SUMMARY ON THE ICASSP 2022 MULTI-CHANNEL MULTI-PARTY MEETING TRANSCRIPTION GRAND CHALLENGE

Fan Yu[1,3], Shiliang Zhang[1], Pengcheng Guo[3], Yihui Fu[3], Zhihao Du[1], Siqi Zheng[1], Weilong Huang[1], Lei Xie[3],*
Zheng-Hua Tan[5], DeLiang Wang[6], Yanmin Qian[7], Kong Aik Lee[8], Zhijie Yan[1], Bin Ma[2], Xin Xu[4], Hui Bu[4]

[1]Speech Lab, Alibaba Group, China [2]Speech Lab, Alibaba Group, Singapore
[3]AISHELL Foundation [4]Beijing Shell Shell Technology Co., Ltd., Beijing, China
[5]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[6]Department of Computer Science and Engineering,
Center for Cognitive and Brain Sciences, The Ohio State University, USA
[7]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[8]Aural and Language Intelligence Department, Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

The ICASSP 2022 Multi-channel Multi-party Meeting Transcription Grand Challenge (M2MeT) focuses on one of the most valuable and the most challenging scenarios of speech technologies. The M2MeT challenge has particularly set up two tracks, speaker diarization (track 1) and multi-speaker automatic speech recognition (ASR) (track 2). Along with the challenge, we released 120 hours of real-recorded Mandarin meeting speech data with manual annotation, including far-field data collected by 8-channel microphone array as well as near-field data collected by each participants' headset microphone. We briefly describe the released dataset, track setups, baselines and summarize the challenge results and major techniques used in the submissions.

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

1. INTRODUCTION

Who spoke what at when is the major aim of rich transcription of real-world multi-speaker meetings. Despite years of research [1, 2, 3], meeting rich transcription is still considered as one of the most challenging tasks in speech processing due to free speaking styles and complex acoustic conditions, such as overlapping speech, unknown number of speakers, far-field attenuated speech signals in large conference rooms, noise, reverberation, etc. As a result, tackling the problem requires a well-designed speech system with multiple related speech processing components, including but not limited to front-end signal processing, speaker identification, speaker diarization and automatic speech recognition (ASR).

The recent advances of deep learning has boosted a new wave of related research on meeting transcription, including speaker diarization [4, 5, 6], speech separation [7, 8, 9] and multi-speaker ASR [10, 11, 12]. The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge (M2MeT) [1] was designed with the aim to provide a common evaluation platform and a sizable dataset for Mandarin meeting transcription [13]. Along with the challenge, we made available the *AliMeeting* dataset to the participants, which contains 120 hours real meeting data recorded by 8-channel directional microphone array and headset microphone. Two tracks are

particularly designed. Track 1 is speaker diarization, in which participants are tasked with addressing the "who spoke when" question by logging speaker-specific speech events on multi-speaker audio data. Track 2 focuses on transcribing multi-speaker speech that may contain overlapped segments from multiple speakers.

This paper summarizes the challenge outcomes. Specifically, we give a brief literature overview on speaker diarization and multi-speaker ASR in Section 2. Section 3 reviews the released dataset and the associated baselines. Sections 4 and 5 discuss the outcome of the challenge with major techniques and tricks used in submitted systems. Section 6 concludes the paper.

2. RELATED WORKS

Speaker diarization and multi-speaker ASR in meeting scenarios have attracted increasing attention. For speaker diarization, conventional clustering-based approaches usually contain a speaker embedding extraction step and a clustering step where the input audio stream is first converted into speaker-specific representation [14], followed by a clustering process, such as Variational Bayesian HMM clustering (VBx) [15], which aggregates the regions of each speaker into separated clusters. The clustering based approach is ineffective to recognize the overlapped speech without additional modules, because it assumes that each speech frame corresponds to only one of the speakers. Therefore, resegmentation was used by [16] to handle the overlap segments. However, overlapping speech detection (OSD) [17] is also a challenging task itself in most situations. Compared with majority studies that work on two-talker telephony conversations, there is a recent trend to handle more challenging speaker dairization scenarios in complicated talking and acoustic environments [18, 19, 20, 21]. For the multi-speaker meetings recorded with a microphone array from distance, the scenario considered in this challenge, speaker dairization becomes more challenging as speaker overlaps happen more frequently and sometimes several speakers speak at the same time in a conference discussion.

