To purify the segments in $\mathbf{S}_i$, we employ the quality-aware masking (QM) method. For each $\mathbf{s}_{i,j}$, we first calculate the corresponding acoustic feature sequence $(\mathbf{x}_{i,j,t})_{t=1}^{T_{i,j}}$, where $T_{i,j}$ is the segment length. Then we feed them into the pre-trained TS-VAD model $\mathcal{F}_{\text{TS-VAD}}(\cdot)$ along with the corresponding speaker embedding $\mathbf{e}_i$ which is obtained from the speaker priors:

$$\left(p_{i,j,1}, \ldots, p_{i,j,T_{i,j}}\right) = \mathcal{F}_{\text{TS-VAD}}\left(\mathbf{x}_{i,j,1}, \ldots, \mathbf{x}_{i,j,T_{i,j}}, \mathbf{e}_i\right) \tag{4}$$

where $p_{i,j,t} \in (0,1)$ is the posterior probability for speaker i at frame t in segment $\mathbf{s}_{i,j}$. The confidence level of the pre-trained TS-VAD model on the quality of segment $\mathbf{s}_{i,j}$ can be reflected in $\mathbf{p}_{i,j} = [p_{i,j,1}, \ldots, p_{i,j,T_{i,j}}]^{\mathrm{T}}$. A large value of $p_{i,j,t}$ indicates that the pre-trained TS-VAD model is highly confident about the presence of speaker i at frame t. Conversely, when $p_{i,j,t}$ is small, the pre-trained TS-VAD model lacks confidence in the presence of the speaker i, and there may be other speakers at frame t. Therefore, we can measure the quality of each segment through the pre-trained TS-VAD model as shown in Fig. 1. To purify the segment $\mathbf{s}_{i,j}$, unreliable frames that most likely belong to other speakers should be masked. To achieve this, a threshold is required to determine whether a frame is of poor quality. Therefore, we propose a dynamic threshold that can change with the overall quality of the segment $\mathbf{s}_{i,j}$:

$$\tau_{i,j} = \min\left(\frac{1}{T_{i,j}}\sum_{t=1}^{T_{i,j}} p_{i,j,t}, \alpha\right) \tag{5}$$

where $\alpha > 0$ is a pre-defined hyper-parameter. We can get the corresponding mask $\mathbf{m}_{i,j}$ by comparing each element in $\mathbf{p}_{i,j}$ with $\tau_{i,j}$ as:

$$m_{i,j,t} = \begin{cases} 1 & p_{i,j,t} >= \tau_{i,j} \\ 0 & p_{i,j,t} < \tau_{i,j} \end{cases} \tag{6}$$

As demonstrated in Eqs. (5) and (6), if the overall quality of the speech segment $\mathbf{s}_{i,j}$ is relatively poor, the threshold $\tau_{i,j}$ will be reduced, and the corresponding quality assessment standard will be looser. This guarantees that enough speech segments are retained for simulation and the extremely poor-quality segments are masked. Through applying the mask $\mathbf{m}_{i,j}$ on segment $\mathbf{s}_{i,j}$, we can attain the masked segment as:

$$\mathbf{s}'_{i,j} = \mathbf{s}_{i,j} \odot \mathbf{m}_{i,j} \tag{7}$$

where $\odot$ is element-wise multiplication. Moreover, to further improve the purity of the adaptation data, we conduct a second round of data cleaning as shown in Fig. 1 by dropping the segments whose most parts are judged to be of poor quality. We

Algorithm 1 Procedure of quality-aware masking

Initialization: Non-overlap Segments $\mathbf{S}_i$

for all segment $\mathbf{s}_{i,j}$ in $\mathbf{S}_i$ **do**
  First Round Data Cleaning:
    Estimate the posterior probability $\mathbf{p}_{i,j}$ for segment $\mathbf{s}_{i,j}$ through pre-trained TS-VAD model using Eq. (4).
    Calculate the threshold $\tau_{i,j}$ through Eq. (5).
    Obtain mask $\mathbf{m}_{i,j}$ by comparing all elements in $\mathbf{p}_{i,j}$ with $\tau_{i,j}$ as in Eq. (6).
    Mask the segment $\mathbf{s}_{i,j}$ through Eq. (7) to obtain the masked segment $\mathbf{s}'_{i,j}$.
  Second Round Data Cleaning:
    Calculate $r_{i,j}$ and $\mu_{i,j}$ using Eqs. (8) and (9).
    if $r_{i,j} >= \mu_{i,j}$ **then**
      Drop the segment $\mathbf{s}'_{i,j}$.
    end if
end for

