THE SECOND MULTI-CHANNEL MULTI-PARTY MEETING TRANSCRIPTION CHALLENGE (M2MET 2.0): A BENCHMARK FOR SPEAKER-ATTRIBUTED ASR

Yuhao Liang[1,2], Mohan Shi[3], Fan Yu[2], Yangze Li[1], Shiliang Zhang[2], Zhihao Du[2], Qian Chen[2],
Lei Xie[1], Yanmin Qian[4], Jian Wu, Zhuo Chen, Kong Aik Lee[5,6], Zhijie Yan[2], Hui Bu[7]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, China
[2]Speech Lab of DAMO Academy, Alibaba Group, China,
[3]NERC-SLIP, University of Science and Technology of China (USTC), China
[4]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[5]ICT Cluster, Singapore Institute of Technology, Singapore
[6]Institute for Infocomm Research, A*STAR, Singapore
[7]Beijing Shell Shell Technology Co., Ltd., Beijing, China

ABSTRACT

With the success of the first Multi-channel Multi-party Meeting Transcription challenge (M2MeT), the second M2MeT challenge (M2MeT 2.0) held in ASRU2023 particularly aims to tackle the complex task of *speaker-attributed ASR (SA-ASR)*, which directly addresses the practical and challenging problem of "who spoke what at when" at typical meeting scenario. We particularly established two sub-tracks. The fixed training condition sub-track, where the training data is constrained to predetermined datasets, but participants can use any open-source pre-trained model. The open training condition sub-track, which allows for the use of all available data and models without limitation. In addition, we release a new 10-hour test set for challenge ranking. This paper provides an overview of the dataset, track settings, results, and analysis of submitted systems, as a benchmark to show the current state of speaker-attributed ASR.

Index Terms— M2MeT 2.0, Alimeeting, Meeting Transcription, Multi-speaker ASR, Speaker-attributed ASR

1. INTRODUCTION

Despite years of research, meeting transcription accuracy still faces significant challenges, including but not limited to overlapping speech, an unknown number of speakers, far-field attenuated speech signals, noise, reverberation, and other factors that can degrade transcription performance. The ICASSP2022 Multi-Channel Multi-Party Meeting Transcription (M2MeT) challenge [1, 2] has played a crucial role in the development of Mandarin meeting transcription technology by addressing the challenge of speech overlap in actual meetings. The challenge consists of two distinct tasks: speaker diarization and multi-speaker automatic speech recognition (ASR). The former involves identifying *who spoke when*

in the meeting, while the latter aims to transcribe speech from multiple speakers. In the second M2MeT challenge (M2MeT 2.0), these two tasks are merged into a single speaker-attributed task.

The M2MeT 2.0 challenge presents two key differences from its predecessor, the first M2MeT. First, the evaluation metric in the first M2MeT challenge was speaker-independent, meaning that transcription could be determined, but the corresponding speaker was not identified. To overcome this limitation and advance current multi-talker ASR systems, the M2MeT 2.0 challenge introduces the *speaker-attributed ASR (SA-ASR)* task. This task not only transcribes the speech but also assigns speaker labels to each transcription. Specifically, we introduce the concatenated minimum permutation character error rate (cpCER) metric to evaluate the performance of the submitted systems. The cpCER is proposed for Mandarin particularly, which is defined similarly to the concatenated minimum permutation word error rate (cpWER) [3]. Second, unlike other related challenges such as Computational Hearing in Multisource Environments (CHIME) [3], Multimodal Information Based Speech Processing (MISP) [4], and M2MeT [1, 2], the M2MeT 2.0 challenge offers participants the freedom to utilize any open-source pre-trained model, which is typically prohibited in these challenges. This flexibility aims to explore the feasible industrial application of academic research proposed in previous studies, utilizing various open-source pre-trained models trained on a large amount of data for the SA-ASR task.

The M2MeT 2.0 challenge consists of two sub-tracks: 1) The fixed training condition track, is designed to enable reproducible research in this field by providing a fixed set of training data, open-source pre-trained models, and evaluation criteria. 2) The open training condition track, aims to bench-

mark state-of-the-art performance in speaker-attributed ASR by allowing participants to use their own data and training techniques.

2. RALATED WORKS

The SA-ASR task [5] involves identifying multiple speakers and transcribing overlapped speech within a single session. One common approach to address this cocktail party challenge is to use speaker diarization to identify the active regions of different speakers. Then, a single-talker ASR system with a speaker separation module can be used to transcribe speech from the known active regions. Alternatively, an end-to-end multi-talker ASR system can be used to transcribe speech and assign speaker labels simultaneously based on corresponding speaker information provided by the diarization system.