The advances of deep learning have shed light on the problem. As a typical solution, recent end-to-end neural diarization (EEND) [5] and its variants [6] have replaced the individual submodules in traditional speaker diarization systems mentioned above with one neural network that directly provides the overlap-aware diarization results. More promisingly, thanks to the advances of speaker embedding extraction [22, 14], target speaker vocie actvity

---

detecion (TS-VAD) [23, 24] was proposed to judge target speaker's activeness for each speech frame, which can estimate multiple speakers at the same time, leading to a promising solution to handle overlapped speech.

With the development of deep learning, end-to-end neural approaches have rapidly gained prominence in the speech recognition community [25]. However, ASR in complicated scenarios such as meetings is still not a solved problem with challenges including complex acoustic conditions, unknown number of speakers and overlapping speech. In other words, the challenges mentioned above in speaker diarization also exist in multi-speaker ASR. Besides multi-condition training and data augmentation, speech enhancement [26] and separation [27, 28] are considered as remedy to the complex acoustic conditions and multiple speakers. Speech enhancement that explicitly addresses the background noise has been widely studied [18, 29, 30], where multi-channel signals can be adopted if microphone array is deployed for audio recording. Speech separation and joint-training with ASR under permutation invariant training (PIT) scheme were studied to achieve a high-performance ASR system for overlapped speech [31]. Designing an end-to-end system that directly outputs multi-speaker transcriptions seems a straightforward solution to multi-talker ASR, such as the multi-channel input multi-speaker output (MIMO) approach [32, 33] and the end-to-end unmixing, fixed-beamformer and extraction (E2E-UFE) system [34]. Conditional chain [35] was also proposed to solve the PIT problem of the number of speakers is unknown. However, the above approaches rely on complicated joint training of front and back-end models or re-designing a complicated neural architecture. The SOT method [12] does not change the original ASR network structure designed for single speaker. Instead, it only introduces a special symbol to represent the speaker change. Moreover, speech recognition and diarization for unsegmented multi-talker recordings with speaker overlaps was discussed in the recent JSALT workshop to further promote reproducible research in this field [36].

Different from relevant datasets that have been released before [18, 37, 38], AliMeeting released in this challenge and Aishell-4 [39] are currently the only publicly available meeting datasets in Mandarin. Specifically, AliMeeting has more speakers and meeting venues, while particularly adding multi-speaker discussions with a high speaker overlap ratio.

3. DATASETS, TRACKS AND BASELINES

As described in our challenge evaluation plan [13], AliMeeting contains 118.75 hours [2] of speech data in total. The training set (Train) and evaluation set (Eval) are first released to participants for system development, with 104.75 and 4 hours of speech, respectively, with manual transcription and timestamp. During the challenge ranking period, the 10 hours test set (Test) is released for scoring. Specifically, the Train, Eval and Test sets contain 212, 8 and 20 meeting sessions respectively, and each session consists of a 15 to 30-minute discussion by 2-4 participants. To highlight speaker overlap, the sessions with 4 participants account for 59%, 50% and 57% sessions in Train, Eval and Test, respectively. For Train and Eval sets, we provide the 8-channel audio recorded from the microphone array in far-field as well as the near-field audio from the participant's headset microphone, while the Test set only contains the 8-channel far-field audio.

The challenge consists of two tracks, namely speaker diarization (track 1) and multi-speaker ASR (track 2), measured and ranked on the Test set by Diarization Error Rate (DER) and Character Error Rate (CER) respectively. For both tracks, we also set up two

---

[2]Hours are calculated in single channel of audio.

---

sub-tracks. For the constrained data sub-track, system building for both tracks are restricted to AliMeeting [13], Aishell-4 [39] and CN-Celeb [40], while for the unconstrained data track, participants can use any data set publicly available.

We release baseline systems along with the Train and Eval data for quick start and reproducible research. For the 8-channel data of AliMeeting recorded by microphone array, we select the first channel to obtain *Ali-far*, and adopt CDDMA beamformer [41, 42] on 8-channel data to generate *Ali-far-bf*. We use prefix *Train-**, *Eval-** and *Test-** to denote generated data associated with Train, Eval and Test sets. For example, *Test-Ali-far-bf* means the beamformed data for the Test set.

