



# Spatial-aware Speaker Diarization for Multi-channel Multi-party Meeting

Jie Wang<sup>1</sup>, Yuji Liu<sup>3,4</sup>, Binling Wang<sup>2</sup>, Yiming Zhi<sup>2</sup>, Song Li<sup>1</sup>, Shipeng Xia<sup>2</sup>,  
Jiayang Zhang<sup>3,4</sup>, Feng Tong<sup>3,4</sup>, Lin Li<sup>1\*</sup>, Qingyang Hong<sup>2\*</sup>

<sup>1</sup>School of Electronic Science and Engineering, Xiamen University, China

<sup>2</sup>School of Informatics, Xiamen University, China

<sup>3</sup>College of Ocean and Earth Sciences, Xiamen University

<sup>4</sup>Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Minister of Education

{lilin, qyhong}@xmu.edu.cn

## Abstract

This paper describes a spatial-aware speaker diarization system for the multi-channel multi-party meeting. The diarization system obtains direction information of speaker by microphone array. Speaker-spatial embedding is generated by x-vector and s-vector derived from superdirective beamforming (SDB) which makes the embedding more robust. Specifically, we propose a novel multi-channel sequence-to-sequence neural network architecture named discriminative multi-stream neural network (DMSNet) which consists of attention superdirective beamforming (ASDB) block and Conformer encoder. The proposed ASDB is a self-adapted channel-wise block that extracts the latent spatial features of array audios by modeling interdependencies between channels. We explore DMSNet to address overlapped speech problem on multi-channel audio and achieve 93.53% accuracy on evaluation set. By performing DMSNet based overlapped speech detection (OSD) module, the diarization error rate (DER) of cluster-based diarization system decrease significantly from 13.45% to 7.64%.

**Index Terms:** speaker diarization, overlapped speech detection, DMSNet, ASDB, multi-channel

## 1. Introduction

As the applications of speech signal processing becoming more and more popular, the technologies, such as automatic speech recognition (ASR), speaker diarization are facing many challenges in real-world scenarios. In particular, meeting scenario is almost the most challenging and valuable because of its complexity and diversity, including overlapped speech, robustness of speaker embedding, reverberation and unknown number of speakers, etc.

Speaker diarization has traditionally been addressed as a single-channel problem, in which multi-channels of audios are transformed into mono signal by speech enhancement techniques such as beamforming methods [1][2][3]. Multi-channel audio samples are collected by microphone array that has abundant spatial location information which is beneficial to OSD task [4] and speaker representation [5]. After performing beamforming, the information will decay or even lost in back-end process of speaker diarization system including OSD and clustering. In this paper, to optimize this problem, we design a spatial-aware multi-channel multi-party speaker diarization system for meeting scenario.

\*Corresponding authors.

Table 1: Details of AliMeeting training and evaluation data [6]. The audio samples were collected in various room with different size.

	Train	Eval
Duration(h)	104.75	4.00
Session	212	8
Room	12	5
Participant	456	25
Overlap Ratio(Avg.)	42.27%	34.20%

We evaluated our diarization system on a sizeable real-recorded Mandarin meeting corpus called AliMeeting which was provided by the Multi-channel Multi-party Meeting Transcription Challenge (M2MeT) [6]. M2MeT challenge focuses on addressing the diarization problem in real-world multi-speaker meetings.

