NEURAL SPEAKER DIARIZATION USING MEMORY-AWARE MULTI-SPEAKER EMBEDDING WITH SEQUENCE-TO-SEQUENCE ARCHITECTURE

Gaobin Yang[1]*, Maokui He*[1]*, Shutong Niu*[1]*, Ruoyu Wang*[1]*, Yanyan Yue*[2]*, Shuangqing Qian*[2]*, Shilong Wu*[1]*, Jun Du*[1]*, Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, China
[2]iFlytek Research, Hefei, China
[3]Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

We propose a novel neural speaker diarization system using memory-aware multi-speaker embedding with sequence-to-sequence architecture (NSD-MS2S), which integrates the strengths of memory-aware multi-speaker embedding (MA-MSE) and sequence-to-sequence (Seq2Seq) architecture, leading to improvement in both efficiency and performance. Next, we further decrease the memory occupation of decoding by incorporating input features fusion and then employ a multi-head attention mechanism to capture features at different levels. NSD-MS2S achieved a macro diarization error rate (DER) of 15.9% on the CHiME-7 EVAL set, which signifies a relative improvement of 49% over the official baseline system, and is the key technique for us to achieve the best performance for the main track of CHiME-7 DASR Challenge. Additionally, we introduce a deep interactive module (DIM) in MA-MSE module to better retrieve a cleaner and more discriminative multi-speaker embedding, enabling the current model to outperform the system we used in the CHiME-7 DASR Challenge. Our code will be available at `https://github.com/liyunlongaaa/NSD-MS2S`.

Index Terms— CHiME challenge, speaker diarization, memory-aware speaker embedding, sequence-to-sequence architecture

1. INTRODUCTION

Speaker diarization is the task of tagging an audio recording with labels that indicate "who spoke when" [1]. High-quality diarization outcomes can be beneficial for numerous speech-related tasks, such as generating meeting summaries, analyzing phone conversations, transcribing dialogues, and so on [2]. However, speaker diarization remains challenging in real-world scenarios with varying speaker numbers, adverse acoustic environments, and a large portion of speech overlap.

Conventional clustering-based methods [3], which include voice activity detection (SAD/VAD), speech segmentation, speaker feature extraction (e.g., i-vector [4], d-vector [5], x-vector [6]), speaker clustering [7, 8, 9] and re-segmentation [10], are commonly used in speaker diarization task. Although clustering-based speaker diarization is relatively robust across different domains, they cannot deal well with overlap segments because every segment can only be assigned a single label through clustering.

To tackle this problem above, many excellent methods [11, 12, 13, 14, 15] have been proposed. Recently, target-speaker voice activity detection (TS-VAD) [13][16] was proposed, which used speech features along with speaker embedding as input to predict each speakers activities at each frame. Although TS-VAD has been a great success in many scenarios [17], it still has some unresolved problems. First, the BLSTM-based network architecture leads to slower inference and significantly increases GPU memory usage as decoding duration becomes longer. Second, TS-VAD uses a pre-trained extractor to obtain speaker embeddings (e.g. i-vectors) as input, but the embeddings are often unreliable in real scenarios due to the lack of oracle speaker segments. To solve the computational overhead problem mentioned above, Seq2Seq-TSVAD [15] introduced the sequence-to-sequence architecture and achieved good results. Our previous work NSD-MA-MSE [14] was proposed to solve the problem of unreliable speaker embeddings in real scenarios, which introduced a memory module to facilitate a dynamical refinement of speaker embedding to reduce a potential data mismatch between the speaker embedding extraction and the NSD network. However, the first problem mentioned above remains unsolved in NSD-MA-MSE.

In order to further address both of the TS-VAD problems mentioned above simultaneously, in this paper, we propose a novel neural speaker diarization system for CHiME-7 referred to NSD-MS2S, which first integrates the strengths of MA-MSE and Seq2Seq architecture perfectly. NSD-MS2S processes acoustic features and multi-speaker embeddings separately to avoid dimensional expansion. It then combines these two components through a decoder to predict voice activities for the target speakers, resulting in a significant improvement in both efficiency and performance compared to NSD-MA-

MSE [14]. Then, we introduce a simple and efficient method of input features fusion to further reduce the computational overhead required in the decoder of the transformer and then use a multi-head attention mechanism to capture features at different levels. NSD-MS2S achieved a macro DER of 15.9% on the CHiME-7 EVAL set, representing a relative improvement of 49% over the official baseline system and enabling us to secure first place in the main track of the CHiME-7 DASR Challenge. Additionally, in order to better retrieve multi-speaker embedding from the memory module, we introduce a deep interactive module (DIM) in MA-MSE module. Multi-scale feature fusion of acoustic features and speaker embedding basis vectors can retrieve a cleaner and more discriminative multi-speaker embedding than the original MA-MSE module, making the single model results outperform the system we used in the CHiME-7 DASR Challenge.

