



Leveraging Self-Supervised Learning Based Speaker Diarization for MISP 2025 AVSD Challenge

Zeyan Song^{1,2}, Tianchi Sun^{1,2}, Ronghui Hu^{1,2}, Kai Chen^{1,2}, Jing Lu^{1,2}

¹Key Laboratory of Modern Acoustics, Nanjing University, China

²NJU-Horizon Intelligent Audio Lab, Horizon Robotics, China

{zeyan.song, tianchi.sun, ronghui.hu}@smail.nju.edu.cn, {chenkai, lujing}@nju.edu.cn

Abstract

This paper presents the submission of our team to the audio-visual speaker diarization (AVSD) track of the Multimodal Information Based Speech Processing (MISP) 2025 Challenge. The submitted system is adapted from the DiariZen pipeline, with a primary focus on optimizing it for the challenge dataset. The pipeline consists of a WavLM based local end-to-end neural diarization module followed by two different clustering methods. To further refine the results, DOVER-Lap is employed to integrate results across different input channels and clustering methods. Our final submission system achieves a diarization error rate (DER) of 8.33% on the evaluation set, representing a relative improvement of 46.3% compared to the baseline and ranking 3rd in the AVSD track of this challenge.

Index Terms: MISP Challenge, speaker diarization

1. Introduction

Speaker diarization is the task of dividing an audio recording into segments based on speaker identity [1]. This task plays a vital role in various real-world applications such as meeting transcription [2] and telephone conversation analysis [3]. The AVSD track of the MISP 2025 Challenge [4, 5] focuses on audio-visual speaker diarization. However, we find it hard to improve the performance by introducing the video input, since the mislabeled data constitute a significant obstacle. Therefore, we adopt an audio-only approach for our final system.

Over the years, various methods have been developed to build effective diarization systems. Among them, one of the most commonly used approaches is EEND-VC [6], which first applies end-to-end neural diarization (EEND) to short chunks of the audio recording and then merges the local diarization results through speaker embedding extraction and clustering. This approach leverages the strengths of both clustering-based techniques and EEND, making it particularly effective for handling complex scenarios, including overlapped speech and multi-speaker environments. In this paper, we utilize DiariZen¹ [7], a well-established speaker diarization system following the EEND-VC pipeline, to optimize diarization results on the challenge dataset.

DiariZen is built upon the EEND-VC pipeline by integrating WavLM [8] and Conformer [9] for enhanced performance, which has demonstrated state-of-the-art (SOTA) performance on several speaker diarization datasets including AMI [10] and AISHELL-4 [11]. During training, it predicts local diarization assignments from short speech segments of just a few seconds using EEND. At inference time, it extracts speaker embeddings from these short segments based on the EEND-derived diariza-

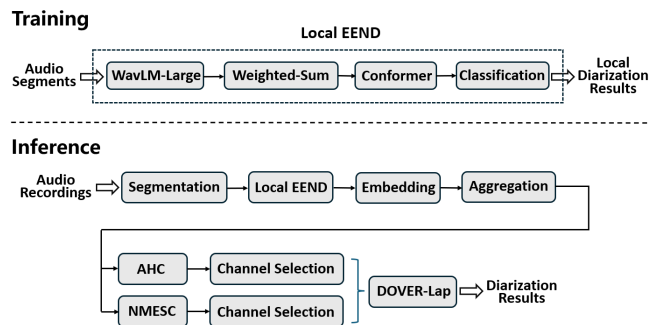


Figure 1: The schematic diagram of the proposed system.

tion labels and then clusters the embeddings to aggregate local results into a global speaker assignment. This approach effectively combines the advantages of EEND for handling short segments with a clustering framework to produce coherent, overall speaker diarization results.

