EEND-EM: End-to-End Neural Speaker Diarization with EM-Network

Beom Jun Woo*, Ji Won Yoon†, Min Hyun Han*, Chan Yeong Moon*, Nam Soo Kim*
* Department of Electrical and Computer Engineering and INMC, Seoul National University, Seoul, Korea
E-mail: {bjwoo, mhhan, cymoon}@hi.snu.ac.kr, nkim@snu.ac.kr Tel/Fax: +82 2-884-1824
† Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
E-mail: jiwonyoon@cau.ac.kr Tel/Fax: +82 2-820-5549

***Abstract*—In recent years, various studies have been conducted to further enhance end-to-end neural speaker diarization (EEND) systems. However most of the methods increase model complexity by requiring additional modules during inference. In this paper, we introduce EEND-EM, a novel end-to-end neural speaker diarization model that integrates an EM algorithm-aware self-distillation method into the EEND framework. Our approach aims to enhance diarization performance by utilizing oracle guidance features derived from the EM algorithm, improving the model's prediction accuracy without increasing model complexity. Through experiments on the LibriMix dataset, EEND-EM demonstrated significant improvements in diarization error rates (DER), particularly in minimizing missed speech (MS) and confusion (CF) metrics, when compared to the baseline EEND-EDA model. Furthermore, attention weight visualizations indicate that the transformer encoders in EEND-EM are trained more effectively.**

I. INTRODUCTION

Speaker diarization aims to assign segments of audio to distinct speakers, effectively aligning the audio data with the corresponding speaker identities to determine 'who spoke when' in a multi-speaker environment. Recent development of technology has increased the demand of speaker diarization in multiple fields, such as speech recognition [1] and speaker verification [2].

Classical cascaded methods approach speaker diarization by first detecting speaker-active frames and then clustering them using speaker embeddings. The number of clusters, which corresponds to the number of speakers, is determined during inference through mathematical techniques like eigenvalue analysis [3] or hierarchical clustering with a preset threshold. Advanced methods such as utilizing improved speaker embeddings [4] or better embedding feature extractor [5] led to increase of clustering-based diarization performance. However, these methods struggle with handling speaker overlaps, as each speech frame is usually assigned to only one speaker. To overcome the limitations, the end-to-end neural speaker diarization (EEND) method has been researched to consider speaker diarization as multi-label classification problem [6], [7] . However both [6] and [7] have limitation of processing situations where the number of speaker is flexible. To address this issue, Horiguchi et al. [8] introduced the EEND-EDA system, employing a sequence-to-sequence approach with an LSTM encoder-decoder network to derive speaker-wise attractors from frame-wise embeddings and estimate speaker existence probabilities from the attractors. This approach helps speaker diarization system to handle varying numbers of speakers.

Research efforts have been made to further enhance the performance of the EEND-EDA system. One approach utilizes speaker-specific prior information, such as voice activity patterns or speaker embeddings [9], [10]. Target Speaker Voice Activity Detection (TS-VAD) improves performance by handling overlapping speech and detecting the speech activities of a set of target speakers [11]. Another method integrates tasks such as speech separation or speech counting. [12], [13] The multitask learning approach suggests that training models together, which are related to the diarization task, can have a complementary effect. However, in terms of model complexity during inference, these methods require additional modules to achieve performance improvements.

In this paper, we adopt EM-Network [14] which proposes a novel self-distillation framework that leverages target information for supervised sequence-to-sequence learning tasks. This framework is designed to improve the prediction accuracy of sequence models by incorporating oracle guidance derived from the target sequence, creating a more accurate latent space. We integrate EM-Network to speaker diarization tasks and show that proposed model EEND-EM can achieve higher performance without additional modules.

II. BACKGROUND

A. EEND-EDA

Our proposed method utilizes the original EEND-EDA [8] where its speaker wise encoder-decoder attractor enables the system to deal with varying number of speakers. EEND-EDA architecture is composed of SA-EEND [7] and LSTM based encoder-decoder. First, the input speech is transformed into Mel-filterbanks. Then the acoustic feature is processed through 2 linear layers and 4 transformer encoder blocks to get frame-wise embeddings. Secondly the embedding goes into LSTM based encoder-decoder which calculates speaker-wise attractor existence probability and produces corresponding attractors. Finally, diarization result is calculated as dot products between the attractors and the frame-wise embeddings. EEND-EDA is trained with 2 types of losses; diarization loss and attractor existence loss.

