Improving Label Assignments Learning by Dynamic Sample Dropout Combined with Layer-wise Optimization in Speech Separation

Chenyang Gao[1], *Yue Gu*[2], *Ivan Marsic*[1]

[1]Rutgers University
[2]Amazon APT

cg694@rutgers.edu, yguam@amazon.com, marsic@rutgers.edu

Abstract

In supervised speech separation, permutation invariant training (PIT) is widely used to handle label ambiguity by selecting the best permutation to update the model. Despite its success, previous studies showed that PIT is plagued by excessive label assignment switching in adjacent epochs, impeding the model to learn better label assignments. To address this issue, we propose a novel training strategy, dynamic sample dropout (DSD), which considers previous best label assignments and evaluation metrics to exclude the samples that may negatively impact the learned label assignments during training. Additionally, we include layer-wise optimization (LO) to improve the performance by solving layer-decoupling. Our experiments showed that combining DSD and LO outperforms the baseline and solves excessive label assignment switching and layer-decoupling issues. The proposed DSD and LO approach is easy to implement, requires no extra training sets or steps, and shows generality to various speech separation tasks.

Index Terms: Speech separation, permutation invariant training, dynamic sample dropout, layer-wise optimization

1. Introduction

Speech separation is a specific type of source separation that focuses on separating human speech sources from overlapping speech signals. Deep learning has demonstrated great success in speech separation [1, 2]. During the training of speaker-independent speech separation models, a commonly recognized challenge known as label ambiguity or permutation problem arises, which pertains to the ambiguity or uncertainty in assigning labels to predictions. This problem is caused by the same nature of the overlapped sound sources where the order of the labels and predictions do not match well during training. To address this issue, permutation invariant training (PIT) was introduced [3, 4], which involves exploring all possible label-assignment pairs and selecting the best pair to update the model. PIT has now become the standard training approach for time-frequency (T-F) domain [5, 6] and time-domain [5, 6, 7, 8, 9, 10, 11, 12, 13, 14] speech separation models. In time-domain approaches, 1-D convolution and transposed convolution are used as trainable front-end to replace the roles of STFT and iSTFT in T-F domain approaches, respectively.

Although PIT has effectively tackled label ambiguity for training speech separation models, it suffers from unstable label assignment switching [15, 16, 17, 18]. This issue arises when a large proportion of label assignments abruptly switch the order in adjacent epochs, resulting in erratic training. The unstable label assignment switching problem is claimed to be caused by the slight difference in pairwise loss during the initial epochs as in [15, 16, 17, 18, 19]. Prior research attempted to address this problem from two main perspectives. Some studies [15, 16] have combined different training strategies with the original PIT approach, including fixing label assignments [15] and fine-tuning a model pre-trained with speech enhancement [16], while others [17, 18, 19] have proposed probabilistic relaxation PIT to prevent the model from being over-confident in the label assignments. However, our experiments show that these approaches are insufficient in resolving this problem, which impedes the model's ability to learn better label assignments. To solve the above issue, we introduce a practical training approach named dynamic sample dropout (DSD) by addressing the issue of excessive label assignment switching. DSD employs a mechanism that considers previous best label assignments and evaluation metrics to identify and exclude the challenging samples that may have a detrimental effect on learning label assignments. Unlike other techniques [15, 16], DSD doesn't require additional data or training steps, so it's applicable in various speech separation settings. Additionally, we combine DSD with layer-wise optimization (LO) [20, 21, 22] to further enhance the model's performance. Through an extensive study of layer-wise optimization, we found that LO reduces layer-decoupling (see Section 5.3), leading to the observed improvement in model performance. Our experiments using LibriMix data demonstrate that the proposed approach outperforms the baseline models by a significant margin in a range of 1.07 to 1.62 dB in SI-SDRi. Our contributions are:

1. We assess and illustrate the limitations of current approaches in addressing excessive label assignment switching.
2. We propose a novel dynamic sample dropout strategy that employs layer-wise optimization to effectively resolve the issues of excessive label-switching and layer-decoupling without the need for additional training samples or steps.
3. We carry out extensive experiments to demonstrate the consistent performance enhancement and generality of the proposed dynamic sample dropout with the layer-wise optimization approach across various speech separation tasks.