To tailor DiariZen specifically for the AVSD track of the challenge and boost performance on its complex, domain-specific data, we introduce several key modifications. First, we retrain DiariZen using only the challenge dataset, ensuring domain relevance. Next, we slightly modify the EEND module to enhance its ability to handle overlapped speech and multiple speakers. In the vector-clustering stage, we experiment with different speaker embedding models and clustering approaches to further refine the ability of the system to distinguish speakers in diverse acoustic conditions. Finally, we apply a range of post-processing techniques including channel selection, DOVER-Lap fusion, and boundary shifting to further improve performance, yielding a more robust overall system for the challenge. Combining the strategies above, our system achieves an 8.33% DER on the evaluation set of the AVSD track.

2. System description

2.1. System overview

Figure 1 illustrates the pipeline of our system. First, we segment a long audio recording into shorter chunks of a few seconds each and apply an end-to-end neural diarization model to these chunks. Next, within each chunk, we identify single-speaker segments to extract speaker embeddings for all detected speakers. Then we aggregate the speaker embeddings from all chunks and apply two clustering methods: agglomerative hierarchical clustering (AHC) and normalized maximum eigengap-based spectral clustering (NMESC) [12]. Finally, we repeat the

¹<https://github.com/BUTSpeechFIT/DiariZen>

pipeline for every available input channel. For each clustering method, we select only those channels which predicted number of speakers meets or exceeds the oracle number of speakers to avoid underestimation. We combine the diarization outputs from all selected channels across both clustering methods using DOVER-Lap [13] to produce the final result.

2.2. EEND module

The EEND module is designed to generate local diarization outputs from short audio segments in an end-to-end manner, leveraging the representational ability of neural networks. It is built on the DiariZen backbone, replacing traditional EEND components with WavLM and Conformer. Specifically, the module comprises a feature extraction block consisting of WavLM, a weighted-sum layer, a linear transformation with layer normalization, followed by a Conformer model, and a classification head composed of a linear layer and a softmax function. Instead of the original DiariZen pipeline, which adopts WavLM-Base+, we opt for WavLM-Large for its stronger capabilities in speech processing tasks [14]. A linear layer following WavLM reduces the feature dimension from 1024 to 256. We then employ four Conformer blocks with 256-dimensional features, ensuring consistency with the output of the feature extraction block.

Powerset loss [15] is employed for EEND model training because it empirically yields better performance. After examining the dataset statistics, we set the maximum number of speakers per chunk to four and the maximum number of overlapping speakers per frame to three, resulting in a total of 15 powerset classes. This configuration helps reduce speaker confusion errors in the DER calculation.

During training, we do not freeze the parameters of WavLM-Large. Instead, we use a smaller learning rate (1×10^{-5}) compared to the other EEND modules (1×10^{-3}). While the parameters of WavLM are fine-tuned from a pretrained model², the other parts in the EEND pipeline including the weighted-sum layer and Conformer blocks are trained from scratch.

We determine the optimal chunk size through experiments, finding that a 16-second window with a 12-second hop size slightly outperforms other configurations. Consequently, we adopt this setup in all subsequent experiments.

Another advantage of the DiariZen pipeline is that it does not require a large amount of training data for EEND, unlike many other approaches that rely on extensive simulated data. As a result, our system is trained on the challenge training set which contains fewer than 120 hours of recorded speech.

2.3. Speaker embedding module

For the speaker embedding module, we use pretrained models³ from the WeSpeaker toolkit [16] without any additional training or fine-tuning. Once local EEND is applied, we isolate single-speaker segments from the diarization results of each short chunk. We then concatenate these segments to extract the corresponding speaker embeddings.

We experiment with three different pretrained speaker models for this challenge. Table 1 presents details on different models and their respective equal error rates (EER) on popular speaker recognition tasks.

Speaker embedding extraction is performed only during the

²<https://github.com/microsoft/unilm/tree/master/wavlm>

³<https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md>

Table 1: EER (%) on VoxCeleb-O or CNCeleb Evaluation Sets for different speaker embedding models trained on VoxCeleb or CNCeleb using the WeSpeaker toolkit (“-CN” and “-Vox” indicate training and evaluation on CNCeleb or VoxCeleb, respectively).