Given diarization result p_t and groundtruth labels y_t, permutation invariant training is done as follows.

$$\mathcal{L}_{\text{pit}} = \min_{s} \sum_{t=1}^{T} \mathcal{H}(p_{s,t}, y_t) \tag{1}$$

where t and s is frame and speaker set respectively. $\mathcal{H}$ refers to binary cross entropy (BCE) which is defined as follows.

$$\mathcal{H}(y_t, p_t) = \sum_{s=1}^{S} \left\{ -y_{s,t} \log p_{s,t} - (1 - y_{s,t}) \log (1 - p_{s,t}) \right\}. \tag{2}$$

Attractor existence loss is calculated with the binary cross entropy (BCE) loss between attractor existence loss probability q and binary label vector l, where S stands for the number of speakers.

$$\mathcal{L}_{\text{attractor}} = \frac{1}{S+1} \mathcal{H}(q, l) \tag{3}$$

B. EM-Network

EM-Network [14] proposes self-distillation method based on mathematical theory to maximize model performance. For current estimates of parameter $\theta^{(t)}$, Q-function calculates the expected value of log-likelihood with respect to the current estimates of the latent variables' z distribution.

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{z|x,\theta^{(t)}}[\log P(x, z|\theta)], \tag{4}$$

where x is input data, and θ is the parameters of the model. After the above E-step, M-step updates θ to maximize the Q-function.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}). \tag{5}$$

Through iteration of above EM steps, parameters that maximize the likelihood of the input is determined.

Equation (6) from EM-Network indicates that the Q-function's minimization is equivalent to minimizing the KL-divergence between the distributions of the EM-Network and the sequence model, thereby making it analogous to the maximization step of the traditional EM algorithm.

$$\begin{aligned} Q(\theta|\phi^{(t)}) &= -D_{KL}\left(P(z|x, y; \phi^{(t)}) \parallel P(z|x; \theta)\right) \\ &\approx -\mathcal{L}_{\text{kd}}(\phi^{(t)}, \theta). \end{aligned} \tag{6}$$

As a result, EM algorithm can be directly applied to sequence to sequence model without any iteration steps, where EM algorithm originally requires. EM-Network proved that addition of distillation loss between teacher model's logit and student model's logit leads to direct optimization of original sequence model, which are speech recognition task and machine translation task. As mentioned in original paper, we also adopt L2 loss instead of KL-divergence loss for training stability.

III. PROPOSED METHOD

Our proposed method starts from the original EEND-EDA [8] framework. The proposed method integrates EM-Network [14] to the speaker diarization task. Firstly, we introduce oracle teacher model that generates oracle guidance feature that combines to original speaker diarization model. Secondly, we explain self distillation-like training scheme so that the original model can learn from oracle guidance.

A. EEND-EM

The most challenging part of designing the EM-Network is to avoid trivial solutions. Simply putting the target into the fusion module makes trivial solution, which means that fusion model in EM-Network copies target itself rather than to use the information from the output embedding of the last transformer encoder. To prevent the trivial solution problem, we analyze label aggregator which corresponds to the post-processing of speaker diarization model.

Diarization result p_t and ground-truth labels y_t are aggregated before calculating loss. Label aggregation divides labeled sequences into overlapping frames, aggregates the labels within each frame based on a majority rule. For this reason, we assumed that giving target speaker label into oracle encoder can be applied same as the original EM-Network.

From this point forward, we refer to the EM-Network as the oracle teacher model and the EEND-EDA part as the student model in the EM aware self-distillation framework. EM-Network is consisted of 3 parts; oracle encoder, sequence model and oracle decoder. Oracle encoder generates oracle guidance feature using speech labels. Unlike original EM-Network oracle encoder structure, EEND-EM oralcle encoder consists of a embedding layer, max pooling and self-attention based transformer encoder layer. Oracle guidance feature is used as key K and value V for input of oracle decoder which is explained below.

$$K, V = OracleEncoder(y_t) \tag{7}$$

For sequence model, we adopt original EEND-EDA framework to our work. After the frame-wise embeddings are generated through stacked transformer blocks, oracle decoder is inserted. Oracle decoder merges the outputs of the sequence model and those of the oracle encoder. The decoder is a transformer decoder that utilizes cross-attention to effectively guide the source sequence information. The cross-attention layer receives the output embeddings of the final transformer encoder as queries and the oracle guidance features as key-value.

$$Q = SA\text{-}EEND(x_t) \tag{8}$$

$$e_{s,t}^{\phi} = OracleDecoder(Q, K, V) \tag{9}$$

Output $e_{s,t}^{\phi}$ is used to calculate the attractor existence probability and diarization result of EM-Network.

