INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Investigation of Spatial-Acoustic Features for Overlapping Speech Detection in Multiparty Meetings

Shiliang Zhang[1], *Siqi Zheng*[1], *Weilong Huang*[1], *Ming Lei*[1], *Hongbin Suo*[1], *Jinwei Feng*[2], *Zhijie Yan*[1]

[1]Speech Lab, Alibaba Group, China
[2]Speech Lab, Alibaba Group, USA

{sly.zsl, zsq174630, yuankai.hwl}@alibaba-inc.com

Abstract

In this paper, we propose an overlapping speech detection (OSD) system for real multiparty meetings. Different from previous works on single-channel recordings or simulated data, we conduct research on real multi-channel data recorded by an 8-microphone array. We investigate how spatial information provided by multi-channel beamforming can benefit OSD. Specifically, we propose a two-stream DFSMN to jointly model acoustic and spatial features. Instead of performing frame-level OSD, we try to perform segment-level OSD. We come up with an attention pooling layer to model speech segments with variable length. Experimental results show that two-stream DFSMN with attention pooling can effectively model acoustic-spatial feature and significantly boost the performance of OSD, result in 3.5% (from 85.57% to 89.12%) absolute detection accuracy improvement compared to the baseline system.

Index Terms: overlapping speech detection, multiparty meeting, spatial spectrum, two-stream DFSMN

1. Introduction

The presence of overlapping speech segments is a common and natural phenomenon in multiparty meetings. This speech event presents a significant challenge to downstream tasks, such as speaker diarization and automatic speech recognition (ASR). For DIHARD-II challenge [1], the use of ground-truth overlap labeling decreased the Diarization Error Rate (DER) from 20.78% to 16.16% as shown in [2]. On the other hand, the underlying assumption of general ASR systems is that speech contains only one single speaker. When encountering multi-party meeting scenario, the performance of ASR system will suffer from significant degradation. Ideally, for meeting scenario, we would like to know "who spoke when" (speaker diarization) [3] and even "who spoke what at when" (speaker-attributed ASR) [4, 5]. Overlapping speech detection (OSD) [6, 7, 8] is one of the key technologies leading to this ideal system.

Over the past few decades, researchers have paid much attention to overlapping speech detection (OSD). One of the first work is [6], which investigates the impact of speaker overlap on diarization error and find that detecting overlapping speech could potentially improve diarization accuracy by 15% relatively. In [7], the overlapping speech detector is an HMM-based Segmenter that operates using various features, such as MFCC, RMS energy, LPC residual energy diarization posterior entropy (DPE). Huijbregts et al. [9] report a detection approach using a single GMM to model overlapping speech, and using Viterbi search for speaker assignment [10] presents the application of convoluted non-negative sparse coding (CNSC) to OSD. Motivated by the observation that significant portion of overlaps in spontaneous conversations take place where the amount of silence is less, an overlap detection based on silence distribution is proposed in [11].

Recently, with the development of deep learning, supervised neural networks have been applied to OSD and achieved promising results. In [12], an OSD system using Long Short-Term Memory (LSTM) recurrent neural networks is proposed. The LSTM is used as a regressor to predict frame-wise overlap scores that are employed to detect segments of overlapping speech. Experimental results on the AMI corpus [13] shown that the LSTM-based OSD system achieves a comparable performance to HMM-based system. Furthermore, the LSTM-based OSD system is integrated into the speaker diarization task to give improvement in diarization error rate in [14]. In [15], the convolutional neural networks (CNNs) are used to detect overlapped speech on independent short time-frames and investigate how the duration of the signal frame influences the accuracy of detection. In [16, 17, 18, 19], a lot of effort is paid to detect the overlapping speech as well as the concurrent speakers in speech. Perception study in [16] show that neural networks based methods can count speakers more accurately by analyzing a considerably shorter recording than human listeners.

Even with these progress, handling overlapping speech is still a challenging and open problem as shown in the DIHARD-I [20], DIHARD-II [1], DIHARD-III [21] and CHIME6 [22] challenges. One of the main obstacles is that it is difficult to obtain and label overlapping speech in real scenes, especially in far-field multi-party meetings. Previous works are mostly conducted using the AMI Corpus [13] or simulated single-channel mixed speech.