2. Related Work

2.1. Label ambiguity and Permutation invariant training

The problem of mono-channel speech separation is formulated as follows. Given a mixture speech signal X containing N speakers: $X = \sum_1^N s_i + n, s_i \in R^t$, where s_i is the clean source of speaker i and n is background noise. The goal is to recover the speech for each speaker from the mixture waveform. Considering a two-speakers case, the label ambiguity problem occurs because the model's predictions r_1 and r_2, and labels s_1 and s_2, could be matched arbitrarily. That is, r_1 could be the estimated recovery of either s_1 or s_2. Consequently, there exists $N!$ possible different combinations of prediction-label pairs; the choice in different combinations is known as the problem of label ambiguity. Permutation invariant training (PIT) [3, 4] has been proposed to solve this label ambiguity during training. In short, PIT traverses all the possible prediction-label pairs and selects only the optimal one to update the model. Although PIT achieves great success in training speech separation models, the unstable label assignment switching during training impairs the model's performance.

2.2. Learning better label assignments

Different approaches have been proposed to overcome unstable label assignment switching in the following two ways.

Improving training strategy—The first group tries to improve the training strategy of speech separation with the original PIT. In [15], the authors proposed a strategy called interrupted and cascaded training. It avoids label assignment switching by fixing the label assignments after the initial PIT training epochs. In [16], the authors proposed using speech enhancement as the pre-training task to stabilize the label assignments.

Improving PIT—The other group tried to modify the original PIT. ProbPIT [17], SinkPIT [18], and soft-minimum PIT [19] were proposed as the probabilistic relaxation version of the original PIT. They used a weighted sum of losses over all possible permutations to avoid the model being over-confident to a specific permutation, showing better learning label assignments ability than the original PIT.

We performed the ablation study to assess if the excessive label assignment switching problem is well-addressed with interrupted and cascaded training strategy [15] and SinkPIT [18]. However, we still observed a sudden excessive switching during training with these approaches, indicating that excessive label assignment switching is not well addressed and still hinders the model from learning better label assignments (see Section 5.1).

3. Dynamic Sample Dropout and Layer-wise Optimization

3.1. Dynamic sample dropout

We introduced our dynamic sample dropout (DSD) training method to overcome the problem of excessive label assignment switching (see Figure 1). During our reproduction of previous methods and baselines, we noticed that the evaluation metrics would abruptly decrease after certain training samples. Our hypothesis for this phenomenon is that challenging training samples can negatively impact the learned label assignments and result in excessive label assignment switching and inconsistent training progress. To validate and resolve this issue, we introduce the DSD training strategy, which dynamically removes filtered training samples based on an evaluation of both the metric and past label assignments. This approach uses a memory bank to keep track of the best evaluation metric and corresponding label assignments for each training sample. The challenging samples are omitted during the corresponding training iterations to maintain stable label assignments. The memory bank is initialized at the first training epoch, where it records the label assignments and evaluation metrics for every sample. In the remaining training epochs, the DSD uses the following criteria to select or dropout the samples at each optimization step:

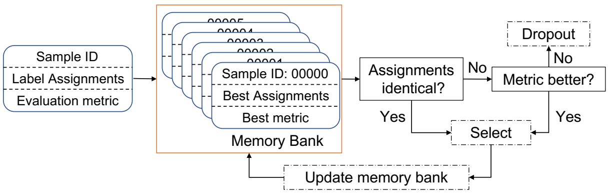

Figure 1: *The proposed dynamic sample dropout.*

1. Select: i) if the current label assignment chosen by PIT is identical to the recorded best assignments; ii) if the label assignment order changed, the evaluation metric relaxed outperforms the recorded best evaluation metric. We update the record for these training samples in the memory bank.
2. Dropout: if the label assignment order changes, the evaluation metric is not relaxed better than the recorded best evaluation metric. The loss for this training sample will drop out.