2. METHODS

The framework of our proposed NSD-MS2S system is shown in Fig.1. The details are introduced as follows.

2.1. Overview of Network

The input of the main network is a set of acoustic features denoted as $\mathbf{X} \in \mathbb{R}^{T\times F}$, where T and F are the length and dimension of log-Mel filter-bank features (FBANKs), respectively. Then, the convolutional layers are used to extract a set of deep features denoted by $\mathbf{F} \in \mathbb{R}^{C\times T\times \frac{F}{2}}$, which is downsampled to $\mathbf{F}' \in \mathbb{R}^{T\times D}$, where C and D are dimensions of channels and features, respectively. The features sequence $\mathbf{F}'$ with positional embedding (PE) is encoded into $\mathbf{E}_{\text{enc}} \in \mathbb{R}^{T\times D}$ by the following speaker detection (SD) encoder, which is a stack of conformer blocks. In addition, $\mathbf{F}'$ and speaker mask matrix $\mathbf{S} \in \mathbb{R}^{N\times T}$ serve as inputs of MA-MSE module, and then the output MA-MSE embedding $\mathbf{E}_M \in \mathbb{R}^{N\times D_M}$ is obtained, where N is the number of speakers and D_M is the dimension of MA-MSE embedding. We concatenate MA-MSE embedding and i-vector to get aggregate embedding $\mathbf{E}_A \in \mathbb{R}^{N\times D}$. We will give specific details about $\mathbf{E}_A$ in Section 2.3.Then, we pass aggregate embedding $\mathbf{E}_A$, decoder embedding $\mathbf{E}_D \in \mathbb{R}^{N\times D}$ and $\mathbf{E}_{\text{enc}}$ along with sinusoidal positional embedding to SD decoder to produce $\mathbf{E}_{\text{dec}} \in \mathbb{R}^{N\times D}$. We will provide a more detailed description of this process in Section 2.2. Finally, the output layer can transform $\mathbf{E}_{\text{dec}}$ into posterior probabilities $\hat{\mathbf{Y}} = [\hat{y}_{\text{nt}}]_{N\times T}$ of voice activities for N speakers.

2.2. Speaker Detection Decoder

The design of the speaker detection (SD) decoder was mainly inspired by [15, 18]. The SD decoder consisting of several SD blocks predicts target-speaker voice activities by considering cross-speaker correlations.

For the forward process of a SD block, first, decoder embedding $\mathbf{E}_D$ and aggregate embedding $\mathbf{E}_A$ pass through their respective multi-layer perception (MLP) to generate within-block representations $\mathbf{E}_D^{Q_1}$, $\mathbf{E}_D^{K_1}$, $\mathbf{E}_D^{V_1}$, $\mathbf{E}_A^{Q_1}$ and $\mathbf{E}_A^{K_1}$. Q, K, and V denote the query, key, and value in the attention mechanism, respectively. If not specifically noted, all the MLP layers map dimension sizes of input features to D. For simplicity, the MLP structure is omitted in Fig.1(c). Then, in order to make the decoder embedding contain speaker information and reduce the subsequent time and space overhead, we fuse the input features and do not increase the feature dimensions, which can be represented by the following equation:

$$\begin{aligned}\mathbf{Q}_1 &= \beta_1 \times \mathbf{E}_D^{Q_1} + (1-\beta_1) \times \mathbf{E}_A^{Q_1} \\ \mathbf{K}_1 &= \beta_2 \times \mathbf{E}_D^{K_1} + (1-\beta_2) \times \mathbf{E}_A^{K_1} \\ \mathbf{V}_1 &= \mathbf{E}_D^{V_1}\end{aligned} \tag{1}$$

