MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

SPATIALLY AWARE SELF-SUPERVISED MODELS FOR MULTI-CHANNEL NEURAL SPEAKER DIARIZATION

Han, Jiangyu; Wang, Ruoyu; Masuyama, Yoshiki; Delcroix, Marc; Rohdin, Johan; Du, Jun; Burget, Lukáš

TR2026-047 April 29, 2026

Abstract

Self-supervised models such as WavLM have demonstrated strong performance for neural speaker diarization. However, these models are typically pre-trained on single-channel recordings, limiting their effectiveness in multi-channel scenarios. Existing diarization systems built on these models often rely on DOVER-Lap to combine outputs from individual channels. Although effective, this approach incurs substantial computational overhead and fails to fully exploit spatial information. In this work, building on DiariZen, a pipeline that combines WavLM-based local end- to-end neural diarization with speaker embedding clustering, we introduce a lightweight approach to make pre-trained WavLM spatially aware by inserting channel communication modules into the early layers. Our method is agnostic to both the number of micro- phone channels and array topologies, ensuring broad applicability. We further propose to fuse multi-channel speaker embeddings by leveraging spatial attention weights. Evaluations on five public datasets show consistent improvements over single-channel baselines and demonstrate superior performance and efficiency com- pared with DOVER-Lap.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

SPATIALLY AWARE SELF-SUPERVISED MODELS FOR MULTI-CHANNEL NEURAL SPEAKER DIARIZATION

Jiangyu Han[1], *Ruoyu Wang*[2], *Yoshiki Masuyama*[3], *Marc Delcroix*[4], *Johan Rohdin*[1], *Jun Du*[2], *Lukáš Burget*[1]

[1]Brno University of Technology, Czechia [2]University of Science and Technology of China, China
[3]Mitsubishi Electric Research Laboratories (MERL), USA [4]NTT, Inc., Japan

ABSTRACT

Self-supervised models such as WavLM have demonstrated strong performance for neural speaker diarization. However, these models are typically pre-trained on single-channel recordings, limiting their effectiveness in multi-channel scenarios. Existing diarization systems built on these models often rely on DOVER-Lap to combine outputs from individual channels. Although effective, this approach incurs substantial computational overhead and fails to fully exploit spatial information. In this work, building on DiariZen, a pipeline that combines WavLM-based local end-to-end neural diarization with speaker embedding clustering, we introduce a lightweight approach to make pre-trained WavLM spatially aware by inserting channel communication modules into the early layers. Our method is agnostic to both the number of microphone channels and array topologies, ensuring broad applicability. We further propose to fuse multi-channel speaker embeddings by leveraging spatial attention weights. Evaluations on five public datasets show consistent improvements over single-channel baselines and demonstrate superior performance and efficiency compared with DOVER-Lap. Our source code is publicly available at `https://github.com/BUTSpeechFIT/DiariZen`.

***Index Terms*—** Multi-channel speaker diarization, DiariZen, self-supervised, WavLM, cross-channel communication

1. INTRODUCTION

End-to-end neural diarization with vector clustering (EEND-VC) [1, 2] has emerged as a key approach to speaker diarization. It integrates local EEND with speaker embedding clustering in a unified framework, enabling scalability to long conversations with many speakers. Recent advances in the local EEND module have been driven by self-supervised models such as WavLM [3–5]. However, these models are typically pre-trained on single-channel audio, which limits their ability to exploit spatial information in multi-channel recordings.

Microphone arrays are increasingly common in meeting applications, where multi-channel spatial information helps distinguish speakers by their relative positions. To exploit such cues, prior studies have incorporated explicit multi-channel features [6–10] or applied attention mechanisms across channels [11–13], yet none have utilized self-supervised models. Another common approach is to run single-channel diarization independently on each microphone and then fuse the outputs with DOVER-Lap [14–16]. While effective, methods based on DOVER-Lap are computationally demanding and fail to fully exploit spatial information. Moreover, most existing studies have been evaluated on limited benchmarks, raising concerns about generalization to diverse microphone array configurations.