Speaker diarization techniques that follow conventional clustering-based approaches usually include two main steps: speaker embedding extraction and clustering. These approaches begin by transforming the input audio stream into a speaker-specific representation, followed by a clustering process like Variational Bayesian HMM clustering (VBx) [6] that groups the regions of each speaker into separate clusters. Clustering-based methods typically assign a single speaker label to each frame, making it challenging for them to handle speech overlap. With the rapid development of deep learning, End-to-End speaker diarization methods like end-to-end neural diarization (EEND) [7] are proposed, leveraging a single neural network to replace the modular cluster-based system. Inspired by the target speaker extraction [8, 9, 10, 11], target speaker voice activity detection (TS-VAD) [12, 13, 14] has been proposed, which can estimate the activity level of each speaker in the presence of overlapping speech, providing a promising solution for speaker diarization.

Single-talker ASR has been extensively investigated in recent years, and various architectures, such as Conformer [15], Branchformer [16], and Paraformer [17], have been proposed to push the limit of the accuracy of ASR.

Meanwhile, there has been a growing interest in multi-talker ASR, which is designed to transcribe speech containing several speakers. One recent approach, called Speaker-Attributed Transformer(SA-Transformer) [18], generates token-level speaker labels during the decoding phase and can directly produce speaker-attributed transcription only using the clustered speaker profile. Moreover, TS-ASR [19] is also a promising approach that utilizes speaker embedding for target speaker extraction (TSE) and achieves good performance when TSE and ASR are jointly trained. By extracting the speaker embedding, TS-ASR can identify the target speaker's voice and enhance the accuracy of the overall ASR system.

The field of rich transcription with speech overlap has undergone extensive research, with advancements facilitated by numerous challenges and open-source datasets [4, 20, 21, 22]. Table 1 outlines the primary datasets utilized in this scenario.

WSJ0-2mix [23] and Libri2Mix [24] datasets involve mixing pairs of utterances from different speakers at random SNRs, making them primarily used for speech separation tasks where there is full overlap. On the other hand, AMI [25], LibriCSS [26], and CHIME-6 [3] datasets are recorded in real rooms. However, the AMI dataset's fixed speaker count of four and poor recording quality limit its practical applications. LibriCSS, similar to Libri2Mix, utilizes the Librispeech corpus to produce speech mixes. However, due to the fixed intonation and pace of reading in Librispeech, there remains a disparity between LibriCSS and real meeting scenarios. CHIME-6 dataset is designed for conference or indoor conversation transcription tasks and accounts for overlapped speech.

Although these datasets significantly contribute to the progress of transcription overlapping speech, they are limited to English. The language barrier poses a challenge in achieving comparable results for non-English languages, such as Mandarin. To overcome this challenge, the AISHELL-4 [27] and AliMeeting [1] datasets have been developed specifically for Mandarin meeting transcription. AISHELL-4 has a lower overlap ratio, while AliMeeting contains intense discussions. Additionally, AliMeeting records the near-field signal of each participant using a headset microphone, ensuring that only the participant's speech is transcribed.

Table 1. Datasets available in the literature in multi-talker speech transcription (OR: overlap rate)

Dataset	Hours	#SPK	Devices	OR (%)
WSJ0-2mix [23]	43	129	Simu	Full
Libri2Mix [24]	292	1252	Simu	Full
AMI [25]	100	190	Headset mic, 8-ch mic array	<10
LibriCSS [26]	10	40	7-ch mic array	0-40
CHIME-6 [3]	35	26	4-ch mic array	40
AISHELL-4 [27]	120	61	8-ch mic array	19
AliMeeting [1]	129	481	Headset mic, 8-ch mic array	42

3. DATASET AND TRACKS

AliMeeting [1, 2], AISHELL-4 [28], and CN-Celeb [29] corpus are adopted as our training data, which is the same as the first M2MeT challenge. The AliMeeting dataset is a collection of multi-talker conversation recordings in a meeting setting, comprising a total of 118.75 hours of speech data. The dataset is split into 104.75 hours for training (*Train*), 4 hours for evaluation (*Eval*), and 10 hours for testing (*Test*). The AliMeeting corpus includes both far-field overlapped audios and corresponding near-field audios, which exclusively record and transcribe single-speaker speech. To evaluate the submitted systems, an additional 10 hours of audio data,