realize this process by comparing the rate of poor-quality parts in $\mathbf{s}_{i,j}$ with a dynamic threshold. The former is defined as:

$$r_{i,j} = 1 - \frac{\sum_{t=1}^{T_{i,j}} m_{i,j,t}}{T_{i,j}} \tag{8}$$

The dynamic threshold is defined as:

$$\mu_{i,j} = \min\left((1 - \tau_{i,j}) + \beta, \gamma\right) \tag{9}$$

where β and γ are pre-defined hyper-parameters. $1 - \alpha + \beta < \gamma < 1$ is used to limit the value of $\mu_{i,j}$. When $r_{i,j} >= \mu_{i,j}$, we directly drop the segment $\mathbf{s}'_{i,j}$. The purpose of Eq. (9) is to provide a looser screening strategy for poor-quality segments (i.e., $\tau_{i,j} < \alpha$) by using a larger $\mu_{i,j}$. After the second round of data cleaning, the remaining segments can be used to simulate the adaptation data. The procedure of QM process is summarized in Algorithm 1.

3.2. Knowledge distillation

Although the QM process can purify the segments, some uncleaned error regions still remain. Therefore, we employ the knowledge distillation (KD) method to alleviate the over-fitting problem in the model adaptation. During the KD process, the teacher model should have robust performance. Therefore, we don't employ the pre-trained TS-VAD model and instead use the QM-TS-VAD model which has undergone one round of adaptation as the teacher model. The student model has the same structure as the teacher one, which is initialized with the pre-trained TS-VAD model. For speaker i at frame t, there are two classes in TS-VAD outputs, namely speech and no-speech classes. Assuming the estimated values for these two classes before softmax in the teacher model are $q^1_{i,t}$ and $q^2_{i,t}$, and the corresponding values in the student model are $k^1_{i,t}$ and $k^2_{i,t}$. We calculate the soft targets from the teacher model and the corresponding softmax values from the student model as:

$$w^d_{i,t} = \frac{\exp\left(q^d_{i,t}/T\right)}{\sum_{d=1}^{2}\exp\left(q^d_{i,t}/T\right)} \tag{10}$$

$$z^d_{i,t} = \frac{\exp\left(k^d_{i,t}/T\right)}{\sum_{d=1}^{2}\exp\left(k^d_{i,t}/T\right)} \tag{11}$$

where $w^d_{i,t}$ is soft target. $z^d_{i,t}$ is the corresponding softmax value for student model under the same temperature T. Then we can obtain the Kullback-Leibler divergence based distillation loss:

$$L_2 = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\left[w^1_{i,t}\log\left(\frac{z^1_{i,t}}{w^1_{i,t}}\right) + w^2_{i,t}\log\left(\frac{z^2_{i,t}}{w^2_{i,t}}\right)\right] \tag{12}$$

The overall loss is obtained by incorporating L_1 and L_2 as:

$$L = (1 - \lambda) \times L_1 + \lambda T^2 \times L_2 \tag{13}$$

where λ is a hyper-parameter controlling the weights of two losses. By employing this overall loss, the adapted model can learn more information during the adaptation process, which can alleviate the over-fitting problem to some extent.

4. Experiments and Result Analysis

4.1. Experimental setup

The training set of the pre-trained TS-VAD model was drawn from Switchboard-1 Release 2 (LDC97S62) [26], Voxconverse [27] DEV set, AMI corpus [28], and the simulated multi-speaker conversations from the Librispeech dataset [29]. The total training set contains about 2500-hour single-channel audio data sampled at 16 kHz. Similar to [21], we also evaluated the performance on eight domains[1] in DIHARD-III [6] evaluation corpus. The considerable variations across these domains impose strict requirements on diarization systems.

For TS-VAD, we extracted 100-dim i-vectors following the pipeline in Kaldi[2] as the speaker embeddings. We set the output speaker number M to 8, which can cover most domains in the evaluation set. Moreover, for the special conversational telephone speech (CTS) domain which is up-sampled from 8 kHz and includes only two speakers, a two-speaker TS-VAD model was used, and the original 8 kHz Switchboard [26] training data was employed. During pre-training, we trained the 8-speaker TS-VAD model on the 2500-hour dataset for 2 epochs, which is consistent with [21]. We also trained the 2-speaker TS-VAD model for 20 epochs. In the adaptation process, we used the VBx [13] as the CSD system in non-overlap segment generation. We simulated the adaptation data through the open source pipeline[3] as in [21, 30]. For a 10-minute utterance, we simulated about 4-hour adaptation data. We only performed one iteration and ran one epoch in this iteration, which effectively improves the efficiency compared with [21]. We used Adam [31] to optimize the model. The batch size was set to 8. All hyper-parameters were obtained from the DEV set of the DIHARD-III corpus. The constant α and β were set to 0.5 and 0.1, respectively. γ was set to 0.7. The λ and T were set to 0.1 and 10, respectively. The diarization error rate (DER) [32] was utilized as the metric without forgiveness collar. The oracle VAD information was used for all systems in our experiments.