After forward propagation of EM-Network, EEND-EDA is forwarded using same input x_t. EEND-EDA shares its parameter with EM-Network. For each epoch, both model

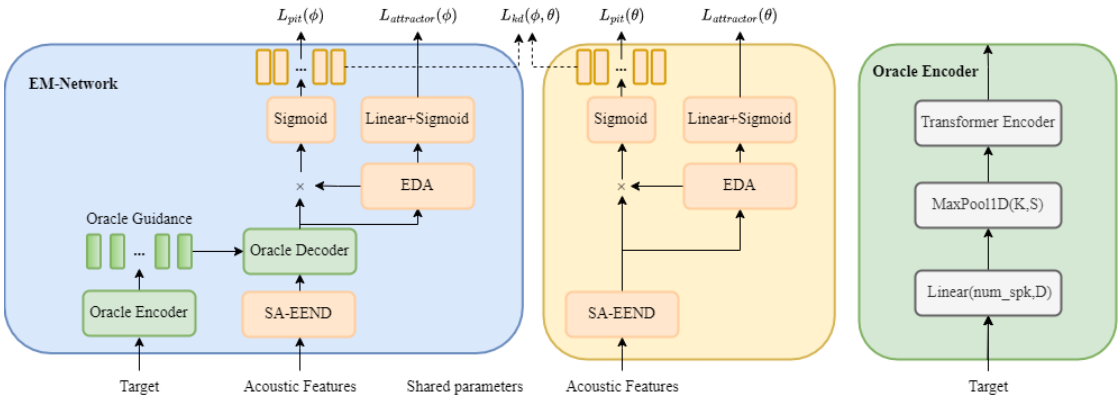


Fig. 1: Model architecture of EEND-EM. Left blue box is architecture of EM-Network. Middle yellow box is overview of EEND-EDA. Green box shows how oracle encoder is composed of. D stands for embedding number. K and S in MaxPool1D block stands for kernel and stride size respectively. All EEND-EDA modules (orange boxes) in EEND-EM share parameters.

is forwarded to predict each diarization result and attractor existence probability.

$$p^{\phi}_{s,t}, q^{\phi}_{s,t} = EM\text{-}Network(x_t) \tag{10}$$
$$p^{\theta}_{s,t}, q^{\theta}_{s,t} = EEND\text{-}EDA(x_t) \tag{11}$$

where parameter has relationship of $\theta \in \phi$. During inference and validation, we exclusively utilize EEND-EDA model. EM-Network in EEND-EM is only used while training.

B. EM algorithm aware training

Let EM-Network's parameter and original speaker diarization model's parameter be ϕ and θ respectively. There are 3 types of losses for training the proposed model. We adopt original loss equation (1) and (3) for both model. PIT loss for each model is $\mathcal{L}_{\text{pit}}(\phi)$ and $\mathcal{L}_{\text{pit}}(\theta)$. Total PIT loss is as followed.

$$\mathcal{L}_{\text{pit}} = \mathcal{L}_{\text{pit}}(\phi) + \mathcal{L}_{\text{pit}}(\theta) \tag{12}$$

Also attractor existence loss is expressed as $\mathcal{L}_{\text{attractor}}(\phi)$ and $\mathcal{L}_{\text{attractor}}(\theta)$ respectively. Total attractor existence loss is as followed.

$$\mathcal{L}_{\text{attractor}} = \mathcal{L}_{\text{attractor}}(\phi) + \mathcal{L}_{\text{attractor}}(\theta) \tag{13}$$

Distillation loss is calculated using l_2 loss between the diarization result of EM-Network and original sequence model.

$$\mathcal{L}_{\text{kd}}(\phi,\theta) = \sum_t (p^{\phi}_t - p^{\theta}_t)^2 \tag{14}$$

Total loss will be as follows. λ is scheduled distillation loss variable to warmup the training of each model parameter ϕ and θ.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pit}} + \mathcal{L}_{\text{attractor}} + \lambda * \mathcal{L}_{\text{kd}}(\phi,\theta) \tag{15}$$

IV. Experiments and Result

A. Training and test datasets

We tested the performance of EEND-EM with the LibriMix dataset [15], which contains training and test mixtures from LibriSpeech [16] train-clean100 and test-clean samples mixed with WHAM! [17] at a 16kHz sampling rate. The dataset includes two-speaker (Libri2Mix) and three-speaker (Libri3Mix) mixtures, consisting of 58 hours/11 hours and 40 hours/11 hours of training/test sets, respectively. We used the min mode during experiment to benchmark our results against previous studies.