called *Test-2023*, is incorporated specifically for testing purposes. *Test-2023* comprises 10 sessions that were recorded in 5 different rooms and include 58 speakers. It is crucial to highlight that none of the speakers in *Test-2023* overlap with those in the AliMeeting corpus.
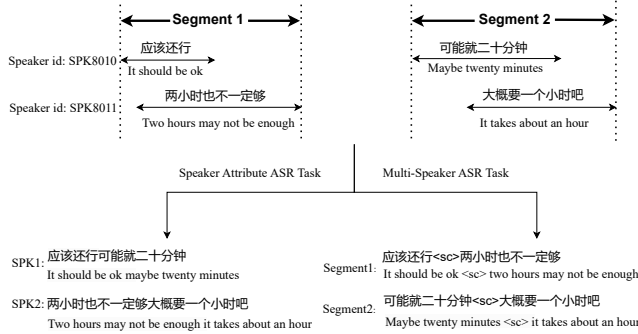
Fig. 1. The main difference between the speaker-attributed ASR task (M2MeT 2.0) and the multi-speaker ASR task (M2MeT).

This challenge introduces the speaker-attributed ASR task which poses a unique challenge of transcribing speech from multiple speakers and assigning a speaker label to the transcription simultaneously. Figure 1 illustrates the difference between the speaker-attributed ASR task and the multi-speaker ASR task. The speaker-attribution ASR task groups transcriptions from the same speaker together, while the multi-speaker ASR task combines overlapping sentences spoken by different speakers. Sub-track 1 restricts participants from using constrained datasets, while sub-track 2 allows the use of any dataset, including private ones. Any open-source pre-trained models are available for both of these two sub-tracks. The accuracy of a speaker-attributed ASR system is evaluated using the cpCER. The calculation of cpCER involves three steps. The first step is to concatenate the reference and hypothesis transcriptions from each speaker in chronological order. This produces a single transcription within one session. Next, the character error rate (CER) is calculated between the concatenated reference and hypothesis transcriptions, and this process is repeated for all possible speaker permutations. Finally, the permutation with the lowest CER is selected as the cpCER for that session.

To provide a clear and concise description of the cpCER computation, we illustrate it in Algorithm 1. To run the algorithm, it is necessary to provide both the ground truth and hypothesis transcriptions for a given session, which are arranged in chronological order. In cases where the length of Y and H is not equal, we employ padding with blank transcriptions to ensure that both sets have the same length.

4. SYSTEM DESCRIPTION

4.1. Baseline system

We release an E2E SA-Transformer baseline built on FunASR [30] toolkit for easy and reproducible research. The

---

Algorithm 1: Computation of cpCER

input : Ground truth: $\{Y_1, Y_2, ..., Y_S\}$, hypothesis of different speakers: $\{H_1, H_2, ..., H_{\hat{S}}\}$. S is the oracle speaker number and $\hat{S}$ is the predicted speaker number

output: The $cpCER$ of given session

1   $Y \leftarrow \{Y_1, Y_2, ..., Y_S\}$;
2   $H \leftarrow \{H_1, H_2, ..., H_{\hat{S}}\}$;
3   $mindistance \leftarrow INFINITY$;
4   **foreach** *permutation of H* **do**
5      $distance \leftarrow 0$;
6   **for** $i \leftarrow 1$ **to** $\max(S, \hat{S})$ **do**
7         $distance \leftarrow$ $distance + editdistance(H[i], Y[i])$;
8   **end**
9   **if** $distance < mindistance$ **then**
10         $mindistance \leftarrow distance$;
11   **end**
12   **end**
13   $totaltoken \leftarrow \sum_{i=0}^{S} length(Y[i])$;
14   $cpCER \leftarrow mindistance/totaltoken \times 100\%$

---

model architecture is illustrated in Figure 2. It comprises an ASR block and a speaker block to carry out ASR and token-level speaker identification. The ASR block is represented as

$$H^{asr} = \text{AsrEncoder}(X), \quad (1)$$
$$o_n = \text{AsrDecoder}(y_{[1:n-1]}, H^{asr}, \bar{d}_n). \quad (2)$$

