MULTI-CHANNEL SPEAKER DIARIZATION USING SPATIAL FEATURES FOR MEETINGS

Naijun Zheng[1], *Na Li*[2], *JianWei Yu*[2], *Chao Weng*[2], *Dan Su*[2], *XunYing Liu*[1], *Helen Meng*[1]

[1] The Chinese University of Hong Kong, [2] Tencent AI Lab

ABSTRACT

Speaker identification for overlapped speech presents a great challenge for speaker diarization tasks in meeting scenarios. In order to overcome such challenges, several overlap-aware resegmentation methods based on deep learning have been integrated into speaker diarization systems. In this paper we propose two multi-channel diarization systems which have enhanced capability in detecting overlapped speech and identify speakers via learning spatial features. The first system applies a multi-look strategy to train networks without given the speakers' direction of arrival(DOA), and the other system estimates the DOA of target speakers based on existing diarization results. Both systems aim to estimate the voice activity of speakers in different directions to handle overlapped speech. Experimental results on the AMI corpus show that the relative improvements of both systems can reach 9.4% and 18.1% in term of diarization error rate (DER) against an overlap-aware single-channel system with a BeamformIt front-end.

***Index Terms*—** speaker diarization, direction of arrival, overlapped speech, multi-look, multi-channel

1. INTRODUCTION

Speaker diarization aims to determine "who spoke when" for a given utterance. It has a practical value in extracting speech of a specific speaker from spoken conversations such as meetings, interviews, broadcast news, etc. [1]. In most cases, well-trained speaker embedding extractors and clustering methods can afford the task [2]. However, overlapped speech mixing different speakers' information presents great challenges for both speaker identification [3] and diarization [4] tasks, especially in meeting scenarios.

To address these problems in overlapped speech, a popular approach is to train an overlapped speech detector to estimate the overlapped regions [5, 6]. Based on the estimated overlapped regions, there are mainly two conventional methods that assign a second speaker for overlapped speech: (1) VBx-2nd method [4] which uses the speaker labels with second largest probability for assignment; (2) the heuristic method [7] which finds the closest but different speaker along the time axis for assignment. However, using only acoustic features is often not sufficient for accurate identifications of speakers in overlapped speech.

Since the locations of speakers can be used for discrimination, several previous efforts have turned their focus to multi-channel speech for speaker diarization. When the microphone array signals are being processed by beamforming techniques to obtain enhanced single-channel signals [8, 9], spatial features (e.g. DOA) can be estimated and combined with acoustic features for clustering [10, 11]. Certain statistic methods are also applied to utilize spatial features, such as Kalman filters [10] and Hidden Markov Model (HMM) clustering [12], to track the speaker's location for diarization. However, only few previous work implements deep learning on multi-channel speaker diarization systems. In [13], a real-time speaker diarization system is enhanced by incorporating spatial information. In [14], a multi-channel target-speaker voice activity detection approach is applied by combining diarization outputs from different channels.

In this paper, we present two deep learning based diarization systems for multi-channel meeting scenarios. One system uses a multi-look strategy which gives several fixed look directions, covering the panorama, to allow the network to learn specific spatial information automatically. The other system tracks the DOA of target speakers and uses estimated DOA to obtain activity information about the speakers. To the best of our knowledge, this is among the first efforts to present end-to-end overlapped speech detection networks for multi-channel speech. The proposed systems utilize spatial information to enable the assignment of more than 2 speakers for overlapped speech and offer increased capability in speaker identification than the conventional methods. When compared with a single-channel, overlap-aware system with a BeamformIt front-end [8], the two proposed systems can provide 9.4% and 18.1% relative improvements in term of DER based on the AMI corpus[15]. We also investigate the robustness of the approaches on an out-of-domain evaluation set.

The rest of the paper is organized as follows: Section 2 introduces the network structure for multi-channel overlapped speech detection. Section 3 and Section 4 describe the algorithms for overlapped speaker assignment in the two systems respectively. The description of the diarization system and datasets are given in section 5. The experimental results of our systems are analysed in section 6. Finally, conclusions are drawn in section 7.

