# CROSS-CHANNEL ATTENTION-BASED TARGET SPEAKER VOICE ACTIVITY DETECTION: EXPERIMENTAL RESULTS FOR THE M2MET CHALLENGE

Weiqing Wang[1], Xiaoyi Qin[2], Ming Li[1,2,*]

[1]Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA
[2]Data Science Research Center, Duke Kunshan University, Kunshan 215316, PR China
ming.li369@duke.edu

## ABSTRACT

In this paper, we present the speaker diarization system for the Multi-channel Multi-party Meeting Transcription Challenge (M2MeT) from team DKU_DukeECE. As the highly overlapped speech exists in the dataset, we employ an x-vector-based target-speaker voice activity detection (TS-VAD) to find the overlap between speakers. Firstly, we separately train a single-channel model for each of the 8 channels and fuse the results. In addition, we also employ the cross-channel self-attention to further improve the performance, where the non-linear spatial correlations between different channels are learned and fused. Experimental results on the evaluation set show that the single-channel TS-VAD reduces the DER by over 75% from 12.68% to 3.14%. The multi-channel TS-VAD further reduces the DER by 28% and achieves a DER of 2.26%. Our final submitted system achieves a DER of 2.98% on the AliMeeting test set, which ranks 1st in the M2MET challenge. In this challenge, our team is denoted as A41.

*Index Terms*— Target-speaker voice activity detection, Multi-channel speaker diarization

## 1. INTRODUCTION

Meeting scenario is one of the most valuable scenarios for speech technologies and becomes increasingly important as the meeting is taking place every day. However, such a scenario is also one of the most challenging scenarios due to the highly overlapped speech and far-field signals with noise and reverberation. In addition, a fixed microphone array is usually equipped as the recording device in such a scenario, providing the multi-channel signals which do not only increase the difficulty but also improve the performance.

As the conventional clustering-based speaker diarization system assumes that a segment only contains one speaker, it is not easy to recognize the overlapped speech without additional modules. Therefore, previous research focuses on reducing the error brought by overlapped speech, including the speech separation [1] as pre-processing, the target-speaker voice activity detection (TS-VAD) [2] as post-processing, and the end-to-end neural diarization (EEND) as post-processing [3] or directly producing the overlap-aware diarization results [4].

Recently, the cross-channel attention has been successfully employed for multi-channel speech signals in speech enhancement [5], speech separation [6, 7], speech recognition [8] and speaker diarization [9] and achieves good results on these tasks. Cross-channel attention can learn the non-linear contextual relationship across channels both within and across time steps. The outputs of the cross-channel attention are usually fused by global average pooling or concatenation for subsequent processing.

In this paper, we present our speaker diarization system on the Alimeeting dataset for M2MeT challenge [10, 11]. Considering that the average speech overlap ratio is over 30% in this dataset, overlap detection is important to reduce the DER. Since TS-VAD achieved the state-of-the-art (SOTA) results in several diarization tasks [12, 13], we select it as our post-processing module for overlap detection and diarization results refinement. In addition, we employ cross-channel self-attention to fuse the speaker embedding of different channels during TS-VAD training to improve the performance further.

The rest of this paper is organized as follows: Section 2 introduces the speaker embedding training. Section 3 presents the details of the diarization system. Section 4 gives the experimental results and discussion. Finally, Section 5 concludes this paper and presents the future work.

## 2. DEEP SPEAKER EMBEDDING EXTRACTION

### 2.1. Dataset

The experiments are conducted on the CN-Celeb dataset [14] and AliMeeting dataset [10]. We adopt the two-stage training method to optimize the speaker embedding model. In the pre-training stage, the CN-Celeb dataset is employed as the training set. Next, the model is separately fine-tuned on two different datasets, one is the AliMeeting training set, and another is the combination of the CN-Celeb and AliMeeting training set. As the AliMeeting dataset does not provide the single-speaker utterances, we select all non-overlapped speech segments for speaker embedding training, where the segments shorter than 2 seconds are dropped. Finally, we build a trial set from the AliMeeting evaluation set to evaluate the performance of the speaker embedding system. The trial set contains 10692 trials from 25 speakers.