In this work, we investigate the overlapping speech detection problem in real far-field multi-party meetings. To carry out this research, we recorded 175 real multi-party meetings using an 8-microphone array. We will describe the details of data corpus in Sec.3.1. Additional to common acoustic features, we propose to utilize the spatial information provided by microphone array signal processing. As shown in later sections, the additional spatial information has been proven to be effective. Moreover, we propose a two-stream DFSMN [23] to jointly model acoustic and spatial feature. Instead of conducting frame-level OSD, we propose to perform segment-level OSD. The original speech signal is divided into segments using an NN-VAD. In order to model speech segments with variable length, we present an attention pooling method. Experimental results show that our proposed OSD system can achieve remarkable improvement, which improves detection accuracy from 85.57% to 89.12%.

2. System Description

In this section, we describe our proposed overlapping speech detection (OSD) system. We first describe the design of microphone array and the array signal processing algorithm to generate the spatial spectrum and enhanced speech. And then we move on to discuss segment-level overlapping speech detection

http://dx.doi.org/10.21437/Interspeech.2021-747

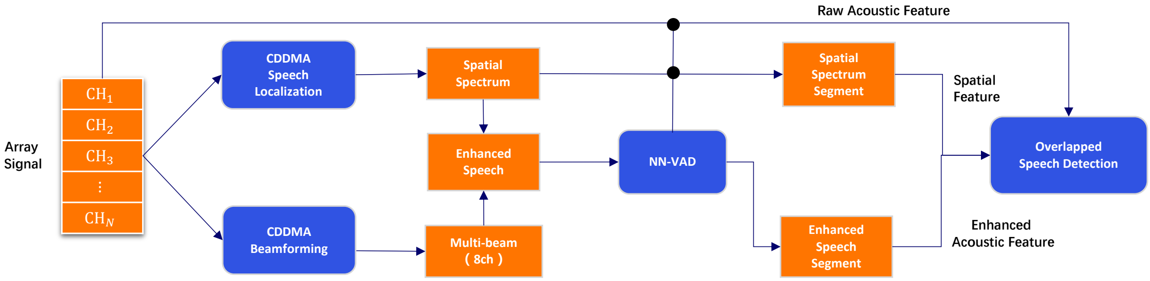

Figure 1: *Illustration of the proposed overlapped speech detection system.*

problem and how to generate the speech and spatial segments with NN-VAD. Finally, we present the proposed two-stream DFSMN based OSD system. Figure 1 is the illustration of the proposed system.

2.1. Beamforming and Source Localization

We follow the design of the set of differential beamformers similar to the multi-beam system in our previous works [24][25]. More specifically, the look-direction of each beamformer is uniformly distributed around a circle to cover the entire space, resulting in the spatially separated output signals of the beamformers.

In order to achieve better spatial-filtering performance, we utilize our recently proposed circular differential directional microphone array (CDDMA) design [26], where the microphone array is based on a uniform circular array with M directional microphones depicted in Figure 2. All the directional elements are uniformly distributed on a circle and directions are pointing outward. The CDDMA beamformer is given as below:

$$\mathbf{h}_{cddma}(\omega) = \mathbf{R}^H(\omega, \boldsymbol{\theta})[\mathbf{R}(\omega, \boldsymbol{\theta})\mathbf{R}^H(\omega, \boldsymbol{\theta})]^{-1}\mathbf{c}_{\boldsymbol{\theta}}. \quad (1)$$

where the vector $\mathbf{c}_{\boldsymbol{\theta}}$ of size $N \times 1$ defines the acoustic properties of the beamformer such as beampattern; the constraint matrix $\mathbf{R}(\omega, \boldsymbol{\theta})$ of size $N \times M$ is constructed by the directional microphone steering vector which exploits the acoustics of microphone elements. As proven in [26], the CDDMA-beamformer demonstrates significant advantages over the conventional method in terms of white noise gain (WNG) and directivity factor (DF), two commonly used performance measures for differential beamforming. WNG not only measures the efficacy to suppress spatially uncorrelated noise for beamformers [27], but also measures the robustness of beamformer [28]. And DF quantifies how the microphone array performs in the environment of reverberation[29].

