J. Shanghai Jiao Tong Univ. (Sci.), 2024
https://doi.org/ 10.1007/s12204-024-2715-2

Multi-Frame Cross-Channel Attention and Speaker Diarization Based Speaker-Attributed Automatic Speech Recognition System for Multi-Channel Multi-Party Meeting Transcription

XU Luzhen[1] (许露真), *YAN Haoyin*[1] (严浩尹), *HE Maokui*[1] (何茂奎), *GUO Zixian*[1] (郭子娴),
ZHOU Yeping[2] (周叶萍), *LIU Peiqi*[2] (刘沛奇), *ZHANG Jie*[1*] (张　结), *DAI Lirong*[1] (戴礼荣)
(1. Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China; 2. China Merchants Bank, Shenzhen 518048, Guangdong, China)

Abstract: This paper describes a speaker-attributed automatic speech recognition (SA-ASR) system submitted to the multi-channel multi-party meeting transcription challenge, which aims to address the "who spoke what" problem. We align the serialized output training-based multi-speaker ASR hypotheses and speaker diarization (SD) results to obtain speaker-attributed transcriptions. We use a pre-trained multi-frame cross-channel attention (MFCCA) model as the ASR module. We build a cascade system which includes a pre-trained speaker overlap-aware neural diarization and target-speaker voice activity detection model as the SD module. Decoding and alignment strategies are further used to improve the SA-ASR performance. Our proposed system outperforms the baseline with a relative improvement of 40.3% in terms of concatenated minimum-permutation character error rate on the AliMeeting dataset, which ranks top-3 on the fixed sub-track.

Keywords: multi-channel multi-party meeting transcription, speaker-attributed automatic speech recognition (SA-ASR), serialized output training, speaker diarization, concatenated minimum-permutation character error rate
CLC number: TN912.34　　**Document code:** A

0 Introduction

Speaker-attributed automatic speech recognition (SA-ASR) is dedicated to answering the question "who spoke what" in multi-party meeting scenarios[1]. It is expected by SA-ASR to transcribe the recorded speech signal, which might contain multiple speakers with overlapping segments, as well as to assign the speaker labels to each recognized word with an unknown number of speakers. Compared to the independent multi-speaker ASR[2] and speaker diarization (SD), SA-ASR is more natural to be applied in real-world multi-speaker environments[3]. Permutation invariant training (PIT) is a typical method for the end-to-end SA-ASR system[4]. However, the maximum number of speakers that the model can handle is constrained by the number of decoders. Besides, duplicated hypotheses in the different outputs might occur, as the outputs are independent in PIT.

In order to overcome these drawbacks, serialized output training (SOT) was proposed in Ref. [5], which introduces a special symbol to indicate the speaker change, allowing SOT-based models to have no limitations on the maximum number of speakers and avoiding the duplicated hypotheses. SOT has been successfully applied to multi-speaker ASR tasks in multi-party scenarios[6-7]. In Ref. [8], frame-level diarization with serialized output training (FD-SOT) which combines the SOT-based ASR and frame-level SD, was proposed by using the aligned timestamps to obtain speaker-attributed transcriptions. The FD-SOT method forms the basis of the proposed system, where we further investigate the impact of the involved multi-speaker ASR and diarization models as well as the post-processing step in this work.

To advance the current state-of-the-art in multi-talker ASR, the multi-channel multi-party meeting transcription (M2MeT2.0) challenge proposes a speaker-attributed ASR task, comprising two sub-tracks: fixed and open training conditions. The focus of this work is on the former sub-track, where participants

Received: 2023-12-19　**Accepted:** 2024-01-05
Foundation item: the National Natural Science Foundation of China (No. 62101523), the Joint AI Laboratory of CMB-USTC (No. FTIT2022058), and the USTC Research Funds of the Double First-Class Initiative (No. YD2100002008)

*E-mail: jzhang6@ustc.edu.cn

are required to use only the fixed-constrained data (i.e., AliMeeting[6], AISHELL-4[9] and CN-Celeb[10]) for system development. The usage of any additional data is strictly prohibited. However, participants can use open-source pre-trained models from third-party sources, such as Hugging Face and ModelScope, provided that the utilized models have to be clearly explained. Furthermore, a new test2023 set comprising around 10 hours of meeting data is used for scoring and ranking in the challenge. Near-field audio, transcriptions and oracle timestamps of this test set are not given, so there is no oracle speaker profile for the test2023 dataset.