Model	EER%
ResNet34-LM-CN	6.492
ResNet34-LM-Vox	0.723
ResNet293-LM-Vox	0.447

inference stage. Specifically, the local window size for inference remains 16 seconds, while the hop size is reduced to 1.6 seconds to ensure finer granularity in speaker representation.

2.4. Clustering module

We explore two clustering methods after extracting speaker embeddings: AHC and NMESC. Through our experiments, we find AHC consistently produces lower DER than NMESC because NMESC is more prone to inaccurately estimating the number of speakers, often leading to overestimation or underestimation in the development and evaluation sets of the challenge.

The AHC implementation closely follows DiariZen which is adopted from the pyannotate toolkit⁴. Additionally, we integrate the NMESC pipeline from the NeMo toolkit⁵ into our system to explore an alternative clustering approach for speaker diarization.

2.5. Post-processing module

The post-processing module for this challenge consists of three key components: predicted speaker number based channel selection, DOVER-Lap fusion on the selected channels and clustering methods, and shifting the hypothesized diarization results. These steps aim to refine the diarization output by selecting the most reliable channels, improving speaker assignment through late fusion, and aligning the final results for better accuracy.

2.5.1. Channel selection

Despite the availability of 8-channel audio in the challenge, we only use the first channel in our training process. After training the local EEND model and performing inference, followed by speaker embedding extraction and clustering, we observe unstable performance on both the development and evaluation sets. We attribute this instability to the difficulty in accurately estimating the number of speakers, as each audio recording in development and evaluation sets contains 4 to 8 speakers, and the speaker embedding models are not further fine-tuned on this specific dataset. The lack of fine-tuning may have limited the ability of the model to distinguish between similar speaker characteristics, leading to a higher speaker confusion rate and ultimately resulting in poorer DER performance.

To address these challenges, we try applying DOVER-Lap across all channels as well as using multi-channel inputs for local EEND, but neither approach yields stable results. Since the evaluation set provides the oracle number of speakers for each

⁴<https://github.com/pyannotate/pyannotate-audio>

⁵<https://github.com/NVIDIA/NeMo>

recording, we investigate whether our system accurately predicts them. Unfortunately, in most channels, the model either overestimates or underestimates the actual speaker count.

Further analysis reveals that underestimation is more detrimental than overestimation. When the system underestimates the number of speakers, it merges distinct speakers into a single cluster, resulting in numerous misassigned segments. This substantially increases the speaker confusion rate in DER. Overestimation, on the other hand, has a less severe impact on overall performance.

Consequently, we refine our use of multi-channel information by running inference on all channels with our single-channel pipeline, then selectively retaining only those channels that correctly predict or overestimate the number of speakers. We apply DOVER-Lap to fuse the diarization results of these selected channels rather than fusing all channels indiscriminately. This strategy yields more stable performance and effectively mitigates the risks associated with underestimation.

2.5.2. DOVER-Lap

DOVER-Lap is a technique for combining diarization outputs from overlap-aware diarization systems, which has proven effective in various challenges including VoxSrc [17] and DISPLACE [18]. Its primary goal is to merge the outputs from multiple diarization systems or channels into a single, coherent result, ultimately improving robustness and accuracy.

After running inference on all channels, we selectively retain only those channels that do not underestimate the number of speakers, then apply DOVER-Lap on these selected channels. This approach produces lower DER than single-channel results and significantly outperforms the strategy of applying DOVER-Lap across all channels.

To further enhance performance, on the evaluation set, we run inference on all channels using both AHC and NMESC, thereby obtaining two diarization outputs per channel. We utilize the same speaker embeddings for both clustering methods, while the EEND module remains fixed. Finally, we perform DOVER-Lap on the selected channels from both clustering outputs, leading to additional improvements in overall performance.

