Effective Speaker Diarization Leveraging Multi-task Logarithmic Loss Objectives

Jhih-Rong Guo[1], Tien-Hong Lo[1], Yu-Sheng Tsao[2], Pei-Ying Lee[1], Yung-Chang Hsu[2], and Berlin Chen[1]
[1]National Taiwan Normal University, Taiwan
[2]EZ-AI, Taiwan
{jhihrong, teinhonglo, 60947089s, berlin}@ntnu.edu.tw,
{sam, mic}@ez-ai.com.tw

Abstract

End-to-End Neural Diarization (EEND) has undergone substantial development, particularly with powerset classification methods that enhance performance but can exacerbate speaker confusion. To address this, we propose a novel training strategy that complements the standard cross entropy loss with an auxiliary ordinal log loss, guided by a distance matrix of speaker combinations. Our experiments reveal that while this approach yields significant relative improvements of 15.8% in false alarm rate and 10.0% in confusion error rate, it also uncovers a critical trade-off with an increased missed error rate. The primary contribution of this work is the identification and analysis of this trade-off, which stems from the model adopting a more conservative prediction strategy. This insight is crucial for designing more balanced and effective loss functions in speaker diarization.

Keywords: speaker diarization, powerset classification, loss function, ordinal log loss, Pyannote

1 Introduction

Speaker diarization is the task of determining "who spoke when" in a recording with multi-speaker. Clustering-based (Wang et al., 2018; Landini et al., 2022; Garcia-Romero et al., 2017) are typically structured as a pipeline of modules, including Voice Activity Detection (VAD), speaker embedding extraction, and a clustering algorithm. While clustering-based approaches can evolve with advancements in speaker embedding and clustering algorithms, its inherent limitation of assigning only a single speaker to each frame still prevents it from performing well on overlapped speech.

Although some studies (Bullock et al., 2020; Charlet et al., 2013) have attempted to mitigate the inherent limitation of clustering-based by using methods such as Overlapped Speech Detection (OSD). However, the additional modules may exacerbate the problem of error propagation within the pipeline. To address the problem of overlapping speech, End-to-End Neural Diarization (EEND) (Fujita et al., 2019a,b; Horiguchi et al., 2020) was proposed. This approach trains a single neural network to directly output the diarization result, thereby removing the potential for error propagation. Furthermore, EEND formulates diarization as a multi-label classification task, which enables it to process overlapped speech. Nevertheless, its direct application to longer audio recordings is impractical due to memory requirements and degraded performance when handling more than four speakers.

The EEND-VC framework, introduced by Kinoshita et al. (2021), ingeniously integrates clustering-based with EEND, bypassing the challenges of standard EEND by applying the EEND model to shorter chunks. Nevertheless, a significant hurdle for most EEND-related methods is the immense amount of training data they require, typically requiring thousands of hours, which necessitates a dependency on simulated data. Consequently, the mismatch between these simulated datasets and the target domain typically requires further model adaptation. To enable training directly on real data, the Pyannote framework (Bredin, 2023) applies EEND to even shorter chunks, enabling the assumption that only a few speakers are present within each chunk. This approach significantly reduces the data dependency, making it feasible to train the EEND model directly on real data.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 140–145
November 20-22, 2025

Recent advancements building upon the Pyannote framework have delivered superior performance in speaker diarization. These improvements are largely attributed to key strategies such as switching speaker diarization from multi-label to powerset multi-class classification problem (Plaquet and Bredin, 2023) and leveraging pre-trained Self-Supervised Learning (SSL) models with more robust encoder like the Conformer (Han et al., 2025; Plaquet et al., 2025). However, while the powerset formulation offers significant advantages over multi-label methods, it can also exacerbate issues related to speaker confusion. Consequently, enhancing the ability of model to classify speakers accurately within powerset remains a valuable area for future research.