In order to improve the transcription performance, in this work we employ a pre-trained multi-frame cross-channel attention (MFCCA) model① based on SOT, to improve the performance of the multi-speaker ASR. For SD, we train a target-speaker voice activity detection (TS-VAD) model[11] to cope with the issue of unknown speaker numbers and apply post-processing techniques in the decoding stage. We further employ a pre-trained speaker overlap-aware neural diarization (SOND) model② to obtain initial diarization results prior to iterating the TS-VAD model to achieve better diarization solutions[12]. The transcriptions obtained by multi-speaker ASR decoding are aligned with the diarization results based on timestamps. Experimental results show that the proposed system can achieve a minimum-permutation character error rate (cpCER) of 24.82% on the test2023 dataset of the fixed sub-track. The achieved cpCER ranks top-3 on the fixed sub-track of this M2MeT2.0 challenge.

The rest of this paper is organized as follows. Section 1 presents the detailed description of the submitted system. Section 2 introduces the experimental setup, followed by results in Section 3. Finally, Section 4 concludes this work.

1 System Description

Figure 1 illustrates the overall submitted SA-ASR system for the fixed sub-track of the M2MeT2.0 challenge from our team. Proposed SA-ASR system is based on the combination of SD results and SOT-based ASR transcriptions via the alignment of timestamps, which is referred to as FD-SOT. We utilize a pre-trained MFCCA model as the recipe for multi-channel multi-talker ASR. As for the SD module, a cascade pipeline is built, which is composed of clustering-based SD, the pre-trained SOND model, a TS-VAD model, and a post-processing strategy. Next, we will explain each component in detail.

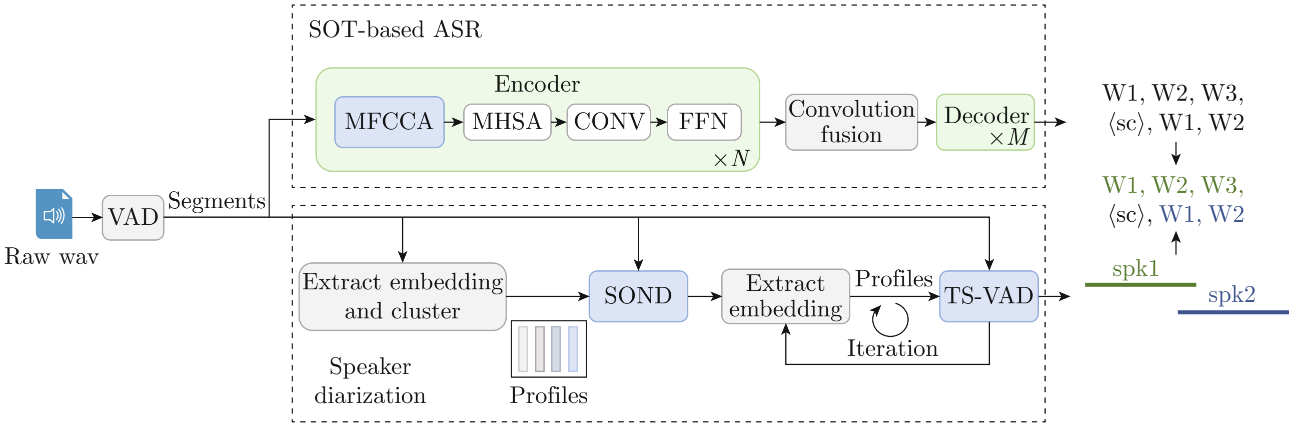

Fig. 1 Proposed multi-channel SA-ASR system for M2MeT2.0 challenge, which includes the MFCCA-based ASR, SOND and TS-VAD-based SD, and the alignment-based decoding components.