We adopt the Kaldi-based diarization system from the CHiME-6 challenge as the baseline system for track 1. The diarizaiton module includes speaker embedding extractor and clustering. DER is scored with collar size of 0 and 0.25 second, but the challenge ranking is based on the 0.25 second collar size. The speaker diarization results for the baseline system are shown in Table 1.

Table 1. Speaker diarization results on Eval and Test in DER (%).

Testing data	Collar size = 0	Collar size = 0.25
Eval-Ali-far	**24.52**	**15.24**
Eval-Ali-far-bf	24.67	15.46
Test-Ali-far	**24.95**	**15.60**
Test-Ali-far-bf	25.16	15.79

We use a Conformer-based [43] ASR model as our single speaker baseline (ConfomerA), which is trained by *Train-Ali-near* set using ESPnet [44]. We adopt Serialized Output Training (SOT) [12] to recognize speech from multiple speakers containing overlapped speech, generating transcriptions of multiple speakers one after another. The baseline results of multi-speaker ASR are shown in Table 2. Note that here SOT and SOT_bf are trained by *Train-Ali-near* and *Train-Ali-far* respectively. Compared with the single-speaker conformer model (ConfomerA), the two SOT multi-speaker models have obtained significant improvement on the Eval and Test sets, where SOT_bf achieves superior performance.

More details on the data arrangements, tracks and baseline results can be referred to the challenge evaluation plan paper [13].

Table 2. Multi-speaker ASR results on Eval and Test in CER (%).

Testing data	ConfomerA	SOT	SOT_bf
Eval-Ali-far	49.0	30.8	34.3
Eval-Ali-far-bf	45.6	33.2	**29.7**
Test-Ali-far	50.4	32.4	35.9
Test-Ali-far-bf	46.3	33.9	**30.9**

4. SUMMARY ON TRACK 1 - SPEAKER DIARIZATION

Finally 14 teams submitted their results to track 1 and the DER for the top 8 teams is summarized in Table 3. Observing the performance by the number of speakers, we can see that in general, the DER increases with the number of speakers in meeting sessions. For most teams, there are clear performance gaps between 2- and 3-speaker sessions and between 3- and 4-speaker sessions. The winner goes to team A41 [45] which achieves the lowest DER of 2.98%, surpassed the official baseline (15.60%) with a large margin. Interestingly, their system works equally well on both 3- and 4-speaker sessions. There are two key techniques ensuring their superior performance: using TS-VAD [23] to find speaker overlap and employing cross-channel self-attention [46] to further improve performance. Table 3 also summarizes the major techniques used by the top 8

9157

Table 3. Top 8 ranking teams in terms of DER in track 1 and their major techniques.

Team Code	Main Approach			Data Augmentation			Front-End processing					Post-Processing			DER (%)			
	VBx Clustering	OSD	TS-VAD	Noise&Reverb.	Speed Perturb.	Data Simulation	Beamforming	DOA	WPE	Separation	Neural Front-End	Re-Clustering	Oracle VAD Proofreading	DOVER-Lap	2-Speaker	3-Speaker	4-Speaker	Average
A41	✓		✓	✓	✓	✓					✓			✓	0.15	5.19	4.37	2.98
V52	✓	✓	✓	✓		✓	✓	✓				✓		✓	0.62	3.51	7.06	3.98
X22	✓		✓	✓		✓			✓				✓	✓	0.44	3.67	7.28	4.05
C16	✓	✓		✓	✓			✓	✓					✓	0.88	5.20	13.58	7.23
Q36	✓	✓		✓	✓	✓	✓		✓	✓		✓		✓	0.51	4.27	15.14	7.58
K56	✓	✓	✓	✓										✓	2.30	9.11	16.35	9.73
S76	✓	✓		✓			✓	✓			✓			✓	0.65	4.97	19.81	9.80
Q19	✓	✓		✓	✓							✓		✓	2.48	6.26	18.88	10.32
Baseline	✓						✓								3.00	6.36	30.59	15.60

teams, namely effective main approach, data augmentation strategy, front-end processing as well as post-processing. We will highlight these key techniques in the following.