For the sound source localization task, we utilize the SRP-PHAT algorithm [30] with the CDDMA-beamformer described above. At first, we assume that the microphone array signals at n^{th} frame are received as:

$$\mathbf{x}\left(\omega, \theta\right) = \left[x_1, x_2, \; \cdots \; x_M\right]^T, \quad (2)$$

where the superscript ^{T} represents the transpose operator, $\omega = 2\pi f$ is the angular frequency, f is the temporal frequency, θ is the incident angle of the signal, and x_m represents the signal of each microphone. Then, for each candidate of incident angle θ, we design each corresponding CDDMA-beamformer to target at the direction of θ, denoted as $\mathbf{h}_{cddma}(\omega, \theta)$, and we calculate the transient steering response power(SRP) at n^{th} frame as below:

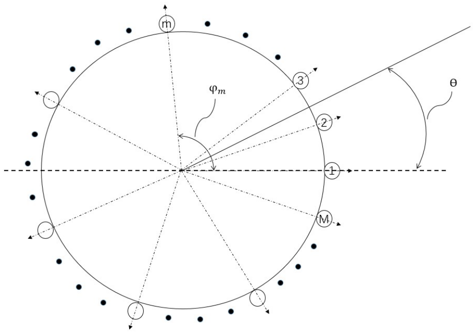

Figure 2: *Uniform circular array with directional microphones*

$$\mathbf{P}_n\left(\theta\right) = \int_{-\infty}^{+\infty} \left|\mathbf{x}\left(\omega, \theta\right)^H \mathbf{h}_{cddma}(\omega, \theta)\right|^2 d\omega. \quad (3)$$

In practice, we will take a recursive smoothing of SRP to obtain the estimate as below:

$$\hat{\mathbf{P}}_n\left(\theta\right) = \alpha\hat{\mathbf{P}}_{n-1}\left(\theta\right) + (1-\alpha)\mathbf{P}_n\left(\theta\right) \quad (4)$$

where α is called the forgetting factor, ranging from zero to one. The locale estimate of incident angle for current frame is given by:

$$\hat{\theta} = \underset{\theta}{argmax}\; \mathbf{P}\left(\theta\right) \quad (5)$$

Based on the estimate of SRP at each frame, we can form a spatial spectrum as below:

$$\mathcal{B}\left(\theta, n\right) = \hat{\mathbf{P}}_n\left(\theta\right), \forall\, n \in \mathbb{N}^+ \quad (6)$$

2.2. Segment-level Overlapping Speech Detection

In real applications, speech signals are usually split into segments with varied lengths using VAD. Instead of performing frame-level or fixed-length OSD, we try to perform varied length segment-level OSD. The enhanced speech is divided into segments using a NN-VAD. Moreover, the time tags are used to split the spatial feature as well as raw speech. As a result, we can get segment-level enhanced acoustic feature, spatial feature, as well as the raw acoustic feature. For each segment, the label is overlapping speech or non-overlapping speech. Overlapping means there is multiple-speaker co-occurrence in the speech segment. The details of data corpus and how to get the label will be introduced in Section 3.1.

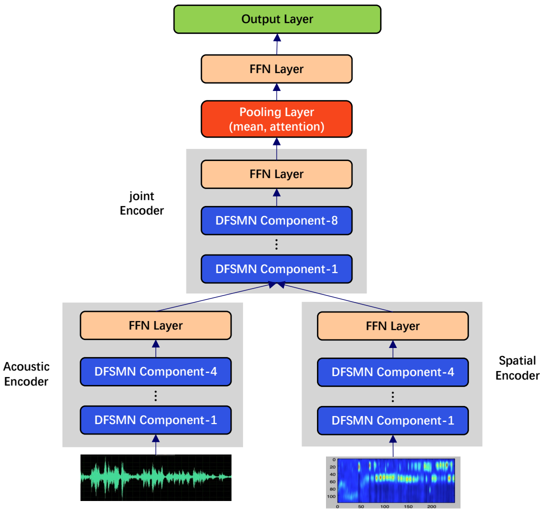

Figure 3: *Illustration of two-stream DFSMN for overlapped speech detection.*

2.3. Two-stream DFSMN

In previous works, various neural architectures have been proposed for OSD task, such as LSTM [12, 13], CNN [15] and TCN [18]. In this work, we propose the deep feed-forward sequential memory networks (DFSMN) [23] for OSD. Previous works in [23, 31, 32, 33] have shown the effectiveness of DFSMN for sequence modeling. Thereby, our purpose is not to compare the effectiveness of different network structures for OSD task, but to investigate how to effectively model acoustic and spatial features. Specially, we come up with a two-stream DFSMN with pooling (mean or attention) layer for jointly modeling acoustic-spatial features.