In this work, we extend DiariZen [5], a speaker diarization pipeline built on EEND-VC that leverages WavLM [3] and Conformer [17] to strengthen the local EEND module. In the local EEND stage, we adapt the pre-trained single-channel WavLM to multi-channel diarization by introducing learnable channel communication modules into its early layers to capture spatial information. Our framework supports various channel communication mechanisms such as ChannelAttention (ChAtt) [11, 12] and Transform-Average-Concatenate (TAC) [18], and is agnostic to both the number of channels and the microphone array topology. To further improve efficiency, we incorporate pruned WavLM [19, 20], which retains strong diarization performance while significantly lowering model complexity. Finally, in the clustering stage, we propose a spatial-attention fusion strategy that effectively integrates multi-channel speaker embeddings, thereby improving clustering performance.

We comprehensively evaluate our method on five public datasets: AMI [21, 22], AISHELL-4 [23], AliMeeting [24], NOTSOFAR-1 [25], and CHiME-6 [26]. The results show consistent benefits over the single-channel baseline and demonstrate superior performance and efficiency compared with the channel fusion using DOVER-Lap. In addition, our EEND model based on pruned WavLM achieves performance comparable to the unpruned model while substantially reducing computational overhead. Moreover, fusing speaker embeddings using multi-channel spatial attention weights in the clustering stage yields significant improvements on CHiME-6, approaching the state of the art [27] with a much simpler and more efficient system that requires no speech separation modules.

2. MULTI-CHANNEL EXTENSIONS

In this section, we extend our DiariZen [5] speaker diarization system, derived from the pyannote pipeline [28,29]. The system follows a two-stage EEND-VC approach: in the first stage, an EEND model detects speaker activities within local overlapping windows; in the second stage, speaker embeddings are extracted from the speech of each detected speaker in each window, and these embeddings are then clustered to map local speaker identities to global ones.

We first introduce the EEND model used in the DiariZen system and explain how we extend it for multi-channel processing. In Section 2.2, we then describe how multi-channel input can be leveraged to obtain more informative speaker embeddings for the second stage.

2.1. Multi-channel EEND

The multi-channel extension of the EEND model in the DiariZen system is illustrated in Figure 1, where LN denotes LayerNorm. Considering only the path from the first-channel input (bold arrows)

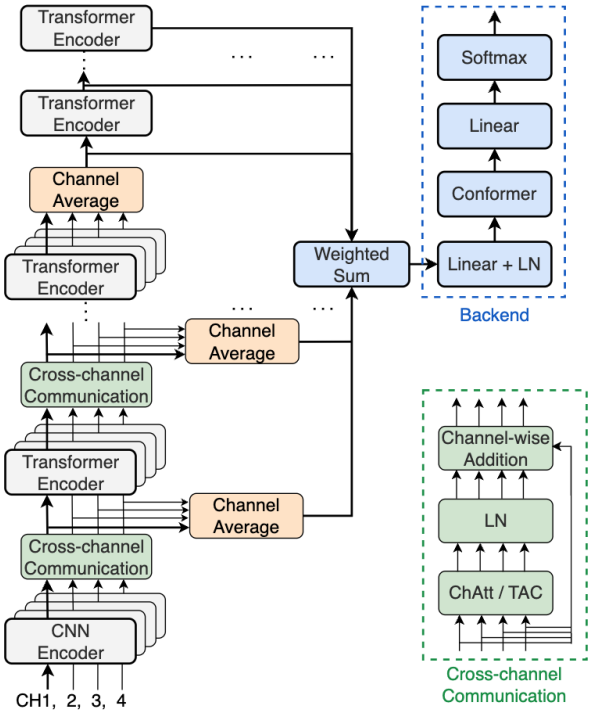

Fig. 1. Framework of the multi-channel WavLM extension with four input microphones and its application to speaker diarization.

and the processing blocks with bold outlines (ignoring the *Channel Average* and *Cross-channel Communication* blocks) recovers the original single-channel EEND model on which the extension is based. This model uses pre-trained WavLM [3] (gray *Encoder* blocks) as a front-end. The output sequences of all WavLM layers are combined via a SUPERB-style weighted sum [30] into a single sequence, which is fed into a Conformer-based backend [17] predicting speaker activities as powerset classes [29]. Further details on the single-channel model and its training can be found in [5].