One of the difficulties in AliMeeting is overlapped speech detection. As shown in Table 1, The average speech overlap ratio of Train and Eval sets are 42.27% and 34.76%, respectively, which make this task more challenging. Overlapped speech not only cause false prediction of speakers number when clustering, but also increases the DER of system directly. In [7], a long short-term memory (LSTM) neural network has been applied to OSD. Furthermore, in order to detect the overlapped speech of speakers, many investigations have been carried out [8][9][10]. H Bredin et al. [11] applied a Bi-LSTM based OSD named PyanNet to address single-channel overlapped speech. However, the audios of AliMeeting is 8-channel far-field recording with rich spatial information, and the performance of Bi-LSTM based OSD still have room to improve. To tackle this challenging problem, we propose DMSNet for overlapped speech detection. DMSNet is a novel multi-channel sequence-to-sequence architecture used for sequence labeling task which consists of an ASDB block that extract spatial feature and a Conformer [12] encoder. Compared with Bi-LSTM based OSD model, DMSNet reduces Detection Error Rate (DetER) [13] from 42.57% to 32.47% on AliMeeting evaluation set. What's more, the audios of AliMeeting are collected from different conference rooms which are full of noise and reverberation. It is worth noting that the speakers of meeting are required to remain in the same position during recording. In order to make full use of spatial information provided by the microphone array, we fused the spatial-vector (s-vector) and x-vector of speakers to make the similarity matrix of speakers more robust. We performed direction-of-arrival (DOA) technology to extract s-vector. After stacking our methods, the DER of diarization system was

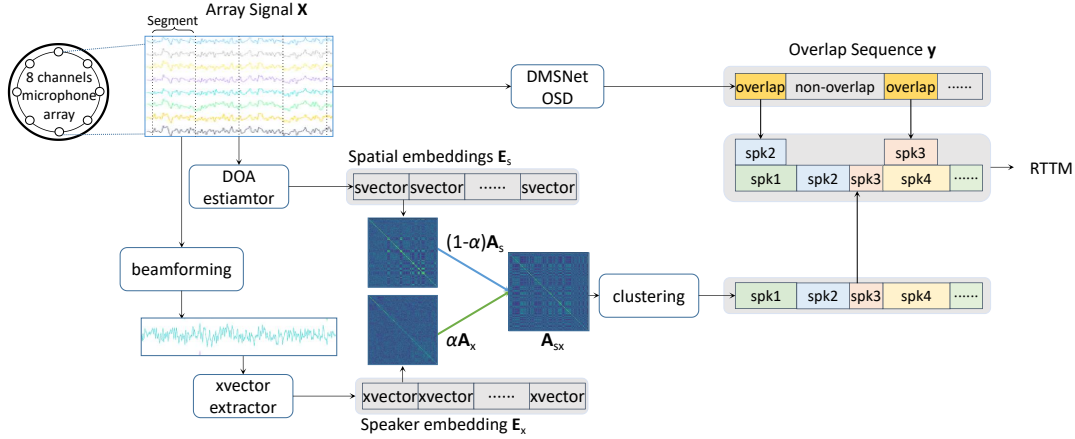


Figure 1: *Spatial-aware speaker diarization system overview. The Rich Transcription Time Marked (RTTM) is the output of speaker diarization system.*

reduced to 7.64% on AliMeeting evaluation set.

The rest of this paper is organized as follows. Section 2 gives the overview of our speaker diarization system. In Section 3, we describe training set and evaluation set of each module. Experimental setup and results are shown in Section 4. Finally, we conclude our work in Section 5.

## 2. System overview

As shown in Figure 1, our speaker diarization system consists of speaker embedding extractor, spatial embedding extractor (DOA estimator), clustering modules and OSD. We segment audio samples according to the oracle VAD label provided by M2MeT. We will describe each module in detail as follows.

### 2.1. Speaker embedding

The ResNet34-SE [14] is employed as the x-vectors extractor with additive margin softmax loss [15] (AAM-softmax) which learns a segment-level representation from the input acoustic feature. The dimension of x-vector is reduced from 256 to 128 by Linear Discriminant Analysis (LDA). In our system, the input is 81-dimensional log-mel filter-bank (FBank) extracted from the original 16kHz audio with a window size of 25ms and a 10ms shift.