4.1. Main Approach

With the assumption that each speech frame corresponds to only one speaker, a clustering-based speaker diarization system is incapable of handling overlapped speech without additional modules. Since AliMeeting has a high ratio of speaker overlap, it is beneficial to adopt effective methods to reduce the error brought by the overlapped speech. The top three teams all employ TS-VAD to find the overlap between speakers while end-to-end approach, e.g., EEND, is not considered. We believe that this is because the single-speaker speech segments in meeting recordings can be effectively used (through clustering) to obtain speaker embedding as the initial input for the TS-VAD model that has been proven consistently effective for handling overlapped speech in the literature. Instead of using the original TS-VAD that takes i-vector as target-speaker embedding, the winner team A41 [45] uses the deep speaker embedding extracted by ResNet [47] to detect the target-speaker. Moreover, with the premise that different acoustic features are complementary, the second-place team V52 [48] proposes a multi-level feature fusion mechanism for TS-VAD, and the fusion between spatial-related and speaker-related features leads to 2% absolute DER reduction on Eval set. Some teams adopt approaches to improve the clustering-based algorithm itself. For example, it is effective to use overlap speech detection (OSD) to divide oracle VAD segments into single speaker segments and overlapped speech segments. Moreover, estimating the direction of arrival (DOA) to distinguish different speakers by the corresponding spatial information is proven to be beneficial. Team Q36 [49] demonstrates that re-assigning speaker labels to the overlapping segments by a speech seperation method can lead to 14.32% relative DER reduction on Eval set (7.47% to 6.40%).

4.2. Data Augmentation

Since the size of the released training data is relatively small, data augmentation is adopted by most teams. For example, noise augmentation and reverberation simulation are generally used, which improves the robustness of the model modestly. Simulated room impulse response (RIR) is used to convolve with the original speech to generate data with reverberation. To further augment the training samples, Team A41, C16 and Q36 adopt the amplification and tempo (change audio playback speed but do not change its pitch) to audio signals. Moreover, as speaker overlap is salient in the data, several teams create an extra simulated dataset based on Alimeeting and CN-celeb. In detail, utterances from different speakers are randomly selected from these data, and then combined with an overlap ratio from 0 to 40%. It is also worth noticing that the winner team A41 simulates data in an online manner in order to obtain more

diverse data and stronger model robustness.

4.3. Front-End Processing

Front-end processing approaches, such as dereverberation, beamforming and speech enhancement, have proven to be effective for downstream tasks dealing with far-field speech. In the challenge, team X22 [50], C16 and Q36 adopt the weighted prediction error (WPE) based on long-term linear prediction for dereverberation, leading to an absolute 0.7% DER reduction on the Eval set. Moreover, the relevant experiments from team X22 show that the offline dereverberation mode is more effective than the online mode. Interestingly, team Q36 found that using multi-channel WPE is harmful to OSD while it is beneficial for speaker clustering and speech separation. Effective adoption of spatial information, including beamforming [51], is also mainly considered by the participants. In particular, team S76 proposes a novel architecture named discriminative multi-stream neural network (DMSNet) for overlapped speech detection. Instead of adopting beamforming, the winner team A41 employs cross-channel self-attention to integrate multi-channel signals, where the non-linear spatial correlations between different channels are learned and fused.

4.4. Post-Processing

Since the challenge does not restrict on the computation workload and system fusion, most teams employ the DOVER-Lap [52] to fuse multiple effective models. The improvement from DOVER-Lap fusion depends on the number and type of models, and the relative DER reduction ranges from 2% to 15%. Note that although conventional VBx clustering is not as good as TS-VAD, but it brings extra gain after model fusion. Re-clustering is also an effective method for conventional clustering-based speaker diarization, which is applied to further refine the number of speakers by combining the very similar clusters according to their cosine distances. Since our challenge provides oracle VAD, Team X22 fuses the results with oracle VAD by deleting wrong speech segments and labeling the silent segments, leading to 22.2% relative DER reduction on Eval set (13.04% to 10.14%).