where β_1 and β_2 are learnable parameters, we expect the model itself to decide what information it needs. $\mathbf{Q}_1$, $\mathbf{K}_1$, and $\mathbf{V}_1$ sequentially go through layernorm (LN) and multi-attention (MA) to extract features at different levels, and then we get within-block features $\mathbf{E}_F \in \mathbb{R}^{N\times D}$. Then, we transform $\mathbf{E}_F$, $\mathbf{E}_A$ and $\mathbf{E}_{\text{enc}}$ to within-block representations $\mathbf{E}_F^{Q_2}$, $\mathbf{E}_A^{Q_2}$, $\mathbf{E}_{\text{enc}}^{K_2}$ and $\mathbf{E}_{\text{enc}}^{V_2}$ via MLP layer. Next, we obtain the queries, keys, and values before the second LN & MA layer by the following function:

$$\begin{aligned}\mathbf{Q}_2 &= \beta_3 \times \mathbf{E}_F^{Q_2} + (1-\beta_3) \times \mathbf{E}_A^{Q_2} \\ \mathbf{K}_2 &= \mathbf{E}_{\text{enc}}^{K_2} + \mathbf{PE} \\ \mathbf{V}_2 &= \mathbf{E}_{\text{enc}}^{V_2}\end{aligned} \tag{2}$$

where $\mathbf{PE}$ is sinusoidal positional embedding and β_3 is learnable parameter. Then, the output of the second LN & MA layer is passed through the feed-forward network (FFN) to generate the next decode embedding. Finally, $\mathbf{E}_{\text{dec}}$ is obtained and sent to the output layer to predict target-speaker voice activities. The output layer is composed of a linear layer plus a sigmoid activation function, which can determine the length of decoding.

2.3. MA-MSE with Deep Interactive Module

The memory-aware multi-speaker embedding (MA-MSE) module can retrieve a clean and discriminative multi-speaker embedding from memory through a simple additive attention mechanism. In [14], the memory component is the core of the MA-MSE module, which is composed of many speaker embedding basis vectors extracted from additional datasets. Specifically, speaker embedding basis vectors can be obtained by clustering the speaker embeddings (e.g., i-vectors or x-vectors) and taking the cluster centers. Before feeding the features $\mathbf{F}'$ into MA-MSE module, we use a clustering-based approach to get the speaker activity $0/1$ mask $\mathbf{S} \in \mathbb{R}^{N\times T}$ on each frame. The features $\mathbf{F}'$ and mask $\mathbf{S}$ are multiplied to get the selected audio features $\mathbf{F}_S = \left[\mathbf{F}_S^1, \mathbf{F}_S^2, ..., \mathbf{F}_S^N\right] \in \mathbb{R}^{N\times D}$ for each speaker and a simple additive attention mechanism is

