THE CUHK-TENCENT SPEAKER DIARIZATION SYSTEM FOR THE ICASSP 2022 MULTI-CHANNEL MULTI-PARTY MEETING TRANSCRIPTION CHALLENGE

Naijun Zheng[1], Na Li[2], Xixin Wu[1], Lingwei Meng[1], Jiawen Kang[1], Haibin Wu[3]
Chao Weng[2], Dan Su[2], Helen Meng[1]

[1]The Chinese University of Hong Kong
[2]Tencent AI Lab
[3]National Taiwan University

ABSTRACT

This paper describes our speaker diarization system submitted to the Multi-channel Multi-party Meeting Transcription (M2MeT) challenge, where Mandarin meeting data were recorded in multi-channel format for diarization and automatic speech recognition (ASR) tasks. In these meeting scenarios, the uncertainty of the speaker number and the high ratio of overlapped speech present great challenges for diarization. Based on the assumption that there is valuable complementary information between acoustic features, spatial-related and speaker-related features, we propose a multi-level feature fusion mechanism based target-speaker voice activity detection (FFM-TS-VAD) system to improve the performance of the conventional TS-VAD system. Furthermore, we propose a data augmentation method during training to improve the system robustness when the angular difference between two speakers is relatively small. We provide comparisons for different sub-systems we used in M2MeT challenge. Our submission is a fusion of several sub-systems and ranks second in the diarization task.

Index Terms— M2MeT, speaker diarization, overlapped speech, multi-channel, feature fusion

1. INTRODUCTION

Speaker diarization is the task of determining "who spoke when" given a long audio [1]. It is a crucial component for meeting transcription which aims to solve the problem of "who speaks what at when" in real-world multi-speaker conversations. Recently, speaker diarization for meeting scenarios has attracted increasing research enthusiasm. However, the lack of large, public real meeting datasets has been a major obstacle for advancing diarization technologies. To address this problem, the M2MeT challenge [2] offers a sizeable corpus for research in speaker diarization and multi-speaker ASR.

This paper describes the CUHK-TENCENT diarization system submitted to Track 1 of the M2MeT challenge. Since overlapped speech spreads throughout the multiple speaker conversations [3], we focused our efforts on how to handle overlapped speech in the diarization task. We employed a VBx system with an overlapped speech detector (OSD) as our baseline. To identify speakers more accurately for overlapped speech, we modified the conventional TS-VAD system [4] by replacing the speaker detection layer with self-attention mechanism to better predict target speaker's activity. Since the spatial information associated with speakers' locations can also be used for distinguishing speakers, we estimated the direction of arrival (DOAs) of target speakers and employed our proposed target-DOA system [5] to detect their activities. Based on the assumption

that there is valuable complementary information between acoustic features, spatial-related and speaker-related features, we propose a multi-level feature fusion mechanism based target-speaker voice activity detection (FFM-TS-VAD) system to improve the performance of the conventional TS-VAD system and our previous target-DOA system. Furthermore, we propose a data augmentation strategy during training to improve the system robustness when the angular difference between two speakers is relatively small. Experimental results on evaluation set demonstrate the superiority of FFM-TS-VAD system compared to other systems.

The rest of the paper is organized as follows: Section 2 introduces the M2MeT challenge. Section 3 presents the overview of our submitted systems. Section 4 describes the proposed FFM-TS-VAD system with feature fusion and the data augmentation strategy. The experimental setup is given in section 5. The results are discussed in section 6. Finally, conclusions are drawn in section 7.

2. M2MET CHALLENGE DATASET

The M2MeT challenge provides a sizeable real-recorded Mandarin meeting corpus called AliMeeting for speaker diarization task (Track1) and ASR task (Track2). The corpus provides far-field data recorded by 8-element microphone arrays and near-field data collected by headset microphones respectively, where only far-field data is provided for testing. The AliMeeting corpus contains 240 meetings in total, where 212 meetings (104.75 hours) for training, 8 meetings (4 hours) for evaluation and 20 meetings (10 hours) for testing. Large number of speakers, various meeting venues and high overlap ratios present huge challenges in real meeting scenarios. The average speech overlap ratios of training and evaluation sets reach 42.27% and 34.76% respectively. For far-field recording, the microphone-speaker distance ranges from 0.3 to 5.0 meters. For each session, the number of speakers ranges from 2 to 4 and the speakers are required to stay in the same positions during recording.