### 2.2. Data Augmentation

We perform online data augmentation [15] with MUSAN dataset [16]. We only augment the speech with ambient noise for the background additive noise, as noise with speech is not allowed in this challenge. In addition, 40,000 simulated room impulse responses (RIRs) from small and medium rooms are used for the reverberation [17]. To further enrich the training samples, we adopt the amplification and tempo (change audio playback speed but do not change its pitch) to audio signals.

### 2.3. Deep Speaker Embedding Model

We employ the ResNet34 [18] as our speaker embedding model, and the structure of the model is the same as that in [19]. The encoding layer is based on global statistic pooling (GSP), and the dimension

ICASSP 2022

of the speaker embedding layer is 128. The ArcFace [20] with a margin of 0.2 and softmax prescaling of 32 is used to train the speaker embedding model. During the evaluation phase, cosine similarity is employed for scoring.

## 3. MULTI-CHANNEL SPEAKER DIARIZATION

### 3.1. Dataset

We only use the Alimeeting dataset in the speaker diarization system. For the clustering-based system, the model is trained on the training set and evaluated on the evaluation set. For the TS-VAD model, we create a simulated dataset from the Alimeeting training set, and the simulation process is as follows:

1. As each speaker in the AliMeeting dataset has a unique identification, we select all non-overlapped speech for each speaker from the AliMeeting training set for simulation.

2. Extract the labels from the transcript of the AliMeeting training set and remove all silence regions.

3. During the training stage, the simulated data is generated in an online manner, where we randomly choose a segment of the label and fill the active region with the continuous non-overlapped speech segments.

The more detailed simulation process can be found in [13]. Finally, the Alimeeting evaluation set is adopted as the validation and evaluation set.

### 3.2. Clustering-based System

The clustering-based system is the same as that in [21], where we use an LSTM-based network to extract the affinity matrix. First, we perform the uniform segmentation on all speech regions with a length of 1.28s and a shift of 0.64s. Next, the speaker embedding is extracted from all segments and augmented with Diac-augmentation [22] with a probability of 0.8. The speaker embedding sequences are then fed to the network for training. The training details can be found in [21, 13]. Finally, we employ spectral clustering to get the initialized diarization result, which is the input of the TS-VAD system.

### 3.3. Single-channel TS-VAD

Unlike the original TS-VAD [2] that takes i-vector as target-speaker embedding, we use the deep speaker embedding extracted by ResNet to detect the target-speaker. Figure 1 shows the architecture of our single-channel TS-VAD model. First, a ResNet, which has the same architecture as the speaker embedding model, extract the frame-level speaker embeddings. Next, the target-speaker embeddings are concatenated with the frame-level speaker embeddings. An encoder layer separately extract the detection state of each target-speakers. These detection states are concatenate and process by a BiLSTM to find the relationship between each speaker. Finally, a linear layer with a sigmoid function produce the final decision for each target speaker. The encoder can be a multi-layer BiLSTM or a Transformer encoder that can encode the contextual information across time, and the Transformer encoder is adopt in out experiments.

The training step is as follows:

1. Copy the parameters of the pre-trained speaker embedding model to the front-end model in the TS-VAD model. We keep the front-end model frozen and only train the back-end model
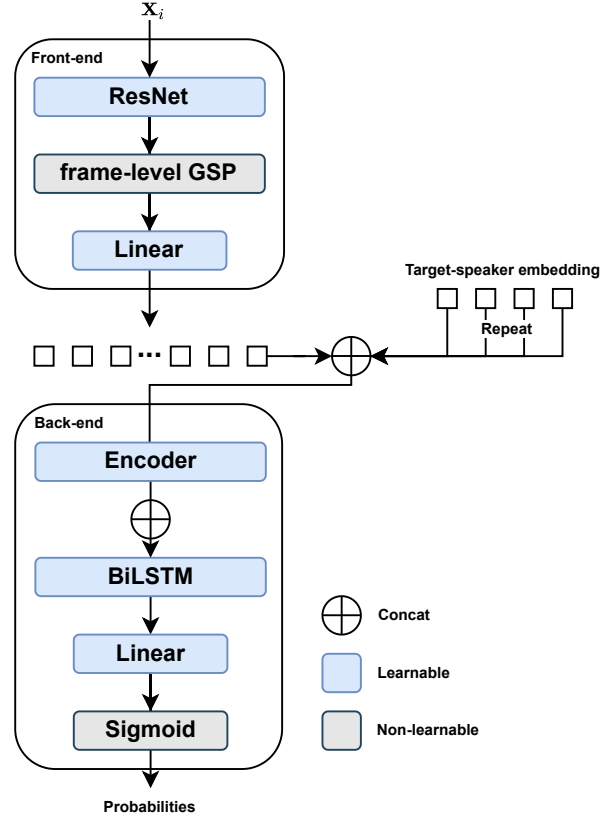


**Fig. 1**. The architecture of the single-channel TS-VAD model

on the simulated data for 10 epochs with a learning rate of $10^{-4}$.

