SPATIALLY-AUGMENTED SEQUENCE-TO-SEQUENCE NEURAL DIARIZATION FOR MEETINGS

Li Li [1,2], *Ming Cheng* [3], *Juan Liu*[1], *Ming Li* [1,2,**]

[1] School of Artifcial Intelligence, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3] School of Computer Science, Wuhan University, Wuhan, China

lili_a0@163.com, ming.cheng@whu.edu.cn, liujuan@whu.edu.cn, ming.li.cuhksz@gmail.com

ABSTRACT

This paper proposes a Spatially-Augmented Sequence-to-Sequence Neural Diarization (SA-S2SND) framework, which integrates direction-of-arrival (DOA) cues estimated by SRP-DNN into the S2SND backbone. A two-stage training strategy is adopted: the model is first trained with single-channel audio and DOA features, and then further optimized with multi-channel inputs under DOA guidance. In addition, a simulated DOA generation scheme is introduced to alleviate dependence on matched multi-channel corpora. On the AliMeeting dataset, SA-S2SND consistently outperforms the S2SND baseline, achieving a 7.4% relative DER reduction in the offline mode and over 19% improvement when combined with channel attention. These results demonstrate that spatial cues are highly complementary to cross-channel modeling, yielding good performance in both online and offline settings.

***Index Terms*—** Speaker Diarization, Online Speaker Diarization, Sequence-to-Sequence Neural Diarization

1. INTRODUCTION

Speaker diarization aims to answer the "who-spoke-when" question [1], serving as a fundamental pre-processing step for downstream tasks (e.g., speech recognition) [2]. Despite notable progress, speaker diarization in meetings remains challenging due to overlapping speech, unreliable speaker embeddings, reverberation, etc.

Early studies on speaker diarization focus on modularized pipelines [3, 4], which segment audio and cluster them by speaker embedding similarity. Assuming each segment contains only one speaker, these methods perform poorly on overlaps. End-to-End Neural Diarization (EEND) addresses this by formulating diarization as a multi-label prediction task, achieving greater robustness [5, 6]. More recently, Target-Speaker Voice Activity Detection (TSVAD) combines modular and neural methods with strong performance [7–9], while Sequence-to-Sequence Neural Diarization (S2SND) further advances online diarization [10]. However, most approaches still rely solely on acoustic embeddings, which are often unreliable in real meetings. In contrast, spatial cues provide an orthogonal source of information, since speakers typically occupy different physical locations. This raises a key question: how can spatial cues from multi-channel recordings improve diarization?

In multi-channel processing, spatial cues can be integrated in three ways: 1) speech enhancement such as beamforming to generate cleaner inputs [11, 12], though they may often introduce distortions that harm speaker discrimination; 2) channel-fusion or attention modules that aggregate signals [13], but typically perform blind fusion rather than true localization; and 3) explicit features like direction-of-arrival (DOA) estimates [14]. Among them, explicit DOA provides direct directional evidence to separate simultaneous speakers and thus holds greater promise for meeting diarization.

To effectively integrate DOA cues into diarization, two challenges must be solved: robust extraction of high-quality DOA under noise, reverberation, and multi-source conditions, and effective fusion of these cues into diarization models for both online and offline scenarios. Traditional localization methods (e.g., statistical filtering [14], HMM-based clustering [15], steered response power (SRP) [16]) have been studied, but recent deep learning approaches show greater potential [17, 18]. In particular, SRP-DNN [18] learns direct-path phase differences and builds an enhanced SRP spectrum with iterative peak detection, achieving robust multi-speaker DOA estimation in adverse conditions. We therefore adopt SRP-DNN for DOA extraction and S2SND [10] as the diarization backbone.

Building on these observations, we propose SA-S2SND (Spatially-Augmented S2SND). It injects DOA cues from SRP-DNN as explicit auxiliary inputs into the S2SND backbone, enhancing discriminability and improving performance in both online and offline modes. To improve generalization and decouple spatial cues from specific arrays, we introduce a simulated DOA strategy: real multi-channel speech is paired with estimated DOA, while simulated multi-channel speech (from single-channel data) is paired with simulated DOA. This reduces reliance on large-scale multi-channel corpora and improves adaptability to diverse arrays and conditions. Our main contributions are as follows:

- We propose SA-S2SND, which integrates DNN-derived DOA as explicit spatial input to S2SND for online and offline diarization.
- We design a simulated-DOA method that decouples spatial cues from array design, enabling effective use of spatial information without large multi-channel corpora.
- We validate SA-S2SND on the AliMeeting dataset, showing consistent DER improvements over S2SND baselines in both modes.