In this paper, we use the cross entropy loss as the main objective function and introduce an ordinal log loss (Castagnos et al., 2022) that considers distances between different classes as an auxiliary objective. Because we believe that using cross entropy loss alone makes it difficult for the model to learn the relationship between different classes during training (e.g., {1} and {1, 2} both contain speaker 1). Although speaker diarization is typically evaluated using nominal metrics, we contend that strategic incorporation of a distance-aware objective function can be beneficial. We call this hybrid objective function as the Multi-task Logarithmic Loss (Multi-task Log Loss). This combination has been proven effective in ordinal classification (Kasa et al., 2024).

2 Methodology

2.1 Multi-task Log Loss

Since speaker diarization is a task primarily evaluated using nominal metrics, we employ the cross entropy loss ($\mathcal{L}_{CE}$) as main objective function:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} p_i log(\hat{p}_i), \tag{1}$$

where N represents the number of classes, and p_i is 1 if class $i \in \{1, 2, ..., N\}$ is the ground-truth class and 0 otherwise. Assuming that class j is the ground-truth label, cross entropy loss can be simplified to $-log(\hat{p}_j)$, where $\hat{p}_j$ denotes the probability for class j as predicted by the model.

To guide the model to learn the relationships between different classes, we incorporate an ordinal log loss ($\mathcal{L}_{OLL}$) as an auxiliary objective function. This approach utilizes a distance matrix to define the distance between classes, where each class represents a unique combination of speakers. The loss is formulated as follows:

$$\mathcal{L}_{OLL} = -\sum_{i=1}^{N} log(1 - \hat{p}_i) d(j, i)^{\alpha}, \tag{2}$$

where $d(j, i)$ is the distance between class j and class i, which is defined by the distance matrix D and scaled by a hyperparameter α.

The multi-task log loss ($\mathcal{L}_{MLL}$) is composed of cross entropy loss and ordinal log loss:

$$\mathcal{L}_{MLL} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{OLL}, \tag{3}$$

where λ is a hyperparameter that respectively determine the weight of the contribution of ordinal log loss to the overall loss.

2.2 Distance Matrix

In ordinal tasks, a distance matrix can be readily constructed from explicitly defined relationships between classes. However, such ordinal relationships are absent in speaker diarization. Therefore, we propose to construct a distance matrix based on the set-theoretic relationships between the different speaker combinations. The distance D_{ij} between any two speaker sets, s_i and s_j, is defined by their symmetric difference, which counts the number of speakers present in one set but not the other. This can be formulated as the size of their union minus the size of their intersection:

$$D_{ij} = |s_i \cup s_j| - |s_i \cap s_j|, \tag{4}$$

where $S = \{s_1, s_2, ..., s_N\}$ represents the set of powerset classes. Assuming each segment contains $C = 3$ speakers and a maximum of $K = 2$ overlapping speakers, the number of powerset class is $N = 7$:

- $\emptyset$ for non-speech frames;
- $\{1\}$, $\{2\}$ and $\{3\}$ for one speaker;
- $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$ for two speaker.

For example, the distance between the class representing speakers 1 and 2, $s_i = \{1, 2\}$, and the class representing only speaker 1, $s_j = \{1\}$, would be $D_{ij} = |\{1, 2\} \cup \{1\}| - |\{1, 2\} \cap \{1\}| = |\{1, 2\}| - |\{1\}| = 2 - 1 = 1$. This intuitively means there is one speaker difference between the two classes. Therefore, when $C = 3$ and $K = 2$, the distance matrix D is:

$$D = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 0 & 2 & 2 & 1 & 1 & 3 \\ 1 & 2 & 0 & 2 & 1 & 3 & 1 \\ 1 & 2 & 2 & 0 & 3 & 1 & 1 \\ 2 & 1 & 1 & 3 & 0 & 2 & 2 \\ 2 & 1 & 3 & 1 & 2 & 0 & 2 \\ 2 & 3 & 1 & 1 & 2 & 2 & 0 \end{pmatrix} \tag{5}$$

2.3 Speaker Diarization Pipeline

We adopt the same three-stage pipeline as Pyannote, which proceeds sequentially through three main components:

1. Segmentation: The input audio is first split into overlapping short segments, and End-to-End Neural Diarization (EEND) is applied to each segment to produce local diarization results.