The criterion for "relaxed better" is defined as:

$$M_{cur} * (1 + sgn(M_{cur}) * \epsilon) > M_{best} \tag{1}$$

where M_{cur} represents the current evaluation metric and M_{best} represents the best evaluation metric, ϵ is a relaxation factor. The sgn function ensures fair sign comparison for negative evaluation metrics. The relaxation step enables DSD to tolerate samples that result in a slightly worse evaluation metric but switch the label assignments. DSD discards challenging samples that may disrupt the learned label assignments during the training process, thereby maintaining a stable label assignment switching ratio. Instead of discarding, an alternative approach is to persist with the previously recorded best label assignments for these challenging samples. Specifically, we use the best-recorded label assignments stored in the memory bank to recalculate the loss for these challenging samples. The *reorder* operation insists on the best-recorded label assignments for these challenging samples and allows them to still participate in the training process. We refer to this approach as DSD (*reorder*) to distinguish it from DSD (*dropout*).

3.2. Layer-wise optimization

We further combined the proposed dynamic sample dropout (DSD) with layer-wise optimization (LO) to enhance the learning of label assignments. LO was introduced for efficient inference in previous studies, such as in [20, 21, 22], where intermediate layers are trained directly with the target, allowing for early-exit strategies to save inference time. Specifically in speech separation, for a model with N repeated sequential modeling blocks, the intermediate outputs from each layer have the same shape, and they are used to reconstruct the clean target. Layer-wise optimization computes a loss term for each layer and sums them up. The layer-wise optimization for speech separation is shown as follows:

$$\mathscr{L} = \frac{1}{N} \sum_{i=1}^{N} w_i * \text{PIT}(\tilde{S}_i, S) \tag{2}$$

where $\tilde{S}_i$ is the reconstructed source from intermediate layer i, S is the target source, and w_i is a weighted scalar that controls each loss's contribution term to the final loss. In addition to the inference efficiency, the layer-wise optimized model also outperforms the model that is only optimized by a single loss term from the last layer, as shown in [22]. Instead of giving a subjective explanation for the improvement, we conducted experiments and analyses to demonstrate how the improved gradient flow affects the behavior of each intermediate layer, by comparing the dissimilarity between the label assignment switching ratio curve of intermediate layers and the last layer. We used the same approach as in [22], where it shares the weights for the mask estimation network and decoder for intermediate outputs generation. We believe this discrepancy is the actual reason for the performance improvement by using LO. Our aim is to increase the popularity of layer-wise optimization in the speech separation community with this new insight.

4. Experimental Setup and Implementation

We used DPTNet [8] as our baseline model for its competitive performance in various speech separation and enhancement tasks [8, 23, 24]. The DPTNet is a time-domain masking-based model that uses dual-path processing [7], which segments audio into short segments and then sequentially applies the intra and inter-procedure. The dual-path processing allows the model to handle the long-range dependencies in long audios. We used the default configuration as in [8] to construct the separation model, where the number of the improved transformer is set as six. We used a kernel size of 16 and strided with 8 in encoder/decoder as in [6, 7]. The entire model contains 2.7 M trainable parameters.

We employed the open-source LibriMix dataset [25], derived from the Librispeech dataset [26], for our experiments. We conducted experiments with various subsets, including different numbers of speakers (Libri2Mix and Libri3Mix) and varying conditions (clean or noisy). The results were reported on the minimum version of the corresponding test sets, following previous studies [25].

We built the model using the Asteroid toolkit [27]. To ensure a fair comparison, the model was trained for 200 epochs with a batch size of 24 in all different experiment settings, consistent with previous work [16]. The Adam optimization algorithm [28] was used with an initial learning rate of 1e-3, and the gradients were clipped with a maximum L_2 norm of 5. The patience for halving the learning rate was set to 10 for the first 80 epochs and 5 for the remaining epochs. The audio segments were divided into 3-second segments for both training and validation. The scale-invariant signal-to-distortion (SI-SDR) metric was used as the training objective [29] and the evaluation metric in DSD. We set the $w_i = \frac{\text{layer index}}{\text{total blocks}}$ in LO. We evaluated the improved signal quality using the SI-SDRi and SDRi metrics.

5. Experiment Results and Analysis

5.1. Baseline experiments with PIT and its variants

We conducted experiments in three settings: i) PIT; ii) SinkPIT; and iii) Interrupted & Cascaded (only PIT-(fix) step), using the LibriMix train-100 subset with the *sep_clean* task. SinkPIT demonstrates the best performance (Table 1), achieving 15.28 dB in SI-SDRi compared to the original PIT (15.13 dB). In contrast to the results in [15], PIT-(fix) shows the worst results with a difference ranging from 14.59 to 14.98 dB in SI-SDRi, depending on the value of *L*.