2. MULTI-CHANNEL OVERLAPPED SPEECH DETECTOR

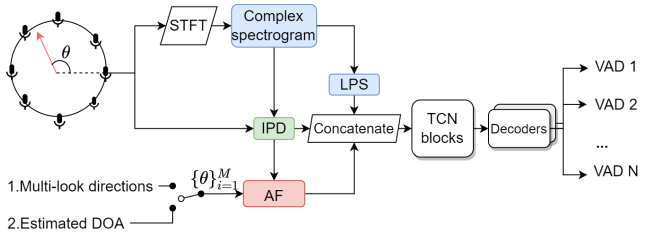

Fig. 1. Multi-channel overlapped speech detector.

Figure 1 shows the network structure for overlapped speech detection with a microphone array, where several features can be extracted from the signals. The logarithm power spectrum (LPS) is obtained from the magnitude of the spectrograms from the waveform of each channel. To obtain the spatial features, we first indicate several microphone pairs $\{\bar{m} = (m_1, m_2)\}$. Then the inter-channel phase difference (IPD) can be computed as follows:

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

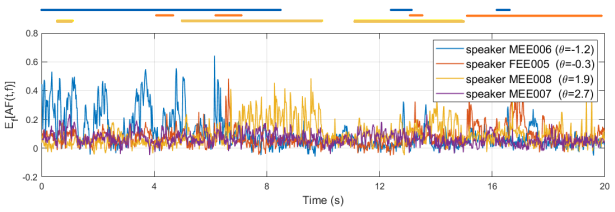

Fig. 2. AF energy distribution of different speakers along the time axis (first 20 seconds of ES2002c meeting). (The top part represents the reference diarization annotation.)

where $\angle Y_m$ denotes the phase of the complex spectrogram obtained from the m-th channel. The location-guide angle feature (AF) [16] gives more specific spatial information about the sound intensity from the look direction θ, which is computed as:

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

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. As shown in Figure 2, the AF has high correlation with the speaker activity along the time axis.

The LPS, IPD and AF are concatenated as the input to the temporal convolutional network (TCN) blocks, which have been applied in the successful Conv-Tasnet [17]. The decoders are composed of transposed convolution layers, and N output ports give the activity information of speakers in different directions, which are denoted as $\hat{\mathbf{y}}^{\text{vad}} = \{\hat{y}_i^{\text{vad}}\}_{i=1}^N$. Finally the activity of the overlapped speech $\hat{y}^{\text{osd}}$ can be determined from the second largest value, i.e., $\hat{y}^{\text{osd}} = \max_{2nd}\{\hat{y}_i^{\text{vad}}\}_{i=1}^N$.

However, in the above process, M look directions $\{\theta\}_{i=1}^M$ need to be given as the prior information. As shown in Figure 1, we apply two approaches to provide the look directions: (1) One uses the multi-look strategy [18] which gives several look directions covering the panorama and allow the network to learn the specific spatial information automatically; and (2) referred as target-DOA, uses the estimated DOA of N target speakers to obtain $\hat{\mathbf{y}}^{\text{vad}}$.

3. MULTI-LOOK SYSTEMS

The multi-look approach does not require the accurate DOA to compute AF. Instead, we select $M = 4$ look directions on the horizontal plane to cover the panorama, that is, $\theta \in \{0, 0.5\pi, \pi, 1.5\pi\}$.

3.1. Training loss function of the network

Since we do not restrict the order of the outputs for speakers, permutation invariant training (PIT) is applied to compute the loss function. Given the reference VAD of speakers $\mathbf{y}^{\text{vad}}$, it can be written as

$$loss = \min_{\text{perm}(\hat{\mathbf{y}}^{\text{vad}})} \text{BCE}(\text{perm}(\hat{\mathbf{y}}^{\text{vad}}), \mathbf{y}^{\text{vad}}) + \alpha * \text{BCE}(\hat{\mathbf{y}}^{\text{osd}}, \mathbf{y}^{\text{osd}}) \quad (3)$$

where $\text{perm}(\cdot)$ is the permutation operation for PIT, the first component loss is the binary cross entropy (BCE) loss of VAD for each speaker, and the second component loss expects the network to focus more on overlapped speech detection with weight α.

3.2. Resegmentation algorithm for overlapped speech

The speaker assignment algorithm for multi-look systems is motivated by the heuristic method [7], and we call it *heuristic++*. The underlying idea is that the speech activity detected from the same output port within a short duration is more likely to originate from the same direction (speaker). The activity regions $\mathbf{VAD}^{est}$ are first computed based on $\hat{\mathbf{y}}^{\text{vad}}$ with a threshold. Then, to incorporate the overlaps into an existing initialization diarization result $\mathbf{D}^{init}$ (e.g, the result from a VBx baseline system [7]), the resegmentation algorithm applies the following three steps.