2. METHODS

In this section, we introduce the method with two components: (1) revisiting direction-of-arrival (DOA) estimation from multi-channel inputs using SRP-DNN [18]; and (2) integrating these DOA cues into the sequence-to-sequence neural diarization (S2SND) backbone [10] to form our proposed SA-S2SND framework.

2.1. DOA Estimation with SRP-DNN

The first component is the direction-of-arrival (DOA) estimation module. We adopt SRP-DNN [18], which estimates robust DOAs for multiple sources by learning direct-path inter-channel phase differences (DP-IPDs) with a causal CRNN. This model is lightweight (0.86M) and easy to retrain for different array configurations. For

**Indicates the corresponding author.

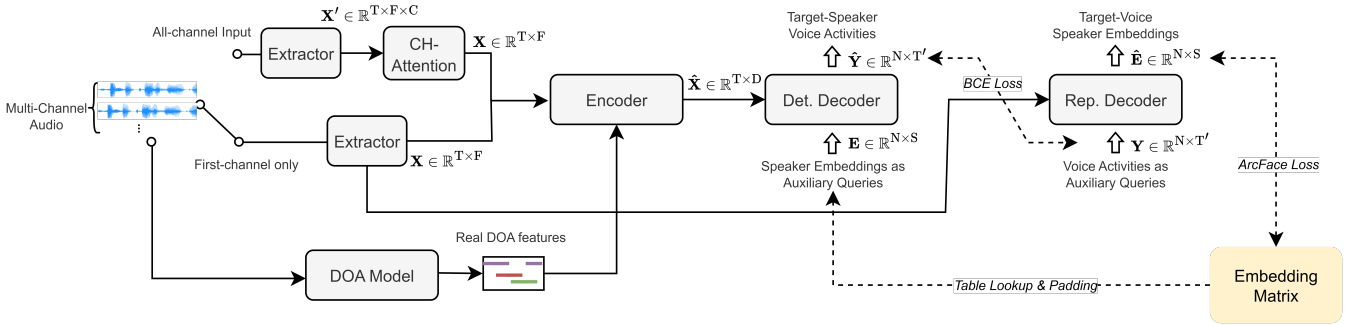

Fig. 1. Spatially-Augmented Sequence-to-Sequence Neural Diarization (SA-S2SND) framework. *Det.* and *Rep.* denote the abbreviations of detection and representation, respectively.

the k-th source, DOA is represented as $\boldsymbol{\theta}_k = \left[\theta_k^{\text{ele}}, \theta_k^{\text{azi}}\right]^T$, with elevation $\theta_k^{\text{ele}} \in [0, \pi]$ and azimuth $\theta_k^{\text{azi}} \in [-\pi, \pi)$.

Specifically, given M microphones, the network extracts multi-channel short-time features (log-magnitude and phase of microphone pairs) and outputs a *summed DP-IPD* vector $\widehat{R}_{mm'}(n)$ for each frame n. The training target is a weighted sum of direct-path IPD vectors:

$$R_{mm'}(n) = \sum_{k=1}^{K} \beta_k(n)\, r_{mm'}(\boldsymbol{\theta}_k(n)), \tag{1}$$

where $r_{mm'}(\boldsymbol{\theta})$ is the DP-IPD vector of candidate DOA and $\beta_k(n) \in [0, 1]$ is the activity probability of source k. This avoids permutation issues, and the network minimizes the MSE between $\widehat{R}_{mm'}(n)$ (prediction) and $R_{mm'}(n)$. From $\{\widehat{R}_{mm'}(n)\}$, an SRP-style spatial spectrum is constructed:

$$P'(\boldsymbol{\theta}; n) = \frac{2}{M(M-1)F} \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} \Re\{\widehat{R}_{mm'}(n)^{\mathsf{H}} r_{mm'}(\boldsymbol{\theta})\}, \tag{2}$$

where F is the number of frequency bins.