Following the work on multi-channel speaker verification [31], we extend EEND as follows. Several early WavLM layers (stacked gray blocks in Figure 1) are run in parallel on each channel (with parameters shared). After each such layer, the outputs are averaged across channels (*Channel Average* blocks, in pale orange). These per-layer averaged sequences are then used in the same SUPERB-style combination as in the single-channel case to provide the input to the back-end. To reduce computation, the later WavLM layers operate directly on the aggregated input in a single-channel manner.

To exploit spatial cues encoded in the multi-channel input, we add trainable *Cross-channel Communication* blocks after each multi-channel encoder layer. We examine two variants of such blocks, both invariant to the number of channels and the microphone array configuration. In the first variant, ChannelAttention (ChAtt) [11, 12], information is exchanged across channels using a standard multi-head self-attention [32] layer, followed by LayerNorm and a residual connection. In the second variant, the multi-head self-attention is replaced by the Transform-Average-Concatenate (TAC) mechanism [18], while retaining LayerNorm and the residual connection. For each frame, TAC applies a shared linear layer with PReLU activation to the input from each channel; the transformed representations are then averaged across channels to obtain a global vector, which is concatenated with each original channel input and projected back to the original dimensionality by a linear layer.

To train the multi-channel EEND model, all parameters are initialized from the pre-trained single-channel model, except for those of the newly introduced Cross-channel Communication blocks. These blocks are initialized as identity mappings to preserve the behavior of the pre-trained model at the start of training. Specifically, the scale and bias parameters in the LayerNorm are set to zero, forcing the block output to zero and thereby passing its input directly to the output through the residual connection. This allows the subsequent fine-tuning on multi-channel data to begin from a stable single-channel baseline while gradually learning to exploit spatial information across microphone channels.

2.2. Spatially Attentive Speaker Embeddings

The previous section described the multi-channel extension of the EEND model, used in the first stage of the DiariZen EEND-VC pipeline. Ideally, speaker embedding extraction in the second stage should also leverage multi-channel input—a topic of our ongoing research—but in this work we use a single-channel extractor and process each channel separately. In Section 3.3.2, we analyze strategies to combine speaker embeddings across channels for optimal performance, including the two approaches proposed here.

These approaches exploit spatial attention weights from a pre-trained ChannelAttention module to select or combine embeddings. Let $\mathbf{S} \in \mathbb{R}^{T\times H\times C\times C}$ be the attention weights from a selected layer, with T frames, H heads, and C channels. Averaging across frames and heads yields a global representation $\mathbf{S}_{\text{g}} \in \mathbb{R}^{C\times C}$, as shown in Figure 3. We then average $\mathbf{S}_{\text{g}}$ over rows (queries) to obtain channel weights $\hat{\mathbf{S}}_{\text{g}} \in \mathbb{R}^{C}$. The per-channel embeddings for each speaker in each local window are then combined across channels, either by (i) selecting the highest-weighted channel (*attentive argmax*) or (ii) computing a weighted average (*attentive weighted fusion*). This procedure requires no additional training, as it directly uses intermediate representations from the pre-trained ChannelAttention module.

3. EXPERIMENTS

3.1. Datasets

We train and evaluate our system on a compound dataset combining five public datasets: AMI [21, 22], AISHELL-4 (AIS-4) [23], AliMeeting (Ali) [24], NOTSOFAR-1 (NSF-1) [25], and CHiME-6 (CH-6) [26]. Since AISHELL-4 lacks a development set, 10% of its training data from each room is randomly selected for validation and the rest for training. For CHiME-6, we apply WPE [33] and BeamformIt [34] to each array, generating beamformed audio and reducing the number of channels from 24 to 6.

3.2. Configurations

Our EEND models are based on WavLM Base+ and its pruned variant with 80% sparsity, both initialized from the pretrained DiariZen models [20]. We insert Cross-channel Communication modules into the first four layers of WavLM models, with input and hidden dimensions of 768 and 256, respectively, and employ 8 heads for the ChannelAttention mechanism. For model training and inference, we use the same hyper parameters as in [5]. VBx [35] is used as the clustering method to determine the speaker mappings between the EEND local windows. We measure performance with diarization error rate (DER), applying a 0.25-second collar for CHiME-6 and none for other datasets. We also report the macro-averaged DER to reflect

overall performance across datasets. Our source code and detailed configurations are publicly available[1].