There are two sub-tracks in the speaker diarization task. In sub-track 1, the system is required to be built from fixed constrained data, i.e., AliMeeting, AISHELL-4 [6] and CN-Celeb [7] corpus. In sub-track 2, extra data is allowed for system training. More details about the challenge data and evaluation metrics can be found in [2].

3. SYSTEM OVERVIEW

For this challenge, we submitted a system that fuses of four types of sub-systems. A brief description of each type of sub-system is shown as follows:

3.1. Baseline system

We used a conventional pipeline system composed of the VBx clustering modules [8] and an overlapped speech detector as our baseline system. Specifically, the BeamformIt [9] technique was applied to covert multi-channel signals into single-channel signals. Based on the (estimated/oracle) VAD bounds, the voiced segments were sliced into sub-segments with a fixed-length window of 1.5 seconds and a sliding overlap of 1.25 seconds. Then a speaker embedding extractor based on an 101-layer Res2Net [10] was used to extract embeddings using the inputs of 40-dimensional logarithmic Mel FBank features. Given the embeddings, the spectrum clustering (SpC) method followed by the VBx algorithm [11] was employed to perform clustering. A post-processing step called re-clustering was applied to further refining the number of speakers by combining the very similar clusters according to their cosine distances [8].

Since the conventional pipeline system cannot handle overlapped speech, a deep network composed of SincNet convolution layers and LSTM layers [12] was trained for VAD and overlapped speech detection (OSD) simultaneously. The heuristic method [8] was used to handle the overlapped speech by finding a different but closest speaker along the time axis as the second speaker.

3.2. Modified TS-VAD system

The TS-VAD system proposed by Rao et al. [4] used the i-vectors as the anchors to identify the speakers within overlapped speech, where the i-vectors were estimated based on the initial diarization annotations from the baseline system. In the original network, bi-directional LSTM (BLSTM) layers were used to learn the temporal relationship between the i-vectors and the acoustic features. Considering the strong capability of self-attention mechanism to capture the global speaker characteristics in diarization systems [13, 14], we added an encoder layer after the convolutional layers to obtain the acoustic feature embeddings. The following two BLSTM layers used for speaker detection were replaced with another two encoder layers. The encoder layers are the same as the transformer encoders [15] but without positional encoding. Two i-vector extractors were trained using 64-dimensional logarithmic Mel FBank features and 40-dimensional MFCC features, respectively.

3.3. Target-DOA system

Since the direction of speakers' sound can also be used for discriminating speakers, we proposed a Target-DOA system using the spatial features as inputs. Firstly, the DOAs of target speakers were estimated according to the initial diarization annotations. Then, the angle features (AF) [16, 17] were computed based on the DOAs and multi-channel signals, the corresponding computation process is given in Section 4.1. Finally, the magnitude spectrum and AF were concatenated as the input to the network consisting of temporal convolutional network (TCN) blocks [18].

3.4. FFM-TS-VAD system

To better utilize the complementary across acoustic features, spatial-related and speaker-related features, we propose a new FFM-TS-VAD system where a multi-level feature fusion mechanism was developed. A detailed description for FFM-TS-VAD system is displayed in Section 4
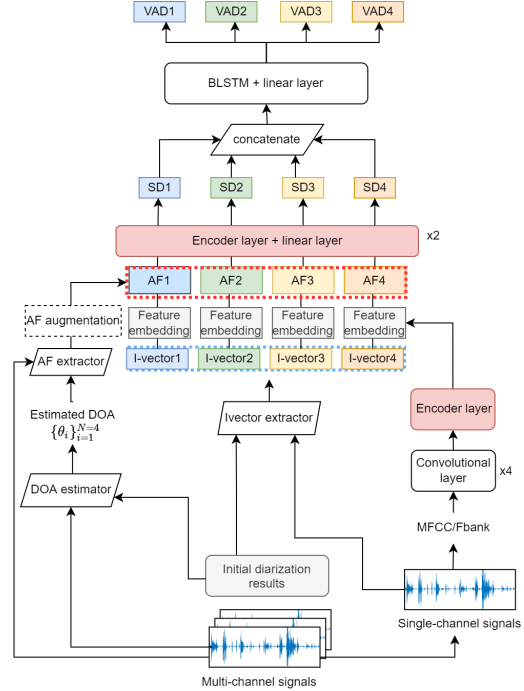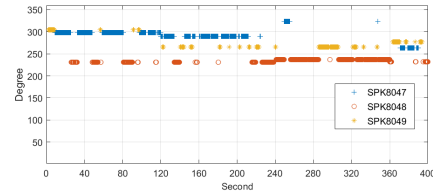

Fig. 1. Diagram of the FFM-TS-VAD network.