2.5.3. Shifting

In this challenge, DER is calculated with a 0-second collar, making the evaluation especially sensitive to minor timing offsets. We observe that shifting the entire diarization result by 0.03 seconds improves performance on both development and evaluation sets. A possible explanation is that the oracle Rich Transcription Time Marked (RTTM) files are generated using the close-talk microphone timestamps of each speaker, while the predictions of our system are derived from far-field microphones. This mismatch between near- and far-field devices, combined with possible hardware latency, leads to a slight misalignment between the reference annotations and system predictions.

By shifting the predicted diarization results earlier, we align the estimated boundaries more closely with the reference. Our experiments demonstrate that bringing the hypothesized segment boundaries closer to the true onset and offset times effectively reduces the overall DER.

Table 2: Performance comparison of different WavLM versions in EEND module on the development set. “Pretrained” refers to a system that directly utilizes the pretrained checkpoints of DiariZen trained on external datasets. For all experiments, we adopt ResNet293-LM pretrained by WeSpeaker on VoxCeleb as the speaker embedding model and use AHC for clustering. (FA: False Alarm, MS: Missed Detection, SC: Speaker Confusion)

System	FA%	MS%	SC%	DER%
Base+ Pretrained	11.52	1.63	0.86	14.01
Base+	3.85	3.50	1.53	8.88
Large	3.47	3.32	0.45	7.24

3. Experiments

3.1. Evaluation of EEND

In DiariZen, WavLM-Base+ is used within the EEND module and has demonstrated competitive performance on various benchmarking datasets. However, despite the superior modeling capacity of WavLM-Large, it was not evaluated in the original experiments of DiariZen. In this work, we compare the two models on the challenge dataset and find that substituting WavLM-Large for WavLM-Base+ achieves better diarization performance.

Table 2 presents the results on the development set. The first row uses pretrained checkpoints from DiariZen, which were trained on AMI [10], AISHELL-4 [11] and AliMeeting [19], but the domain mismatch between these datasets and the challenge dataset reduces overall performance. Next, when we train WavLM-Base+ directly on the challenge dataset, there is a marked improvement in DER on the development set. Replacing WavLM-Base+ with WavLM-Large yields an even further reduction in DER. A detailed breakdown of False Alarm (FA), Missed Detection (MISS) and Speaker Confusion (SC) shows that WavLM-Large outperforms WavLM-Base+ on all three metrics, highlighting the stronger modeling capabilities of WavLM-Large for the diarization task. Meanwhile, more emphasis should be placed on FA and MISS, as further analysis reveals that SC on the development set is highly dependent on whether the speakers are correctly identified in session M028, a factor closely tied to speaker embedding extraction and clustering hyperparameters. Based on these observations, we adopt WavLM-Large as the core EEND module in all subsequent experiments.

3.2. Evaluation of speaker models

To evaluate different speaker models within our pipeline, we utilize pretrained models from the WeSpeaker toolkit, which were originally trained on either the VoxCeleb or CNCeleb dataset. However, the model trained on CNCeleb performs significantly worse compared to its VoxCeleb counterpart.

Table 3 compares the performance of three speaker embedding models, demonstrating that deeper ResNet architectures trained on the VoxCeleb dataset generally achieve better DER on the development set. Based on these findings, we select ResNet293-LM trained on VoxCeleb as the speaker embedding model for subsequent experiments.

3.3. Evaluation of post-processing

Table 4 illustrates how various post-processing techniques impact DER on both the development and evaluation sets. The first

Table 3: Performance comparison of different speaker embedding models on the development set. We adopt WavLM-Large for EEND and AHC for clustering. (“-CN” and “-Vox” indicate models pretrained on CNCeleb or VoxCeleb, respectively)