3.3. Results and Discussion

3.3.1. Performance under oracle clustering

Because the proposed multi-channel EEND extensions affect only the local EEND output, we first evaluate them with oracle clustering. Specifically, instead of performing clustering, the speakers detected in the local windows are assigned global (ground-truth) speaker labels so as to minimize the overall DER. Results are shown in Table 1.

As a baseline, we use single-channel systems, where **audio from the first channel** is used for both training and evaluation. For ChannelAttention (ChAtt) and Transform-Average-Concatenate (TAC), we use the first four channels to simplify the experiments and accelerate validation, while still retaining spatial information from multiple, physically distinct microphones. A checkmark denotes the use of the pruned WavLM Base+ model with 80% sparsity [20]; otherwise, the unpruned model is applied. Our earlier work [19, 20] showed that up to 80% of WavLM parameters can be removed through structured pruning, yielding substantial speedups without degrading diarization accuracy. However, its effectiveness in multi-channel processing has not been studied prior to this work.

As shown in Table 1, the pruned WavLM performs slightly worse under the single-channel condition, but this degradation disappears in multi-channel scenarios. Although extending pre-trained models to multi-channel has typically been computationally demanding, our results suggest that the pruned model can be effectively applied, substantially reducing computational complexity. We therefore adopt the pruned WavLM for all subsequent experiments.

Across most of datasets, different cross-channel communication methods achieve comparable performance, all outperforming the single-channel baseline. Therefore, we adopted ChannelAttention for the remaining experiments due to its simplicity and intuitive mechanism for multi-channel communication.

3.3.2. Performance under VBx clustering

While previous experiments demonstrate the effectiveness of our multi-channel EEND extensions, they are restricted to oracle clustering with four channels. To evaluate performance under more realistic conditions, Table 2 presents results using VBx [35] for speaker embedding clustering, where all microphone channels are utilized.

The first section of Table 2 reports baseline results using a single-channel EEND model trained on recordings from **randomly selected microphone channels**. The first line shows results with only the first channel as input. The second line corresponds to a widely used multi-channel diarization strategy [15, 16], in which both stages of the EEND-VC pipeline are applied separately to each channel, and the channel-wise diarization outputs are then combined with DOVER-Lap [14]. The third line (Average probs & embs) corresponds to a system where EEND is applied separately to each channel, but the output probabilities are averaged across channels to form a fused output. This fused output is then used as the frame-level speaker activity reference for extracting embeddings from each channel, which are subsequently averaged across channels for clustering. This simple averaging outperforms DOVER-Lap on CHiME-6, exposing its limitations in some conditions.

The following results are obtained with the multi-channel EEND model using ChannelAttention, which produces a single diarization output. However, speaker embeddings are still extracted separately from each channel. For the line "ChAtt, DOVER-Lap", clustering is applied independently to each channel, yielding per-channel global diarization outputs that are then combined with DOVER-Lap. This approach clearly outperforms the single-channel baseline, with large gains on AliMeeting and CHiME-6, but offers no extra advantage over the other two baselines except on AliMeeting. In the next line, speaker embeddings are averaged across channels before clustering, yielding a single global diarization output. This reduces the DER on CHiME-6 to 28.5, demonstrating the benefit of incorporating spatial information into the embeddings through simple aggregation.

The last two lines of Table 2 show results for two multi-channel approaches introduced in Section 2.2: *attentive argmax* for selecting embeddings, and *attentive weighted fusion* for combining embeddings. Relative to the DOVER-Lap baseline in the second row, the attentive argmax approach shows substantial gains on AliMeeting and CHiME-6 while performing comparably on the other datasets. It is also more efficient, since local EEND and speaker embedding clustering are performed only once rather than per channel. On CHiME-6, attentive weighted fusion further reduces DER to 27.5, approaching the performance of top-ranked systems [27], while maintaining a simpler design without speech separation.