1.1 SOT

In order to guide the reader, we first introduce the SOT method[4] in this section. The SOT technique is able to overcome the limitation of the maximum number of speakers by modeling the dependencies between different speaker output sequences in an efficient and straightforward style[5]. During training, in order to recognize multiple utterances from different speakers, the SOT serializes the recognition results of different speakers into a long sequence, where a special token ⟨sc⟩ is inserted as a delimiter between different sentences to concatenate the transcriptions of different utterances. Note that the SOT also sorts the reference labels according to their start times, which means "first-in, first-out" (FIFO). The experiments show that the FIFO-based SOT scheme achieves a better CER than the PIT, which depends on the calculations of all permutations[5]. Therefore, the SOT is used as the training strategy in the multi-speaker ASR module of

①https://www.modelscope.cn/models/NPU-ASLP/speech_mfcca_asr-zh-cn-16k-alimeeting-vocab4950

②https://www.modelscope.cn/models/damo/speech_diarization_sond-zh-cn-alimeeting-16k-n16k4-pytorch our proposed system.

1.2 FD-SOT Framework

By integrating the multi-speaker hypotheses of SOT-based ASR model and the SD results, we propose to adopt FD-SOT system[8] as our basic framework, which aims to obtain transcriptions with speaker attributes by aligning the timestamps. First, based on the output of SD, the number of utterances is estimated using audio sentence segmentation, denoted as a. Let the number of speech segments in the SOT-based ASR output be denoted as b. In the case $a = b$, the alignment is performed directly based on the sequential order of the segments. In the case $a > b$, we discard the $a - b$ segments with the shortest durations from the TS-VAD output and then proceed with alignment. Conversely, in the case $a < b$, we discard the $b - a$ shortest transcription texts from ASR model and align the remaining segments. Finally, the speakers from SD and the transcriptions from SOT-based ASR are matched in a chronological order[13]. The benefit of this framework is that it enables a comprehensive and synchronized audio representation with both speaker identification and speech content information being taken into account.

1.3 MFCCA Model

In the proposed FD-SOT system, the multi-speaker ASR module is specifically a pre-trained MFCCA model[14]. For multi-channel ASR, the previously proposed cross-channel attention mechanisms have some problems, such as limited ability to extract fine-grained channel information[15] or only considering channel information for the current time step[16]. To overcome these limitations, we use the MFCCA-based conformer model for our SOT-based ASR system.

MFCCA effectively incorporates the channel context between adjacent frames to improve the modeling capacity for both frame-level and channel-level contextual information, e.g., see Fig. 2. The ith head of MFCCA is calculated as follows:

$$\boldsymbol{Q}_i = \bar{\boldsymbol{X}}\boldsymbol{W}_i^q + (\boldsymbol{b}_i^q)^{\mathrm{T}} \in \mathbf{R}^{T\times C\times D}, \tag{1}$$

$$\boldsymbol{K}_i = \bar{\boldsymbol{X}}_{\mathrm{cc}}\boldsymbol{W}_i^k + (\boldsymbol{b}_i^k)^{\mathrm{T}} \in \mathbf{R}^{T\times (2F+1)C\times D}, \tag{2}$$

$$\boldsymbol{V}_i = \bar{\boldsymbol{X}}_{\mathrm{cc}}\boldsymbol{W}_i^v + (\boldsymbol{b}_i^v)^{\mathrm{T}} \in \mathbf{R}^{T\times (2F+1)C\times D}, \tag{3}$$

$$\boldsymbol{H}_i = \mathrm{softmax}\Big(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{D}}\Big)\boldsymbol{V}_i \in \mathbf{R}^{T\times C\times D}, \tag{4}$$

