CHiME-7 Challenge Track 2 Technical Report: Purification Adapted Speech Enhancement System

Jaehoo Jang, Myoung-Wan Koo†

Sogang University, Korea

jeahoo4128@sogang.ac.kr, mwkoo@sogang.ac.kr

Abstract

This technical report presents our implementation details of the speech enhancement system and provides experimental results on the UDASE task in the CHiME-7 challenge. In the domain-adapted RemixIT[1] pipeline, we introduce two significant modifications. Firstly, we incorporate a speech purification technique at the pipeline when conducting self-supervised learning for the student model. This technique predicts the frame-level SNR of the pseudo-target speech and utilizes them as weights for the discrepancy function between the pseudo-target speech and the student model's estimated speech. As a second modification, we replace the Sudo-rm-rf[2] architecture with the Mossformer[3], which incorporates convolution-augmented joint local and global self-attention mechanisms. It performs fully-computed self-attention on local chunks and utilizes linearized low-cost self-attention over the entire sequence. We demonstrate the superior performance of our approach compared to the baseline.

Index Terms: speech enhancement, noise suppression, domain adaptation, CHiME-7 challenge

1. Introduction

Speech enhancement systems that utilize supervised learning primarily rely on the methodology of extracting clean speech through a masking network [2, 4, 5, 6]. These systems are effective in terms of improving speech quality and noise suppression. However, if only unlabeled noise mixtures are available without clean source speech, it's impossible to train such systems. Thus, designing an unsupervised or self-supervised speech enhancement system that efficiently uses these noise mixtures remains a big challenge.

In light of this, the CHiME-7 challenge aims to improve the noise suppression performance on the in-domain speech by utilizing both an unlabeled in-domain CHiME-5[7] dataset and a labeled out-of-domain Librimix[8] dataset. RemixIT pipeline is a baseline provided by the challenge organizers. In this system, the fully-supervised teacher model is trained using Librimix. Then, CHiME-5 data is fed into the frozen teacher model, which outputs pseudo-target speeches and noise waveforms. These are used to create noise-permuted bootstrapped mixtures, which are then provided to the student model for self-supervised learning. Additionally, the parameters of the student model can be transferred back to the teacher model for continuous refinement at the end of each epoch.

Our main focus is to enhance the performance of the baseline system by implementing two key modifications. One is the speech purification technique and another is applying Mossformer architecture. In the next chapter, we explain a detail of our approach.

2. Methodology

2.1. Speech purification

The initial application of the speech purification method in terms of a self-supervised learning scheme was proposed in the work by [9]. In previous work, it is assumed that the frames of noise mixture with high SNR are almost identical to frames of clean speech. And the target speech is assumed to potentially contain noise. The characteristic of target speech in [9] is similar to the pseudo-target speech output by the teacher model in the RemixIT pipeline. This is due to the fact that a teacher model, trained on out-of-domain (OOD) data, cannot produce perfectly clear speech for the target domain. Therefore, during the training of the student model, we can use the corresponding method.

The core principle of speech purification is to design it in a way that allows frames of the pseudo-target speech with a comparatively high signal-to-noise ratio (SNR) to have more impact on the discrepancy function. Specifically, we calculate the weight of the frame-wise SNR for the pseudo-target speech and then multiply it with the frame-wise segmental signal-to-noise ratio (segmental SNR) proposed by [9].

The frame-wise SNR is output from an independent neural network referred to as SNR predictor. This network takes a pseudo-target speech as input and estimates SNR on a frame-by-frame basis. and the SNR predictor's output logits pass a sigmoid function to convert the weights of each frame.

The segmental SNR loss is detailed in equation (1). The J, H, and N respectively denote the number of frames, hop size, and frame size, while j refers to the index of a specific frame. w_i represents the Hann window function of length N, $\overline{s}$ denotes the pseudo-target speech (not the bootstrapped mixture), and $\overline{r}$ is the residual vector between $\overline{s}$ and the estimated speech. p_j denotes the weight for the j-th frame. Ultimately, we combine the segmental SNR loss with the SI-SDR, which is used as the loss function in the original pipeline, with equal weights.

$$\overline{SegSNR} = -\frac{1}{J}\sum_{j=0}^{J-1} p_j [10 \log_{10} \frac{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj}\overline{s}_i)^2}{\sum_{i=Hj}^{Hj+N-1}(w_{i-Hj}\overline{r}_i)^2}] \tag{1}$$

2.2. Mossformer architecture

Mossformer[3] is a transformer-based model specifically designed for monaural speech separation. It incorporates a gated single-head transformer architecture and enhanced joint self-attention with convolution. It achieves state-of-the-art results on the WSJ0-2/3mix[10] and WHAM!/WHAMR![11, 12] datasets. The model consists of an encoder-decoder structure with convolutional layers and a masking net. The encoder converts input speech data into feature vectors using Conv1D lay-

†Corresponding Author

		LibriCHiME-5	CHiME-5		
Model	Type	SI-SDR	OVR-MOS	BAK-MOS	SIG-MOS
Sudo-rm-rf (baseline)	Supervised	9.39	2.81	3.54	3.23
	RemixIT	11.70	2.86	3.65	3.28
	RemixIT$_{\text{vad}}$	11.57	2.85	3.66	3.27
Sudo-rm-rf$_p$	RemixIT$_{\text{vad}}$	12.42	2.88	**3.71**	3.33
Mossformer	Supervised	10.63	2.88	3.52	3.39
	RemixIT	12.42	**2.90**	3.60	**3.39**
	RemixIT$_{\text{vad}}$	**12.58**	2.84	3.48	3.35