3.3.3. Analysis of inference time

In Figure 2, we compare the inference time of different methods across stages on a single A5000 GPU. Inference uses a batch size of 32 and the AMI EN2002a recording, about 2143 seconds long with eight channels. The macro DER over all datasets is also reported. Compared with the single-channel baseline, DOVER-Lap substantially improves performance but incurs considerable computational overhead. Our method is consistently more efficient than DOVER-Lap, particularly with the attentive argmax approach. Among the three stages, speaker-embedding extraction dominates computational cost. Extracting embeddings only from the microphone channel with the highest attention score can significantly reduce this overhead. These findings indicate that our method provides a more practical and efficient alternative to DOVER-Lap.

Table 1. Performance comparison under oracle clustering. Multi-channel systems use first 4 channels for training.

System	WavLM		DER (%)					Macro
	Pruned	Params	AMI	AIS-4	Ali	NSF-1	CH-6	
Single-channel	-	94.4M	13.5	8.9	12.5	14.2	24.9	14.8
	✓	18.8M	13.3	9.3	12.7	14.2	25.7	15.0
ChAtt	-	94.4M	13.1	9.2	12.2	14.0	22.8	14.3
ChAtt	✓	18.8M	12.9	9.0	11.8	13.8	22.9	14.1
TAC	✓	18.8M	12.8	8.9	12.0	14.1	22.9	14.1

Table 2. Performance comparison under VBx clustering. Multi-channel systems use all available channels for training.

System	DER (%)					Macro
	AMI	AIS-4	Ali	NSF-1	CH-6	
Single-channel	15.3	11.3	15.0	17.7	33.8	18.6
DOVER-Lap	14.7	10.9	13.5	17.1	30.9	17.4
Average probs & embs	14.9	11.0	14.0	17.5	28.8	17.2
ChAtt, DOVER-Lap	14.8	11.0	12.8	17.4	31.3	17.5
ChAtt, average embed.	14.9	11.1	12.9	17.6	28.5	17.0
ChAtt, att. argmax	14.9	11.0	12.8	17.5	29.5	17.2
ChAtt, att. weighted fusion	14.8	11.2	12.8	17.4	27.5	16.7

[1] https://github.com/BUTSpeechFIT/DiariZen

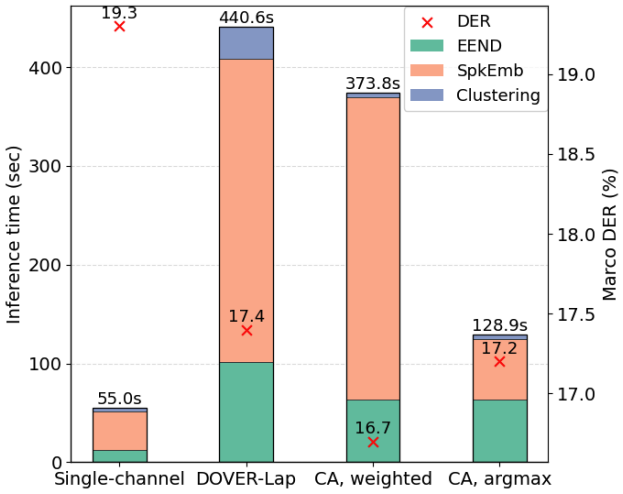

Fig. 2. Inference time at different stages for various methods. The results are averaged over five runs. CA denotes ChannelAttention.

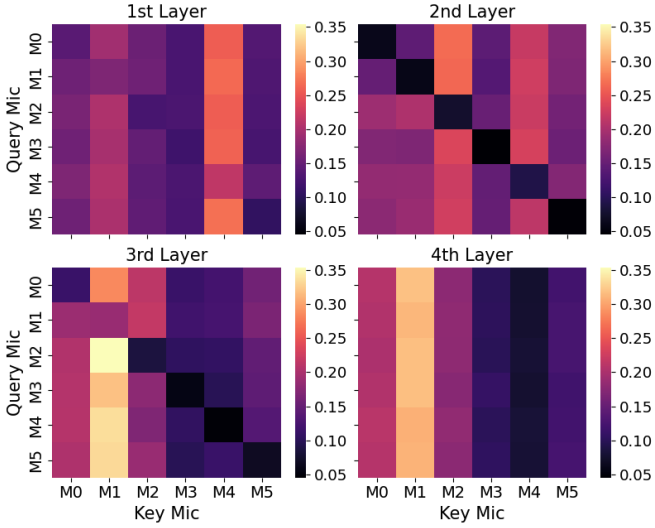

Fig. 3. Layer-wise channel attention weights averaged over frames and attention heads on CHiME-6 (S21, 100th chunk).