where, $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$ and $\boldsymbol{V}_i$ denotes the query, key and value matrices for the ith head; $\boldsymbol{H}_i$ is the scaled dot-produce attention which is applied to the query, key and value; $\boldsymbol{W}^*$ and $\boldsymbol{b}^*$ ($* = q$, k, v) are learnable parameters; T, C and D stand for time, channel and feature dimensions, respectively; F is the number of local context frames to be concatenated at each time step. A single channel feature input is $\bar{\boldsymbol{X}}$, so a C-channel input can be defined as $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{X}}_0, \bar{\boldsymbol{X}}_1, \cdots, \bar{\boldsymbol{X}}_{C-1})$. $\bar{\boldsymbol{X}}_{\mathrm{cc}} = (\bar{\boldsymbol{X}}_{\mathrm{cc}}^0, \bar{\boldsymbol{X}}_{\mathrm{cc}}^1, \cdots, \bar{\boldsymbol{X}}_{\mathrm{cc}}^T)$, where $\bar{\boldsymbol{X}}_{\mathrm{cc}}^t$ $(t = 1, 2, \cdots, T)$ is the concatenation of the context frames at time step t, given by $\bar{\boldsymbol{X}}_{\mathrm{cc}}^t = (\bar{\boldsymbol{X}}^{t-F}, \cdots, \bar{\boldsymbol{X}}^t, \cdots, \bar{\boldsymbol{X}}^{t+F}) \in \mathbf{R}^{(2F+1)C\times D}$.

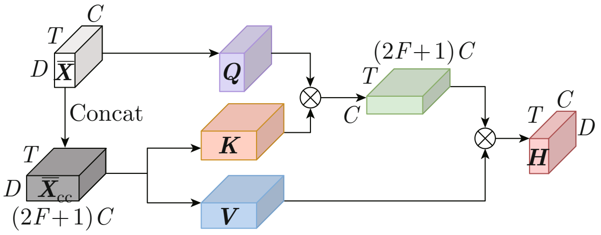

Fig. 2 Structure of the MFCCA procedure.

The MFCCA-based ASR model follows the typical conformer structure, including the MFCCA, multi-headed self-attention (MHSA), convolution (CONV) and feed forward network (FFN) modules. The MFCCA module takes multi-channel feature information as input and concatenates the channel information from several preceding and subsequent frames. The number of the concatenated frames F at each time step is a parameter which can balance the model performance and computational cost[14]. To better integrate the high-dimensional representations from the encoder output and reduce the potential loss of channel-specific information caused by directly reducing the channel dimensions, multi-layer convolution module is employed to gradually decrease the number of channels. Considering the number of channels and microphone array geometry, the MFCCA-based ASR model employs a channel masking strategy[14], where the multi-channel inputs are randomly masked to improve the model robustness.

1.4 Speaker Diarization

We employ the clustering-based SD to resolve the initial results. As for the speaker profile preparation phase, we first utilize a pre-trained x-vector speaker verification model[③] on speech segments with a fixed window length and window shift to extract speaker embeddings[12]. The speaker verification model consists of a ResNet34 backbone network for frame-level speaker feature extraction, global statistic pooling and multiple fully connected layers. Subsequently, spectral clustering is then utilized to classify the extracted embedding vectors, and the embeddings within the same cluster are averaged to obtain the profile of each speaker. As clustering-based SD techniques are unable to handle overlapping speech and do not directly minimize speaker identification errors, hybrid systems were proposed to combine clustering algorithms with neural network-based SD models, such as the typical TS-VAD model[17] and SOND model[12]. Likewise, we then consider to use the speaker profiles generated by the clustering-based SD model and speech features (e.g.,

[③] https://modelscope.cn/models/damo/speech_xvector_sv-zh-cn-cnceleb-16k-spk3465-pytorch

MFCC and Fbank) to predict the activity probability of each speaker in this work.

The SOND model uses a power-set encoding approach[12] to model SD as a single-label classification (SLC) task, which consists of a speech encoder, speaker encoder, context-dependent scorer, context-independent scorer and a speaker combination network for predicting power-set encodings. In SLC, different combinations of speakers are represented by unique power-set encodings, and the prediction of these encodings is used to explicitly model the inter-speaker correlations and overlapping speech. Hence, the SOND model allows for a better correlation between speakers as well as explicitly modeling overlapping speech. In the proposed cascade SD system, the SOND model is thus used as the first-stage diarization component prior to the TS-VAD model, such that the inter-speaker correlations and an explicit modeling of overlapping speech are both incorporated.