The architecture of two-stream DFSMN is as shown in Figure3, which consists of acoustic-encoder, spatial-encoder, joint-encoder, pooling layer and softmax output layer. The acoustic and spatial encoders are used to convert acoustic and spatial features into deep representations, respectively. These outputs are concatenated before fed into a joint-encoder. Given acoustic feature ($\mathbf{X}_a$) and spatial feature ($\mathbf{X}_s$) of a segment, the output of joint-encoder can be denoted as:

$$\mathbf{H} = f_{as}(f_a(\mathbf{X}_a); f_s(\mathbf{X}_s)). \tag{7}$$

Here, f_a, f_s and f_{as} denote the transformations of acoustic-encoder, spatial-encoder and joint-encoder, respectively.

In this work, we adopt the same architecture for acoustic and spatial encoders, which consists of four DFSMN components and a FFN layer. For the joint-encoder, the only difference is that it consists of eight DFSMN components. The FFN layer consists of a nonlinear transformation with ReLU activation and a linear transformation layer. Detailed description of DFSMN can be found in [23]. Here, we give a briefly review of the DFSMN component that the operation of the ℓ-th DFSMN component take the following form:

$$\mathbf{h}_t^\ell = \max(\mathbf{W}^\ell \mathbf{m}_t^{\ell-1} + \mathbf{b}_t^\ell, 0) \tag{8}$$

$$\mathbf{p}_t^\ell = \mathbf{V}_t^\ell \mathbf{h}_t^\ell + \mathbf{v}_t^\ell \tag{9}$$

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell} \mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell} \mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \tag{10}$$

Table 1: *Number of overlapping and non-overlapping segments.*

	Training set	Evaluation set
Overlapping	211351	6271
Non-overlapping	63410	1866

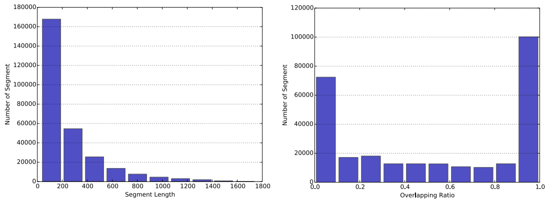

Figure 4: *Distribution of the data corpus. Left) length distribution of training dataright) Overlapping ratio distribution of training data.*

Here, $\mathbf{h}_t^\ell$ and $\mathbf{p}_t^\ell$ denote the outputs of the ReLU layer and linear layer respectively. $\mathbf{m}_t^\ell$ denotes the output of the ℓ-th memory block. N_1^ℓ and N_2^ℓ denotes the look-back and look-ahead orders of the ℓ-th memory block, respectively. s_1 is the stride factor of look-back filter and s_2 is the stride of look-ahead filter. For all the DFSMN-components in this work, we use ReLU layer with 1024 units, linear layer with 256 units, memory block with $N_1^\ell = 10$, $N_2^\ell = 10$, $s_1 = 1$ and $s_2 = 1$.

The output of the joint-encoder, denoted as $\mathbf{H}$ in Eq. 7 is fed into the *pooling layer*. In order to transform a variable length sentence into a fixed size vector representation, in addition to the commonly used mean pooling, we propose an attention pooling layer. For input $\mathbf{H} \in \mathbb{R}^{dxT}$, the operation of attention layer takes the following form:

$$\mathbf{A} = \mathrm{softmax}(\mathbf{W}_2 ReLU(\mathbf{W}_1\mathbf{H} + \mathbf{b_1}) + \mathbf{b_2}) \tag{11}$$

Here, $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_T), \mathbf{a_i} \in \mathbb{R}^{dx1}$ is the attention coefficients. We use the vector weights rather than scalars, which can control every element in all hidden vectors $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T)$. And then, we achieve the fixed size vector representation by using a weighted summation of the hidden vectors, as following:

$$\mathbf{v} = \sum_{i=1}^{T} \mathbf{a}_i \odot \mathbf{h}_i. \tag{12}$$

The fixed-size vector representation is then fed into the FFN layer and output layer. There are two units in the output layer, which denote the classification of overlapping speech and non-overlapping speech respectively.