In the ASR block, the AsrEncoder converts the given acoustic feature X into a series of hidden embeddings H^{asr}. The AsrDecoder then produces the output distribution o_n step-by-step. To generate each token, the AsrDecoder takes in the history token $y_{[1:n-1]}$, the hidden embeddings H^{asr}, and the weighted speaker profile $\bar{d}_n$. Compared to the other encoder-decoder-based ASR models, our model differs in its use of the weighted speaker profile $\bar{d}_n$, computed in the speaker block. This profile is employed to bias the transcription towards a specific speaker, which enhances the model's ability to identify individual speakers within a session. The posterior probability of token i at the n-th decoding step is formulated as

$$Pr(y_n = i | y_{[1:n-1]}, s_{[1:n]}, X, D) = o_{n,i}. \quad (3)$$

On the other hand, the speaker block is denoted as

$$H^{spk} = \text{SpeakerEncoder}(x), \quad (4)$$
$$q_n = \text{SpeakerDecoder}(y_{[1:n-1]}, H^{spk}, H^{asr}), \quad (5)$$
$$\beta_{n,k} = \frac{\exp(\cos(q_n, d_k))}{\sum_j^K \exp(\cos(q_n, d_j))}, \quad (6)$$
$$\bar{d}_n = \sum_{k=1}^{K} \beta_{n,k} d_k. \quad (7)$$

The speaker encoder takes the input acoustic feature X and produces the speaker embedding H^{spk}, which has the same shape as the ASR embedding H^{asr} and represents the speaker's unique characteristics. During each decoding step n, the SpeakerDecoder uses $y_{1:n-1}$, H^{spk}, and H^{asr} to generate a speaker query q_n. Based on this query, a cosine distance-based attention weight $\beta_{n,k}$ is calculated for every profile d_k in D. These weights can be viewed as posterior probabilities for the n-th token being attributed to the k-th speaker, taking into account all previous estimations, as well as relevant information from the input X and the speaker profiles in D. The $\beta_{n,k}$ can be represented as

$$Pr(s_n = k|y_{[1:n-1]}, s_{[1:n-1]}, X, D) = \beta_{n,k}. \qquad (8)$$

To extract speaker embeddings and initialize the SpeakerEncoder, we leveraged a pre-trained x-vector extractor from ModelScope trained on CN-Celeb. To train our E2E SA-ASR system, we employed a two-stage training strategy. In the first stage of training the E2E SA-ASR system, we trained a standard Conformer for the ASR task, which was utilized in the second stage to initialize the ASR block. In this second stage, both the ASR and speaker losses were incorporated to fine-tune the model. By using Eqs. 3 and 8, the joint posterior of token Y and speaker S, optimized to be maximized during training, is represented as

$$Pr(Y, S|X, D) = \prod_{n=1}^{N} Pr(y_n|y_{[1:n-1]}, s_{[1:n]}, X, D) \qquad (9)$$
$$\times Pr(s_n|y_{[1:n-1]}, s_{[1:n-1]}, X, D).$$

During the training phase, speaker embeddings were extracted from audio solely containing one speaker to generate speaker profiles, utilizing the oracle time stamp. However, during the decoding phase, when the oracle speaker label was absent, we turned to spectral clustering for providing the speaker profile.
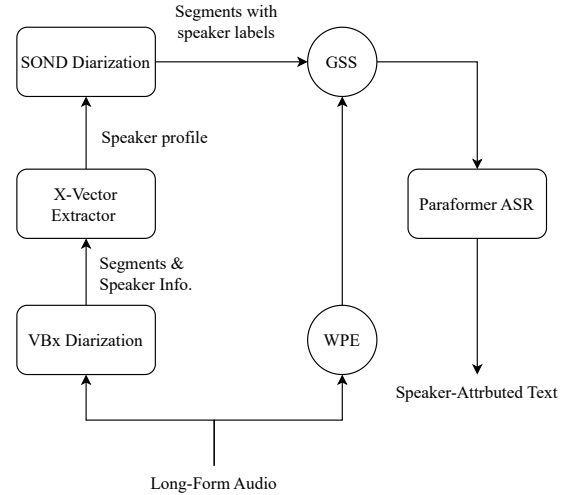
4.2. Official modular system