We then employ the TS-VAD model[④] to tackle the unknown number of speakers[11]. The TS-VAD module in the SD system performs individual detection of speech activity for each speaker, implicitly addressing the issue of speech overlaps. The TS-VAD model uses multi-channel information. It capitalizes on results from multiple channels (8 channels are treated as 8 systems), which are then fused via DOVER-Lap[11]. Considering that the number of speakers is varying[18], the number of output nodes M is selected as the maximum number of speakers in any recording in the training set. In our system, the number of output nodes M is empirically set to be 4. Based on the oracle profiles during training and the clustering-based diarization during decoding, the number of existing speakers in each speech segment is estimated, say $\hat{M}$. Similarly to Ref. [18], we can then perform:

(1) If $\hat{M} = M$, the trained TS-VAD model can be directly applied to the recordings.

(2) If $\hat{M} > M$, M speakers who have the longest non-overlapping speaking duration are chosen from $\hat{M}$ in the initial diarization output, and the remaining speakers are discarded.

(3) If $\hat{M} < M$, $\hat{M}$ output nodes are assigned to the "test" speakers, while $M - \hat{M}$ nodes are assigned to dummy speakers chosen randomly from the training set. These dummy speakers are abandoned when generating the final diarization output.

As such, the number of speakers in the input can be fixed during the training and decoding steps.

1.5 Decoding Procedure

The decoding process can be summarized as following steps. First, the long waveform is segmented into speech segments by the VAD provided by the organizer, and then we send the x-vectors and speech segments to the SOND model to produce more reliable diarization results. In order to achieve a better performance, we regenerate more accurate speaker embeddings using single-speaker speech segments from diarization and send them to the TS-VAD model for 3 additional iterations to obtain the final diarization results. Finally, the short-duration diarization utterances are removed before being integrated with the SOT-based ASR hypotheses for refinement. The effectiveness of such operations for post-processing will be analyzed in Section 3.

2 Experimental Setup

2.1 Data Preparation

ASR The pre-trained MFCCA-based multi-channel multi-talker ASR model is trained on AliMeeting, AISHELL-4, and Ali-simu. The AliMeeting dataset is a challenging Mandarin dataset, which consists of 104.75 h for training (Train), 4 h for evaluation (Eval), and 10 h as test set (Test). Each session was recorded by an 8-channel annular microphone array (Ali-far) and a single-channel headset microphone (Ali-near), lasting 15 to 30 min with 2 to 4 speakers. AISHELL-4 is also a Mandarin speech corpus recorded by an 8-channel annular microphone array, but each session contains 4 to 8 speakers. Ali-simu is a simulated training dataset using Ali-near and consists of 600 h recordings. Each utterance covers 2 to 4 speakers with an overlapping ratio ranging from 15% to 40%.

SOND The speaker embedding extractor of SOND model is pre-trained on the CN-Celeb corpus, which is a sizable speaker recognition dataset. Volume amplification, tempo perturbation, noise addition (from the MUSAN dataset[19]), and reverberation simulation are performed to augment the training samples. After pre-training, the embedding extractor is fine-tuned on AliMeeting. The SOND model is first pre-trained with a simulated dataset created from AliMeeting and then fine-tuned on real segments.

TS-VAD Our TS-VAD model is trained on AliMeeting_Train_RAW, AliMeeting_Train_WPE, CN-Celeb_simu and AISHELL-4. AliMeeting_Train_RAW contains real meeting data from Ali-far. Instead of directly employing the manual label from the transcription as the final training label, we use the force alignment technique, which ensures more precise frame-level training targets and eliminates the silence of manual label segments brought by small pauses in

[④] The TS-VAD model mainly includes a speaker detection component comprising a two-layer bidirectional LSTM with projection (BLSTMP) splices acoustic features which are extracted by 4 convolutional layers from raw Fbanks along with the i-vectors, and produces M spliced outputs. These outputs are then passed through a one-layer BLSTMP, to produce M 2-class outputs corresponding to the speech and silence probabilities.