Step 1: Align the $\mathbf{VAD}^{est}$ into a common space according to $\mathbf{D}^{init}$. Within each short duration of length d_t, select a permutation order of $\mathbf{VAD}^{est}$ that best matches the existing result $\mathbf{D}^{init}$ by computing their BCE. Based on the permutation order, assign the speaker labels from $\mathbf{D}^{init}$ to the matched active regions in $\mathbf{VAD}^{est}$.

Step 2: After processing the whole recording, search for the remaining unlabeled but active regions and assign them with the nearest speaker labels from the same output port.

Step 3: If there exists overlapped regions from different ports but labeled with the same speaker, re-allocate the overlapped regions with the nearest but different speaker labels along the time axis.

Figure 3 gives an example for our algorithm, where the circled numbers ahead of each assignment denote the step in the algorithm.

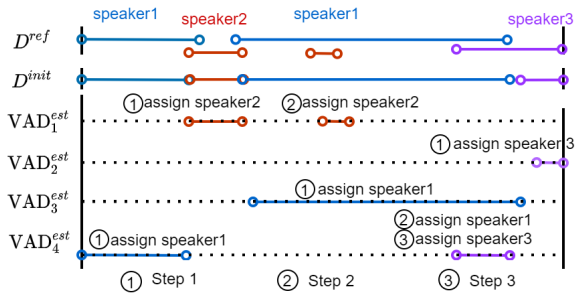

Fig. 3. An example for *heuristic++* algorithm, where D^{ref} refers to reference annotations. Note that the overlapped region in VAD_4^{est} was first assigned to Speaker 1 at Step 2 but re-allocated to Speaker 3 at Step 3 to avoid the duplication with the region in VAD_3^{est}.

4. TARGET-DOA SYSTEMS

In target-DOA systems, accurate DOA information of each speaker are required when computing the AF. To track the DOA, we first use the multiple signal classification (MUSIC) method [19, 20] to estimate DOA with a 2-second slide window. Then, we apply an existing diarization result (e.g, the result from a VBx baseline system) to label the active regions for different speakers. Figure 4 shows the estimated DOA along the time axis based on the reference diarization annotations, where each color represents a different speaker. It can be found that the speakers did not always stay at the same places but changed locations during the meeting.[1] Meanwhile, the estimated DOA often has undesirable burr which may be caused by noise, overlaps or short active duration. Thus, during evaluation, we filter out the active segments of short duration , silence and overlapped speech (if detected) according to the existing diarization results to obtain the reliable estimation of DOA. Finally, the estimated DOA is averaged for every minute to track speakers.

During training, we use a loss computation similar to Eq(3) except that no $\text{perm}(\cdot)$ is applied, and the output $\hat{\mathbf{y}}^{\text{vad}}$ should have correct activity information according to $\hat{\mathbf{y}}^{\text{vad}}$ with the given DOA, i.e., $loss = \text{BCE}(\hat{\mathbf{y}}^{\text{vad}}, \mathbf{y}^{\text{vad}}) + \alpha * \text{BCE}(\hat{\mathbf{y}}^{\text{osd}}, \mathbf{y}^{\text{osd}})$.

[1] As observed from recorded videos, the participants might exchange seats or walk to the blackboard for discussing.

Table 1. Meetings selected from the AMI corpus

	Meeting	Num.	Dur. (h)	Overlap ratio(%)
Training set	ES2002, ES2005, ES2006, ES2007, ES2008, ES2009, ES2010, ES2012, ES2013, ES2015, ES2016, EN2001b, EN2003, EN2004a, EN2005a, EN2006, EN2009	53	30.25	15.41
Development set	ES2003, ES2011	8	3.93	11.86
Evaluation set	ES2004, ES2014, EN2002	12	7.14	19.00
Evaluation set (out-of-domain)	IS1008, IS1009	8	3.37	10.59

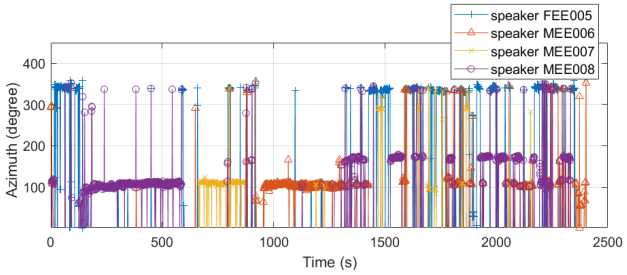

Fig. 4. Estimated DOA (ES2002c meeting).