To localize multiple sources, SRP-DNN applies an iterative detection-and-removal (IDL) strategy for each frame n: 1) obtain $\widehat{R}_{mm'}(n)$ via the CRNN; 2) construct $P'(\boldsymbol{\theta}; n)$ and find the dominant peak as candidate DOA; 3) estimate its weight, namely the energy ratio of $\widehat{R}_{mm'}(n)$; 4) remove the source's contribution from $\widehat{R}_{mm'}(n)$ and repeat until the predefined threshold is reached. This yields cleaner spatial spectra and more reliable DOA estimates under reverberant, noisy, and multi-speaker conditions, providing explicit spatial cues for the diarization backbone.

2.2. Proposed SA-S2SND

In this work, we propose SA-S2SND, extending sequence-to-sequence diarization (S2SND) [10] and its multi-channel variant (MC-S2SND) [19] with cross-channel attention. Although these models support both online and offline diarization, they lack explicit spatial information that essential for separating simultaneous speakers in meetings. SA-S2SND addresses this by injecting DOA cues from SRP-DNN into the backbone.

2.2.1. Architecture

Fig. 1 shows the SA-S2SND architecture. It follows the S2SND backbone [10], which performs joint speaker detection and representation with an extractor, an encoder, and two coupled decoders.

The extractor is a ResNet [20] with segmental statistical pooling (SSP) [9], producing frame-level embeddings. A Conformer encoder [21] models long-range dependencies and outputs contextualized features $\hat{X}$. On top, two symmetric decoders operate: 1) the representation decoder (Rep.) uses extractor outputs and voice-activity queries to generate target embeddings $\widehat{E}$; 2) the detection decoder (Det.) uses encoder features $\hat{X}$ and embedding queries to predict activities $\widehat{Y}$. These decoders form an inverse pair, conditioning embeddings on activities and vice versa, thus avoiding clustering or PIT assignment.

SA-S2SND further incorporates spatial cues into this backbone. For each block, the DOA model outputs active-speaker azimuths, $\theta_k^{\text{azi}} \in \mathbb{R}^{T'' \times \hat{N}}$ with $\hat{N} \leq 2$. Since SRP-DNN produces azimuths in $[-180°, 180°)$ at $5°$ resolution, we construct a DOA matrix $\mathbf{O} \in \mathbb{R}^{T'' \times A}$, where each row encodes the probability of activity at a given azimuth, aligned with β in Eq. (1). Because SRP-DNN runs at coarser resolution (T'') than frame-level embeddings (T), $\mathbf{O}$ is upsampled by nearest-neighbor interpolation, projected to dimension D, and fused with acoustic embeddings via residual addition:

$$\mathbf{X} = \mathbf{X} + \text{Linear}_{\mathbb{R}^A \to \mathbb{R}^D}(\texttt{interpolate}(\mathbf{O}))/\sqrt{D}, \tag{3}$$

where D is the hidden dimension. This enriches encoder representations with directional priors, analogous to positional encoding, enabling the model to distinguish speakers that overlap temporally but originate from different spatial locations.

In the multi-channel case, a cross-channel attention module is applied to extractor outputs, following MC-S2SND [19]. While MC-S2SND aggregates frame-level speaker embeddings from different channels without preserving explicit spatial information. In contrast, SA-S2SND injects DOA cues as directional information, providing complementary spatial evidence and further improving performance.

2.2.2. Training Process

We divide training into two parts: *Part A* trains the model with single-channel audio assisted by multi-channel DOA features; *Part B* upgrades it to multi-channel audio (via cross-channel attention) with multi-channel DOA features. The model is trained on fixed-length blocks. Two loss terms are used jointly:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{NT'} \sum_{n,t'} \left[y_{n,t'} \log \hat{y}_{n,t'} + (1 - y_{n,t'}) \log(1 - \hat{y}_{n,t'}) \right] \tag{4}$$

$$\mathcal{L}_{\text{Arc}} = \frac{1}{N} \sum_{n} -\log \frac{e^{s\cos(\theta_n + m)}}{e^{s\cos(\theta_n + m)} + \sum_{i \neq S_n} e^{s\cos\theta_i}} \tag{5}$$

and the total loss is $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Arc}}$. A learnable embedding matrix E_{all} and a non-speech embedding are used to form input enrollment queries via table lookup; absent slots are padded accordingly during mini-batching.