System	FA%	MS%	SC%	DER%
ResNet34-LM-CN	3.47	3.32	13.84	20.63
ResNet34-LM-Vox	3.47	3.32	1.25	8.04
ResNet293-LM-Vox	3.47	3.32	0.45	7.24

row presents the results under the *max-speaker-per-frame=2* setting during EEND training, meaning the model can handle at most two overlapping speakers in a single frame. We include the first row to maintain consistency with our previous analyses of the local EEND and speaker models, which are conducted under the same configuration. This allows for a direct comparison between previous findings and the post-processing experiments presented here. Under this setting, DER is 7.24% on the development set and 9.30% on the evaluation set. However, we later observe that setting *max-speaker-per-frame=3* as row 2 depicts significantly improves DER on the evaluation set (8.99%). Meanwhile, DER on the development set rises under this setting, likely because its overlap ratio is only about 0.64%, leaving almost no scenario with three simultaneous speakers. In contrast, the evaluation set features a higher overlap ratio, making the ability to model three concurrent speakers more advantageous. Note that neither row 1 nor row 2 includes post-processing, so the results showed in row 2 effectively serve as the baseline for comparing subsequent configurations, all of which adopt *max-speaker-per-frame=3*.

Rows 2 to 6 in Table 4 demonstrate how channel selection and DOVER-Lap influence the final DER performance. In row 2, we use single-channel input with AHC clustering, achieving 7.52% DER on the development set and 8.99% DER on the evaluation set. Rows 3 and 4 apply DOVER-Lap to all channels in one case and to selected channels in the other, both using AHC-based clustering. The results show that including all channels significantly worsens the DER on both the development and evaluation sets, while restricting DOVER-Lap to selected channels improves the single-channel baseline by 0.14% on the development set and 0.31% on the evaluation set. The channel selection depends solely on whether the predicted number of speakers matches or exceeds the oracle speaker count for each recording. As a result, a channel that accurately predicts speakers during development set inference may not provide the same level of accuracy for the evaluation set. Therefore, the selection process is conducted independently for the development and evaluation sets.

Rows 5 and 6 replace AHC with NMESC for clustering speaker embeddings, but the local EEND and speaker embedding extraction configurations remain unchanged. Under single-channel conditions, NMESC performs worse than the AHC baseline due to an increase in SC. However, applying channel selection and DOVER-Lap effectively mitigates this issue, leading to more than 1% improvement in SC and a significant reduction in overall DER. Notably, on the development set, NMESC does not underestimate the number of speakers, allowing us to retain all channels for diarization.

We find that applying DOVER-Lap across all channels for each audio recording, which performs “late fusion” on multiple diarization outputs, results in worse performance than the single-channel baseline. We attribute this to the instability in

Table 4: Ablation study of post-processing techniques on the development (Dev) and evaluation (Eval) sets. We adopt WavLM-Large for EEND and use ResNet293-LM as the speaker embedding model.

System	Dev				Eval
	FA%	MS%	SC%	DER%	DER%
EEND <i>max-spk-per-frame=2</i>	3.47	3.32	0.45	7.24	9.30
AHC (ch1)	3.86	3.16	0.50	7.52	8.99
AHC (all)	3.77	3.15	1.25	8.18	-
AHC (selected)	3.79	3.13	0.46	7.38	8.68
NMESC (ch1)	3.86	3.16	1.58	8.59	-
NMESC (all/selected)	3.78	3.16	0.55	7.49	-
AHC+NMESC	3.77	3.14	0.45	7.36	8.61
Shifting	3.58	2.95	0.45	6.98	8.33

SC, which is caused by underestimating the number of speakers on certain channels in both clustering methods. This underestimation introduces large discrepancies in speaker assignments across channels, disrupting the alignment process within DOVER-Lap. To address this issue, we adopt a “Channel Selection + DOVER-Lap” approach, choosing only channels that can accurately predict or overestimate the speaker count during inference. This strategy ensures a more consistent speaker assignment among the selected channels, thereby strengthening the alignment process and ultimately reducing SC. As a result, we observe significant improvements in DER on both the development and evaluation sets.