2. Embedding: Based on the local diarization information, speaker embeddings are extracted from speech segments corresponding to each speaker.

3. Clustering: The extracted speaker embeddings are grouped using a clustering algorithm to map speakers across all segments and generate the final global speaker diarization result.

For the segmentation stage within our pipeline, we retrain the model by adopting the EEND framework proposed by Han et al. (2025). As depicted in Figure 1, the architecture first extracts features from an audio input using a pre-trained WavLM model. The feature outputs from each layer are subsequently combined through a weighted sum with learnable parameters to create a fused representation. This representation then undergo a projection layer and layer normalization before being fed into the Conformer. Finally, another linear layer as the classifier, producing logits for the N output classes. During training, all parameters of the pre-trained WavLM backbone are kept frozen.

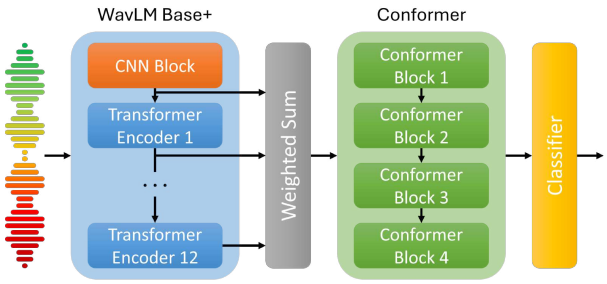

Figure 1: The architecture of EEND model.

3 Experiments

To ensure that experimental results are reproducible, we will conduct model training and evaluation using the DiariZen toolkit[1], which is driven by AudioZen and Pyannote 3.1.

3.1 Baseline

In this paper, we conduct a comparative analysis against a model trained exclusively with cross entropy loss. To ensure a fair evaluation, we maintained a consistent model architecture and configuration for all experiments, with the exception being the additional hyperparameters introduced by the multi-task log loss.

3.2 Datasets

We use AMI, AliMeeting, and AISHELL-4 as datasets for model training and evaluation, with total durations of 98.38, 126.34, and 120.25 hours respectively.The detailed duration for each dataset is presented in Table 1.

Table 1: A summary of the datasets (hrs.)

Dataset	Train	Dev	Test
AMI	79.65	9.67	9.06
AliMeeting	111.36	4.21	10.78
AISHELL-4	97.39	10.14	12.73
Compound	288.40	24.01	32.56

3.3 Evaluation Metrics

For evaluation, we employ Diarization Error Rate (DER), which is the sum of three error types: Missed Error Rate (MER), the percentage of speech time that is incorrectly labeled as non-speech; False Alarm Rate (FAR), the percentage of non-speech time incorrectly labeled as speech; and Confusion Error Rate (CER), the percentage of speech time assigned to the wrong speaker.

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

Table 2: A comparison of speaker diarization performance on the AMI, AliMeeting, and AISHELL-4 datasets for EEND model trained with cross entropy loss ($\mathcal{L}_{CE}$) versus multi-task log loss ($\mathcal{L}_{MLL}$).

	AMI			**AliMeeting**			**AISHELL-4**		
	MER	**FAR**	**CER**	**MER**	**FAR**	**CER**	**MER**	**FAR**	**CER**
$\mathcal{L}_{CE}$ **(baseline)**	**9.08**	3.94	4.46	**8.58**	3.07	7.13	**2.96**	4.29	3.41
- 250ms collar	6.87	1.95	2.58	4.54	0.87	5.63	1.21	1.71	2.36
$\mathcal{L}_{MLL}$ **(proposed)**	10.51	**3.12**	**4.03**	9.25	**2.55**	**6.30**	3.62	**3.79**	**3.16**
- 250ms collar	8.07	1.46	2.30	5.09	0.70	4.95	1.52	1.48	2.19

Table 3: An ablation study on the performance of the multi-task log loss with varying weights (λ). This comparison highlights three scenarios: (1) $\lambda = 0.5$, which yields the best performance, alongside (2) $\lambda = 0.3$ and (3) $\lambda = 0.7$, which represent cases with a lesser and greater influence from the ordinal log loss, respectively.