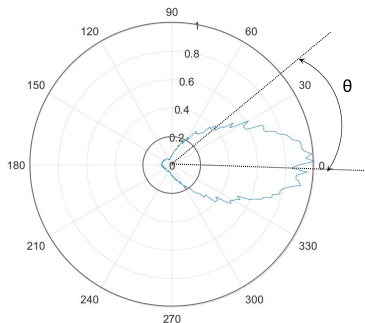


Figure 2: *A diagram of circular array microphones.*

### 2.2. Spatial embedding

The spatial embeddings are extracted by a DOA estimator. As shown in Figure 2, the look-direction of each beamformer is uniformly distributed around a circle, so as to generate the spatial separation representations of the beamformer. Superdirective beamforming (SDB) [16] algorithm is used to extract spatial representation named s-vectors on sliding windows. We set the window length and window shift to 1s and 0.5s respectively, which is consistent with speaker embedding extractor. The signal of the  $C$ -channel microphone array is denoted as  $\mathbf{x}_c(t)(c=1,2,\dots,C)$ . The final output of the beamformer can be expressed as:

$$\mathbf{Y}(\omega) = \sum_{c=1}^C \mathbf{h}_c^*(\omega) \mathbf{x}_c(\omega) \quad (1)$$

Eq.1 can be simplified into the following matrix form:

$$\mathbf{Y}(\omega) = \mathbf{H}^H(\omega) \mathbf{X}(\omega) \quad (2)$$

where  $\mathbf{H}(\omega) = [\mathbf{h}_1(\omega), \mathbf{h}_2(\omega), \dots, \mathbf{h}_C(\omega)]^T$  is the beamforming filter of the array,  $\mathbf{X}(\omega) = [\mathbf{x}_1(\omega), \mathbf{x}_2(\omega), \dots, \mathbf{x}_C(\omega)]^T$  is the received signal, and superscript ' $H$ ' represents conjugate transpose. Given  $\mathbf{H}(\omega)$ , the response in space domain of the array is:

$$\mathbf{B}(\omega, \theta) = \mathbf{H}^H(\omega) \mathbf{d}(\omega, \theta) \quad (3)$$

where  $\theta$  indicates direction,  $\mathbf{d}(\omega, \theta)$  is the steering vector of the array. For suppressing the noise that is outside the steering direction, the idea of SDB is to maximize the directivity factor. Combined with no distortion constraint in the desired direction, SDB is transformed into an optimization problem:

$$\min_{\mathbf{H}(\omega)} \mathbf{H}^H(\omega) \mathbf{R}_{\text{NN}} \mathbf{H}(\omega) \quad s.t. \mathbf{H}^H(\omega) \mathbf{d}(\omega, \theta_0) = 1 \quad (4)$$

where  $\theta_0$  is look-direction of target speaker. The filter coefficient can be obtained as:

Table 2: The topology of DMSNet architecture.

Layer	Output Size
Input Layer	$L \times C$
ASDB Block	$T \times D \times 1$
Conformer	$T \times D \times 1$
FC1	$T \times 128$
FC2	$T \times 2$
Softmax	$T \times 1$

$$\mathbf{H}(\omega) = \frac{\mathbf{R}_{\text{NN}}^{-1}(\omega)\mathbf{d}(\omega, \theta_0)}{\mathbf{d}^H(\omega, \theta_0)\mathbf{R}_{\text{NN}}^{-1}(\omega)\mathbf{d}(\omega, \theta_0)} \quad (5)$$

where  $\mathbf{R}_{\text{NN}}(\omega)$  is the covariance matrix of noise. The dimension of filter is  $(N, C, K)$  in DOA estimator. In our experiment, according to the number of audio channels, we set  $C=8$ . The order of FIR in the SDB was  $K=128$  in our experiment. We also set the number of FIR filter  $N$  to 120 (a value optimized experimentally). These filters in DOA estimator were artificially designed as  $N$  depthwise convolution with 128 kernels size. As shown in Figure 1, when the 8-channel raw audio is fed to the estimator, single-channel signals from  $N$  directions can be obtained. By normalizing the energy of signals in  $N$  directions, we extract spatial embedding (s-vector) of segmented speech. The spatial embedding  $\mathbf{e}_s = \{e_{s1}, e_{s2}, \dots, e_{sN}\}$  denotes the energy distribution of the different look-direction beamformer, where  $e_{sn}$  represents the probability of speaker in  $n$  direction. By increasing  $N$ , higher resolution can be obtained, and we find that suitable value of  $N$  will improve the robustness of s-vector.

