The NWPU-ByteAudio System for CHiME-7 Task 2 UDASE Challenge

Zihan Zhang[1,2]*,Runduo Han*[1]*,Ziqian Wang*[1]*,Xianjun Xia*[2]*, Yijian Xiao*[2]*, Lei Xie*[1*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]ByteDance, China

zhzhang@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

This paper describes the NWPU-ByteAudio system for CHiME-7 Task 2 - unsupervised domain adaptation for conversational speech enhancement (UDASE). To better make use of the in-domain mixture data, we improve the self-supervised learning (SSL) approach RemixIT with MetricGAN discriminator, resulting in an updated version called *RemixIT-G*. Under the RemixIT-G framework, we take Uformer+ as the speech enhancement model, which is an improved version of Uformer updated with the MetricGAN discriminator as well. We also apply an unsupervised noise adaptation model to generate noisy speech in the target domain. A perceptual contrast stretching (PCS) method is used to further improve the auditory perception quality of the enhanced speech. Our approach has achieved an SI-SDR of 12.95 and an OVRL-MOS of 3.07 in the CHiME-7 task 2 evaluation set.

Index Terms: Speech enhancement, unsupervised domain adaptation, RemixIT, MetricGAN

1. Introduction

In recent years, deep neural network (DNN) based speech enhancement has achieved superior performance over the traditional signal processing based methods [1]. However, the widely adopted supervised learning of neural models requires a large number of paired data for training. Since it is not feasible to capture paired noisy and clean speech in real-world scenarios, noisy speech is usually simulated by mixing clean speech and noise. This leads to a mismatch between the simulated training data and the real-world data in real applications [2]. *Unsupervised domain adaptation* has been recently proposed to solve this problem [2, 3, 4, 5].

The CHiME-7 Challenge UDASE task aims to use unlabeled data to overcome the performance drop caused by domain mismatch for speech enhancement models trained on simulated data. In other words, it focuses on improving neural speech enhancement models with the help of in-domain unlabeled data and out-domain labeled data.

In this challenge, we submitted a *self-supervised learning* (SSL) approach based on RemixIT [4]. RemixIT follows a continuous teacher-student learning scheme, effectively leveraging the in-domain mixture data to train a model matching the target domain. Moreover, we explore the efficacy of MetricGAN [6] discriminator in our approach. Specifically, we use Uformer [7] as the backbone of the teacher network but update it with MetricGAN+ [8] for better generalization capabilities. Similarly, we improve RemixIT by incorporating MetricGAN-U [9] during the training of the student network. To get better speech and noise estimates from the pre-trained teacher using unlabeled data, we adopt unsupervised noise adaptation by data simulation [10]. Moreover, to make use of the in-domain noise for model training, a voice activity detection (VAD) module [11] is used to automatically extract the noise segments from the CHiME-5 training set. Lastly, to further enhance the auditory quality of the enhanced speech, we apply perceptual contrast stretching (PCS) [12] during training and decoding.

*: Corresponding author.

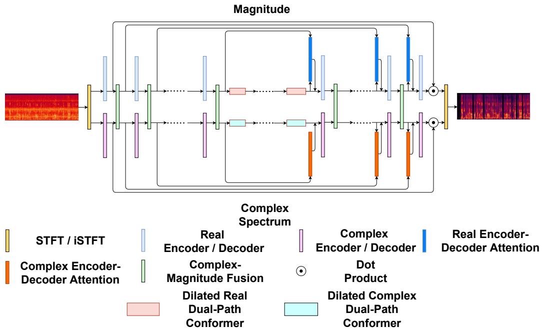

Figure 1: *The overall architecture of Uformer [7].*

2. Proposed approach

We first introduce our neural speech enhancement model and then describe the improved self-supervised learning scheme. Unsupervised noise adaptation and perceptual contrast stretching are finally introduced.