	Compound			
	DER	**MER**	**FAR**	**CER**
baseline	15.47	**6.65**	3.74	5.08
(1)	**15.23**	7.51	**3.15**	**4.57**
(2)	15.59	7.57	**3.15**	4.88
(3)	15.54	7.74	3.22	4.58

Table 4: A comparison of the performance with different distance between the non-speech and speaker-active classes within the distance matrix. The conditions are as follows: (1) represents the original configuration, (4) sets the distance to 2 (i.e., $d(\emptyset, i) = 2$ and $d(j, \emptyset) = 2$), and (5) sets the distance to 4.

	Compound			
	DER	**MER**	**FAR**	**CER**
baseline	15.47	**6.65**	3.74	5.08
(1)	**15.23**	7.51	**3.15**	**4.57**
(4)	15.69	6.95	3.72	5.02
(5)	15.80	6.80	3.82	5.19

3.4 Experimental Setups

The EEND model was trained on the compound training set and validated on the compound development set, using a pre-trained WavLM Base+ model[2] as a frozen feature extractor. We set the maximum number of speakers to $C = 4$ and the maximum overlapping speakers to $K = 2$. The input audio was divided into 8-second segments with a 6-second hop size. The model was trained for a maximum of 100 epochs using the AdamW optimizer with a learning rate of 1×10^{-3} and a batch size of 64. Early stopping with a patience of 10 epochs was applied based on the validation loss. The hyperparameter α for ordinal log loss was set to 1.5. In the subsequent diarization pipeline, speaker embeddings were extracted using the ResNet34-LM[3], followed by Agglomerative Hierarchical Clustering (AHC) to produce the final output.

[2]https://huggingface.co/microsoft/wavlm-base-plus

[3]https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

3.5 Results

The diarization performance of the EEND models, trained with either the conventional cross entropy loss or our proposed multi-task log loss, is detailed in this section. Table 2 presents a comprehensive comparison of the two models across the AMI, AliMeeting, and AISHELL-4, evaluated under two conditions: with no forgiveness collar (rows 1 and rows 3) and with a 250ms forgiveness collar (rows 2 and rows 4). On the compound dataset, our proposed multi-task log loss achieves relative improvements of 15.8% in FAR and 10.0% in CER compared to the baseline. These gains, however, are accompanied by a notable regression in MER, which leads to only a marginal improvement in the overall DER from 15.47% to 15.23%. This outcome suggests a fundamental trade-off: our ordinal-aware loss function effectively guides the model to be more precise in identifying speakers and avoiding false speech detection, but it does so by adopting a more conservative behavior. We will further explain this in subsequent experiments.

To determine the optimal contribution of our auxiliary objective, we conducted an ablation study on its weight, λ, with results shown in Table 3. The findings indicate that a weight of $\lambda = 0.5$ yields the best overall DER. While different weights modulate the balance between MER, FAR, and CER, the study reinforces the previously observed trade-off, where a lower FAR and CER consistently correlate with a higher MER compared to the baseline.

We hypothesize that in segments of high uncertainty, the model prefers to predict non-speech to minimize the penalties associated with incorrect speaker assignments, thus increasing the MER. To further investigate the cause of the elevated MER and validate our hypothesis regarding the model's conservative behavior, we performed a targeted analysis by modifying the distance between the non-speech class and all speaker-active classes in the distance matrix. The results, presented in Table 4, reveal a clear and direct relationship. As the distance from the non-speech class is increased (conditions (4) and (5)), the MER shows a corresponding improvement. However, this improvement comes at the cost of a gradual regression in both FAR and CER. This experiment confirms our hypothesis: a smaller distance incentivizes the model to predict non-speech in uncertain segments as a low-penalty alternative, leading to more missed errors. Conversely, a larger distance forces the model to make more definitive—and consequently, more error-prone—classifications among speaker-active classes.