5. SUMMARY ON TRACK 2 - MULTI-SPEAKER ASR

Twelve teams submitted their results to track 2 and the CER for the top 5 teams is summarized in Table 4. Similar to the observation in track 1, CER sharply increases with the number of speakers in the meeting sessions, mainly due to the high speaker overlap ratio in meetings with more speakers. The winner team R62 [53] obtained the lowest average CER of 18.79% with over 12% absolute CER reduction as compared with the baseline. The superior performance comes from a SOT-based multi-speaker ASR system with large-scale data simulation. Moreover, system fusion is also beneficial as reported by the winner team. In their approach, the standard conformer-based joint CTC/Attention Conformer [43] and

9158

Table 4. Top 5 ranking teams in terms of CER in track 2 and their major techniques.

Team Code	Main Approach			Data Augmentation					Front-End Processing					Post-Processing	CER (%)			
	SOT	Conformer	U2++	Noise&Reverb.	Speed Perturb.	Pitch Shifting	Data Simulation	Spec Aug.	Beamforming	Front-&Back-Joint Model	WPE	DOA	LM	Model Fusion	2-Speaker	3-Speaker	4-Speaker	Average
R62	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	8.39	18.04	27.43	18.79
B24	✓	✓		✓	✓		✓	✓	✓	✓			✓	✓	10.49	19.80	29.50	20.81
X18	✓	✓			✓				✓	✓					10.29	18.57	35.20	23.15
G34	✓	✓		✓	✓			✓	✓		✓		✓	✓	11.69	38.12	30.42	24.81
M57	✓	✓		✓	✓			✓					✓		12.26	23.38	39.74	26.79
Baseline	✓	✓						✓	✓				✓		14.60	31.57	43.68	30.90

U2++ [54] model with a bidirectional attention decoder are fused with clear performance gain. The main approaches, data augmentation strategies, front- and back-end processing are summarized in Table 4.

5.1. Main Approach

The 5 top teams all adopt the SOT approach [12] similar to our multi-speaker baseline system, resulting in over 15% CER reduction compared with single speaker ASR system on the Eval set. The SOT method has an excellent ability to model the dependencies among outputs for different speakers and no longer has a limitation on the maximum number of speakers. Undoubtedly, the Conformer architecture [43], which models both local and global context of speech, is used by all teams. It should be noted that besides Conformer, the winner team R62 [53] also uses the recent U2++ [54] structure, where a bidirectional attention decoder is used to integrate information from both directions at inference. As a result, the fusion of the two models brings 8.7% relative CER reduction on the Eval set.

5.2. Data Augmentation

Similar to track 1, various data augmentation tricks were widely adopted in track 2. Noise augmentation, reverberation simulation, speed perturbation and SpecAugmentation are the mainstream methods with stable performance improvement. According to the report provided by second-place team B24 [55], relative CER reduction of 13.5% can be achieved by multi-channel multi-speaker data simulation as compared with the baseline trained using Train-Ali-far. Compared with speaker diarization, data simulation for multi-speaker ASR is more complex, which needs to consider various factors such as speaker turn and conversation duration. Thus fine-grained data simulation is essential to ensure consistent performance gain. For example, the simulation on speaker overlapping ratio should be reasonable, including the coverage of extreme cases like sudden (very brief) interruption from another speaker. It is also worth noticing that the winner team R62 makes substantial efforts in data augmentation and simulation. Finally, they expand the original training data to about 18,000 hours, which achieves 9.7% absolute CER reduction compared with the baseline system.

5.3. Front-End Processing

Similar to track 1, the classical front-end processing techniques in far-field speech recognition, including beamforming, dereverberation and DOA, are also adopted in track 2 with performance gain. Specifically, beamforming is used by most teams and WPE-based dereverberation is considered by two teams, while DOA estimation of target speaker is only used by the winner team R62 among the top 5 teams. Front-end and back-end joint modeling using neural networks is also considered by the second- and third-place team (B24 and X18). With the premise that optimizing front-end and back-end separately will lead to sub-optimal performance, joint modeling will make the whole system to be optimized under the final metric. Team B24 and X18 both take multi-channel signal as the input of a neural front-end and then cascade the front-end with the back-end Conformer ASR model. The whole neural architecture is then jointed trained. According to the report from B24, joint modeling leads to 13.3% relative CER reduction (from 24.0% to 20.8%) on Eval set.