Part A: Single-channel model + Multi-channel DOA (Stages 1–3). The model takes single-channel audio, while DOA features are provided in parallel: SRP-DNN estimates DOA for real recordings, and pseudo-DOA is generated online for simulated mixtures using VAD with random DOA assignment and perturbation.

- **Stage 1.** Initialize a ResNet extractor pretrained on speaker verification, freeze it, and train encoder/decoders on simulated mixtures with pseudo-DOA (LR=*1e-4*).
- **Stage 2.** Unfreeze the extractor and train on 80% simulated (pseudo-DOA) + 20% real (SRP-DNN DOA), jointly optimizing acoustic and spatial cues.
- **Stage 3.** Fine-tune the entire single-channel SA-S2SND with a reduced LR of *1e-5*.

Part B: Multi-channel model + Multi-channel DOA (Stages 4–5). The model is upgraded with cross-channel attention to process multi-channel audio, while DOA is still provided by SRP-DNN.

- **Stage 4.** Add cross-channel attention on extractor outputs; freeze prior parameters and train only this branch. (LR=*1e-4*).
- **Stage 5.** Unfreeze all modules and jointly fine-tune the full multi-channel SA-S2SND with DOA (LR=*1e-5*).

This two-part schedule clarifies the scope: Stages 1–3 train a single-channel model with DOA support, while Stages 4–5 extend it to multi-channel with injected DOA, ensuring a consistent path from acoustic-only to spatially-augmented modeling.

2.2.3. *Inference Process*

The inference of SA-S2SND follows the block-wise sliding-window scheme of S2SND [10]. Each block has left, chunk, and right contexts with lengths L_{left}, L_{chunk}, and L_{right}, giving $L = L_{\text{left}} + L_{\text{chunk}} + L_{\text{right}}$. Setting chunk shift to L_{chunk} yields latency $L_{\text{chunk}} + L_{\text{right}}$.

On top of the block-wise process, SA-S2SND incorporates DOA features: SRP-DNN azimuth cues are interpolated and fused with encoder outputs to align spatial and temporal information. The detection decoder takes a fixed embedding matrix $E = [e_{\text{pse}}, E_{\text{buf}}, E_{\text{non}}]$ (pseudo-speaker, buffered speakers, and non-speech paddings), and outputs $\hat{Y} = [\hat{y}_{\text{pse}}, \hat{Y}_{\text{buf}}, \hat{Y}_{\text{non}}]$. The representation decoder extracts updated embeddings $\hat{E} = [\hat{e}_{\text{pse}}, \hat{E}_{\text{buf}}, \hat{E}_{\text{non}}]$. Buffer management with quality weighting accumulates reliable embeddings and registers new speakers once a threshold is exceeded.

After the first (online) pass, the final embedding buffer enables a second-pass offline decoding, where acoustic and DOA-enhanced representations jointly improve offline results. The inference follows S2SND [10], with DOA injection as the only modification.

3. EXPERIMENTAL SETUP

3.1. Datasets

For simulated data, we use VoxCeleb2 [22] (1M utterances, 6,112 speakers) with on-the-fly mixture generation. Real data come from AliMeeting [23] (104.75h train, 4h eval, 10h test) containing 8-channel far-field array and headset recordings. We use the far-field array signals after dereverberation with NARA-WPE [**]. Additionally, we also use a compound dataset for real data: Alimeeting [23], DIHARD III [24], VoxConverse [25], MISP2022 [26]. The model is trained on the training set, validated on the evaluation set, and tested without oracle VAD or collar tolerance.