Table 1: *Performance on the sep_clean test set. "L" indicates the number of epochs trained with PIT.*

Method	SI SDRi/SDRi
PIT= DSD ($\epsilon = +\infty$)	15.13/15.53
SinkPIT	15.28/15.70
(PIT)-(fix) *L=1*	14.98/15.39
(PIT)-(fix) *L=5*	14.59/15.00
(PIT)-(fix) *L=10*	14.65/15.05
(PIT)-(fix) *L=25*	14.99/15.40
(PIT)-(fix) *L=50*	14.84/15.24
(PIT)-(fix) *L=75*	14.93/15.34
(PIT)-(fix) *L=100*	14.79/15.19
(PIT)-(fix) *L=150*	14.92/15.31
DSD (*reorder*, $\epsilon = 0.0$)	15.46/15.90
DSD (*dropout*,$\epsilon = 0.0$)	15.62/16.07
DSD (*dropout*,$\epsilon = 0.1$)	15.75/16.20
DSD (*dropout*,$\epsilon = 0.2$)	15.59/16.05
DSD (*dropout*,$\epsilon = 0.5$)	15.39/15.83
LO	15.84/16.27

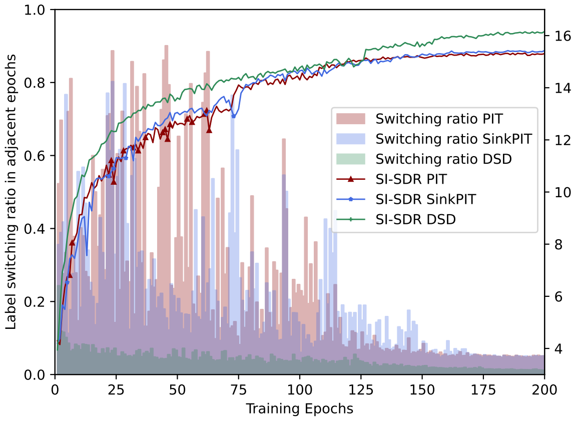

Figure 2: *Label assignments switching ratio versus validation SI-SDR for PIT, SinkPIT, and DSD.*

To better understand the training process, we analyzed the label assignment switching ratio curve as follows. We selected the label assignment using PIT for each training sample after every epoch and defined the fraction of samples with different label assignments in adjacent epochs as the label assignment switching ratio. Figure 2 shows the label assignment switching curve against validation SI-SDR for PIT and SinkPIT. We noted that a large fraction of label assignments switched in adjacent epochs, causing instability in the training path and a decline in SI-SDR, as indicated by the markers on the SI-SDR curves. This observation also sheds light on the reason behind the ineffectiveness of PIT-(fix); excessive label assignment switching in the initial PIT procedure cannot guarantee that the model learns promising label assignments, thus fixing the label assignments in later steps cannot enhance the performance (*L*=1 outperformed *L*=100). These findings suggest that excessive label assignment switching hinders the model's ability to learn better label assignments, and existing approaches like Interrupted & Cascaded and SinkPIT do not effectively solve this issue.

5.2. Dynamic sample dropout

We evaluated the proposed dynamic sample dropout (DSD) approach using the LibriMix train-100 subset with the *sep_clean* task. It is important that when $\epsilon = +\infty$, the DSD strategy becomes identical to PIT because it will accept all of the data. As shown in Table 1, we observe a significant improvement in performance compared to the baseline when we used DSD (15.75 versus 15.13 in SI-SDRi). We then performed an ablation study comparing DSD (*reorder*) and the original DSD (*dropout*) and found that DSD (*dropout*) outperformed DSD (*reorder*), indicating that *dropout* is a more efficient method in dealing with challenging samples. We conjectured that DSD (*reorder*) still involved challenge samples in training, which impeded learning better label assignments. Moreover, the speech separation performance initially improved and then decreased as the relaxation factor was increased, suggesting that relaxation is more effective in handling the criteria for dropping training samples. We also evaluated the label assignment switching curve of the proposed DSD method (as shown in Figure 2). The label assignment switching ratio of the model trained with DSD became significantly more stable compared to the original PIT and SinkPIT. The faster convergence rate and improved performance further highlighted the importance of maintaining a stable label assignment switching ratio for the model to learn better label assignments. Additionally, in the most stringent