3. Experiments

3.1. Corpus

We use the microphone array (as shown in Figure 2) to record 175 real meetings. These meetings took place in 20 different rooms. For each meeting, we placed 1 to 3 microphone arrays in different positions according to the size of the meeting room. Each of the meeting lasts about 30 minutes and has 4 participants. For each meeting, we give participants a topic, and then let them talk freely around the topic. Each participant uses a headset microphone to record near-field speech. We manually annotate the near-field speech of each participant to get the transcriptions and timestamps. Each participant's near-field marked

Table 2: *Detailed experimental results for various models with different input features and pooling method. SPA denotes the spatial feature generated by the SRP-PHAT algorithm.*

Features	Model	Mean Pooling				Attention Pooling			
		ACC(%)	F1	P	R	ACC(%)	F1	P	R
Raw Acoustic Feature	DFSMN	85.57	0.907	0.900	0.914	87.45	0.920	0.904	0.937
Enhanced Acoustic Feature	DFSMN	84.54	0.892	0.891	0.893	86.81	0.917	0.891	0.944
SPA	DFSMN	-				79.62	0.876	0.822	0.939
Raw Acoustic Feature+SPA	DFSMN	-				84.18	0.899	0.883	0.916
Raw Acoustic Feature+SPA	Two-stream DFSMN	-				**89.12**	**0.931**	**0.911**	**0.952**

timestamp will be further processed to obtain the far-field microphone array timestamp. We can then generate the segment-level overlapping speech label based on them.

We split the 175 recorded meetings into training (170 records) and evaluation (5 records) sets. Both training and evaluation sets are processed with the CDDMA beamforming (as in section 2.1) and then divided into segments with the NN-VAD. The number of overlapping and non-overlapping segments in training and evaluation sets are as listed in Table 1. During recording, we require the overlapping speech ratio of each meeting to be around 30%. However, in the final recording annotation was returned, we found that the proportion of frame-level overlapping speech was about 40%. After we split the speech into segments , we find that the segment-level overlapping ratio of training data is about 76%. Even though the overlap ratio is higher than regular meetings, it does not affect the research on overlapping speech detection (OSD). The distributions of the segment length and overlap ratio are shown in Figure 4. Overlap ratio is defined by the proportion of duration of multiple concurrent active speakers in the entire segment.

3.2. Experimental Setup

The first experiment in this work is to investigate different features as input to the overlapped speech detection (OSD) module. As shown in Figure 1, we will compare the raw acoustic feature, enhanced acoustic feature and spatial feature. For the raw speech, we choose the first channel from the 8-channel array signal. The acoustic feature is the 80-dimensional log-mel filterbank (FBK) energies computed on 25ms window with 10ms shift. We stack the consecutive frames within a context window of 7 (3+1+3) to produce the 560-dimensional features and then down-sample the inputs frame rate to 60ms. The spatial feature is the 120-dimensional spatial spectrum, which is also down-sample to the frame rate of 60ms. We use the standard DFSMN with 12 DFSMN-components. The detailed configurations of DFSMN-component are the same as described in Section 2.3.

The second experiment is to evaluate the proposed bi-stream DFSMN with acoustic-spatial feature for OSD. For comparison, we also trained a DFSMN based baseline system, using spliced acoustic-spatial features as the input.

3.3. Results and Discussions

Table 2 shows the detailed experimental results for various systems. We report the overall classification accuracy (overlapping and non-overlapping) and the F-score of the overlapping detection on the evaluation set. Comparison of different features with DFSMN model for OSD shown that the raw acoustic feature achieve the best performance. The purpose of speech enhancement is to extract the target speaker, which has an inhibitory effect on the rest of the speakers. As a result, it is not suitable for OSD task. The individual spatial spectrum features are not robust to background noise. Since the overlap is not present in the whole segment, using attention pooling can get better results than mean pooling. Figure 5 is the visualization of spectrogram, spatial spectrum and attention coefficients for a segment in evaluation set. It demonstrates that the attention pooling learns to pay more attention to overlap region in the speech segment.