In addition to the E2E SA-Transformer baseline, we also develop an official modular system based on pre-trained models that serves as a strong alternative. With the goal of encouraging research and development in the field of speaker-attributed ASR, we are dedicated to releasing this system in the near future. The modular baseline is illustrated in Figure 3. Our Front-End process leverages two methods to enhance the input audio: weighted prediction error (WPE), a well-known dereverberation technique, and guided source separation (GSS), which utilizes prior segmentation information to assist the separation process. To obtain speaker-labeled segments for GSS, we employ a speaker diarization module. The input long-form audio is first processed by a VBx diarization system to obtain an initial diarization output. Using this output, an x-vector extractor generates a speaker profile for each session from non-overlapping speech segments. The speaker profiles are then utilized by a pre-trained
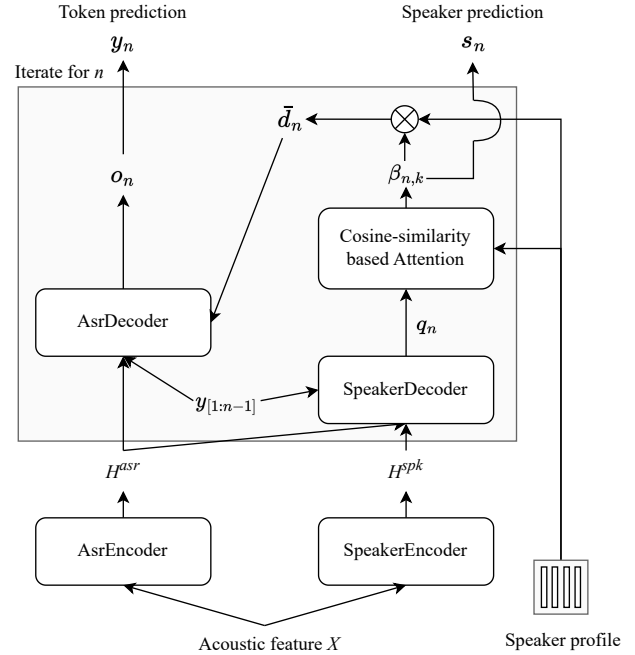
Fig. 2. The SA-Transformer [18] baseline system uses in M2MeT 2.0 challenge.

speaker overlap-aware neural diarization (SOND) [31] model to generate the final diarization output for GSS. Following GSS, we segment the audio based on the diarization output and feed it into a Paraformer for single-talker ASR with speaker labels.

Fig. 3. The official modular system basing on open-source pre-trained models.

5. EXPERIMENTAL SETUP

In this section, we describe the settings of the SA-Transformer baseline and the modular system based on the pre-train models.

5.1. SA-Transformer baseline

The ASR Encoder is composed of 12 Conformer layers, each equipped with 4-head multi-head attention (MHA). The MHA and feed-forward network (FFN) dimensions are set to 256 and 2048, respectively. The ASR Decoder is made up of 6 transformer layers. The Speaker Encoder is initialized with a pre-trained ResNet x-vector extractor, but with an additional linear layer that transforms the embedding dimension into 256. The Speaker Decoder comprises 3 transformer layers. The speaker loss weight is set to 0.5, while the CTC loss weight is set to 0.3.

5.2. Official modular system

We have implemented VBx diarization in our modular speaker-attributed system, following the workflow from the first M2MeT challenge baseline. To extract x-vectors, we use the same extractor as in SA-Transformer. When generating speaker profiles, we discard segments shorter than 0.5 seconds. Our system utilizes the pre-trained SOND and Paraformer models, which are both open-sourced and can be accessed on ModelScope. The train set of AliMeeting is utilized to fine-tune the pre-trained paraformer model for 50 epochs with a learning rate of 0.00005.

6. RESULTS AND ANLYSIS

In this section, we provide a comprehensive analysis of various systems used in this challenge. The major techniques used and evaluation results of each team is shown in Table 2. A total of 30 teams registered for this challenge and 8 of them submitted their results. Almost all teams use constraint data to build their system except C17, so they submit the same result to the two sub-tracks. While we initially released an end-to-end SA-ASR baseline, most participants opted for a modular system approach due to the long training time and limited performance with restricted data for end-to-end approaches. Moreover, with the availability of open-source pre-trained models, developing a modular system is more straightforward and yields satisfactory results. Consequently, we examine different modules, data augmentation techniques, and post-processing methods to ascertain their effectiveness in improving the overall system performance.

6.1. Speaker modules

The speaker modules play a crucial role in the speaker-attributed ASR system by providing time-stamp and speaker profiles for subsequent ASR module. The input audio is typically divided into shorter segments, and then a speaker embedding extractor, such as ResNet [32], or CAM++ [33], is employed to encode the audio into an embedding that captures the speaker characteristics. The speaker embedding is used to identify the speaker and corresponding speech segment throughout the ASR process.