5.4. Post-Processing

As reported by several teams, the contribution from language modeling (LM), either n-gram or neural LM, is very weak. This is mainly because the building of LM is only restricted to the transcripts of the training data while using extra text data is prohibited according to the challenge rule. Most teams employ model fusion which brings absolute improvement ranging from 10% to 15% on the Eval set as compared with the baseline. For example, the winner team R62 has eventually fused 7 models by simple ROVER, including 3 Conformer models and 4 U2++ models, trained with different configurations of data. Other fusion tricks include LM rescoring for single speaker and multi-speaker ASR models (team G34) and model averaging from different training stages (team B24).

6. CONCLUSIONS

This paper briefly describes the setup of the ICASSP 2022 multi-channel multi-party meeting transcription challenge (M2MeT) and summarizes the outcomes of the challenge, highlighting the major techniques used by the top performing teams. We conclude this paper with listing the following major findings. With limited a mount of data to train systems, data augmentation and simulation are effective for both speaker darization and multi-speaker ASR. Likewise, system fusion is another important trick with steady performance gain if system computational resource is not constrained. Front-end processing techniques are also beneficial for far-field scenarios including meeting transcription – the task at hand. But uniquely for meetings like the AliMeeting data, speaker overlap should be explicitly addressed. For speaker diarization, TS-VAD is still the superior approach to handle speaker overlap. By using the above-mentioned methods and tricks, the diarization error rate has been lowered to 3% on AliMeeting. For multi-speaker ASR, Conformer is still the state-of-the-art (single-speaker) ASR model used by most teams and Serialized Output Training is the easy-to-use approach to explicitly consider speaker overlap. Front-end and back-end joint modeling using neural networks is also a promising solution that deserves future investigation. The best performing system in the ASR track achieves 18.79% character error rate given the limited training data.

7. REFERENCES

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

[2] 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.

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

[4] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, et al., "A review of speaker diarization: Recent advances with deep learning," *Proc. CSL*, vol. 72, pp. 101317, 2022.

[5] 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*. IEEE, 2019, pp. 296–303.

[6] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Proc. INTERSPEECH*. ISCA, 2020, pp. 269–273.

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

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

[9] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*. IEEE, 2017, pp. 246–250.

[10] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. INTERSPEECH*. ISCA, 2017, pp. 2456–2460.

[11] Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *Proc. TASLP*, vol. 26, no. 1, pp. 184–196, 2017.

[12] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, et al., "Serialized output training for end-to-end overlapped speech recognition," in *Proc. INTERSPEECH*. ISCA, 2020, pp. 2797–2801.

[13] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al., "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," *arXiv preprint arXiv:2110.07393*, 2021.

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

[15] 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*, vol. 71, pp. 101254, 2022.

[16] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *Proc. ICASSP*. IEEE, 2015, pp. 4794–4798.

[17] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, et al., "Pyannote. audio: neural building blocks for speaker diarization," in *Proc. ICASSP*. IEEE, 2020, pp. 7124–7128.

[18] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments*, 2020.

[19] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*. ISCA, 2018, pp. 1561–1565.

[20] Neville Ryant, Kenneth Church, Christopher Cieri, et al., "The second dihard diarization challenge: Dataset, task, and baselines," in *Proc. INTERSPEECH*. ISCA, 2019, pp. 978–982.

[21] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, et al., "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[22] Ahilan Kanagasundaram, Robert Vogt, David Dean, Sridha Sridharan, and Michael Mason, "I-vector based speaker recognition on short utterances," in *Proc. INTERSPEECH*. ISCA, 2011, pp. 2341–2344.

[23] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, et al., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. INTERSPEECH*. ISCA, 2020, pp. 274–278.

[24] Maokui He, Desh Raj, Zili Huang, Jun Du, Zhuo Chen, and Shinji Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," *arXiv preprint arXiv:2108.03342*, 2021.

[25] Jinyu Li, "Recent advances in end-to-end automatic speech recognition," *arXiv preprint arXiv:2111.01690*, 2021.

[26] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH*. ISCA, 2021, pp. 2472–2476.

[27] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," in *Proc. TASLP*. IEEE, 2019, pp. 1256–1266.

[28] Yuzhou Liu and DeLiang Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. TASLP*, vol. 27, pp. 2092–2102, 2019.

[29] Jahn Heymann, Lukas Drude, Christoph Boeddeker, et al., "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *Proc. ICASSP*. IEEE, 2017, pp. 5325–5329.