### 2.3. Later fusion

In the embedding fusion module, we perform late fusion method [5] to construct a separate similarity matrix  $\mathbf{A}_{sx}$ . We score the cosine similarity of embeddings in pairs to get similarity matrix. We then use the following formula to yield the fused similarity matrix which act as the input of clustering module:

$$\mathbf{A}_{sx} = a\mathbf{A}_x + (1-a)\mathbf{A}_s \quad (6)$$

where  $\mathbf{A}_x$  and  $\mathbf{A}_s$  denoted the cosine similarity matrix of  $x$ -vector and  $s$ -vector, respectively.

### 2.4. Clustering

In this stage, we perform normalized maximum eigengap spectral clustering (NME-SC) [17] to obtain speaker labels of segments. NME-SC can estimate the number of clusters adaptively.

### 2.5. Overlapped speech detection

The duration of overlapped speech accounts for a large proportion in AliMeeting corpus. In our system, we take two-stage OSD method to solve overlapped speech problem. The first stage is to detect overlapped region of speech. Then, in the second stage, we use the output of heuristic algorithm [18] to obtain secondary speaker labels. The secondary speaker labels of non-overlapped region will be removed. In the first stage, we propose a novel neural network architecture named DMSNet to detect overlapped speech of multi-channel audio. DMSNet is a multi-channel sequence-to-sequence architecture which can address the sequence labeling task which matches a raw audio  $\mathbf{X}$

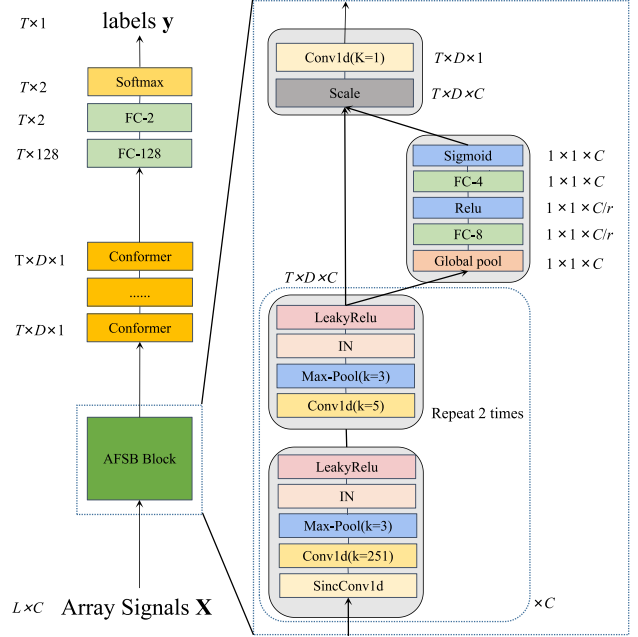


Figure 3: DMSNet is a multi-channel sequence-to-sequence architecture used for sequence labeling task.  $L$  is the length of raw audio,  $C$  is the number of channel and  $T$  is the frame number of each chunked audio. The parameter  $r$  is a reduction ration for controlling the computational cost of SE-Block. IN denote Instance Normalization layer.

$= \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_C\}$  to the corresponding label sequence  $\mathbf{y}$  where  $\mathbf{x}_c = \{x_{c1}, x_{c2}, \dots, x_{cL}\}$  is of  $L$  length raw waveform and  $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$  is  $T$  frames label. The labeling principle is that  $y_t=0$  if there is non-overlapped at time step  $t$  and  $y_t=1$  if there is overlapped. We use pyannote.audio toolkit [11] to built this module.