Further improvements are obtained by applying DOVER-Lap to all channels selected under the same principle in both AHC and NMESC pipelines as depicted in row 7, yielding a 7.36% DER on the development set and 8.61% on the evaluation set. This strategy arises from our observation that NMESC demonstrates more stable performance than AHC on single-channel audio, prompting us to combine both clustering methods for a more robust solution. Finally, incorporating the shifting technique further reduces FA and MISS, resulting in a final DER of 6.98% on the development set and 8.33% on the evaluation set. Overall, these post-processing enhancements collectively provide an improvement of approximately 0.66% over the baseline without post-processing, highlighting the effectiveness of our methods in diarization tasks.

4. Conclusions

This paper presents our submission to the AVSD track of the MISP 2025 Challenge. Our approach is based on the EEND-VC pipeline, adapted from Diarize, which comprises a WavLM-Large and Conformer based local EEND module, a ResNet based speaker embedding extraction module, and AHC/NMESC based clustering. We also employ a set of post-processing techniques in three stages. First, we apply channel selection to retain only those channels that accurately predict or overestimate the number of speakers. Next, we perform DOVER-Lap on these selected channels with different clustering methods to fuse their outputs. Finally, we shift the diarization results, refining the segment boundaries and improving overall performance. Although we do not utilize an audio-visual strategy, our audio-only approach achieves 8.33% DER on the evaluation set, representing a 46.3% relative improvement over the baseline and ranking 3rd in the AVSD track of this challenge.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 12274221) and the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant No. 2024CSJGG1100).

6. References

- [1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.
- [2] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Peer, X. Xiao, B. M. Elizalde, N. Kanda, X. Wang, S. Shaer, S. Yagev, Y. Asher, S. Sivasankaran, Y. Gong, M. Tang, H. Wang, and E. Krupka, "Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," in *Interspeech 2024*, 2024, pp. 5003–5007.
- [3] L. Serafini, S. Cornell, G. Morrone, E. Zovato, A. Brutti, and S. Squartini, "An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings," *Computer Speech & Language*, vol. 82, p. 101534, 2023.
- [4] H. Chen, C.-H. H. Yang, J.-C. Gu, S. M. Siniscalchi, and J. Du, "MISP-Meeting: A real-world dataset with multimodal cues for long-form meeting transcription and summarization," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025, pp. 1–14.
- [5] M. Gao, S. Wu, H. Chen, J. Du, C.-H. Lee, S. Watanabe, J. Chen, S. S. Marco, and O. Scharenborg, "The multimodal information based speech processing (misp) 2025 challenge: Audio-visual diarization and recognition," 2025. [Online]. Available: <https://arxiv.org/abs/2505.13971>
- [6] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7198–7202.
- [7] J. Han, F. Landini, J. Rohdin, A. Silnova, M. Diez, and L. Burget, "Leveraging self-supervised learning for speaker diarization," *arXiv preprint arXiv:2409.09408*, 2024.
- [8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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.
- [9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.
- [10] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.
- [11] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Interspeech 2021*, 2021, pp. 3665–3669.
- [12] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.
- [13] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 881–888.
- [14] Z. Chen, N. Kanda, J. Wu, Y. Wu, X. Wang, T. Yoshioka, J. Li, S. Sivasankaran, and S. E. Eskimez, "Speech separation with large-scale self-supervised learning," in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [15] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Interspeech 2023*, 2023, pp. 3222–3226.
- [16] S. Wang, Z. Chen, B. Han, H. Wang, C. Liang, B. Zhang, X. Xiang, W. Ding, J. Rohdin, A. Silnova *et al.*, "Advancing speaker embedding learning: Wespeaker toolkit for research and production," *Speech Communication*, vol. 162, p. 103104, 2024.
- [17] J. Huh, J. S. Chung, A. Nagrani, A. Brown, J.-w. Jung, D. Garcia-Romero, and A. Zisserman, "The vox celeb speaker recognition challenge: A retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [18] S. Baghel, S. Ramoji, S. Jain, P. R. Chowdhuri, P. Singh, D. Vijayaseenan, and S. Ganapathy, "Summary of the displace challenge 2023-diarization of speaker and language in conversational environments," *Speech Communication*, vol. 161, p. 103080, 2024.
- [19] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.