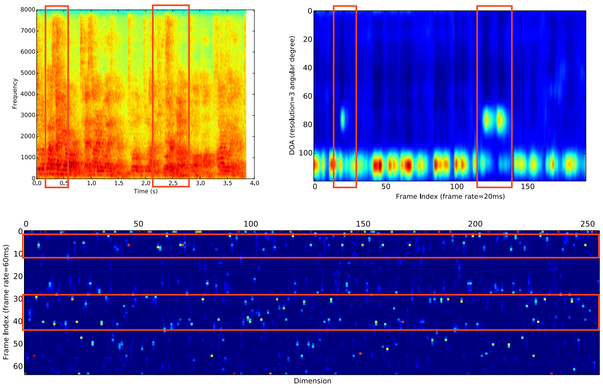

Figure 5: *From top to bottom are a)spectrogram; b)spatial spectrum of SRP-PHAT; c)visualization of attention coefficients in attention pooling layer.The red boxes represent the area of overlapping speech.*

For the proposed two-stream DFSMN based OSD system, it achieve an 89.12% classification accuracy. Compared with splicing acoustic and spatial features and then using DFSMN for modeling, using two-stream DFSMN can significantly improve the performance. Moreover, the experimental results also show the effectiveness of combining acoustic and spatial features for OSD.

4. Conclusions

In this paper we propose an overlapping speech detection system for real multiparty meeting. We focus on the segment-level OSD problem and come up with an attention pooling method to transform varied length speech segments into fixed-size vector representation. Moreover, we explore the combination of acoustic features and spatial features obtained from array signal processing for OSD. We come up with a two-stream DFSMN to jointly model acoustic-spatial features and experimental results show remarkable performance improvement. Our work is only a preliminary exploration of the benefit of array signal processing technology to OSD problems. How to make better use of the information provided by signal processing is worthy of further exploration.

5. References

[1] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "Second dihard challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.

[2] Z. Zajíc, M. Kunešová, M. Hrúz, and J. Vaněk, "Uwb-ntis speaker diarization system for the dihard ii 2019 challenge," *arXiv preprint arXiv:1905.11276*, 2019.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[4] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[5] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6630–6634.

[6] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 683–686.

[7] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. 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.

[8] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[9] M. Huijbregts, D. A. van Leeuwen, and F. Jong, "Speech overlap detection in a two-pass speaker diarization system," 2009.

[10] J. T. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 340–344.

[11] S. H. Yella and F. Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[12] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[13] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88. Citeseer, 2005, p. 100.

[14] N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging lstm models for overlap detection in multi-party meetings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5249–5253.

[15] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning." in *INTERSPEECH*, 2017, pp. 1198–1202.

[16] ——, "Overlapped speech detection and competing speaker counting—humans versus deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 850–862, 2019.

[17] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2018.

[18] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," in *INTERSPEECH 2020*, 2020.

[19] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guérin, "Multichannel crnn for speaker counting: an analysis of performance," *arXiv preprint arXiv:2101.01977*, 2021.

[20] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," *2018, tech. Rep.*, 2018.

[21] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[22] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[23] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[24] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 437–444.

[25] S. Zheng, W. Huang, X. Wang, H. Suo, J. Feng, and Z. Yan, "A real-time speaker diarization system based on spatial spectrum," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, Canada, June 6-11*, 2021.

[26] W. Huang and J. Feng, "Differential beamforming for uniform circular array with directional microphones," in *INTERSPEECH 2020*.

[27] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[28] J. Benesty, J. Chen, and C. Pan, *Fundamentals of differential beamforming*. Springer, 2016.

[29] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Wiley Online Library, 2018.

[30] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000.

[31] M. Bi, H. Lu, S. Zhang, M. Lei, and Z. Yan, "Deep feed-forward sequential memory networks for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4794–4798.

[32] S. Zhang, M. Lei, Y. Liu, and W. Li, "Investigation of modeling units for mandarin speech recognition using dfsmn-ctc-smbr," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7085–7089.

[33] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "San-m: Memory equipped self-attention for end-to-end speech recognition," *arXiv preprint arXiv:2006.01713*, 2020.