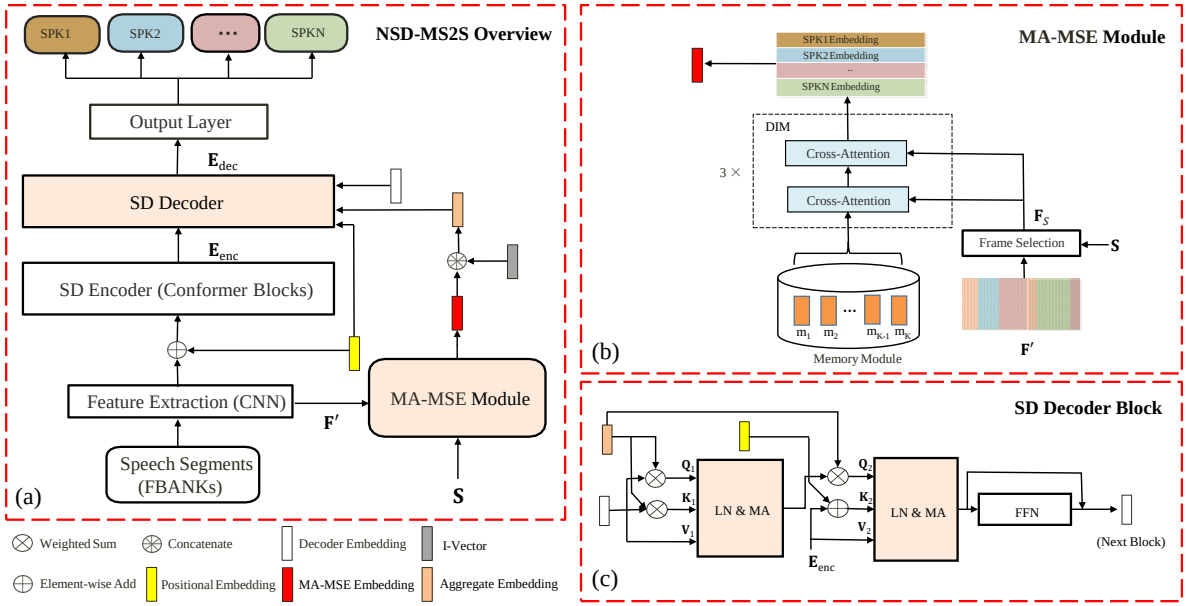

Fig. 1: The proposed NSD-MSE framework. (a) is the overview of NSD-MS2S architecture. (b) and (c) are schematic diagrams of the structures of MA-MSE Module and SD Decoder Block, respectively.

used to choose the speaker embedding bases of the memory that are most similar to the current speech segment.

After the CHiME-7 DASR Challenge, we found that if the structure of the MA-MSE module is not well-designed it may significantly impair performance in complex acoustic scenarios. In addition, a too simple mechanism also limits performance improvement. Based on this, we introduce the deep interactive module (DIM), which replaces the additive attention mechanism with a dot-product attention mechanism and deepens the number of interaction layers. This multi-scale feature fusion approach can better extract cleaner, more discriminating multi-speaker embedding from memory module.

The DIM is composed of 3 DIM blocks, each consisting of two cross-attention structures along the feature dimension. Given that all speaker embedding basis vectors in memory module are represented by $\mathbf{M} \in \mathbb{R}^{K \times D_M}$, where K is the number of vectors. In first DIM block, the input features of n-th speaker $\mathbf{F}_S^n$ and $\mathbf{M}$ are used and calculated by the following function:

$$\mathbf{H}_1^n = \text{Softmax}\left(\frac{\left(\mathbf{F}_S^n \mathbf{W}_1^{n,q}\right)\left(\mathbf{M}\mathbf{W}_1^{n,k}\right)^T}{\sqrt{d_m}}\right)\mathbf{M} \quad (3)$$

where $\mathbf{W}_1^{n,q} \in \mathbb{R}^{D \times D}$ and $\mathbf{W}_1^{n,k} \in \mathbb{R}^{D_M \times D}$ are the learnable weights, the scaling $\sqrt{d_m}$ is for numerical stability. Then, the output of DIM block is calculated by:

$$\mathbf{H}_2^n = \text{Softmax}\left(\frac{\left(\mathbf{F}_S^n \mathbf{W}_2^{n,q}\right)\left(\mathbf{H}_1^n \mathbf{W}_2^{n,k}\right)^T}{\sqrt{d_m}}\right)\mathbf{H}_1^n \quad (4)$$

where $\mathbf{W}_2^{n,q} \in \mathbb{R}^{D \times D}$ and $\mathbf{W}_2^{n,k} \in \mathbb{R}^{D_M \times D}$ are the learnable weights. After that, $\mathbf{H}_2^n$ and $\mathbf{F}_S^n$ are passed to the next DIM block. Finally, the MA-MSE embedding $\mathbf{E}_M$ is obtained. $\mathbf{E}_M$, serving as essential supplementary speaker information, will be concatenated with the current speaker's i-vector to generate aggregate embedding $\mathbf{E}_A$.

2.4. Loss Function

With the input acoustic features set $\mathbf{X}$ and speaker mask matrix, NSD-MS2S predicts speech/silence probabilities for each of the N speakers with $\hat{\mathbf{Y}} = [\hat{y}_{\text{nt}}]_{N \times T}$. We adopt the binary cross-entropy (BCE) loss of multiple speakers as the learning objective:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}[y_{\text{nt}} \log(\hat{y}_{nt}) + (1 - y_{\text{nt}}) \log(1 - \hat{y}_{\text{nt}})] \quad (5)$$

3. EXPERIMENT

3.1. CHiME-7 DASR Challenge

In CHiME-7 Challenge [19], Distant Automatic Speech Recognition (DASR) with Multiple Devices in Diverse Scenarios is the main track participants must always submit. The entire training data originates from two parts, one is the CHiME-7 DASR Official data (e.g., CHiME-6, DiPCo, and Mixer 6 Speech), and the other is the external data allowed under official rules. More details on how we organized and simulated the training data can be found in our technical report [20].

Table 1: Performance comparison of different methods on CHiME-7 DEV and EVAL set (collar = 0.25 s).