[**] https://github.com/fgnt/nara_wpe

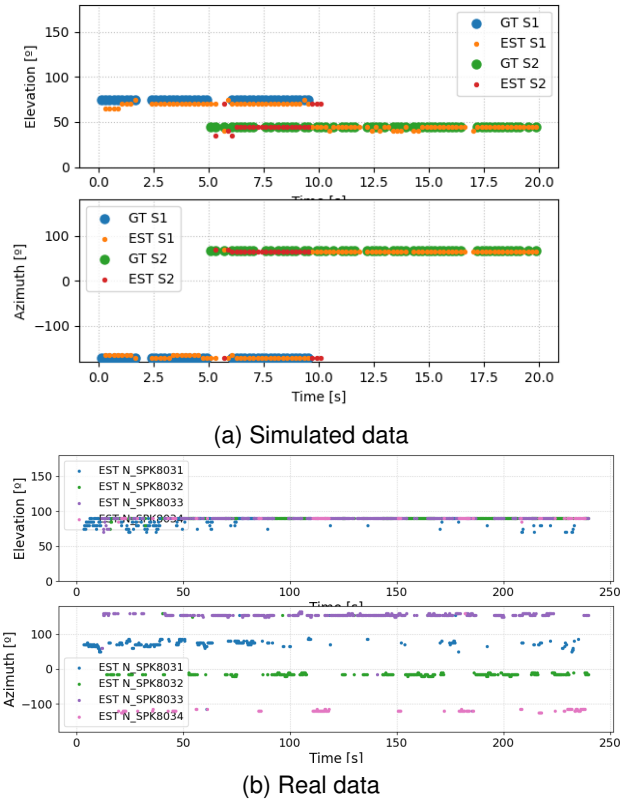

Fig. 2. Illustration of DOA estimation results for (a) simulated data, (b) real Alimeeting data.

Table 1. Overlap ratio of more than two speakers across datasets.

Dataset	Ratio (%)
AliMeeting	6.0
AISHELL-4	0.5
AMI	3.0

3.2. Network Configurations

The network has five components: extractor, channel-attention, encoder, DOA model, and two decoders. We implement S2SND-Small and S2SND-Medium, both using a ResNet-34 [20] extractor. The small version uses residual widths {32,64,128,256}, and the medium {64,128,256,512}. The channel-attention module is a 2-block Transformer [27] with 512-dim 8-head attention and 1024-dim feed-forward layers. The encoder is a Conformer [21] with a kernel size of 15, and both decoders follow the speaker-wise decoder of S2SND [10]. And the two model variants share the same 4-block structure but differ in their hidden dimensions: S2SND-Small (256-dim, 8-head attention; 512-dim FFN) and S2SND-Medium (384-dim attention; 768-dim FFN), totaling 16.56M and 45.96M parameters, respectively. The single- and multi-channel versions differ only by the added channel-attention module. The DOA model (0.86M) is used only for estimation and not updated during training.

3.3. Training and Inference Details

For training, the audio is segmented into 8s windows with 2s overlaps, which are then normalized and converted into 80-dimensional log-Mel filterbanks (25-ms frame length, 10-ms shift). The output resolution is 10 ms [10], with a maximum of $N = 30$ speakers. To ensure permutation invariance, input embeddings and activities

Table 2. Performance on Alimeeting test sets with various model size and inferring conditions. The Diarization Error Rates (DERs) are reported without Oracle VAD and collar tolerance. Ch. Num. means the number of channels for S2SND.

ID	Model Size	Ch. Num.	Usage of DOA	Methods	1-2 SPKs		2+ SPKs		Total	
					Online DER (%)	Offline DER (%)	Online DER (%)	Offline DER (%)	Online DER (%)	Offline DER (%)
E1	Small(16.56M)	1	✗	S2SND	6.41	5.41	20.72	17.58	16.03	13.59
E2	Small(16.56M+0.86M)	1	✓	SA-S2SND	6.35	5.40	19.75	16.10	15.35	12.59
E3	Small(18M)	8	✗	S2SND	5.86	5.29	19.24	16.44	14.85	12.79
E4	Small(18M+0.86M)	8	✓	SA-S2SND	5.90	5.01	16.33	13.68	12.93	10.84
E5	Medium(45.96M)	1	✗	S2SND	6.23	5.46	18.57	16.06	14.53	12.59
E6	Medium(45.96M)	1	✗	S2SND *	6.84	5.59	17.41	14.13	13.94	11.33
E7	Medium(45.96M+0.86M)	1	✓	SA-S2SND *	5.88	5.28	16.33	12.95	12.92	10.40

* We use the compound dataset. The lowest online and offline DERs are highlighted by the gray background.

are randomly shuffled with reassigned labels. Data augmentation includes Musan noise [28] and RIR reverberation [29]. Optimization uses AdamW with BCE and ArcFace losses [30] ($s = 32, m = 0.2$). Experiments run on two RTX-A6000 GPUs.