3.3.4. Analysis of attention scores

Since our method achieves the largest gains on CHiME-6, we further analyze the channel attention weights using the 100th chunk of session S21 from the CHiME-6 evaluation set. As shown in Figure 3, results indicate that different layers exhibit distinct behaviors: in the second layer, the model tends to ignore the current microphone, whereas in the fourth layer all queries assign higher weights to the first three channels. Moreover, we observe that attention patterns vary across chunks and sessions, indicating that the system adapts to speaker-position changes by exploiting multi-channel spatial cues.

3.3.5. Analysis of microphone dependence

Our earlier experiments demonstrate substantial multi-channel improvements on AliMeeting and CHiME-6, while gains on the other three datasets are more modest. To understand this, we examine two systems: a single-channel model trained on randomly selected microphone audio, and ChannelAttention trained on all available

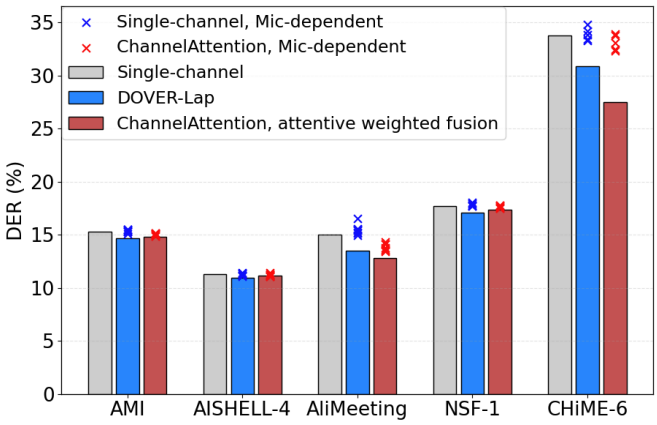

Fig. 4. Performance comparison of single-channel (Mic-dependent) and multi-channel methods. NSF-1 denotes NOTSOFAR-1 dataset.

microphones. At inference, the single-channel model is applied separately to each microphone channel to obtain microphone-dependent (Mic-dependent) results. For ChannelAttention, similar Mic-dependent results are derived by clustering speaker embeddings extracted from each microphone individually.

Figure 4 compares the Mic-dependent results across all channels (red and blue cross markers, $\times$) with the single-channel system, DOVER-Lap fusion, and ChannelAttention with attentive weighted fusion. Substantial variance in Mic-dependent DERs is observed for AliMeeting and CHiME-6, likely due to their recording configurations. For instance, CHiME-6 microphones are distributed across multiple rooms, leading to pronounced performance gaps. This variability highlights the value of spatial information and explains the observed gains. By comparison, AMI, AISHELL-4, and NOTSOFAR-1 exhibit more consistent recording conditions, which limits the potential benefits of multi-channel modeling.

4. CONCLUSION

In this work, we proposed an efficient and general approach for extending pre-trained single-channel EEND models to multi-channel speaker diarization. By inserting channel communication modules into the early layers and initializing them to preserve the pre-trained model's behavior, our method captures spatial cues without disrupting original representations. We further introduced strategies to combine speaker embeddings using spatial attention weights, requiring no additional training. Importantly, the approach is agnostic to the number of microphone channels and array topology. Experiments on five public datasets show that our method consistently outperforms the single-channel baseline and surpasses DOVER-Lap, while being significantly more efficient. On CHiME-6, it achieves strong performance without relying on speech separation. These results highlight the practicality and scalability of the proposed framework for real-world multi-channel diarization.

5. ACKNOWLEDGEMENTS

The work was supported by Ministry of Education, Youth and Sports of the Czech Republic (MoE) through the OP JAK project "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications" (ID:CZ.02.01.01/00/23_020/0008518). Computing on IT4I supercomputer was supported by MoE through the e-INFRA CZ (ID:90254).

6. REFERENCES

[1] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *Proc. ICASSP*. IEEE, 2021, pp. 7198–7202.

[2] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Proc. Interspeech*, 2021, pp. 3565–3569.