4.2. Experimental results

Table 1 lists the DER performance comparison among different systems on the evaluation set. Based on this table, several observations could be made. Firstly, although the TS-VAD model can handle overlapping speech, the more robust VBx method can achieve better performance on different domains, except for the CTS where TS-VAD is trained on matched telephone

[1]The eight domains are BROADCAST, COURTROOM, MAP TASK, CLINICAL, SOCIOLINGUISTIC LAB, SOCIOLINGUISTIC FIELD, CTS and MEETING. The detailed introduction is in [6].

[2]https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5

[3]https://github.com/jsalt2020-asrdiar/jsalt2020_simulate

Table 1: *DERs (%) comparison among different systems on the eight domains of evaluation set from DIHARD-III Challenge. ITS-VAD means iterative TS-VAD. BROADC. is BROADCAST. COURT is COURTROOM. SOC. is SOCIOLINGUISTIC.*

System \ Domain	BROADC.	COURT	MAP TASK	CLINICAL	SOC.LAB	SOC.FIELD	CTS	MEETING
VBx	4.22	3.08	3.41	11.08	6.04	8.05	14.20	33.20
TS-VAD	5.56	3.88	4.75	15.87	6.71	9.45	6.41	31.13
ITS-VAD [21]	4.46	3.07	3.20	10.03	3.81	7.10	6.11*	28.17
QM-TS-VAD	4.26	**3.06**	1.67	9.93	3.76	6.86	5.90	27.79
QM-TS-VAD + KD	**4.23**	3.12	**1.59**	**9.81**	**3.69**	**6.67**	**5.85**	**27.48**

* The original number in [21] is 9.71. Here we use a better two-speaker model as mentioned in Section 4.1, which leads to better performance.

dataset and MEETING domain whose overlap ratio is quite high. This indicates that the performance of the pre-trained TS-VAD model is unstable when encountering domain variability. Secondly, employing the iterative unsupervised adaptation method (ITS-VAD) in [21] has led to significant and consistent performance improvements across eight domains. This indicates that the ITS-VAD approach can effectively improve the model by leveraging adaptation data. Thirdly, after applying the quality-aware mask (i.e., QM-TS-VAD), almost all eight domains have different degrees of improvements compared with the ITS-VAD method, especially in the MAP TASK domain where the relative improvement is 47.8%. This implies that the quality-aware masking can efficiently purify the adaptation data and improve the model performance. Moreover, note that for ITS-VAD, the number of iterations in some domains may be greater than one [33], while the proposed QM-TS-VAD only requires one iteration, which significantly improves the efficiency of the adaptation process. Furthermore, apart from the COURT domain, the remaining seven domains show performance improvements after introducing the KD method (QM-TS-VAD + KD), indicating that the KD method can alleviate the over-fitting problem caused by misleading data that have not been masked in the QM process. Besides, the proposed techniques add minimal computation time, mainly from the data generation (about 300 seconds per utterance) and model refinement (about 2-4 minutes per utterance) processes.

4.3. Analysis on quality-aware masking process

As illustrated in Table 1, the improvement brought by the QM process varies in different domains. For instance, in the COURT domain, the application of the QM process brings minimal improvement (from 3.07% to 3.06%), whereas in the MAP TASK domain, the QM process can bring nearly 50% relative improvement (from 3.20% to 1.67%). To analyze the reason for this phenomenon, we conduct a statistic analysis when applying the QM on the COURT and MAP TASK domains. Fig. 2 shows the histograms of relative frequency (%) for posterior probabilities (PP) estimated by the pre-trained TS-VAD model and the corresponding thresholds on MAP TASK (MT) and COURT domains. As both domains have relatively good speaker priors, the majority of posterior probabilities are close to 1, where the QM process will not be applied. Therefore, we only consider the posterior probabilities less than 0.5. As shown in Fig. 2 (a), the posterior probability distribution of MT is relatively uniform, indicating that the pre-trained model lacks confidence in a considerable portion of data. Consequently, the distribution of corresponding dynamic thresholds is also relatively uniform as shown in Fig. 2 (b), which can filter out a considerable number of segments that may contain potential interferences. On the contrary, possibly due to the pre-training dataset containing corpus with dialog styles similar to the COURT domain, the pre-trained TS-VAD model is very confident in the COURT domain. The corresponding posterior probabilities under 0.5 are concentrated around 0, as shown in Fig. 2 (c). This makes the corresponding dynamic thresholds also around 0, as in Fig. 2 (d), resulting in a small portion of the data being masked, thereby leading to minimal changes in performance. However, setting a minimum value for the dynamic thresholds would mask a considerable amount of data on COURT, leading to a decrease in performance as well. Therefore, the improvements achieved by the QM process in a relatively matched scenario are relatively small, which is consistent with our intuition.