Fig. 2. The track of the DOA estimates in R8008_M8013 meeting.

4. PROPOSED FFM-TS-VAD SYSTEM

Figure 1 shows the diagram of the proposed FFM-TS-VAD system. Features of different levels, such as the acoustic-related feature embedding, location-guide angle feature (AF) and speaker-related i-vector, are fused as the inputs. The speaker detection (SD) information is then processed by two encoder layers and a linear layer as that in the modified TS-VAD system. Finally, the activity outputs are produced by feeding the concatenated SDs to an BLSTM layer followed by a linear layer.

4.1. Input features

The input include three types of features, i.e., the acoustic features, the AF and the i-vector. The following introduces the computation details of the used AF and i-vector.

4.1.1. Computation of AF

We first estimate the DOAs of target speakers, and then compute AF based on the estimated DOAs and multi-channel signals. To obtain accurate DOAs of the speakers, the time distribution of speakers' activities is a prerequisite. During the training stage, we use the reference annotations to obtain speaker activity information. During the

evaluation stage, we use the initial diarization annotations from our baseline system. According to the annotations, we collect the active segments of each speaker and filter out the overlapped speech and the short-duration segments. Since the speakers in M2MeT dataset were required to remain in the same positions, we then concatenate all the speech segments of the target speaker to estimate the DOAs. Given the geometric information of the microphone array, FRI-based DOA estimation algorithm [19] is implemented for the estimation.

Given the DOA θ and multi-channel signals $\mathbf{x}(t)$, we first compute the logarithmic power spectrum (LPS) $\mathbf{X}(t, f)$ from the magnitude of the spectrograms from the waveform of channels. Then, we indicate several microphone pairs $\{\bar{m} = (m_1, m_2)\}$ and compute the inter-channel phase difference (IPD) as follows:

$$\text{IPD}^{\bar{m}}(t, f) = \angle X_{m_1}(t, f) - \angle X_{m_2}(t, f), \tag{1}$$

where $\angle X_m$ denotes the phase of the complex spectrogram obtained from the m-th channel. Finally, AF can be computed as:

$$AF_\theta(t, f) = \sum_{\bar{m}} \cos(\angle v_\theta^{\bar{m}}(f) - \text{IPD}^{\bar{m}}(t, f)), \tag{2}$$

where $\angle v_\theta^{\bar{m}}(f) = 2\pi f \Delta^{\bar{m}} \cos(\theta^{\bar{m}})/c$ denotes the phase differences between the selected microphone pair for direction θ at frequency f, $\Delta^{\bar{m}}$ is the distance between the selected microphone pair, $\theta^{\bar{m}}$ is the relative angle between the look direction θ and the microphone pair $\bar{m}$, and c is the sound velocity. Given $\theta_i, i = 1, ..., N$ from i-th speaker, AF_{θ_i} contains more specific spatial information about the sound intensity than the estimated DOA θ_i at time t. In other words, if the dominating sound is from θ, then AF_θ will be close to 1.

4.1.2. Computation of i-vectors

The extraction of i-vectors was based on a gender-independent universal background model (UBM) with 512 mixtures and a total variability matrix with 100 total factors. 64-dimensional FBank and 40-dimensional MFCC were used to train the UBMs and total variability matrices, respectively.

4.2. Data augmentation strategy for AF

When spatial information is used for speaker diarization, the system performance degrades a lot in presence of those cases where different speakers are very close. Figure 2 shows the tracks of the DOA along the time axis, where SPK8047 and SPK8049 had very close DOA. Since close DOA led to similar AF, it misled the network and caused it to lose the capability to discriminate the close speakers.

Considering the complementary information between the AF and i-vectors, we aim to develop a network that can adaptively select reliable features for the difficult cases above. To increase the data for such cases based on the training set, it is necessary to exploit a data augmentation strategy during the training stage. The conventional method of using near-field recordings and artificial reverberation impulse to simulate multi-channel signals have high computational costs and may incur mismatch with the real-recorded data. We propose an efficient approach to simulate spatial features of closed speakers during training.