The performance of the diarization model is typically measured by the diarization error rate (DER), which is impacted by both speaker activation region prediction and speaker identification accuracy. However, in the context of

speaker-attribute ASR tasks, the speaker activation region is not always necessary. As a result, we present the results of speaker counting in Table 3. Teams X27 and M42 both employ TS-VAD with different speaker embedding extractors. By utilizing the clustering result as initialization to further enhance diarization accuracy, team X27 achieves DER of 3.64% on the *Test* set. Team C17 and V29 opted for spectral clustering and VBx clustering, respectively. Due to the low overlap ratio in the *Test-2023* set, cluster-based diarization methods can also perform well on speaker counting. Team C31 and the official modular system use the SOND model and achieved an accuracy of 80% and 100% in speaker counting, respectively. The main difference is that the official modular system uses the VBx to split out the single-talker speech part which can produce more accurate speaker profiles. Benefiting from the accuracy of speaker counting, the official modular system achieves DER of 1.51% on the *Test-2023* set.

6.2. Front-end

Out of all the participating teams, only the top 2 teams use the front-end process. The winning team, X27, implements the weighted prediction error (WPE) and weighted delay-and-sum acoustic beamforming (BeamformIt) techniques, which have proven effective in recognizing far-field audio. Team M42 utilizes a Conformer-based Metric GAN (CMGAN) [34] to separate multi-talker segments into single-talker segments and compared the effectiveness of separate training and joint training strategies. Joint training of the speech separation module and ASR module results in a significant 4.2% reduction in the character error rate (cpCER) on the *Test-2023* setting (26.60% → 22.40%). The official modular system adopts WPE and GSS. GSS is proven to be an effective method in various challenges when the time stamp can be easily obtained. In the official modular system, it results in a 1.80% cpCER reduction compared to the process of WPE and BeamformIt (11.98% → 10.18%).

6.3. ASR module

Multiple ASR models are explored by the participants, including U2++ [35], Paraformer [17], MFCCA [36], and SA-Transformer [18]. In their experiment, Team M42 utilizes the U2++ and Paraformer models and assessed their performance by comparing the cpCER obtained after joint training with the speaker separation module. The results indicate that Paraformer outperforms U2++ with a lower cpCER of 20.67% as compared to 22.40%. Notably, the top four teams processed the audio into single talker segments, which could be effectively handled using a pre-trained standard ASR model.

Team X27 [37] successfully transforms a single-talker ASR model into a target speaker model [38] by incorporating the target speaker's embedding into a fully connected layer and injecting it with an element-wise product between the encoder layers. This approach enabled the model to accurately recognize the target speaker's speech with high precision.

Table 2. Results of top 5 ranking teams in terms of cpCER and their major techniques. Team code O represents the official modular system and Team code B represents the SA-Transformer baseline system

Team Code	Tracks		Data		Spk-Emb Extractor			Diarization				Front-End				ASR					Data Augmentation						Post-prosessing		cpCER(%)			
	Sub-track 1	Sub-track 2	Constraint Data	WenetSpeech	ResNet-34	ResNet-101	CAM++	TS-VAD	SOND	Spectral Clustering	VBx Clustering	WPE	CMGAN	BeamformIt	GSS	U2++	Paraformer	MFCCA	SOT	SA-Transformer	Speed Perturb.	Noise & Reverb.	Pitch Shifting	Data Simulation	Audio Codec	SpecAugment	Diarization System Fusion	ASR System Fusion	2-Speaker sessions (6)	3-Speaker sessions (6)	4-Speaker sessions (8)	Average
O	✓	✓	✓		✓	✓		✓			✓	✓			✓		✓									✓			9.01	7.34	10.31	8.84
X27	✓	✓	✓		✓			✓		✓		✓		✓		✓					✓		✓			✓	✓	✓	11.81	9.58	12.52	11.27
M42	✓	✓	✓				✓	✓		✓			✓			✓	✓				✓	✓	✓	✓	✓	✓		✓	18.34	16.99	20.86	18.64
C17		✓	✓	✓		✓					✓					✓										✓			27.01	22.98	17.7	22.83
V29		✓	✓		✓					✓							✓									✓			20.44	18.25	33.05	23.52
C31	✓	✓	✓		✓			✓										✓	✓	✓						✓			21.72	19.01	34.97	24.82
B	✓	✓	✓		✓					✓									✓	✓						✓			35.95	40.49	49.4	41.55

Table 3. The speaker counting accuracy (%) on 20 sessions in *Test-2023* set. e_{spk} and o_{spk} represent the estimated speaker number and oracle speaker number, respectively

Team code	$e_{spk}<o_{spk}$	$e_{spk}=o_{spk}$	$e_{spk}>o_{spk}$
Official	0	100	0
X27	0	90	10
M42	0	100	0
C17	10	90	0
V29	0	95	5
C31	15	80	5
Baseline	10	50	40