[3] S. Chen, C. Wang, Z. Chen, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] N. Tawara, M. Delcroix, A. Ando, et al., "NTT speaker diarization system for CHiME-7: multi-domain, multi-microphone end-to-end and vector clustering diarization," in *Proc. ICASSP*, 2024, pp. 11281–11285.

[5] J. Han, F. Landini, J. Rohdin, et al., "Leveraging self-supervised learning for speaker diarization," in *Proc. ICASSP*, 2025, pp. 1–5.

[6] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 257–264.

[7] S. Araki, M. Fujimoto, K. Ishizuka, et al., "A DOA based speaker diarization system for real meetings," in *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2008, pp. 29–32.

[8] T. Hager, S. Araki, K. Ishizuka, et al., "Handling speaker position changes in a meeting diarization system by combining doa clustering and speaker identification," in *Proc. IWAENC*, 2008, vol. 106.

[9] N. Zheng, N. Li, J. Yu, et al., "Multi-channel speaker diarization using spatial features for meetings," in *Proc. ICASSP*, 2022, pp. 7337–7341.

[10] T. Cord-Landwehr, T. Gburrek, M. Deegen, et al., "Spatio-spectral diarization of meetings by combining tdoa-based segmentation and speaker embedding-based clustering," in *Proc. Interspeech*, 2025, pp. 5223–5227.

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

[12] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *Proc. ICASSP*, 2022, pp. 9171–9175.

[13] S. Wu, J. Du, M. He, et al., "Semi-supervised multi-channel speaker diarization with cross-channel attention," in *Proc. ASRU*, 2023, pp. 1–8.

[14] D. Raj, L. Garcia-Perera, Z. Huang, et al., "Dover-lap: A method for combining overlap-aware diarization outputs," in *Proc. SLT*, 2021, pp. 881–888.

[15] N. Kamo, N. Tawara, A. Ando, et al., "Microphone array geometry-independent multi-talker distant asr: NTT system for dasr task of the CHiME-8 challenge," *Computer Speech & Language*, vol. 95, pp. 101820, 2026.

[16] R. Wang, J. Du, S. Niu, et al., "Three-stage modular speaker diarization collaborating with front-end techniques in the CHiME-8 NOTSOFAR-1 challenge," *Computer Speech & Language*, p. 101863, 2025.

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

[18] Y. Luo, Z. Chen, N. Mesgarani, et al., "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, 2020, pp. 6394–6398.

[19] J. Han, F. Landini, J. Rohdin, et al., "Fine-tune before structured pruning: Towards compact and accurate self-supervised models for speaker diarization," in *Proc. Interspeech*, 2025, pp. 1583–1587.

[20] 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.

[21] J. Carletta, S. Ashby, et al., "The AMI meeting corpus: A pre-announcement," in *Proc. MLMI*, 2005, pp. 28–39.

[22] W. Kraaij, T. Hain, M. Lincoln, et al., "The AMI meeting corpus," in *Proc. Int. Conf. Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.

[23] Y. Fu, L. Cheng, S. Lv, et al., "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proc. Interspeech*, 2021, pp. 3665–3669.

[24] F. Yu, S. Zhang, Y. Fu, et al., "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, 2022, pp. 6167–6171.

[25] A. Vinnikov, A. Ivry, A. Hurvitz, et al., "Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," in *Proc. Interspeech*, 2024, pp. 5003–5007.

[26] S. Watanabe, M. Mandel, J. Barker, et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[27] R. Wang, M. He, J. Du, et al., "The ustc-nercslip systems for the chime-7 dasr challenge," in *Proc. CHiME*, 2023.

[28] H. Bredin, "pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. Interspeech*, 2023, pp. 1983–1987.

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

[30] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, et al., "Superb: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[31] L. Mošner, R. Serizel, L. Burget, et al., "Multi-channel extension of pre-trained models for speaker verification," in *Proc. Interspeech*, 2024, pp. 2135–2139.

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

[33] L. Drude, J. Heymann, C. Boeddeker, et al., "NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Proc. ITG Symposium Speech Communication*, 2018, pp. 1–5.

[34] 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.

[35] F. Landini, J. Profant, M. Diez, et al., "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.