2.1. Uformer+: improved Uformer with MetricGAN

We choose *Uformer* as our enhancement model. Uformer [7] is a Unet-based dilated dual-path conformer network working in both complex and magnitude domains for simultaneous speech enhancement and dereverberation. As shown in Fig. 1, Uformer has two distinct branches – the magnitude branch and the complex branch. The primary focus of the magnitude branch is the suppression of noise, while the complex branch serves as an auxiliary module to compensate for the possible loss of spectral details and the phase mismatch.

In the challenge, we apply MetricGAN [6] in the training of the Uformer model to improve its generalization ability. The main idea of MetricGAN is to mimic the behavior of a target evaluation function such as PESQ [13] with a neural network and thus such a non-differentiable metric can be used during model training. Specifically, we use MetricGAN+ [8], an improved version, to predict PESQ or DNS-MOS [14]. We call our updated model *Uformer+*.

2.2. RemixIT-G: improved RemixIT with MetricGAN

RemixIT [4] is a self-supervised learning (SSL) strategy based on pseudo-labeling and continual training. It remixes the clean speech and noise estimated by a pre-trained teacher model to obtain the pseudo-labels which are then used to train a student

Table 1: *Results on reverberant LibriCHiME-5 dev set*

Model	Predicted metric	SI-SDR (dB)
Unprocessed	-	6.57
Sudo rm-rf	-	8.23
Uformer+	PESQ	**8.83**
Uformer+	DNS-MOS	8.66

model. The teacher model is constantly updated with the weight of the student model during training.

We use Uformer+ introduced in Section 2.1 as the teacher model, which is pre-trained on the out-of-domain data. To further improve performance, we add the MetricGAN-U [9] discriminator in RemixIT to learn the remixed pseudo-labels. This updated version is named as *RemixIT-G*.

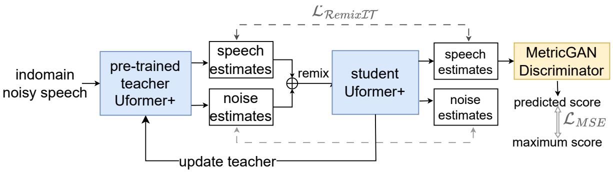

Figure 2: *RemixIT-G: improved RemixIT with MetricGAN.*

The structure of RemixIT-G is shown in Fig. 2. To predict PESQ without clean labels, we pre-train a MetricGAN discriminator on out-of-domain data. In fact, the quality net is trained simultaneously with the teacher model. The enhanced speech of the student model is fed into the discriminator which predicts PESQ. We calculate the loss with the maximum score, which can make the student model to focus on the improvement of perceived speech quality to obtain better performance.

2.3. Unsupervised noise adaptation

We use a data-simulation-based method called UNA-GAN [10] to generate noisy speech in the target domain. The structure of UNA-GAN is shown in Fig. 3. The generator is used to map clean speech to noisy speech. The discriminator determines whether the input is real or simulated. In this design, the generator learns to incorporate noise, which matches the target domain, into clean speech. To prevent the generator from generating too much noise and overwriting the clean speech, contrast learning is used to maximize the mutual information between the paired clean and noisy magnitude spectra.

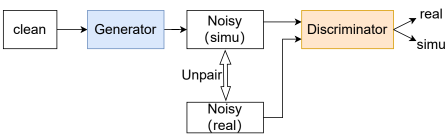

Figure 3: *The structure of UNA-GAN.*

2.4. Perceptual contrast stretching

Perceptual contrast stretching (PCS) [12] is derived based on the critical band importance function. It stretches the contrast of target features in the data based on a set of auditory weights. The weights are designed according to the critical band importance [15]. As a pre-processing step, PCS is applied to the input noisy speech during training. Likewise, PCS is also applied to the enhanced speech after inference as a post-processing step.

3. Experiments

3.1. Dataset

For supervised training, we use speech signals from LibriSpeech [16] to generate clean labels. To cope with the possible 1-3 speakers, we choose to generate clean labels for 1, 2, and 3 speakers in the ratio of 70%, 20%, and 10%, respectively. Noise signals are from WHAM! [17] noise dataset and CHiME-5 train set. We adopt a voice activity detection (VAD) model [11] to automatically extract the in-domain noise from the unlabeled CHiME-5 train set.