Team C31 implements the frame-level diarization with SOT (FD-SOT) [19] multi-talker ASR system to mitigate the issue of overlap in the ASR module. This multi-talker ASR system combines diarization and ASR results by adhering to the first-in first-out property of SOT. By splitting the ASR result using speaker change symbols and assigning speaker labels using diarization results, FD-SOT ensures one-to-one correspondence. Furthermore, the MFCCA multi-channel solution is adopted as their ASR model, which effectively leverages fine-grained channel-wise information at each time step. It should be noted that MFCCA achieves state-of-the-art level performance with limited data from AliMeeting. However, the conversion of multi-talker ASR transcripts to speaker-attributed ASR transcripts remains a challenging task. The official modular system adopts Paraformer as its ASR model. We use the AliMeeting dataset to fine-tune the pre-trained model and achieve the cpCER of 8.84%, while before fine-tuning, the cpCER is 10.18%.

6.4. Data augmentation and post-processing

Data augmentation methods are only adopted by the top 2 teams. One such team, X27, uses the non-overlapping segments in Aishell-4 and CN-Celeb as simulation data to train their TS-VAD model. They have also applied speed perturbation to the training data of their TS-ASR model. Furthermore,

Dover-lap is employed to merge the diarization result from different channels and ROVER is also employed as the final system fusion method, which results in a 0.24% cpCER reduction on the Test set, reducing it from 17.08% to 16.84%. Team M42 effectively employs a wide range of data augmentation methods, such as speed perturbation, adding noise and reverb, pitch shifting, data simulation, audio codec, and SpecAugment [39]. They have also developed U2++ and Paraformer based speaker-attributed systems, and utilize a system fusion method that depends on whether the audio has undergone speech separation. They discover that the conformer model performs better on audio that goes through speech separation while the paraformer model is more effective on audio without speech separation. Therefore, the processed audio is taken as the input of U2++, while the Paraformer takes other inputs to produce the fused hypothesis. This approach results in 2.03% cpCER reduction on the *Test-2023* set.

7. CONCLUSION

This paper provides an overview of the outcomes of the M2MeT 2.0 challenge, with a focus on the techniques used by the top-ranking teams. Given the limited training data and system-building period, leveraging open-source pre-trained models to construct a modular system is an effective approach. For speaker modules, TS-VAD and SOND are potent methods, but accurate speaker profiles are necessary for optimal performance. Front-end processing methods such as WPE, beeamformIt, CMGAN, and GSS are beneficial for transcribing far-field data, although CMGAN may have a negative effect on ASR accuracy for non-overlapped speech. In the ASR module, SOT-based methods underperform due to limited training data, while single-talker ASR models trained on a large amount of data perform well when speech is well-preprocessed in a modular system. Data simulation is less important than that in the first M2MeT challenge, as pre-trained ASR models provide sufficient initialization for fine-tuning with a small amount of data. The best-performing system achieves 8.84% cpCER given the limited training data of the challenge.

8. REFERENCES

[1] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, et al., "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*. IEEE, 2022.

[2] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Lei Xie, et al., "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. ICASSP*. IEEE, 2022.

[3] Shinji Watanabe, Michael Mandel, Jon Barker, et al., "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. CHiME*, 2020, pp. 1–7.

[4] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, Jia Pan, Jian-Qing Gao, and Cong Liu, "The first multimodal information based speech processing (MISP) challenge: Data, tasks, baselines and results," in *Proc. ICASSP*. IEEE, 2022, pp. 9266–9270.

[5] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," in *Proc. INTERSPEECH*. 2020, pp. 36–40, ISCA.

[6] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Proc. CSL*, p. 101254, 2022.

[7] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. ASRU*. 2019, pp. 296–303, IEEE.

[8] Quan Wang, Hannah Muckenhirn, Kevin W. Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez-Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*. 2019, pp. 2728–2732, ISCA.

[9] Marc Delcroix, Tsubasa Ochiai, Katerina Zmolikova, Keisuke Kinoshita, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. ICASSP*. 2020, pp. 691–695, IEEE.

[10] Yukai Ju, Wei Rao, Xiaopeng Yan, Yihui Fu, Shubo Lv, Luyao Cheng, Yannan Wang, Lei Xie, and Shidong Shang, "TEA-PSE: tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS challenge," in *Proc. ICASSP*. 2022, pp. 9291–9295, IEEE.

[11] Yukai Ju, Shimin Zhang, Wei Rao, Yannan Wang, Tao Yu, Lei Xie, and Shidong Shang, "TEA-PSE 2.0: Sub-band network for real-time personalized speech enhancement," in *Proc. SLT*. 2022, pp. 472–479, IEEE.