As shown in Figure 3, DMSNet consists of attention superdirective beamforming (ASDB) block and Conformer encoder. Inspired by SDB algorithm which designs different artificial filters for channels to enhance the signal, we propose a learnable ASDB block to overcome many shortcomings of the SDB. In ASDB block, the 8-channels raw audio is split into segments by sliding window and the feature of segments is extracted by SincNet [19]. The weights of each channels are self-adapted, which make the module learn the spatial information of speech better. Squeeze-and-excitation (SE) block [20] is applied to learn the weights of each channel and proposed to improve the representational power of a network enabling it to perform dynamic channel-wise feature. Then, we use 1-dimension convolution whose kernel size is  $1 \times 1$  to combine the latent spatial feature of different channels. We also adopt 24 layers Conformer as encoder to capture speaker-spatial information of signal from the entire sequence. Finally, the classification layer including linear and softmax layers output the label. The complete architecture for DMSNet is shown in Figure 3. We train the architecture with binary cross entropy (BCE) loss.

## 3. Data preparation

The dataset was collected by 8-channel microphone array, of which 212 audio samples are training set (Train), 8 audio samples are evaluation set (Eval). The number of speakers within one meeting session ranges from 2 to 4. Our use of training set in this challenge is as follows:

Table 3: Comparison of different OSD architecture. SDB-SincNet denotes that the inputs of SincNet were beamformed by SDB.

ID	Architecture	Extraction block	Encoder	DetER(%)	Accuracy(%)	Precision(%)	Recall(%)
M1	PyanNet [11]	SDB-SincNet	Bi-LSTM	42.57	91.61	85.23	70.06
M2	DMSNet-SDB(ours)	SDB-SincNet	Conformer	35.48	92.97	87.22	75.58
M3	DMSNet-L(ours)	ASDB	Bi-LSTM	36.68	92.69	88.18	73.12
M4	DMSNet(ours)	ASDB	Conformer	<b>32.47</b>	<b>93.53</b>	<b>89.22</b>	<b>76.81</b>

Table 4: Experimental result for various diarization system with different speaker embeddings. The collar size of DER is 0.25s.

Embedding	DER(%)	
	Score overlap	Not score overlap
s-vector	17.13	5.14
x-vector	13.81	0.87
sx-vector	<b>13.45</b>	<b>0.57</b>

- Speaker embedding extractor: We take CN-celeb1 (793 speakers) and CN-Celeb2 (2000 speaker) [21] as the training set containing 2793 speakers in total (the CN-Celeb1-test part is excluded from training).
- Overlapped speech detection: We use AliMeeting training set to train OSD models.

## 4. Experimental result and discussion

### 4.1. OSD methods

The first experiment in this work is to explore the importance of different components in DMSNet. As shown in Table 3, we design different OSD architectures with different extract block and encoder layers. The inputs of models are chunk 2s audio. The 60-dimensional speech features were extracted by SincNet. We perform SDB to enhance the 8-channel audio on AliMeeting evaluation set using PyanNet architecture and DMSNet-SDB architecture. The encoder of DMSNet-L is 2 layer Bi-LSTM with 128-dimension hidden units. Considering that the filter and time delay of each channel is discriminative in SDB method, we make the weight of each channel different in ASDB. In DMSNet-L architecture, we take ASDB as extraction block. DMSNet consists of an ASDB Extraction block and 24 layers Conformer encoder whose topological structure are the same in all OSDs.

We report the overall accuracy (overlapped and non-overlapped) precision, recall and DetER on the evaluation set. DMSNet achieve the best performance among these architectures and ASDB block is a necessary component of DMSNet. The purpose of FBS is to enhance the target speaker signal, while inhabits the rest of the speakers signals. As a result, the spatial information of array signals is lost. ASDB extraction block is based on channel attention to extract the spatial feature of speaker, which is suitable for OSD task. Conformer encoder is good at capturing local and global dependencies of sequence. By performing DMSNet, we achieve 93.53% accuracy and 32.47% DetER on evaluation set.