Method	Set	CHiME-6		DiPCo		Mixer 6		Macro	
		DER	**JER**	**DER**	**JER**	**DER**	**JER**	**DER**	**JER**
x-vectors + SC	DEV	40.32	42.31	24.47	28.97	15.8	23.07	26.86	31.45
	EVAL	36.32	43.39	25.18	35.08	9.53	12.08	23.67	30.18
+ NSD-MA-MSE	DEV	32.27	34.76	21.04	24.01	9.28	12.94	20.86	23.90
	EVAL	32.09	37.61	22.78	31.34	6.21	7.12	20.36	25.35
+ NSD-MS2S	DEV	29.93	33.92	18.22	22.36	9.85	13.08	19.33	23.12
	EVAL	30.50	36.01	21.64	29.83	5.5	6.3	19.21	24.04
+ DIM	DEV	**28.36**	**31.49**	**17.06**	**18.54**	**7.27**	**9.56**	**17.56**	**19.86**
	EVAL	**29.45**	**33.84**	**19.31**	**26.63**	**5.01**	**5.54**	**17.92**	**22.00**

3.2. Implementation Details

In our experiments, top-6 audio channels [19] are selected to perform VAD using a baseline VAD model fine-tuned by CHiME-6 and Mixer 6 data, we use ECAPA-TDNN [21] to extract x-vectors and spectral clustering (SC) as our diatization initialization system.

For our NSD-MS2S system, the input is the 40-dim FBANKs features. All encoder-decoder modules have 6 blocks sharing the same settings: the size of D is 512, attention with 8 heads, and 1024-dim feed-forward layers with a dropout rate of 0.1. Conformer block is designed identically to [22]. The feature fusion coefficient β is all initiated to 0.5. We use two MA-MSE modules whose memory modules consist of 256-dimensional x-vectors and 100-dimensional i-vectors, respectively. All speaker embeddings in the memory module are extracted from the VoxCeleb1 and VoxCeleb2 datasets. It is worth noting that the model we used for the CHiME-7 competition is without the DIM module. The decoding duration of the output layer is set to 8s (T=800) and mixup [23] is used in training. We use Adam with a learning rate of $1e-4$ to optimize the entire model for 6 epochs. We found that even though there is a single-model performance degradation during training, there is still a gain when fusing models from different epochs.

3.3. Results and Analysis

Since Dover-lap [24] works slowly when the number of channels is large, we averaged the posterior probabilities of all the channels output by NSD-MS2S as the final output. Our diarization system is actually a multi-step iterative system in the CHiME-7 DASR Challenge [20], but for a fair comparison, we present the results of different single model systems at the first iteration in the Table 1. Compared to NSD-MA-MSE , NSD-MS2S makes the macro DER drop relatively by 5.6% on EVAL set. Furthermore, DIM enhances the performance of the NSD-MS2S, resulting in a reduction of the macro DER from 19.21% to 17.92% on EVAL set.

Table 2: Computational resource consumption with NSD-MA-MSE and NSD-MS2S.

Method	Parameters (M) ↓	GPU Memory (MB) ↓	Inference Time ↓
NSD-MA-MSE	50.19	5803	1.0
NSD-MS2S	57.53	3019	0.47
+DIM	59.07	3050	0.48

Table 3: Performance comparison of different fusion models on CHiME-7 DEV and EVAL set (collar = 0.25 s).

Method	Set	CHiME-6		DiPCo		Mixer 6		Macro	
		DER	**JER**	**DER**	**JER**	**DER**	**JER**	**DER**	**JER**
Fusion*	DEV	30.34	32.34	18.54	20.20	7.99	10.76	18.95	21.1
	EVAL	29.39	34.13	20.58	28.67	5.01	5.57	18.32	22.79
Fusion†	DEV	26.78	29.45	16.13	17.38	**7.22**	**9.45**	16.71	18.76
	EVAL	28.51	32.63	18.83	25.72	**4.95**	**5.45**	17.43	21.26
Fusion¶	DEV	**25.81**	**27.64**	**15**	**15.92**	8.96	12.27	**16.59**	**18.61**
	EVAL	**25.11**	**28.86**	**16.36**	**22.06**	6.14	6.81	**15.87**	**19.25**

* and † represent the results for NSD-MS2S and NSD-MS2S (+DIM) at the first iteration, respectively.

¶ stands for our NSD-MS2S system result submitted to the main track of CHiME-7 DASR Challenge after final iteration.