The speaker assignment algorithm for the target-DOA systems is similar to the *heuristic*++ in multi-look systems except that at Step 1 the permutation order of $\mathbf{VAD}^{est}$ is selected within the whole duration of the meeting. For convenience, we call it *permutation* algorithm in the following experiments.

5. EXPERIMENTAL SETUP

We use the AMI corpus [15] for experimentation. The training set, development set and evaluation set are divided according to full-corpus partitions[2] which are recorded at the Edinburgh room and consist of at most four speakers in each file. The details (duration and the overlap ratios) of the datasets are shown in Table 1. To evaluate the robustness of the systems, an out-of-domain dataset composed of the meetings recorded at the Idiap room is tested with a different room size and DOA distribution. In both rooms, eight-element circular microphone arrays with 10cm radius are placed in the center of the meeting room table between the speakers. We assume that different speakers are located in different directions.

5.1. VBx baseline diarization system

In the VBx baseline diarization system [7], the segments are extracted by a fixed-length window of 1.5 seconds with a slide overlap of 1.25 seconds, and an 101-layer residual neural network (ResNet)[21] with inputs of 64-dimensional log Mel filter bank features are used for extracting embeddings. VoxCeleb1 and VoxCeleb12 [22] and CN-CELEB [23] are used as the training datasets with data augmentation from the MUSAN and RIR corpus.

For clustering, we apply the spectrum clustering method with the number of clusters set to 4, and the VBx method [2] is followed for refinement without handling overlaps.

5.2. DOA estimation and data augmentation for training set

In order to obtain reliable DOA information while training the the target-DOA systems, we use the reference diarization annotations and the moving track of the speakers observed from the videos to process the estimated DOA based on the MUSIC method. For evaluation, we use existing diarization results to process estimated DOA.

[2]https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

To balance the probability of overlapped speech and non-overlapped speech, we apply data augmentation for multi-channel speech using on-the-fly mode. 50% of the training segments are artificially made by summing two chunks cropped from the same meetings, and the signal-to-signal ratios are sampled between 0 and 5dB. We discard the artificial mixtures having the overlaps from same speakers to avoid spatial information aliasing. The training segments are of fixed-length of 4 seconds.

5.3. Implementation details

We use 257-dimensional LPSs extracted from the spectrograms with 512-length window and 50% hop ratio as input features. IPDs are computed from the 4 microphone pairs, i.e., (1, 5), (2, 6), (3, 7) and (4, 8). The networks are trained on short segments with fixed 4-second length. The number of the look directions M and the output ports N are both set to 4. The configuration of TCN blocks are given as follows: bottleneck size $B = 256$, number of channels $H = 512$ in the convolutional blocks with kernel size $P = 3$. The learning rate is set to 1e-4 with Adam optimizer. ReduceLROnPlateau schedule and early stop are also adopted.

When applying the *heuristic*++ algorithm, the segment duration d_t in Step 1 is set to 4 seconds which is equal to the length of training segments. To obtain stable DOA estimates for target-DOA systems, the duration threshold to filter out short active segments is empirical set to 5 seconds.

6. RESULTS AND ANALYSIS

6.1. Overlapped speech detection

We first evaluate the accuracy of overlapped speech detection. As shown in Table 2, setting $\alpha = 0.5$ for overlapped speech detection loss in Eq(3) can lower the error rates for all systems.

Among the multi-look systems, the accuracy can be gradually improved by concatenating the spatial features (e.g. IPD and AF) with the amplitude features (LPS). For target-DOA systems, it is straightforward to find that the quality of estimated DOA information has a high impact on the detection accuracy. Compared with the systems using reference diarization annotations, diarization results from the VBx baseline system cannot obtain very accurate DOA due to unlabeled overlaps and speaker confusion errors, which leads to high detection error rate. However, we will show later that, even with poor detection performance, the target-DOA systems still bring significant improvement to the existing baseline results on account of the directional input features.

6.2. Speaker diarization with handling overlaps

Table 3 shows the diarization results for both the development set and evaluation set in term of DER and Jaccard error rate (JER), where no forgiveness collar is used. To focus on the overlaps handling, we use oracle VAD in all systems.