[12] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Y. Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. INTERSPEECH*. 2020, pp. 274–278, ISCA.

[13] Dongmei Wang, Xiong Xiao, Naoyuki Kanda, Takuya Yoshioka, and Jian Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *arxiv preprint arxiv:2208.13085*, 2022.

[14] Weiqing Wang, Xiaoyi Qin, and Ming Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *Proc. ICASSP*. 2022, pp. 9171–9175, IEEE.

[15] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*. 2020, pp. 5036–5040, ISCA.

[16] Yifan Peng, Siddharth Dalmia, Ian R. Lane, and Shinji Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*. 2022, pp. 17627–17643, PMLR.

[17] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *Proc. INTERSPEECH*. 2022, pp. 2063–2067, ISCA.

[18] Naoyuki Kanda, Guoli Ye, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "End-to-end speaker-attributed ASR with Transformer," in *Proc. INTERSPEECH*. ISCA, 2021, pp. 4413–4417.

[19] Fan Yu, Zhihao Du, Shiliang Zhang, Yuxiao Lin, and Lei Xie, "A comparative study on speaker-attributed automatic speech recognition in multi-party meetings," in *Proc. INTERSPEECH*. 2022, pp. 560–564, ISCA.

[20] Jonathan G Fiscus, Nicolas Radde, John S Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *Proc. MLMI*. Springer, 2005, pp. 369–389.

[21] Jonathan G Fiscus, Jerome Ajot, Martial Michel, and John S Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. MLMI*. Springer, 2006, pp. 309–322.

[22] Jonathan G Fiscus, Jerome Ajot, and John S Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Proc. MTPH*. 2007, pp. 373–389, Springer.

[23] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*. 2016, pp. 31–35, IEEE.

[24] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," in *arxiv preprint arxiv:2005.11262*.

[25] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, et al., "The AMI meeting corpus," in *Proc. ICMT*, 2005, p. 100.

[26] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*. IEEE, 2020, pp. 7284–7288.

[27] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, et al., "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proc. INTERSPEECH*. ISCA, 2021, pp. 3665–3669.

[28] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen, "AISHELL-4: an open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proc. INTERSPEECH*. 2021, pp. 3665–3669, ISCA.

[29] Y. Fan, J. W. Kang, L. T. Li, K. C. Li, H. L. Chen, S. T. Cheng, P. Y. Zhang, Z. Y. Zhou, Y. Q. Cai, and D. Wang, "Cn-celeb: A challenging chinese speaker recognition dataset," in *Proc. ICASSP*. 2020, pp. 7604–7608, IEEE.

[30] Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang, "Funasr: A fundamental end-to-end speech recognition toolkit," in *Proc. INTERSPEECH*. 2023, ISCA.

[31] Zhihao Du, Shiliang Zhang, Siqi Zheng, and Zhi-Jie Yan, "Speaker overlap-aware neural diarization for multi-party meeting analysis," in *Proc. EMNLP*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, Eds. 2022, pp. 7458–7469, Association for Computational Linguistics.

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

[33] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *arxiv preprint arxiv:2303.00332*, 2023.

[34] Ruizhe Cao, Sherif Abdulatif, and Bin Yang, "CMGAN: conformer-based metric GAN for speech enhancement," in *Proc. INTERSPEECH*, Hanseok Ko and John H. L. Hansen, Eds. 2022, pp. 936–940, ISCA.

[35] Di Wu, Binbin Zhang, Chao Yang, Zhendong Peng, Wenjing Xia, Xiaoyu Chen, and Xin Lei, "U2++: unified two-pass bidirectional end-to-end model for speech recognition," in *arXiv preprint arXiv:2106.05642*.

[36] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yuhao Liang, Zhihao Du, Yuxiao Lin, and Lei Xie, "MFCCA:multi-frame cross-channel attention for multi-speaker ASR in multi-party meeting scenario," in *Proc. SLT*. 2022, pp. 144–151, IEEE.

[37] Xiang Lyu, Yuhang Cao, Qing Wang, Jingjing Yin, Yuguang Yang, Pengpeng Zou, Yanni Hu, and Heng Lu, "Pp-met: a real-world personalized prompt based meeting transcription system," in *arxiv preprint arXiv:2309.16247*.

[38] Takafumi Moriya, Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, and Takahiro Shinozaki, "Streaming target-speaker ASR with neural transducer," in *Proc. INTERSPEECH*. 2022, pp. 2673–2677, ISCA.

[39] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*. 2019, pp. 2613–2617, ISCA.