Table 2: *Similarity between intermediate layers and the last layer. L_1 distance is reported.*

Comparison	PIT	DSD	LO	DSD+LO
1 vs 6	37.50	33.31	27.50	0.83
2 vs 6	29.98	16.73	2.26	0.20
3 vs 6	25.56	4.16	0.13	0.07
4 vs 6	14.78	0.56	0.05	0.03
5 vs 6	6.67	0.37	0.03	0.02

scenario ($\epsilon = 0.0$), the percentage of discarded samples in the training set gradually decreased from 3% to less than 1% over the course of training, pointing to the fact that challenging samples disrupt the learned label assignments during training and result in excessive label assignment switching. For subsequent experiments, we employed $\epsilon = 0.1$ and *dropout* for DSD.

5.3. Layer-wise optimization

We evaluated the performance of the layer-wise optimization (LO) approach on the LibriMix train-100 subset with the *sep_clean task*. The results show that using LO improved the performance (Table 1), achieving 15.84 SI-SDRi and 16.27 SDRi, surpassing the baseline results of PIT and SinkPIT. We analyzed the label assignment switching ratio curve for each layer of the models trained with and without LO. We determined how the improved gradient flow in LO affects the behavior of the intermediate layers by comparing the similarity in the label assignment switching curves of the intermediate layers with that of the last layer. We argue that this similarity somehow reflects the training process of the model. We also indicate this similarity by calculating the L_1 distance between label assignment switching curves. Based on Figure 3 and Table 2, we made the following observations:

1. The curves of the label assignment switching ratio for Layers 2-6 change in tandem in the LO-trained model, while Layer 1 shows a distinct trend (Figure 3, LO.)
2. Unlike LO, where Layers 2-3 showed a similar trend in the changing of label assignment switching ratio curves, only Layers 4-6 are changing in unison in PIT and DSD, even though they were trained without LO (Figure 3, PIT & DSD.)
3. The problem of excessive label assignment switching exacerbates the differences in the label assignment switching ratio curves, with the distance of PIT-trained model being much greater than DSD-trained model (Table 2, PIT & DSD.)

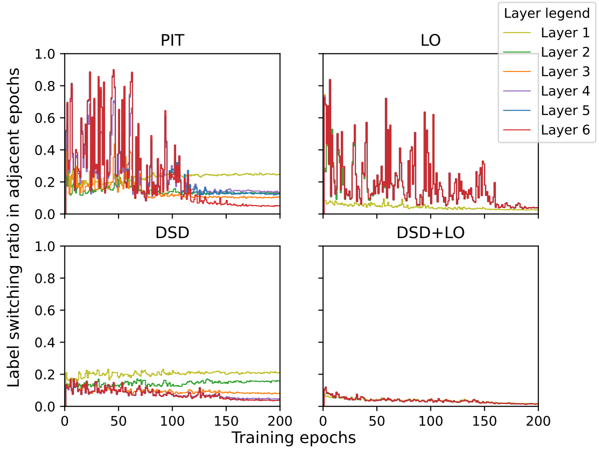

Figure 3: *Layer-wise label assignment switching curve with different training strategies.*

Table 3: *Experiment results on different tasks, results reported on corresponding min version of test sets. For DSD, we set $\epsilon = 0.1$*

Task	DSD	LO	train-100	train-360
2spk-C (PIT)	×	×	15.13/15.50	15.92/16.27
2spk-C	✓	×	15.75/16.20	16.70/17.09
2spk-C	×	✓	15.84/16.27	16.84/17.19
2spk-C (Ours)	✓	✓	16.28/16.75	17.22/17.56
2spk-N (PIT)	×	×	11.64/12.21	-/-
2spk-N	✓	×	12.33/12.92	-/-
2spk-N	×	✓	12.50/13.10	-/-
2spk-N (Ours)	✓	✓	12.79/13.39	-/-
3spk-C (PIT)	×	×	11.92/12.38	13.34/13.76
3spk-C	✓	×	12.54/13.00	14.10/14.53
3spk-C	×	✓	12.58/13.04	14.20/14.64
3spk-C (Ours)	✓	✓	12.99/13.47	14.96/15.40