2. Next, after the back-end model converges, we unfreeze the front-end model and jointly train the whole model for another 10 epochs with a learning rate of $10^{-4}$.

3. Finally, we fine-tune the model on the AliMeeting training set for 200 epochs with a learning rate of $10^{-5}$. The 5 best models with the lowest validation loss are averaged as the final model for evaluation and inference.

The model is optimized by the Adam optimizer with binary cross-entropy loss. The input is 16s chunked waves, and the acoustic feature is 80-dim log Mel-filterbank energies (Fbank) with a frame length of 25ms and a frame shift of 10ms. The dimension of the output is a 4-dimensional vector that represents the existence probabilities for each speaker at each time step as the maximum number of speakers in the AliMeeting dataset is 4. We separately train a model for each channel, producing 8 single-channel TS-VAD models.

These three training steps are very important to obtain a good result in our experiments. We also try to directly train the whole model, but the model cannot converge for a very long time no matter we train from scratch or copy the parameters of the pre-trained speaker embedding model to the front-end ResNet. If we directly train the whole model, the parameters of the front-end model will be changed when the back-end model does not converge, which makes the training difficult.

**Table 1**. The performance of speaker embedding model.

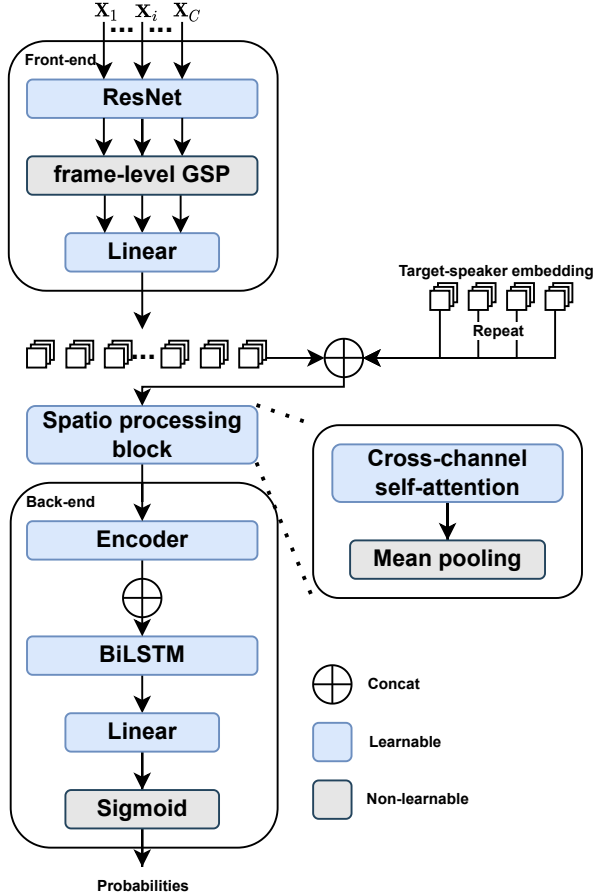| Training data | Training stage | CN-Celeb trial | | AliMeeting trial | |
|---|---|---|---|---|---|
| | | EER[%] | mDCF$_{0.01}$ | EER[%] | mDCF$_{0.01}$ |
| CN-Celeb | Pre-train | 12.65 | 0.6751 | - | - |
| CN-Celeb + AliMeeting | Fine-tune | - | - | **3.199** | **0.353** |
| AliMeeting | Fine-tune | - | - | 3.816 | 0.382 |



**Fig. 2**. The architecture of the multi-channel TS-VAD model

### 3.4. Multi-channel TS-VAD

The training set and training process are the same as that of the single-channel TS-VAD, and the only difference is the architecture and the input of the model. Figure 2 shows the architecture of the multi-channel TS-VAD, where a cross-channel self-attention layer is employed on the concatenated embeddings to learn the cross-correlation between different channels.