Table 2. Overlapped speech detection in term of Miss(%), False alarm(%) and Error(%). *DOA_i* denotes the DOA estimated with initial diarization results from the VBx baseline system and *DOA_r* denotes the DOA estimated with reference diarization annotations.

Overlapped speech detector	Input features	α=0.0						α=0.5					
		Development set			Evaluation set			Development set			Evaluation set		
		MISS	FA	Error	MISS	FA	Error	MISS	FA	Error	MISS	FA	Error
Multi-look	LPS	7.17	1.57	8.74	11.65	1.52	13.17	6.90	1.71	8.61	11.02	1.49	12.51
Multi-look	LPS,IPD	5.81	1.43	7.24	8.25	2.54	10.78	5.49	1.42	6.91	8.05	2.26	10.31
Multi-look	LPS,IPD,AF	4.56	1.94	**6.49**	6.65	1.89	**8.54**	4.77	1.67	**6.44**	6.06	2.33	**8.39**
Target-DOA	LPS,IPD,AF, DOA_i	6.60	1.83	8.42	8.61	2.10	10.72	5.09	2.35	7.44	6.18	3.72	9.90
Target-DOA	LPS,IPD,AF, *DOA_r*	4.53	1.52	**6.05**	5.32	2.56	**7.88**	4.02	1.87	**5.89**	4.63	3.06	**7.69**

Table 3. Diarization results in term of DER(%) and JER(%) with oracle VAD. *Permutation_i* method uses the DOA estimated with the initial diarization results from System 4, *Permutation_r* method uses the DOA estimated on the reference diarization annotations and *Permutation_r_m* method uses the DOA estimated with the reference diarization annotations and movement times of speakers

No.	Use BeamformIt	Overlapped speech detector	Method	Development set					Evaluation set				
				FA	MISS	SC	DER	JER	FA	MISS	SC	DER	JER
1	✓	-	(Baseline1)	0.00	14.84	7.50	22.34	32.35	0.00	22.17	5.70	27.88	33.51
2	✓	Pyannote2.0	VBx-2nd	2.27	10.18	9.19	21.64	32.23	1.66	14.29	9.05	25.00	31.94
3	✓		Heuristic	2.26	10.21	8.77	21.24	31.46	1.66	14.31	7.66	23.63	30.49
4	✗	-	(Baseline2)	0.00	14.84	10.19	25.03	36.59	0.00	22.17	7.64	29.82	36.86
5	✗	Multi-look	VBx-2nd	1.58	7.96	12.79	22.30	35.13	1.26	15.13	10.18	26.58	34.64
6	✗		Heuristic	1.58	7.96	12.34	21.87	34.48	1.26	15.13	9.42	25.83	33.89
7	✗		Heuristic++	2.59	7.06	10.40	20.05	31.75	3.13	8.85	9.44	21.42	29.99
8	✗	Target-DOA	Permutation_i	3.44	7.33	7.67	**18.43**	**27.72**	5.30	8.72	5.35	**19.36**	**24.51**
9	✗	Target-DOA	*Permutation_r*	*4.43*	*6.33*	*5.11*	*15.87*	*23.77*	*4.20*	*6.85*	*3.34*	*14.38*	*19.87*
10	✗		*Permutation_r_m*	*2.81*	*6.04*	*3.03*	*11.87*	*16.83*	*3.99*	*6.76*	*2.69*	*13.43*	*17.57*

We apply beamforming for System 1∼3 with the BeamformIt tool [8] to obtain enhanced single-channel speech. Then, the pyannote2.0 [6] network trained with AMI Mix-Headset data is applied for overlapped speech detection. The VBx-2nd method and heuristic method applied in System 2 and System 3 reduces the miss error rate, but the speaker confusion (SC) error increases at the same time.

In other remaining systems, no BeamformIt is applied to enhance the signals. System 4 uses the first channel of the microphone array for single-channel diarization, and the results are used as the initial diarization results for the following System 5∼8. When comparing the performance of System 1 and System 4, we can find that beamforming can greatly reduce the SC error. However, only using beamforming in the front-end does not make a full use of spatial information in the diarization task. In System 5∼7, the multi-look detectors with spatial information inputs can provide much lower miss error rate, and the proposed *heuristic++* algorithm enables better speaker identification capability, the effectiveness of which is apparent when compared with conventional second speaker assignment methods. For target-DOA detectors, it is surprising to find that specifying the DOA of speakers can obtain an excellent DER reduction. Compared with the overlap-aware resegmentation algorithms in System 4∼7, using *permutation* algorithm with target-DOA detectors can have no degradation but improvement for SC error. In this way, System 8 can obtain relative 13.2% and 18.1% improvement of DER against System 3 using BeamformIt on the development and evaluation sets respectively.