B. Configurations

We used baseline model EEND-EDA from ESPnet [18] [1]. Acoustic feature is generated through 80-dimensional log-mel filterbanks with window size of 25ms and frame shift of 10ms. SA-EEND from Fig. 1 is consisted of 4 transformers with embedding size of 256. EDA module is 1 recurrent neural network with hidden unit size of 256. Oracle Encoder has 3 modules. Linear layer maps number of speakers to embedding size 256. Then transformer encoder is followed by maxpooling layer with kernel size of K and stride size of S. Oracle decoder is a transformer decoder with embedding size of 256. We optimize via Adam with batch size of 64. Warmup learning rate scheduler is applied where warmup step is 30000 and learning rate is 2e-3. The scheduled distillation loss variable λ is set to increase logarithmically from 1e-8 to 1 across epochs. Total epoch is set as 250. Training is done with one NVIDIA GeForce RTX 3090 GPU.

C. Evaluation metric

We assessed speaker diarization performance by measuring the diarization error rate (DER (%)), which includes speaker confusion (SC (%)), false alarms (FA (%)), and missed detections (MS (%)). The collar tolerance is set to 0 seconds.

[1] https://github.com/espnet/espnet/tree/master/egs2/librimix/diar1

TABLE I: Experimental results on a fixed 2-speaker scenario (Libri2Mix) for min mode in terms of DER, FA, and MI (%). † denotes our re-implementation.

Method	DER	FA	MS	CF
SA-EEND†[7]	6.13	3.54	2.11	0.48
EEND-EDA†[8]	5.93	3.48	2.07	0.38
EEND-EM$_{512,128}$	5.61	3.47	1.83	0.31
EEND-EM$_{128,64}$	5.23	3.67	1.34	**0.22**
EEND-EM$_{64,32}$	5.15	3.57	1.35	0.23
EEND-EM$_{32,16}$	**4.98**	**3.45**	**1.29**	0.24

TABLE II: Experimental results on a fixed 3-speaker scenario (Libri3Mix) for min mode.

Method	DER	FA	MS	CF
SA-EEND† [7]	9.05	5.92	2.72	0.41
EEND-EDA† [8]	8.81	5.81	2.62	0.38
EEND-EM$_{128,64}$	7.14	5.21	1.65	0.28
EEND-EM$_{64,32}$	6.93	5.15	1.51	0.27
EEND-EM$_{32,16}$	**6.82**	**5.14**	**1.44**	**0.24**

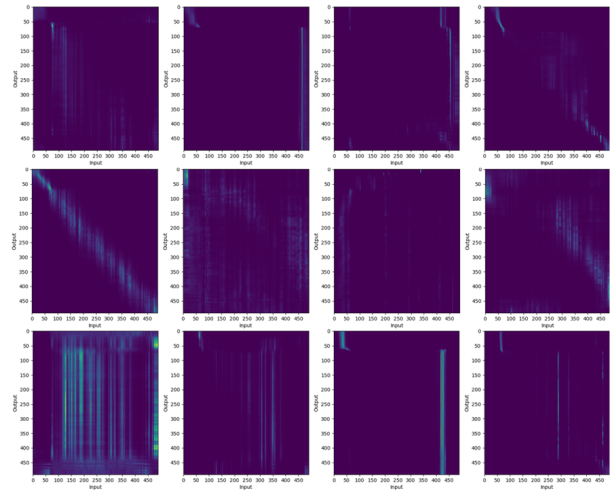
D. Results on the LibriMix dataset

We performed experiments on scenarios with a set number of speakers. The proposed models were assessed under conditions involving 2 speakers and 3 speakers, utilizing the test sets min mode from Libri2Mix and Libri3Mix, respectively. We trained each SA-EEND [7], EEND-EDA [8] and EEND-EM based on ESPNet [18]. From Table I and II, we can see that application of EM-Network for speaker diarization proves to be effective. From the result, we analyze that the more the kernel and stride sizes differ from the original label aggregating settings, the better the model's performance. We can see that relative performance improvement is 19.07% and 29.17% for 2-speaker and 3-speaker scenarios, respectively.