Given $C$-channels of Fbank sequence $\mathcal{X} = (\mathbf{X}_1, ..., \mathbf{X}_i, ..., \mathbf{X}_C)$ and the target-speaker embedding from $N$ target speakers $\tilde{\mathcal{S}} = (\tilde{\mathbf{S}}_1, ..., \tilde{\mathbf{S}}_i, ..., \tilde{\mathbf{S}}_C)$, the corresponding target speaker decision is $\mathcal{Y} = (\mathbf{y}_1, ..., \mathbf{y}_t, ..., \mathbf{y}_T)$, where $\mathbf{X}_i \in \mathbb{R}^{T \times F}$ is of $T$ frames, $\tilde{\mathbf{S}}_i \in \mathbb{R}^{N \times D}$ is the D-dimensional speaker embedding from N target speakers, and $\mathbf{y}_t \in \{0,1\}^N$ is the target speaker decision at time step t with dimension of N. Next, the front-end ResNet takes

the $C$-channels of Fbank sequence as input and produce a frame-level speaker embedding sequence $\hat{\mathcal{S}} = (\hat{\mathbf{S}}_1, ..., \hat{\mathbf{S}}_i, ..., \hat{\mathbf{S}}_C)$, where $\hat{\mathbf{S}}_i \in \mathbb{R}^{T \times D}$ is the frame-level speaker embedding from the $i^{\text{th}}$ channel. Later, the target-speaker embedding is repeated T times and the frame-level speaker embedding is repeated N times. These two speaker embeddings are concatenated at the embedding dimension: $\mathcal{S} = (\mathbf{S}_1, ..., \mathbf{S}_i, ..., \mathbf{S}_C)$, where $\mathbf{S}_i \in \mathbb{R}^{T \times N \times 2D}$.

The cross-channel self-attention takes the concatenated speaker embedding $\mathbf{S}_{\text{in}}$ as input:

$$\mathbf{S}_{\text{in}} = \text{concat}(\mathcal{S}) \in \mathbb{R}^{T \times N \times C \times 2D} \quad (1)$$

$$\mathbf{Q}^i = \mathbf{W}_Q^i \mathbf{S}_{\text{in}} + \mathbf{b}_Q^i \quad (2)$$

$$\mathbf{K}^i = \mathbf{W}_K^i \mathbf{S}_{\text{in}} + \mathbf{b}_K^i \quad (3)$$

$$\mathbf{V}^i = \mathbf{W}_V^i \mathbf{S}_{\text{in}} + \mathbf{b}_V^i, \quad (4)$$

where $\mathbf{Q}^i$, $\mathbf{K}^i$ and $\mathbf{V}^i$ denotes the query, key and value matrices for the $i^{\text{th}}$ head. $\mathbf{W}^i \in \mathbb{R}^{E \times 2D}$ and $\mathbf{b}^i \in \mathbb{R}^E$ are the weight and bias for the $i^{\text{th}}$ head. Next, the scaled dot-produce attention is applied to the query, key and value:

$$\text{Attention}(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i) = \text{softmax}\frac{\mathbf{Q}^i(\mathbf{K}^i)^\mathsf{T}}{\sqrt{n}}\mathbf{V}^i, \quad (5)$$

where $n = E$. Next, a positional-wise feed-forward layer with a ReLU activation is applied to generate the output, where layer norm and residual connections are employed between each layer. The output is denoted by $\mathbf{S}_{\text{out}} \in \mathbb{R}^{T \times N \times C \times 2D}$. Finally, the output is averaged over channel dimension by a global average pooling layer:

$$\mathbf{S}' = \frac{1}{C}\sum_{i=1}^{C} S_{\text{out},i}, \quad (6)$$

where $\mathbf{S}' \in \mathbb{R}^{T \times N \times 2D}$. Finally, the back-end model processes this fused speaker embedding $\mathbf{S}'$ in the same way as the singel-channel TS-VAD processes the concatenated speaker embedding. In our experiments, we employ a 2-layer and 2-head Transformer Encoder as the cross-channel self-attention layer.

### 3.5. Fusion

We employ the DOVER-Lap [23] to fuse the systems that we mentioned above.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

### 4.1. Speaker Embedding

The performance of the speaker embedding model is reported in Table 1. As the model fine-tuned with CNCeleb + AliMeeting training set shows the best performance, it is adopted as the speaker embedding model for the speaker diarization task.