The resemblance and disparity between the label assignment switching ratio curves indicate coherence and incoherence, respectively, in the training directions of each intermediate layer during training. We refer to the issue of intermediate layers having dissimilar switching ratio curves as the “layer-decoupling” problem. Our findings suggest that LO serves as a regularization technique for the training direction of the middle layer, which significantly mitigates the layer-decoupling problem. Nevertheless, the excessive problem of label assignment switching still affects the layer-decoupling problem in the LO-trained model (Layer 1 has a different trend), and further amplifies it (observation #3).

5.4. Combining DSD and LO

To take advantage of both the DSD and LO strategies, we combined them to address the excessive label assignment switching and layer-decoupling problems simultaneously. We first applied the same experiment and the analysis as in Section 5.3. Figure 3 and Table 2 shows that the DSD+LO eliminates both excessive label assignment switching and layer-decoupling problem. And it also leads to a further improvement in the separation performance (in Tale 3). To show the applicability of DSD+LO in general speech separation, we evaluated the performance of the combined DSD+LO approach on various speech separation tasks, including a larger dataset (train-360 subset), noisy conditions (*sep_noisy*), and more speaker scenarios (Lirbi3Mix). The results (Table 3) show that the proposed DSD+LO-trained models outperform the baselines with a margin between 1.07 to 1.62 dB in the SI-SDRi. The DSD+LO-trained models also outperform the DSD- and LO-trained models, showing the complementarity of these two training strategies.

6. Conclusion

We studied the issue of excessive label assignment switching in speech separation and discovered that existing methods were unable to effectively address it. We proposed dynamic sample dropout (DSD) to maintain a stable label assignment switching ratio by removing samples that could negatively impact the learned label assignments. We further introduced layer-wise optimization (LO) to improve separation performance by reducing layer decoupling. By combining DSD and LO, our proposed model outperformed all baselines, effectively addressing both excessive label assignment switching and layer decoupling. The DSD + LO training strategy is easy to implement, requires no extra training sets or steps, and demonstrates strong generality to various single-channel speech separation tasks. We believe that it could be easily employed in multi-channel scenarios because PIT is also widely used in multi-channel settings.

This work is supported by NIH/NLM grant number R01LM011834 and NSF grant IIS-1763827.

7. References

[1] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[5] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[6] ——, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[8] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech 2020*, 2020, pp. 2642–2646.

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[10] Z. Zhang, B. He, and Z. Zhang, "Transmask: A compact and fast speech separation model based on transformer," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5764–5768.

[11] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[12] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020, pp. 1–6.

[13] M. W. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5759–5763.

[14] C. Li, L. Yang, W. Wang, and Y. Qian, "Skim: Skipping memory lstm for low-latency real-time continuous speech separation," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 681–685.

[15] G.-P. Yang, S.-L. Wu, Y.-W. Mao, H.-y. Lee, and L.-s. Lee, "Interrupted and cascaded permutation invariant training for speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6369–6373.

[16] S.-F. Huang, S.-P. Chuang, D.-R. Liu, Y.-C. Chen, G.-P. Yang, and H. yi Lee, "Stabilizing Label Assignment for Speech Separation by Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 3056–3060.

[17] M. Yousefi, S. Khorram, and J. H. Hansen, "Probabilistic Permutation Invariant Training for Speech Separation," in *Proc. Interspeech 2019*, 2019, pp. 4604–4608.

[18] H. Tachibana, "Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 491–495.

[19] M. Yousefi and J. H. Hansen, "Single-channel speech separation using soft-minimum permutation invariant training," *Speech Communication*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639323000675

[20] D. Bralios, E. Tzinis, G. Wichern, P. Smaragdis, and J. L. Roux, "Latent iterative refinement for modular source separation," in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] S. Kim and M. Kim, "Bloom-net: Blockwise optimization for masking networks toward scalable and efficient speech enhancement," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 366–370.

[22] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7164–7175.

[23] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[24] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7098–7102.

[25] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[26] 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)*. IEEE, 2015, pp. 5206–5210.

[27] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proc. Interspeech 2020*, 2020, pp. 2637–2641.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.