the speakers. We obtain targets by a Gaussian mixture model (GMM) with tri3 (LDA+MLLT) alignment. AliMeeting_Train_WPE is a dereverberated version of Ali-far by utilizing the WPE[20] acoustic reflections algorithm. Also, the force alignment targets are taken as training labels therein. CN-Celeb_simu is a simulated data from CN-Celeb. In each session, 2 to 4 speakers from CN-Celeb are randomly selected and their sentences are combined at a 0—40% overlapping ratio. Then we add simulated reverberation and noise from MUSAN for data augmentation. Reliable training labels are obtained by the VAD module. For AISHELL-4, we directly use the given targets for training, since it does not provide near-field data or force alignment on far-field data (i.e., resulting in unreliable targets). Table 1 summarizes the details of all those training data used in our system. Notice that the ASR and SOND models are existing pre-trained models, but the TS-VAD is the model we trained using the data in Table 1.

Table 1 Modules and data summary of the proposed SA-ASR system

Sub-system	Pre-trained model	Data
ASR	√	AliMeeting AISHELL-4
SOND	√	Ali-simu CN-Celeb AliMeeting (with simulated data)
TS-VAD	×	AliMeeting_Train_RAW/WPE CN-Celeb_simu AISHELL-4

Implementation Our SA-ASR system is implemented using PyTorch and the experiments are conducted using the FunASR toolkit. During the training phase of the TS-VAD model, each session is segmented into short segments with a window length of 8 s and a window shift of 6 s. Mixup[21] is performed within each session. The Adam optimizer[22] is used to update the model parameters with a learning rate of 0.000 1. The TS-VAD model is trained using 8 GeForce RTX3090 GPUs.

3 Experimental Results

Table 2 shows the ASR performance on the AliMeeting evaluation set in terms of speaker independent-character error rate (SI-CER)[8]. As the proposed SA-ASR system depends on the pre-trained MFCCA model, we compare the ASR performance under different decoding strategies. The integration of a transformer-based language model (LM) in the decoding phase for the pre-trained MFCCA model can lead to a certain degree of performance improvement. It is clear that setting the beam size to 20 outperforms the case when the beam size is 10. Therefore, we adopt the MFCCA model with a beam size of 20 in combination with the LM during decoding in the sequel. Compared to the ASR module in the baseline FD-SOT[8], the proposed approach can achieve a relative reduction of up to 45.3% in SI-CER.

Table 2 SI-CER performance comparison on the AliMeeting Eval set using different decoding processes

Decoding configuration		SI-CER/%
Beam size	LM	
10	×	16.51
10	√	16.30
20	×	16.47
20	√	**16.25**

The comparison of diarization results is shown in Table 3. We use the oracle speaker profiles as inputs to the diarization model. By employing our SOND+TS-VAD system, it is clear that a noticeable reduction in the diarization error rate (DER)[23] can be achieved compared to using only the SOND model as the diarization component. In the case of using speaker embeddings generated by spectral clustering as inputs to the diarization model, it can be observed that the DER obtained from the TS-VAD model is significantly better than that using only the SOND model. We also evaluate the diarization results obtained by using our SOND+TS-VAD system, leading to a further reduction in the DER compared to the cluster+TS-VAD model. Compared to the SOND model, SOND+TS-VAD can decrease the DER from 16.78% to 11.06%. As a result, we see that the TS-VAD contributes more to SD in this context. For completeness, we additionally compare the cascade choice of TS-VAD+SOND, it is clear that the TS-VAD+SOND performs much worse than SOND+TS-VAD, implying the importance of the cascading rule. Therefore, we will choose SOND+TS-VAD as the SD module in the sequel.

Table 3 DER and cpCER performance on the Alimeeting Eval set using different diarization systems

Diarization system	DER/%	cpCER/%
Oracle+SOND	14.98	42.74
Oracle+SOND+TS-VAD	9.60	38.56
Cluster+SOND	16.78	43.77
Cluster+TS-VAD	12.68	42.65
Cluster+SOND+TS-VAD	**11.06**	**39.77**
Cluster+TS-VAD+SOND	18.90	46.66

We also demonstrate the SA-ASR performance with different diarization systems in Table 3. It can be seen that the cpCER is closely related to the diarization performance. The TS-VAD model is better than the SOND when no short-duration diarization results are deleted. In the case of using spectral clustering, the proposed combination of SOND+TS-VAD not only obtains the lowest DER, but also leads to the best ASR accuracy.