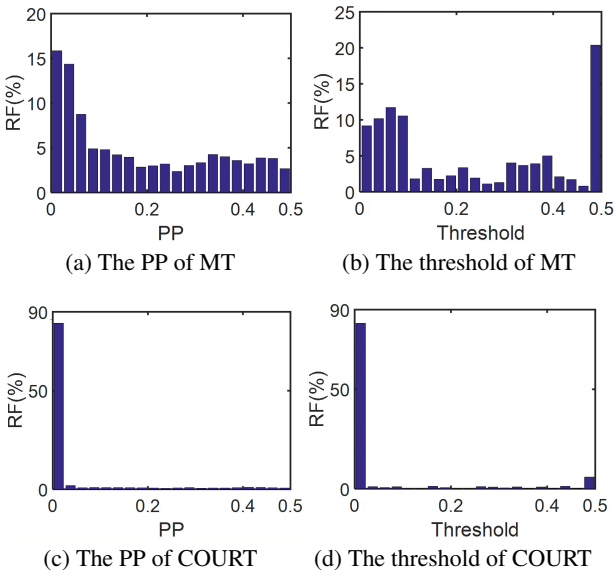

Figure 2: *Relative frequency (RF) statistics for posterior probabilities (PP) estimated by the pre-trained TS-VAD and the corresponding thresholds on MAP TASK (MT) and COURT domains.*

5. Conclusion

In this paper, we propose an unsupervised adaptation approach, called QM-TS-VAD, with quality-aware masking for speaker diarization. Moreover, we introduce a knowledge distillation (KD) technique in addition to QM-TS-VAD. Experiments on the eight domains from DIHARD-III Challenge show that QM can effectively purify the adaptation data, and KD can further alleviate the over-fitting problem. In the future, we will explore speaker priors to better purify the adaptation data.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62171427.

7. References

[1] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] 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," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[3] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, vol. 1, 2018.

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," *2018, tech. Rep.*, 2018.

[5] ——, "Second DIHARD challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.

[6] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," in *Proc. INTERSPEECH 2021*, 2021, pp. 3570–3574.

[7] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[8] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[11] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system." in *Interspeech*, 2007, pp. 1853–1856.

[12] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, 2013.

[13] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101254, 2022.

[14] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7114–7118.

[15] D. Raj, Z. Huang, and S. Khudanpur, "Multi-class spectral clustering with overlaps for speaker diarization," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 582–589.

[16] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[17] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. Interspeech 2020*, 2020, pp. 269–273.

[18] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5064–5068.

[19] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Interspeech 2020*, 2020, pp. 274–278.

[20] Y. Takashima, Y. Fujita, S. Horiguchi, S. Watanabe, L. P. G. Perera, and K. Nagamatsu, "Semi-Supervised Training with Pseudo-Labeling for End-To-End Neural Diarization," in *Proc. Interspeech 2021*, 2021.

[21] Y.-X. Wang, J. Du, M. He, S. Niu, L. Sun, and C.-H. Lee, "Scenario-dependent speaker diarization for DIHARD-III challenge." in *Interspeech*, 2021, pp. 3106–3110.

[22] S. Horiguchi, N. Yalta, P. Garcia, Y. Takashima, Y. Xue, D. Raj, Z. Huang, Y. Fujita, S. Watanabe, and S. Khudanpur, "The Hitachi-JHU DIHARD III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by dover-lap," *arXiv preprint arXiv:2102.01363*, 2021.

[23] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Continual self-supervised domain adaptation for end-to-end speaker diarization," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 626–632.

[24] S.-T. Niu, J. Du, L. Sun, Y. Hu, and C.-H. Lee, "QDM-SSD: Quality-aware dynamic masking for separation-based speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97S62," *Linguistic Data Consortium*, 1993.

[27] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[28] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 01 2005.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[30] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE spoken language technology workshop (SLT)*. IEEE, 2021, pp. 897–904.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[32] "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.

[33] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.