For inference, the system applies block-wise sliding windows with left/chunk/right contexts. And the online latency is 0.8s (0.64s+0.16s). DOA features are fused at the encoder, and final buffers enable both online decoding and offline re-scoring.

4. RESULTS

4.1. DOA Analysis

It is worth noting that the original SRP-DNN was developed for the 12-channel LOCATA dataset [31] and thus not directly applicable to the 8-channel circular array of AliMeeting. To address this, we retrain the CRNN component of SRP-DNN under the AliMeeting configuration. Fig. 2 illustrates DOA estimation results on both simulated LibriSpeech mixtures and real AliMeeting recordings. The simulated data show negligible DOA errors, while real meeting scenarios exhibit clearly separated azimuth estimates across different speakers. In fact, meeting participants are typically seated, resulting in limited elevation variation, while azimuth provides clear and discriminative spatial cues (Fig. 2(b)). Therefore, azimuth is used as the primary spatial feature. Notably, SA-S2SND is not limited to azimuth-only inputs, and elevation can be incorporated without architectural changes when needed.

Furthermore, SRP-DNN's IDL-based detection tracks at most two speakers per frame, we believe it is reasonable in practice, since frames with more than two active speakers are rare—only 6% in AliMeeting, 0.5% in AISHELL-4 and 3% in AMI (see Table 1), and thus have a negligible impact on overall diarization performance.

4.2. Results on Alimeeting Dataset

Table 2 shows SA-S2SND performance under different training conditions on Alimeeting, from which several conclusions can be drawn.

First, adding DOA consistently improves over baselines for both small and medium model: total DER drops by 4.2% online (16.03→15.35) and 7.4% offline (13.59→12.59), with larger gains in multi-speaker cases, showing better robustness under complex conversational conditions. Second, E2 vs. E3 indicates DOA-based decoupling is more effective than channel fusion offline. Third, E3 vs. E4 shows further DER reductions (12.9% online, 15.2% offline), highlighting complementarity: channel-attention captures correlations, while DOA provides explicit spatial cues. As a result, E4 achieves the best performance among small models, with 19.3%/20.3% relative gains over E1 (online/offline). Finally, scaling to the medium model further improves performance (E5 vs. E1), and training with the compound dataset and DOA achieves the overall best results, demonstrating the scalability and effectiveness of the proposed SA-S2SND framework.

Table 3. Comparisons of our methods with others on the Alimeeting test sets (offline).

Methods	DER (%)
Offline	
Diaper [32]	20.70
EEND-EDA [33]	12.30
Pyannote.audio [34]	15.20
EEND-M2F [35]	13.20
EEND-TA + FT [33]	11.41
WavLM-Large [36] †	10.80
S2SND (E6 in Table 2)	11.33
SA-S2SND (E7 in Table 2)	**10.40**

† State-of-the-art by submission.

4.3. Comparison with Other Existing Methods

Table 3 compares our proposed methods with the previous results on the Alimeeting dataset. In the offline scenario, our proposed SA-S2SND obtain the lowest DERs of 10.40%. The WavLM-large method employs a pretrained WavLM with 63.3M parameters and uses a 16s window for both training and inference (DER=15.1% for 8s window). Despite using a shorter 8s window and without pretraining models, our method achieves state-of-the-art performance.

5. CONCLUSIONS

This work proposes SA-S2SND, a novel spatially-augmented diarization system that integrates direction-of-arrival (DOA) cues into a sequence-to-sequence neural diarization. By incorporating DOA features and a staged training strategy, the model unifies both single- and multi-channel inputs, supporting both online and offline inference. Evaluated on the AliMeeting test set, SA-S2SND achieves significant reductions, particularly when enhanced with channel attention. Future work will focus on improving multi-speaker DOA robustness and further enhancing overall system performance.