Given a speaker and the estimated DOA θ_i in a meeting recording, we first compute the corresponding AF_{θ_i}. Imagining that when speaker i and speaker j get closer, their originally independent AF will become more similar. When they get close enough, anyone's sound may lead to high AF. Thus, we use $\max(\cdot)$ operation to simulate this condition, where the AF features from speaker i and speaker

j can be roughly replaced with $\hat{AF}$ as follows:

$$\hat{AF}_{\theta_i}(t, f) = \max(AF_{\theta_i}(t, f), \gamma \cdot AF_{\theta_j}(t, f))$$
$$\hat{AF}_{\theta_j}(t, f) = \max(\gamma \cdot AF_{\theta_i}(t, f), AF_{\theta_j}(t, f)), \tag{3}$$

where the parameter γ controls the similarity degree. In this way, the DOAs of two speakers can also be updated as $\theta_{i,j} \leftarrow (\theta_i + \theta_j)/2 + \theta_n$, where θ_n is a random perturbation of small value.

To give more specific information to the network, we also compute the minimum angular difference of each speaker as follows:

$$\Delta\theta_i = \min_{k \in [1, N], k \neq i} |\theta_i - \theta_k|, \tag{4}$$

which are concatenated with the AF features as input.

4.3. Training loss function of the network

Since the order of the outputs for speakers matches with the order of input DOAs and i-vectors, we do not need to do permutation-invariant training. To pay more attention to the overlapped speech, we use a multi-objective loss function. First, the activities of the speech $\hat{y}^{\text{vad}}$ and overlapped speech $\hat{y}^{\text{osd}}$ can be determined from the maximum value and the second maximum value of $\hat{\mathbf{y}}^{\text{vad}} = \{\hat{y}_i^{\text{vad}}\}_{i=1}^N$ respectively, i.e.,

$$\hat{y}^{\text{vad}} = \max\{\hat{y}_i^{\text{vad}}\}_{i=1}^N, \quad \hat{y}^{\text{osd}} = \max_{2nd}\{\hat{y}_i^{\text{vad}}\}_{i=1}^N \tag{5}$$

Then, the training loss function can be written as

$$loss = \text{BCE}(\hat{\mathbf{y}}^{\text{vad}}, \mathbf{y}^{\text{vad}}) + \alpha * \text{BCE}(\hat{y}^{\text{vad}}, y^{\text{vad}})$$
$$+ \beta * \text{BCE}(\hat{y}^{\text{osd}}, y^{\text{osd}}) \tag{6}$$

where the first component loss is the binary cross entropy (BCE) loss of activities for each speaker, and the remaining loss components expect the network to focus more on VAD and OSD with weights α and β.

The speaker assignment algorithm can be simply achieved by a threshold-based algorithm on $\hat{\mathbf{y}}^{\text{vad}}$. To obtain the final segments, the short burrs and gaps can be simply discarded and filled respectively with a post-processing step [12].

5. EXPERIMENTAL SETUP

5.1. Data preparation

For sub-track 1, we used the AliMeeting, AISHELL-4 and CN-Celeb to train our speaker embedding extractors, as a result the training set contains 3,544 speakers in total. During training OSD and network-based diarization systems, only the AliMeeting and the AISHELL-4 corpora were used.

For sub-track 2, besides the above three corpora, the Voxceleb 1&2 [20, 21] corpora were also used to train the speaker embedding extractor in our baseline system. The other systems remained the same with sub-track 1.

5.2. Data augmentation for training set

During training the speaker embedding extractor, i-vector extractors and TS-VAD networks, 3-fold speed perturbation (x0.9, x1.0, x1.1) was applied to increase the number of speakers. We only performed data augmentation with the noise part in the MUSAN corpus and no reverberation was used.

Table 1. Diarization results in term of DER(%) and JER(%) with oracle VAD for the evaluation set and test set. SpC: Spectrum cluster. FA: false alarm. SC: speaker confusion. $(\cdot)\star$ denotes our baseline system to provide the initial diarization annotations. $(\cdot)^*$ denotes the meeting list which excludes the meetings whose speakers have small angular differences($\leq 45°$).