**Table 2**. The performance of speaker diarization model on the AliMeeting evaluation set. For TS-VAD model, we report the DER of the 3rd round, where the input is the diarization results from the previous round.
* The result is from the MC-TS-VAD model initialized by ground-truth label.

| Model | AliMeeting evaluation set | | | | | AliMeeting test set | |
|---|---|---|---|---|---|---|---|
| | MISS[%] | FA[%] | SpkErr[%] | DER[%] | JER[%] | DER[%] | JER[%] |
| Clustering | | | | | | | |
|     channel 1 | 10.6 | 1.2 | 1.9 | 13.80 | 27.03 | - | - |
|     channel 2 | 10.5 | 1.4 | 1.0 | 12.86 | 24.19 | - | - |
|     channel 3 | 11.0 | 1.2 | 1.0 | 13.27 | 25.56 | - | - |
|     channel 4 | 10.8 | 1.4 | 0.9 | 13.13 | 24.77 | - | - |
|     channel 5 | 10.9 | 1.3 | 1.1 | 13.38 | 25.06 | - | - |
|     channel 6 | 10.7 | 1.3 | 1.0 | 12.97 | 23.86 | - | - |
|     channel 7 | 10.8 | 1.3 | 1.0 | 13.08 | 23.67 | - | - |
|     channel 8 | 10.6 | 1.2 | 0.6 | **12.68** | **23.09** | - | - |
|     DOVER-Lap fusion | 11.6 | 0.8 | 0.8 | 13.23 | 24.89 | - | - |
| SC-TS-VAD | | | | | | | |
|     channel 1 | 2.5 | 1.0 | 0.7 | 4.12 | 12.43 | - | - |
|     channel 2 | 2.6 | 1.1 | 0.6 | 4.26 | 12.82 | - | - |
|     channel 3 | 2.8 | 1.0 | 0.4 | 4.21 | 12.78 | - | - |
|     channel 4 | 2.5 | 0.9 | 0.5 | 3.93 | 12.56 | - | - |
|     channel 5 | 2.6 | 0.9 | 0.4 | 3.95 | 12.08 | - | - |
|     channel 6 | 2.5 | 0.9 | 0.4 | 3.90 | 11.86 | - | - |
|     channel 7 | 2.3 | 0.9 | 0.4 | 3.61 | 11.55 | - | - |
|     channel 8 | 2.4 | 0.8 | 0.3 | 3.49 | 11.28 | - | - |
|     DOVER-Lap fusion | 2.3 | 0.5 | 0.3 | **3.14** | **11.08** | - | - |
| MC-TS-VAD | 1.1 | 1.1 | 0.1 | **2.26** | **8.27** | **2.98** | **9.24** |
| MC-TS-VAD* | 1.1 | 1.1 | 0.1 | 2.32 | 8.34 | - | - |
| Official baseline [11] | - | - | - | 15.24 | - | 15.6 | - |

## 4.2. Speaker Diarization

During the inference stage of the TS-VAD model, we first select all non-overlapped speech regions for each speaker based on the initialized results from the clustering-based system and extract the target-speaker embeddings. Next, we remove all silence regions and break the audio signals into 16s chunked waves with a 4s shift. After obtaining the probabilities from the TS-VAD model, we apply the median filtering with a window size of 7 to smooth the probability sequences. This results can also be the initialization of the next round of TS-VAD inference, and we infer it for 3 rounds.

Table 2 shows the results of all systems on the AliMeeting evaluation set. The missed speaker time (MISS), false alarm speaker time (FA), speaker error time (SpkErr), DER, and JER are reported for the evaluation set. In addition, for the clustering-based system and single-channel TS-VAD (SC-TS-VAD) system, the performance of each channel is reported, and the results of all channels are fused by DOVER-Lap. Results show that the model trained on the data of the 8th channel achieves the lowest DER, and fusion degrades the performance. For the SC-TS-VAD model, the fused system shows the best performance, and the MISS error is significantly reduced compared with the clustering-based method. The multi-channel TS-VAD (MC-TS-VAD) can further reduce the MISS error and achieve the lowest DER of 2.26% on the evaluation set and 2.98% on the test set according to the chanllenge results [11]. We also try to fuse this system with other systems, but the performance always goes worse. Therefore, our submitted system is this single system (MC-TS-VAD) without any fusion.