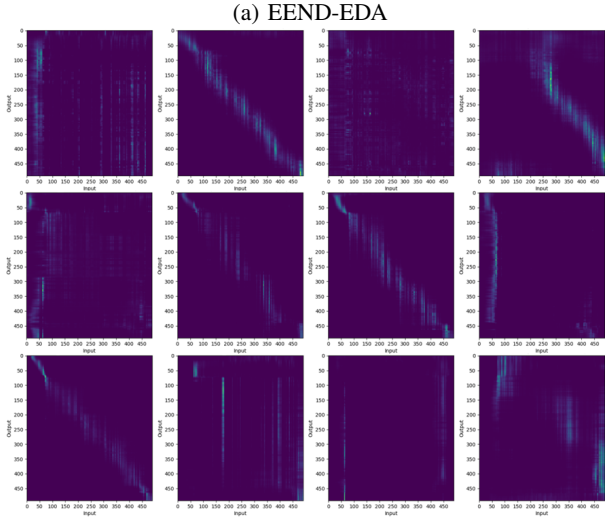
Table I includes experiment of EEND-EM where kernel and stride size is equivalent to label aggregator setting, which is 512 and 128 each. The results indicate that the proposed model yields similar outcomes to the original model, suggesting that the EM-Network produces trivial solutions. As the authors assumed, setting the kernel and stride similar to the label aggregator is ineffective for generating oracle guidance features.

E. Visualization of self-attention heads

To further investigate the effect of EM algorithm-aware training, we plot attention weights for both the baseline and proposed models. Figure 2 shows the first three transformer encoder weights drawn from a Libri2Mix test dataset. The attention weights in the EEND-EDA exhibit varied distribution patterns. Some plots display sparse attention concentrated in specific regions, while others have scattered or diagonal patterns indicating sequential attention. In contrast, the EEND-EM$_{32,16}$ shows more coherent attention patterns, with the diagonal self-attention weights being more pronounced and consistent across different plots. From the results, we observe that improved alignment in attention weights correlates with better speaker-related outcomes.



(a) EEND-EDA



(b) EEND-EM$_{32,16}$

Fig. 2: Visualization of attention weight matrices in the first three encoder block. Each line corresponds to each encoder block with four heads.

V. Conclusions

In this paper, we propose EEND-EM, which integrates an EM algorithm-aware self-distillation method into the EEND framework. The results demonstrate the effectiveness of this approach, particularly in improving the MS and CF metrics. Additionally, the attention plots indicate that the transformer encoder in the proposed model is better trained compared to the baseline EEND-EDA. Finally, we can explore better structures for generating oracle encoder and decoder features to provide improved oracle guidance in the future.

VI. Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

References

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558. DOI: 10.1109/ICASSP.2018.8462661.

[2] C. Xu, W. Rao, J. Wu, and H. Li, "Target speaker verification with selective auditory attention for single and multi-talker speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2696–2709, 2021. DOI: 10.1109/TASLP.2021.3100682.

[3] 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, 2020. DOI: 10.1109/LSP.2019.2961071.

[4] J.-W. Jung, H.-S. Heo, B.-J. Lee, *et al.*, "In search of strong embedding extractors for speaker diarisation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096449.

[5] H.-S. Heo, Y. Kwon, B.-J. Lee, Y. J. Kim, and J.-W. Jung, "High-resolution embedding extractor for speaker diarisation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097190.

[6] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202572807.

[7] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 296–303, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202572979.

[8] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "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, ISSN: 2329-9304. DOI: 10.1109/taslp.2022.3162080. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2022.3162080.

[9] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, *Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization*, 2022. arXiv: 2208.13085 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2208.13085.

[10] Z. Du, S. Zhang, S. Zheng, and Z. Yan, *Speaker overlap-aware neural diarization for multi-party meeting analysis*, 2022. arXiv: 2211.10243 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2211.10243.

[11] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Interspeech 2020*, ser. interspeech$_2$020, ISCA, Oct. 2020. DOI: 10.21437/interspeech.2020-1602. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1602.

[12] S. Maiti, Y. Ueda, S. Watanabe, *et al.*, *Eend-ss: Joint end-to-end neural speaker diarization and speech separation for flexible number of speakers*, 2022. arXiv: 2203.17068 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2203.17068.

[13] J. Ao, M. S. Yıldırım, R. Tao, *et al.*, *Used: Universal speaker extraction and diarization*, 2024. arXiv: 2309.10674 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2309.10674.

[14] J. Yoon, S. Ahn, H. S. Lee, M. Kim, S. Kim, and N. S. Kim, "Em-network: Oracle guided self-distillation for sequence learning," *ArXiv*, vol. abs/2306.10058, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259203399.

[15] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, *Librimix: An open-source dataset for generalizable speech separation*, 2020. arXiv: 2005.11262 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2005.11262.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[17] G. Wichern, J. Antognini, M. Flynn, *et al.*, *Wham!: Extending speech separation to noisy environments*, 2019. arXiv: 1907.01160 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1907.01160.

[18] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456.