No.	Model	Input features	AF aug.	Sub-track1 on Eval.					Sub-track1* on Eval.		Sub-track2 on Eval.		Sub-track1 on Test.	
				FA	MISS	SC	DER	JER	DER	JER	DER	JER	DER	JER
1	SpC+VBx	X-vector	-	0.00	13.10	0.47	13.57	25.10	14.36	25.94	13.60	25.31	13.51	25.73
2⋆	SpC+VBx+OSD	X-vector	-	1.09	4.17	2.84	8.10	20.87	8.25	21.19	8.00	20.80	9.45	21.96
3	TS-VAD	Fbank-based i-vector	-	1.81	2.68	1.27	5.76	14.70	5.98	14.98	5.80	14.68	7.04	15.97
4	TS-VAD	MFCC-based i-vector	-	1.74	2.62	1.10	5.46	14.41	5.67	14.78	5.50	14.32	6.92	15.92
5	Target-DOA	LPS+IPD+AF	-	1.17	3.92	4.15	9.23	17.34	4.78	12.21	9.84	17.90	11.95	20.75
6	FFM-TS-VAD	AF+Fbank-based i-vector	✗	1.04	2.86	0.34	4.24	11.55	3.86	10.82	4.49	11.79	7.77	15.45
7	FFM-TS-VAD	AF+MFCC-based i-vector	✗	1.08	3.10	0.39	4.47	11.83	4.01	11.75	4.53	12.01	6.79	15.03
8	FFM-TS-VAD	AF+Fbank-based i-vector	✓	0.83	2.55	0.26	**3.64**	**10.53**	**3.31**	**10.01**	**3.91**	**10.75**	**5.63**	**12.75**
9	FFM-TS-VAD	AF+MFCC-based i-vector	✓	1.07	2.93	0.35	4.34	11.71	3.77	10.86	4.37	11.75	7.06	14.43
	Dover-Lap	*System 2 9*	-	*0.49*	*2.57*	*0.22*	*3.28*	*10.28*	-	-	*3.38*	*10.63*	*4.60*	*12.39*
	Dover-Lap	*System 2 4, 5*,6*,7*, 8 9*	-	*0.51*	*2.49*	*0.23*	*3.22*	*10.23*	-	-	*3.23*	*10.42*	*3.98*	*11.19*
	Official baseline [2]	-	-	-	-	-	*15.24*	-	-	-	-	-	*15.67*	-

While training the OSD and network-based diarization systems, we applied data augmentation to balance the probability of overlapped speech and non-overlapped speech using on-the-fly mode. 50% of the training segments were artificially made by summing two chunks cropped from the same meetings, and the signal-to-signal ratios were sampled between 0 and 10dB. During training the FFM-TS-VAD systems, 50% of the training segments were artificially augmented with Eq (3) when using the AF augmentation.

5.3. Implementation details

In our baseline systems, the threshold in the re-clustering process was set to 0.7. When computing the IPD and AF, we selected 4 microphone pairs, i.e., (1, 5), (2, 6), (3, 7) and (4, 8) in Eq (1) and Eq (2). All the networks were trained on short segments with a fixed 4-second length, and the number of the output ports N was set to 4. During evaluation, if the actual number of speakers was less than 4, we selected the speaker embeddings from the independent training dataset as 'virtual' speakers and selected directions with large angular differences from the target speakers as 'virtual' DOAs.

We used the transformer layers with 4 multi-heads and 512 units in the modified TS-VAD and FFM-TS-VAD systems while keeping the other configurations same to [4]. The γ in Eq (3) was uniformly sampled within $[0.8, 1]$ and θ_n was randomly sampled from $[0, 45°]$. For the loss function in Eq (5), α and β were manually set to 0.25. For all diarization networks, the learning rate was set to 1e-4 with Adam optimizer. ReduceLROnPlateau schedule and early stop were also adopted.

6. RESULTS AND ANALYSES

In this challenge, our system is denoted as V52 [22]. Table 1 shows the diarization results for the evaluation set in terms of diarization error rate (DER) and Jaccard error rate (JER), where the collar is set to 0.25 seconds. Oracle VAD is used in all sub-tracks.