Table 1: *Overall experiment results of our implemented pipeline system. The baseline system is Sudo-rm-rf. Improved version with speech purification is Sudo-rm-rf$_p$. The remaining is the Mossformer implementation system.*

ers and ReLU activation. These feature vectors are processed by the masking net, while the decoder reconstructs the encoded output back into the original speech format. The masking net includes normalization, positional encoding, pointwise convolutions, and reshaping operations before entering the Mossformer block. The Mossformer block, the core part of the model, incorporates a convolution module, an attentive gating mechanism, and joint local and global single-head self-attention. The convolution module captures local feature patterns using linear layers, SiLU activation functions, and 1D depth-wise convolutions. The attentive gating mechanism improves performance by incorporating attention-based gating, which combines local and global attention to model long-range interactions. During experiments, a large version of Mossformer with 42.1 million parameters was utilized.

3. Experimental setup

To address the CHiME-7 UDASE challenge, we follow the guidelines and utilize three different datasets: CHiME-5 (unlabeled in-domain dataset), Librimix (labeled out-of-domain dataset), and LibriCHiME-5 (labeled dataset resembling the in-domain data). We extract subsets for training, development, and evaluation from each dataset using an official toolkit from the CHiME-7 challenge's github[1]. The model is trained using these subsets in the original format provided by the toolkit.

To implement the pipeline system described in the methodology, we initially used the provided baseline implementation to assess its performance in our experimental setup. And we employed two external tools. The first tool[2] integrates an SNR predictor and the segmental SNR loss implementation. Additionally, we used publicly available pre-trained weights of the SNR predictor from the same github[2] and froze them during training. The pre-trained model was trained using a mixture of utterances from Librispeech[13] and noises from MUSAN[14]. More detailed informations are described in [9]. The second tool we utilized was the Mossformer architecture implementation[3], as described in [3].

Subsequently, we incorporated the aforementioned methods into the RemixIT baseline system separately. In one approach, we applied the purification method with SNR predictor, while in the other, we simply replaced Sudo-rm-rf with Mossformer. The former maintained the same hyperparameters as the baseline setting without setting 200 epochs, while for the latter, we followed the configuration for the large model as described in [3]. During training, we used a learning rate of 1.5e-4, conducted 100 epochs, and employed the Adam optimizer. In both approaches, we maintained the same learning rate throughout the training process. All experiments were conducted on six NVIDIA A100 GPUs with 80 GB of memory.

4. Result

Table 1 shows our experiment results. We used self-supervised learning with two subsets: unlabeled-10s (RemixIT setting) and vad-10s (RemixIT$_{vad}$ setting) from CHiME-5. The baseline Sudo-rm-rf experiment yielded an SI-SDR score of 11.57 using the vad-10s subset. This was achieved by training the models from scratch without altering the provided code by the challenge organizers. As incorporating purification techniques, the SI-SDR score improved to 12.42. Additionally, the corresponding systems achieved the highest BAK-MOS score of 3.71.

Mossformer outperforms Sudo-rm-rf in SI-SDR with an impressive score of 12.58 in RemixIT$_{vad}$ setting. In RemixIT setting, Mossformer also achieved the highest scores, recording 2.90 for OVR-MOS and 3.39 for SIG-MOS, respectively. Despite having significantly more parameters and slower training speeds compared to Sudo-rm-rf, latency is not a constraint in this challenge, so we proposed both system pipelines.

Consequently, we submitted two systems for the challenge. ISDS1 utilized the Mossformer model in the RemixIT setting, trained on unlabeled-10s data. For ISDS2, we employed the Sudo-rm-rf model with the purification method in the RemixIT$_{vad}$ setting.

5. Conclusions

In the CHiME-7 challenge, our speech enhancement system showed significant improvements through two key modifications: speech purification and the integration of the Mossformer architecture. By implementing speech purification techniques and incorporating convolution-augmented self-attention in the Mossformer model, we surpassed the performance of the baseline system. These results underscore the significance of speech purification methods and the advantages of using the innovative Mossformer architecture for noise suppression in speech enhancement tasks.

[1] https://github.com/UDASE-CHiME2023/baseline

[2] https://github.com/IU-SAIGE/pse

[3] https://github.com/modelscope/modelscope

6. References

[1] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1329–1341, 2022. [Online]. Available: https://doi.org/10.1109/JSTSP.2022.3200911

[2] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -rf: Efficient networks for universal audio source separation," in *30th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2020, Espoo, Finland, September 21-24, 2020.* IEEE, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/MLSP49062.2020.9231900

[3] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," *CoRR*, vol. abs/2302.11824, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.11824

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2915167

[5] 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 2020, Barcelona, Spain, May 4-8, 2020.* IEEE, 2020, pp. 46–50. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9054266

[6] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021.* IEEE, 2021, pp. 21–25. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9413901

[7] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 1561–1565. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1768

[8] 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.

[9] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1342–1356, 2022. [Online]. Available: https://doi.org/10.1109/JSTSP.2022.3181782

[10] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016.* IEEE, 2016, pp. 31–35. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7471631

[11] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1368–1372. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2821

[12] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020.* IEEE, 2020, pp. 696–700. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053327

[13] 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, South Brisbane, Queensland, Australia, April 19-24, 2015.* IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[14] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484