As in the proposed FD-SOT system, the transcriptions obtained from the multi-speaker ASR model have to be aligned with the timestamps of the SD results. We found that some interfering speech might be recognized by analyzing the decoding results when target-speaker speech duration is too short, resulting in a lot of insertion errors. Compared with the insertion errors caused by oracle speaker labels that cover all speaker's speech, there are relatively fewer deletion errors resulting from the SD results (i.e., ignoring short speaker utterances). Based on this, we further show the impact of minimum time[5] of diarization utterances in Table 4, where the cpCER results of different diarization systems on the AliMeeting Eval set are included. When the minimum time is greater than 0 s, the SOND model performs better than the TS-VAD. The combination of SOND and TS-VAD outperforms the individual utilization of these two models for almost all values of minimum time. It is clear that the SOND+TS-VAD system performs better than the TS-VAD+SOND in most cases. Furthermore, we can see that all the models achieve the best results around a rough minimum time of 1.3 s. For different diarization systems, the post-processing strategy can bring an absolute cpCER reduction ranging from 9.06% to 12.99% on the Eval set compared with the cases of preserving all short speaker speech.

Table 4 Comparison of cpCER on the AliMeeting Eval using different diarization systems with the different minimum time of diarization utterances

Diarization system	cpCER/%							
	0 s	0.3 s	0.5 s	0.7 s	0.9 s	1.1 s	1.3 s	1.5 s
Oracle+SOND	42.74	38.05	35.68	33.75	32.23	30.99	**29.75**	30.06
Oracle+SOND+TS-VAD	38.56	37.95	35.65	33.28	30.86	29.00	**28.91**	29.27
Cluster+SOND	43.77	38.61	37.13	34.51	32.54	31.63	**31.11**	31.27
Cluster+TS-VAD	42.65	41.65	39.26	36.49	34.52	33.45	**33.10**	33.35
Cluster+SOND+TS-VAD	39.77	38.93	35.90	33.65	31.64	**30.71**	30.75	30.90
Cluster+TS-VAD+SOND	46.66	42.36	39.64	37.18	35.69	34.58	**33.82**	33.84

Table 5 shows the final results of our system on both Eval and Test2023 datasets. We find that on Test2023 the performance first improves with an increase in the duration but then decreases as we have explained in Table 4. The best cpCER is roughly obtained when the minimum time is around 0.7 s. More importantly, the proposed system is far better than the official baseline on both Eval and Test2023 sets.

Table 5 Comparison of cpCER on the AliMeeting Eval and Test2023 datasets based on the different minimum time of diarization utterances

System	Minimum time/s	cpCER/%	
		Eval	Test2023
Baseline	—	53.76	41.55
Ours	0	39.77	26.50
	0.5	35.90	24.88
	0.6	34.77	24.86
	0.7	33.65	**24.82**
	0.8	32.37	25.08
	1.1	**30.71**	—

4 Conclusion

In this work, we presented an improved FD-SOT system, which was submitted to the SA-ASR fixed subtrack of the M2Met2.0 challenge. The ASR component utilizes a pre-trained MFCCA model, while the SD part incorporates the SOND and TS-VAD models given the unknown number of speakers. Decoding strategy for SD and the post-processing technique for frame-level alignment between the SOT-based ASR and diarization results were employed, and the corresponding efficacy was shown experimentally. It was also shown that in the context of multi-speaker ASR, the transcribing accuracy is positively related to the diarization performance. Therefore, it might be possible to improve the SA-ASR performance from these two perspective. In the future, we will dedicate to reducing the decoding

[5] The minimum time controls the amount of short speaker durations. In experiments, we discard the diarization results whose duration is smaller than the minimum time. The minimum time of 0 s means that all diarization results are preserved.

complexity in order to fulfill real-time requirements.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] FISCUS J G, AJOT J, GAROFOLO J S. The rich transcription 2007 meeting recognition evaluation [M]//Multimodal technologies for perception of humans. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 373-389.