6. ACKNOWLEDGEMENT

This research is funded in part by the National Natural Science Foundation of China (62571223) and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

7. GENERATIVE AI USE DISCLOSURE

Generative AI tools were used only for language editing, such as improving clarity and grammar. No substantive content, analysis, or conclusions were generated by AI, and the authors remain fully responsible for this manuscript.

8. REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe *et al.*, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, 2022.

[2] N. Kanda *et al.*, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," in *Proc. INTERSPEECH*, 2020, pp. 36–40.

[3] Q. Wang, C. Downey, L. Wan, Mansfield *et al.*, "Speaker diarization with LSTM," in *Proc. ICASSP*, 2018, pp. 5239–5243.

[4] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, 2022.

[5] Y. Fujita, N. Kanda, S. Horiguchi *et al.*, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. INTERSPEECH*, 2019, pp. 4300–4304.

[6] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[7] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. INTERSPEECH*, 2020, pp. 274–278.

[8] M. Cheng, W. Wang, Y. Zhang, X. Qin, and M. Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *Proc. ICASSP*, 2023, pp. 1–5.

[9] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.

[10] M. Cheng, Y. Lin, and M. Li, "Sequence-to-sequence neural diarization with automatic speaker detection and representation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2719–2734, 2025.

[11] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[12] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer *et al.*, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop*, 2018.

[13] S. Horiguchi, Y. Takashima, P. Garcia, S. Watanabe *et al.*, "Multi-channel end-to-end neural diarization with distributed microphones," in *Proc. ICASSP*, 2022, pp. 7332–7336.

[14] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada *et al.*, "A doa based speaker diarization system for real meetings," in *Hands-Free Speech Communication and Microphone Arrays*, 2008.

[15] J. H. Wong, X. Xiao, and Y. Gong, "Hidden markov model diarisation with speaker location information," in *Proc. ICASSP*, 2021, pp. 7158–7162.

[16] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE transactions on cybernetics*, vol. 46, no. 1, pp. 20–26, 2015.

[17] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[18] B. Yang, H. Liu, and X. Li, "Srp-dnn: Learning direct-path phase difference for multiple moving sound source localization," in *Proc. ICASSP*, 2022, pp. 721–725.

[19] M. Cheng, F. Su *et al.*, "Multi-Channel Sequence-to-Sequence Neural Diarization: Experimental Results for The MISP 2025 Challenge," in *Proc. Interspeech*, 2025, pp. 1898–1902.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[21] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[22] J. S. Chung, A. Nagrani *et al.*, "Voxceleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018, pp. 1086–1090.

[23] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, 2022, pp. 6167–6171.

[24] N. Ryant, P. Singh *et al.*, "The third dihard diarization challenge," in *Proc. INTERSPEECH*, 2021, pp. 3570–3574.

[25] J. S. Chung *et al.*, "Spot the conversation: Speaker diarisation in the wild," in *Proc. INTERSPEECH*, 2020, pp. 299–303.

[26] H. Chen, H. Zhou, J. Du *et al.*, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP*, 2022.

[27] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[30] J. Deng, J. Guo, N. Xue *et al.*, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019.

[31] H. W. Lollmann *et al.*, "The locata challenge data corpus for acoustic source localization and tracking," in *Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 410–414.

[32] F. Landini, M. Diez, T. Stafylakis, and L. Burget, "Diaper: End-to-end neural diarization with perceiver-based attractors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3450–3465, 2024.

[33] S. J. Broughton *et al.*, "Pushing the limits of end-to-end diarization," in *Proc. INTERSPEECH*, 2025, pp. 5218–5222.

[34] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH*, 2023, pp. 3222–3226.

[35] M. Härkönen, S. J. Broughton, and L. Samarakoon, "Eend-m2f: Masked-attention mask transformers for speaker diarization," in *Proc. INTERSPEECH*, 2024, pp. 37–41.

[36] J. Han, P. Pálka, M. Delcroix *et al.*, "Efficient and generalizable speaker diarization via structured pruning of self-supervised models," *arXiv preprint arXiv:2506.18623*, 2025.