In our experiments, we also find that the initialized results are not very important for TS-VAD in this challenge. The reason is that each of the recordings lasts over 30 minutes, which can produce a very long single-speaker segment for each speaker. If the number of speakers is correctly estimated, the TS-VAD model can always achieve the same results after several rounds even the initialized results are different. We also report the Multi-channel TS-VAD results initialized by the ground-truth label. It is interesting to note that even using the ground-truth label as initialization, the results are still similar to the results initialized by clustering-based results. Therefore, we do not explore the multi-channel clustering-based system.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we present our submitted system for the M2MeT challenge. As the dataset contains highly overlapped speech, most of the error of our clustering-based system is from the MISS error. Next, we employ the single-channel TS-VAD model to refine the diarization results and reduce the MISS error. Results show that the DER is significantly reduced by 75%. In addition, we also apply cross-channel attention to the TS-VAD model to further improve the performance by 28%.

In the future, we will train a speaker embedding model on the multi-channel data and evaluate the performance of the diarization system. Also, we will perform some ablation studies on the cross-channel self-attention layer, e.g., applying the cross-channel self-attention on the detection states instead of the speaker embedding or increasing the number of parameters of the cross-channel self-attention.

# 6. REFERENCES

[1] Xiong Xiao, Naoyuki Kanda, Zhuo Chen, Tianyan Zhou, Takuya Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Yu Wu, Jian Wu, et al., "Microsoft Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2020," in *Proc. ICASSP*, 2021, pp. 5824–5828.

[2] 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*, pp. 274–278.

[3] Shota Horiguchi, Paola García, Yusuke Fujita, Shinji Watanabe, and Kenji Nagamatsu, "End-to-end speaker diarization as post-processing," in *Proc. ICASSP*, pp. 7188–7192.

[4] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. Interspeech 2020*, pp. 269–273.

[5] Minh Tri Ho, Jinyoung Lee, Bong-Ki Lee, Dong Hoon Yi, and Hong-Goo Kang, "A Cross-Channel Attention-Based Wave-U-Net for Multi-Channel Speech Enhancement," in *Proc. Interspeech 2020*, pp. 4049–4053.

[6] Dongmei Wang, Zhuo Chen, and Takuya Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," in *Proc. Interspeech 2020*, pp. 339–343.

[7] Dongmei Wang, Takuya Yoshioka, Zhuo Chen, Xiaofei Wang, Tianyan Zhou, and Zhong Meng, "Continuous speech separation with ad hoc microphone arrays," in *2021 European Signal Processing Conference (EUSIPCO)*, pp. 1100–1104.

[8] Feng-Ju Chang, Martin Radfar, Athanasios Mouchtaris, and Maurizio Omologo, "Multi-Channel Transformer Transducer for Speech Recognition," in *Proc. Interspeech 2021*, pp. 296–300.

[9] Shota Horiguchi, Yuki Takashima, Paola Garcia, Shinji Watanabe, and Yohei Kawaguchi, "Multi-channel end-to-end neural diarization with distributed microphones," *arXiv preprint arXiv:2110.04694*, 2021.

[10] 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*, 2022.

[11] 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*, 2022.

[12] Yu-Xuan Wang, Jun Du, Maokui He, Shu-Tong Niu, Lei Sun, and Chin-Hui Lee, "Scenario-Dependent Speaker Diarization for DIHARD-III Challenge," in *Proc. Interspeech 2021*, 2021, pp. 3106–3110.

[13] Weiqing Wang, Danwei Cai, Qingjian Lin, Lin Yang, Junjie Wang, Jin Wang, and Ming Li, "The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge," *arXiv preprint arXiv:2109.02002*, 2021.

[14] 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 *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[15] Weicheng Cai, Jinkun Chen, Jun Zhang, and Ming Li, "On-the-fly data loader and utterance-level aggregation for speaker and language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[16] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484*, 2015.

[17] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

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

[19] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Odyssey 2018*, pp. 74–81.

[20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR 2019*, pp. 4690–4699.

[21] Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization," in *Proc. Interspeech 2019*, 2019, pp. 366–370.

[22] Qiujia Li, Florian L Kreyssig, Chao Zhang, and Philip C Woodland, "Discriminative neural clustering for speaker diarisation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 574–581.

[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)*, pp. 881–888.