In Table 2, we have analyzed the computational efficiency and overhead of the different methods under the same setting: batch is 16, one 12G Tesla V100 on the same machine. The inference time takes the average time to complete three runs of the CHiME-6 DEV set and is benchmarked using NSD-MA-MSE (with 1.0 as the reference). It can be seen that although the number of parameters has increased in NSD-MS2S, the GPU memory footprint and inference speed are superior to NSD-MA-MSE.

Table 3 shows the results of model fusion for 6 different epochs. The single model results of NSD-MS2S with DIM are even better than the fusion models we used in the CHiME-7 DASR Challenge at first iteration (macro DER on EVAL set, 17.92 vs 18.32). However, the fusion models results of NSD-MS2S with DIM at first iteration are still worse than the final iteration we submitted to the CHiME-7 challenge (macro DER on EVAL set, 17.43 vs 15.87). On one hand, this demonstrates the effectiveness of our iterative strategy [20] in enhancing the performance of NSD-S2S, with notable improvements across CHiME-6 set and DiPCo set, except for small performance degradation on the Mixer 6 set. On the other hand, it suggests the potential for surpassing our CHiME-7 champion system.

4. CONCLUSIONS

We presented a novel approach called NSD-MS2S for the diarization of multi-speaker conversations, which provided state-of-the-art results in the CHiME-7 DASR Challenge. Compared to NSD-MA-MSE, NSD-MS2S not only increases the speed and reduces GPU memory consumption, but also improves performance. In the future, we will explore the directions of lighter structure, faster inference, and less resource consumption to facilitate the diarization system further toward practicality.

5. REFERENCES

[1] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, pp. 101317, 2022.

[2] 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.

[3] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[4] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[5] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018, pp. 4879–4883.

[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[7] Kyu J Han and Shrikanth S Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[8] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, 2013.

[9] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with LSTM," in *ICASSP*, 2018, pp. 5239–5243.

[10] Paola García, Jesus Villalba, Hervé Bredin, Jun Du, Diego Castan, Alejandrina Cristia, Latane Bullock, Ling Guo, Koji Okabe, Phani Sankar Nidadavolu, et al., "Speaker detection in the wild: Lessons learned from JSALT 2019," in *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020.

[11] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[12] Zili Huang, Shinji Watanabe, Yusuke Fujita, Paola García, Yiwen Shao, Daniel Povey, and Sanjeev Khudanpur, "Speaker diarization with region proposal network," in *ICASSP*, 2020, pp. 6514–6518.

[13] 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," *INTERSPEECH*, pp. 274–278, 2020.

[14] Mao-Kui He, Jun Du, Qing-Feng Liu, and Chin-Hui Lee, "ANSD-MA-MSE: Adaptive Neural Speaker Diarization Using Memory-Aware Multi-Speaker Embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[15] Ming Cheng, Weiqing Wang, Yucong Zhang, Xiaoyi Qin, and Ming Li, "Target-Speaker Voice Activity Detection via Sequence-to-Sequence Prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] Maokui He, Desh Raj, Zili Huang, Jun Du, Zhuo Chen, and Shinji Watanabe, "Target-Speaker Voice Activity Detection with Improved i-Vector Estimation for Unknown Number of Speaker," in *INTERSPEECH*, 2021, pp. 3555–3559.

[17] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "The third DIHARD diarization challenge," *INTERSPEECH*, pp. 3570–3574, 2021.

[18] Liu Shilong, Li Feng, Zhang Hao, Yang Xiao, Qi Xianbiao, Su Hang, Zhu Jun, and Lei Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," in *ICLR*, 2022, pp. 10–20.

[19] Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola Garcia, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur, "The chime-7 dasr challenge: Distant meeting transcription with multiple devices in diverse scenarios," *arXiv preprint arXiv:2306.13734*, 2023.

[20] Ruoyu Wang, Maokui He, Jun Du, Hengshun Zhou, Shutong Niu, Hang Chen, Yanyan Yue, Gaobin Yang, Shilong Wu, Lei Sun, et al., "The ustc-nercslip systems for the chime-7 dasr challenge," *arXiv preprint arXiv:2308.14638*, 2023.

[21] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *21st Annual conference of the International Speech Communication Association (INTERSPEECH 2020)*. ISCA, 2020, pp. 3830–3834.

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech 2020*, 2020.

[23] Zhang Hongyi and David Lopez-Paz Moustapha Cisse, Yann N. Dauphin, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018, pp. 10–20.

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