[2] YU D, CHANG X K, QIAN Y M. Recognizing multi-talker speech with permutation invariant training [C]//*Interspeech 2017*. Stockholm: ISCA, 2017: 2456-2460.

[3] SHI M, DU Z, CHEN Q, et al. CASA-ASR: Context-aware speaker-attributed ASR [DB/OL]. (2023-05-21). https://arxiv.org/abs/2305.12459

[4] SEKI H, HORI T, WATANABE S, et al. A purely end-to-end system for multi-speaker speech recognition [C]//*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne: ACL, 2018: 2620-2630.

[5] KANDA N, GAUR Y, WANG X F, et al. Serialized output training for end-to-end overlapped speech recognition [C]//*Interspeech 2020*. Shanghai: ISCA, 2020: 2797-2801.

[6] YU F, ZHANG S L, FU Y H, et al. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge [DB/OL]. (2021-10-14). http://arxiv.org/abs/2110.07393

[7] YU F, ZHANG S L, GUO P C, et al. Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge [DB/OL]. (2022-02-08). http://arxiv.org/abs/2202.03647

[8] YU F, DU Z H, ZHANG S L, et al. A comparative study on speaker-attributed automatic speech recognition in multi-party meetings [C]//*Interspeech 2022*. Incheon: ISCA, 2022: 560-564.

[9] FU Y H, CHENG L Y, LV S B, et al. AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario [C]//*Interspeech 2021*. Brno: ISCA, 2021: 3665-3669.

[10] FAN Y, KANG J W, LI L T, et al. CN-CELEB: A challenging Chinese speaker recognition dataset [DB/OL]. (2019-10-31). http://arxiv.org/abs/1911.01799

[11] HE M K, LV X, ZHOU W L, et al. The USTC-Ximalaya system for the ICASSP 2022 multi-channel multi-party meeting transcription (M2MeT) challenge [DB/OL]. (2022-02-10). http://arxiv.org/abs/2202.04855

[12] DU Z H, ZHANG S L, ZHENG S Q, et al. Speaker overlap-aware neural diarization for multi-party meeting analysis [C]//*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi: ACL, 2022: 7458-7469.

[13] SHI M H, ZHANG J, DU Z H, et al. A comparative study on multichannel speaker-attributed automatic speech recognition in multi-party meetings [DB/OL]. (2022-11-01). http://arxiv.org/abs/2211.00511

[14] YU F, ZHANG S L, GUO P C, et al. MFCCA: Multi-frame cross-channel attention for multi-speaker ASR in multi-party meeting scenario [C]//*2022 IEEE Spoken Language Technology Workshop*. Doha: IEEE, 2023: 144-151.

[15] CHANG F J, RADFAR M, MOUCHTARIS A, et al. Multi-channel transformer transducer for speech recognition [C]//*Interspeech 2021*. Brno: ISCA, 2021: 296-300.

[16] WANG W Q, QIN X Y, LI M. Cross-channel attention-based target speaker voice activity detection: Experimental results for the M2met challenge [C]//*ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. Singapore: IEEE, 2022: 9171-9175.

[17] MEDENNIKOV I, KORENEVSKY M, PRISYACH T, et al. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario [C]//*Interspeech 2020*. Shanghai: ISCA, 2020: 274-278.

[18] HE M K, RAJ D, HUANG Z L, et al. Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker [C]//*Interspeech 2021*. Brno: ISCA, 2021: 3555-3559.

[19] SNYDER D, CHEN G G, POVEY D. MUSAN: A music, speech, and noise corpus [DB/OL]. (2015-10-28). http://arxiv.org/abs/1510.08484

[20] YOSHIOKA T, NAKATANI T. Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(10): 2707-2720.

[21] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization [DB/OL]. (2017-10-25). http://arxiv.org/abs/1710.09412

[22] KINGMA D P, BA J. Adam: A method for stochastic optimization [DB/OL]. (2014-12-22). http://arxiv.org/abs/1412.6980

[23] FISCUS J G, AJOT J, MICHEL M, et al. The rich transcription 2006 spring meeting recognition evaluation [M]//Machine learning for multimodal interaction. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 309-322.