We applied beamforming for System 1∼4 with the BeamformIt tool [9] to obtain enhanced single-channel speech. Since System 1 does not handle the overlaps, the DER is mainly from the miss rate. With the OSD applied in System 2, the miss rate decreases a lot, but the speaker confusion (SC) error increases due to inaccurate overlapped speaker identification based on the heuristic algorithm. Compared with above systems, TS-VAD systems (System 3 and 4) can obtain much better performance, especially in SC. We used single iteration to extract the i-vectors and DOAs in System 3∼9, since the initial diarization annotations obtained from System 2 already indicate the regions of overlapped speech.

System 5 uses the spatial information to handle overlapped speech. As discussed earlier, when the angular difference between the speakers is relatively small, the computed AF cannot identify the speakers correctly. To avoid such bad cases, we exclude the meetings[1] whose speakers have small angular differences ($\leq 45°$), which can be detected according to Eq (4). The results on the remaining meetings are listed in the sub-track1* column in Table 1. In this case, the DER of System 3 is better than those of the TS-VAD systems, which shows the high efficiency of spatial information in most situations.

For System 6∼9, we applied the proposed FFM-TS-VAD system using both AF and i-vectors. With feature fusion, the networks can obtain further reduction on the SC error and obtain much more accurate estimation of overlaps compared to TS-VAD and Target-DOA systems. Meanwhile, when AF augmentation strategy is applied, System 8 with AF and FBank-based i-vectors inputs can obtain relative 55.06% improvement of DER against System 2.

We also list the results in sub-track2 in Table 1. Compared with the systems in sub-track1, we only use a different speaker embedding extractor in System 1 and 2 while keeping the other networks unchanged, and similar results can be obtained.

Finally, we applied the Dover-lap algorithm with greedy mapping [23] to do system fusion and the absolute DER on evaluation set can reach 3.28%. When we use the results from sub-track1* for System 5∼6 during fusion, we observe a further reduction on DER which can be reduced to 3.22% finally. For the test set, the DER of our fusion system achieves 3.98%. Compared with the official baseline system [2] whose DER is 15.24% on evaluation set and 15.6% on test set, our fusion system achieves significant improvement.

7. CONCLUSION AND FUTURE WORK

In this paper, we present our system for overlap-aware diarization tasks in the M2MeT challenge. The proposed FFM-TS-VAD system using the multi-level feature fusion mechanism and the data augmentation strategy obtains marked improvement over conventional methods. Processing confusable spatial features extracted from the close speakers in the multi-channel speaker diarization task remains an open problem.

8. ACKNOWLEDGEMENTS

---

[1]R8008_M8013 is excluded, and 7 other meetings are retained.

9. REFERENCES

[1] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*. IEEE, 2022.

[3] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4353–4356.

[4] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri 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*, 2020, pp. 274–278.

[5] Zheng Naijun, Li Na, Yu JianWei, Weng Chao, Su Dan, Liu XunYing, and Meng Helen, "Multi-channel speaker diarization using spatial features for meetings," in *submitted to ICASSP2021*.

[6] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xu Xin, Du Jun, and Jingdong Chen, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," 2021.

[7] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, "Cnceleb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[8] Federico Landini, Ondřej Glembek, Pavel Matějka, Johan Rohdin, Lukáš Burget, Mireia Diez, and Anna Silnova, "Analysis of the but diarization system for voxconverse challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5819–5823.

[9] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[10] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[11] Mireia Diez, Lukáš Burget, Federico Landini, and Jan Černockỳ, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2019.

[12] Hervé Bredin and Antoine Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021.

[13] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *Proc. Interspeech 2019*, pp. 4300–4304, 2019.

[14] Qingjian Lin, Yu Hou, and Ming Li, "Self-attentive similarity measurement strategies in speaker diarization.," in *INTERSPEECH*, 2020, pp. 284–288.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[17] Meng Yu, Xuan Ji, Bo Wu, Dan Su, and Dong Yu, "End-to-end multi-look keyword spotting," in *INTERSPEECH*, 2020.

[18] Yi Luo and Nima Mesgarani, "Conv-tasnet Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[19] Hanjie Pan, Robin Scheibler, Eric Bezzam, Ivan Dokmanić, and Martin Vetterli, "Frida: Fri-based doa estimation for arbitrary array layouts," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 3186–3190.

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[22] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, Yanmin Qian, Kong Aik Lee, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu, "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *Proc. ICASSP*. IEEE, 2022.

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