In summary, our investigation into applying an ordinal-aware loss to the EEND framework has yielded a crucial insight. While the proposed multi-task log loss effectively reduces FAR and CER, its primary contribution is the revelation of a distinct trade-off with the MER. Our experiments, particularly the analysis of the non-speech class distance, provide strong evidence that this trade-off arises directly from the incentive of model to adopt a more conservative prediction strategy under this loss structure. Therefore, the key takeaway from our results is the characterization of this complex behavior. This insight is critical for understanding the implications of incorporating ordinal constraints in powerset speaker diarization.

4 Conclusion

In this paper, we investigated the effect of introducing an ordinal log loss to the training of an EEND model. Our findings demonstrate that equipping the model with distance information between different speaker combination classes effectively enhances performance in terms of FAR and CER, yielding relative improvements of 15.8% and 10.0%, respectively. However, these gains were largely offset by a regression in the MER, which resulted in only a marginal improvement in the overall DER. We further identified that this MER degradation was directly linked to the distance assigned to the non-speech class within our proposed distance matrix. Our experiments confirmed that a smaller distance incentivizes the model to adopt a more conservative prediction strategy in uncertain segments, thereby increasing missed speech errors. Therefore, the key takeaway from our results is the identification and explanation of this complex behavior. This insight is critical for understanding the implications of incorporating ordinal constraints in powerset-based speaker diarization and offers a clear direction for future improvements.

5 Future Work

Based on our findings, we propose two potential improvements. First, the manually defined, set-theoretic distance matrix could be replaced by a data-driven approach. A future direction would be to learn the distances between speaker combination classes directly from the training data. This could yield a distance matrix that is more optimally aligned with the acoustic features of the data and potentially improve the overall balance of the proposed multi-task log loss. Second, to directly counteract the MER regression observed in our experiments, we propose integrating feature fusion techniques that have proven effective for VAD. Inspired by recent findings from Tripathi et al. (2025), who demonstrated that fusing traditional MFCC features with pre-trained model representations can significantly reduce the MER, we plan to explore a similar strategy. A promising approach would be to incorporate a feature fusion module at the input stage of our EEND model.

References

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech*.

Latané Bullock, Hervé Bredin, and Leibny Paola Garcia-Perera. 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *Proc. ICASSP*.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A simple log-based loss function for ordinal text classification. In *Proc. COLING*.

Delphine Charlet, Claude Barras, and Jean-Sylvain Liénard. 2013. Impact of overlapping speech detection on speaker diarization for broadcast news and debates. In *Proc. ICASSP*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-end neural speaker diarization with permutation-free objectives. In *Proc. Interspeech*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. In *Proc. ASRU*.

Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. 2017. Speaker diarization using deep neural network embeddings. In *Proc. ICASSP*.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *Proc. ICASSP*.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. In *Proc. Interspeech*.

Aniket Kasa, Siva Rajesh Goel, Karan Gupta, Sumegh Roychowdhury, Pattisapu Priyatam, Anish Bhanushali, and Prasanna Srinivasa Murthy. 2024. Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. In *Proc. ACL*.

Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara. 2021. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In *Proc. ICASSP*.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language*.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. Interspeech*.

Alexis Plaquet, Naohiro Tawara, Marc Delcroix, Shota Horiguchi, Atsushi Ando, Shoko Araki, and Hervé Bredin. 2025. Dissecting the segmentation model of end-to-end diarization with vector clustering.

Kumud Tripathi, Chowdam Venkata Kumar, and Pankaj Wasnik. 2025. Attention is not always the answer: Optimizing voice activity detection with simple feature fusion. In *Proc. Interspeech*.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker diarization with lstm. In *Proc. ICASSP*.