[30] Yuma Koizumi, Shigeki Karita, Arun Narayanan, Sankaran Panchapagesan, et al., "Snri target training for joint speech enhancement and recognition," *arXiv preprint arXiv:2111.00764*, 2021.

[31] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, et al., "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*. IEEE, 2020, pp. 7284–7288.

[32] Xuankai Chang, Wangyou Zhang, Yanmin Qian, et al., "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. ASRU*. IEEE, 2019, pp. 237–244.

[33] Wangyou Zhang, Christoph Boeddeker, Shinji Watanabe, Yanmin Qian, et al., "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *Proc. ICASSP*. IEEE, 2021, pp. 6898–6902.

[34] Jian Wu, Zhuo Chen, Jinyu Li, et al., "An end-to-end architecture of online multi-channel speech separation," ISCA, 2020, pp. 81–85.

[35] Jing Shi, Xuankai Chang, Pengcheng Guo, Shinji Watanabe, Yusuke Fujita, Jiaming Xu, Bo Xu, and Lei Xie, "Sequence to multi-sequence learning via conditional chain mapping for mixture signals," in *Proc. NIPS*, 2020, vol. 33, pp. 3735–3747.

[36] Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, , et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. SLT*. IEEE, 2021, pp. 897–904.

[37] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, et al., "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, pp. 1–5.

[38] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, et al., "The AMI meeting corpus," in *Proc. ICMT*. Citeseer, 2005, vol. 88, p. 100.

[39] 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.

[40] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, et al., "CN-CELEB: A challenging chinese speaker recognition dataset," in *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[41] Weilong Huang and Jinwei Feng, "Differential beamforming for uniform circular array with directional microphones.," in *Proc. INTERSPEECH*. ISCA, 2020, pp. 71–75.

[42] Siqi Zheng, Weilong Huang, Xianliang Wang, Hongbin Suo, Jinwei Feng, and Zhijie Yan, "A real-time speaker diarization system based on spatial spectrum," in *Proc. ICASSP*. IEEE, 2021, pp. 7208–7212.

[43] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*. ISCA, 2020, pp. 5036–5040.

[44] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, , et al., "ESPnet: End-to-End speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[45] Weiqing Wang, Xiaoyi Qin, and Ming Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for m2met challenge," in *arXiv preprint arXiv:2202.02687*, 2022.

[46] Feng-Ju Chang, Martin Radfar, Athanasios Mouchtaris, and Maurizio Omologo, "Multi-channel transformer transducer for speech recognition," *arXiv preprint arXiv:2108.12953*, 2021.

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

[48] Naijun Zheng, Na Li, Xixin Wu, Lingwei Meng, et al., "The cuhk-tencent speaker diarization system for the icassp 2022 multi-channel multi-party meeting transcription challenge," in *arXiv preprint arXiv:2202.01986*, 2022.

[49] Jingguang Tian, Xinhui Hu, and Xinkang Xu, "Royalflush speaker diarization system for icassp 2022 multi-channel multi-party meeting transcription challenge," in *arXiv preprint arXiv:2202.04814*, 2022.

[50] Maokui He, Xiang Lv, Weilin Zhou, et al., "The ustc-ximalaya system for the icassp 2022 multi-channel multi-party meeting transcription (m2met) challenge," in *arXiv preprint arXiv:2202.04855*, 2022.

[51] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *Proc. TASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[52] Desh Raj, Leibny Paola Garcia-Perera, Zili Huang, Shinji Watanabe, Daniel Povey, et al., "Dover-lap: A method for combining overlap-aware diarization outputs," in *Proc. SLT*. IEEE, 2021, pp. 881–888.

[53] Shuaishuai Ye, Peiyao Wang, Shunfei Chen, et al., "The royalflush system of speech recognition for m2met challenge," in *arXiv preprint arXiv:2202.01614*, 2022.

[54] Di Wu, Binbin Zhang, Chao Yang, et al., "U2++: Unified two-pass bidirectional end-to-end model for speech recognition," *arXiv preprint arXiv:2106.05642*, 2021.

[55] Chen Shen, Yi Liu, Wenzhi Fan, et al., "The volcspeech system for the icassp 2022 multi-channel multi-party meeting transcription challenge," in *arXiv preprint arXiv:2202.04261*, 2022.

9160