6.3. Robustness of the proposed systems

In this section, we evaluate the robustness of the proposed systems, where an out-of-domain evaluation set is recorded in a different room with the training set. As shown in Table 4, the mismatch of the spatial features between the training set and evaluation set cause great degradation for the multi-look system, which reflects the weakness of the data-driven approach in out-of-domain data. However, target-DOA systems still provide robust improvement on the baseline system as long as the estimated DOA is largely correct. Compared with the single-channel system with BeamformIt and pyannote2.0 (Baseline1), target-DOA system can still obtain 10.5% and 22.2% relative improvement of DER and JER respectively.

Table 4. Diarization results in term of DER(%) and JER(%) on the out-of-domain evaluation set with oracle VAD.

Overlaps Detector	Method	Evaluation set (out-of-domain)				
		FA	MISS	SC	DER	JER
-	(Baseline1)	0.00	12.81	2.57	15.38	21.39
Pyannote2.0	Heuristic	1.02	8.21	3.89	13.12	19.79
-	(Baseline2)	0.00	12.81	2.82	15.63	21.77
Multi-look	Heuristic++	3.15	12.48	27.79	43.42	56.58
Target-DOA	Heuristic	1.29	8.08	4.34	13.71	20.59
Target-DOA	Permutation_i	2.01	7.83	1.90	**11.74**	**15.39**

7. CONCLUSION AND FUTURE WORK

In this paper, we present multi-channel systems using multi-look and target-DOA approaches for overlap-aware diarization tasks in meeting scenarios. Several speaker assignment algorithms are investigated for overlapped speech and obtain remarkable improvement against conventional methods. Experimentation shows that there is still plenty of room for improvement when comparing System 8 and System 9∼10 which use reference diarization annotations to estimate DOA. In the future, we will attempt to improve the DOA estimation method and try to use simulated speech for training.

8. ACKNOWLEDGEMENTS

This project is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N).

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] 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.

[3] Naijun Zheng, Na Li, Bo Wu, Meng Yu, JianWei Yu, Chao Weng, Dan Su, XunYing Liu, and Helen Meng, "A joint training framework of multi-look separator and speaker embedding extractor for overlapped speech," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[4] 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.

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

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

[7] Federico Landini, Ondřej Glembek, Pavel Matějka, Johan Rohdin, Lukáš Burget, Mireia Diez, and Anna Silnova, "Analysis of the BUT diarization system for voxconverse challenge," *arXiv preprint arXiv:2010.11718*, 2020.

[8] 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.

[9] Ashish Arora, Desh Raj, Aswin Shanmugam Subramanian, Ke Li, Bar Ben-Yair, Matthew Maciejewski, Piotr Żelasko, Paola Garcia, Shinji Watanabe, and Sanjeev Khudanpur, "The jhu multi-microphone multi-speaker asr system for the chime-6 challenge," *arXiv preprint arXiv:2006.07898*, 2020.

[10] Shoko Araki, Masakiyo Fujimoto, Kentaro Ishizuka, Hiroshi Sawada, and Shoji Makino, "A doa based speaker diarization system for real meetings," in *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2008, pp. 29–32.

[11] Eugene Chin Wei Koh, Hanwu Sun, Tin Lay Nwe, Trung Hieu Nguyen, Bin Ma, Eng-Siong Chng, Haizhou Li, and Susanto Rahardja, "Speaker diarization using direction of arrival estimate and acoustic feature information: The i 2 r-ntu submission for the nist rt 2007 evaluation," in *Multimodal Technologies for Perception of Humans*, pp. 484–496. Springer, 2007.

[12] Jeremy HM Wong, Xiong Xiao, and Yifan Gong, "Hidden markov model diarisation with speaker location information," in *20211 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7158–7162.

[13] Siqi Zheng, Weilong Huang, Xianliang Wang, Hongbin Suo, Jinwei Feng, and Zhijie Yan, "A real-time speaker diarization system based on spatial spectrum," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7208–7212.

[14] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, et al., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," 2020.

[15] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[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] 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.

[18] Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, and Dong Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.

[19] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[20] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

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