### 4.2. Spatial-speaker embedding fusion

The second experiment is designed to investigate different speaker embedding as input to the clustering module. In order to evaluate the robustness of various embeddings, we perform NME-SC for speaker diarization with different embeddings. In our diarization systems, the window length is 1s and the window shift is 0.5s. In this experiment, we use oracle VAD label

Table 5: Detailed experimental results for various diarization system with different OSD module. The collar size of DER is 0.25s.

ID	OSD	MISS(%)	FA(%)	SpkErr(%)	DER(%)
S1	-	12.9	0	0.5	13.45
S2	Oracle	1.8	0.0	4.3	6.10
S3	PyanNet	5.5	<b>0.3</b>	2.9	8.75
S4	DMSNet-SDB	3.6	1.0	3.6	8.17
S5	DMSNet-L	4.2	1.1	<b>2.8</b>	8.13
S6	DMSNet	<b>3.6</b>	0.6	3.5	<b>7.64</b>

and we evaluate the system with different scoring options. As shown in Table 4, the individual spatial embeddings are not robust for clustering. According to Eq. 6, we fused the spatial embeddings and speaker embedding. After tuning on training set, we set the  $a$  to 0.95. By comparing with different scoring options in Table 4, we found that overlapped speech caused the high DER. When we ignored the overlapped speech, the fusion method resulted in 34% relative improvement on evaluation set, which proved the robustness of spatial-speaker embedding.

### 4.3. Speaker diarization systems

The third experiment is to apply the proposed DMSNet OSD for speaker diarization on AliMeeting evaluation set. To investigate the importance of OSD, we also evaluate a speaker diarization system with oracle OSD label. For System 1~6, we use spatial-speaker similarity matrix as input to clustering module. Table 5 shows the results of experimental results of our systems with different OSD modules. System 1 uses NME-SC to assign the most probability speaker label to segment. The result of System 2 proves the importance of OSD module, i.e. there is still a room for improvement on the detection of overlap. Compared with System 3, the result of System 6 suggests that using DMSNet can significantly improve the performance on DER.

## 5. Conclusions

In this paper, we propose a spatial-aware speaker diarization system for multi-channel multi-party recording in real-world meeting. We present two methods to improve the performance of speaker diarization system for multi-channel audio. We also propose a sequence-to-sequence architecture named DMSNet for OSD. DMSNet is a channel-wise neural network architecture which adaptively extracts latent spatial information of array audios by modeling interdependencies between channels. We also perform a DOA estimator to extract spatial embedding of speaker. We demonstrate our speaker diarization on AliMeeting and verify its effectiveness.

## 6. Acknowledgements

Thanks to the National Natural Science Foundation of China (Grant No.61876160 and No.62001405) for funding.

## 7. References

- [1] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.
- [2] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 897–904.
- [3] 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*, vol. 1, 2018.
- [4] W. Chen, E. S. C. Van Tung Pham, E. S. Chng, and X. Zhong, "Overlapped speech detection based on spectral and spatial feature fusion," in *Interspeech*, 2021, pp. 4189–4193.
- [5] W. Kang, B. C. Roy, and W. Chow, "Multimodal speaker diarization of real-world meetings using d-vectors with spatial features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6509–6513.
- [6] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," *arXiv preprint arXiv:2110.07393*, 2021.
- [7] 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.
- [8] V. Andrei, H. Cucu, and C. Burileanu, "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.
- [9] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," in *INTERSPEECH 2020*, 2020.
- [10] S. Zhang, S. Zheng, W. Huang, M. Lei, H. Suo, J. Feng, and Z. Yan, "Investigation of spatial-acoustic features for overlapping speech detection in multiparty meetings," *Proc. Interspeech 2021*, pp. 3550–3554, 2021.
- [11] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.
- [12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [13] H. Bredin, "pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech 2017*, 2017.
- [14] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function," in *Interspeech*, 2019, pp. 2883–2887.
- [15] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [16] C. J. Pan Chao, Huang Gongping, "Microphone array beamforming: an overview," *Journal of Signal Processing*, vol. 36, no. 6, p. 804, 2020.
- [17] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.
- [18] 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.
- [19] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.
- [20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [21] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: 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.