Table 2: *Results on reverberant LibriCHiME-5 eval set*

Model	Training strategy	SI-SDR (dB)
Unprocessed	-	6.59
Sudo rm-rf	fully-supervised	7.80
Sudo rm-rf	RemixIT	9.44
Sudo rm-rf	RemixIT + VAD	10.05
Uformer+	fully-supervised	8.79
Uformer+	UNA-GAN finetune	9.37
Uformer+	RemixIT	12.04
Uformer+	RemixIT-G	**12.95**

Table 3: *Results on CHiME-5 eval subset*

Model	Training strategy	OVRL	BAK	SIG
Unprocessed	-	2.84	2.92	3.48
Sudo rm-rf	fully-supervised	2.88	3.59	3.33
Sudo rm-rf	RemixIT	2.82	3.64	3.26
Sudo rm-rf	RemixIT + VAD	2.84	3.62	3.28
Uformer+	fully-supervised	3.03	3.88	3.35
Uformer+	UNA-GAN finetune	3.05	3.91	3.36
Uformer+	RemixIT	3.04	**3.94**	3.37
Uformer+	RemixIT-G	**3.07**	3.93	**3.39**

We use an online data generation strategy, randomly selecting a signal-to-noise ratio (SNR) to combine clean label and noise signal, while SNR randomly ranges from 0dB to 25 dB. A total of 50,000 room impulse responses (RIRs) are generated using the HYB method [18][1], with random room size 5×3×3 to 8×5×5 and RT60 0.2-1s. Totally 30% of the clean speech is convoluted with RIRs to simulate reverbed signals.

The other part of the paired noisy-clean data is generated by UNA-GAN, which is used to finetune the teacher model. We use CHiME-5 train set as unpaired noisy speech to train the UNA-GAN, thus making it capable of generating in-domain noise, even if the noise overlaps with human speaking and cannot be simply separated using VAD. And for the RemixIT-G training step, we only use the unlabeled CHiME-5 train set to adapt the student.

3.2. Experimental results

To prove the performance of Uformer+ and select the optimal training metric for MetricGAN, we first trained Uformer+ using the LibriMix dataset, following the same data setup as the baseline fully-supervised Sudo rm-rf model [19]. We test the SI-SDR [20] on the reverberant LibriCHiME-5 dev set, the results are shown in Table 1. Compared to the Sudo rm-rf, Uformer+ obtains an absolute improvement of 0.6dB in SI-SDR. For MetricGAN, it appears that the optimal training metric is PESQ.

Table 2 shows the SI-SDR metric on the reverberant LibriCHiME-5 eval set and Table 3 shows the DNSMOS metric on the CHiME-5 eval subset. It is worth mentioning that Uformer is a causal model with 9.46M parameters. We have the following conclusions. First, Uformer+ has a better performance than Sudo rm-rf. The increase of the parameters and the use of MetricGAN have brought benefits. Second, UNA-GAN is capable of learning a clean-to-noisy transformation and adapting the speech enhancement model to the target noise by simulated data. Third, RemixIT brings a large improvement on SI-SDR. Through our improvements, RemixIT-G can also improve the score of DNSMOS with the help of MetricGAN-U.

[1] https://github.com/marianne-m/brouhaha-vad

4. References

[1] P. Ochieng, "Deep neural network techniques for monaural speech enhancement: State of the art analysis," *arXiv preprint arXiv:2212.00369*, 2022.

[2] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[3] K. Saijo and T. Ogawa, "Remix-cycle-consistent learning on adversarially learned separator for accurate and stable unsupervised speech separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4373–4377.

[4] 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 Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[5] K. Saijo and T. Ogawa, "Self-remixing: Unsupervised speech separation via separation and remixing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[7] Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, and L. Xie, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7417–7421.

[8] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[9] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7412–7416.

[10] C. Chen, Y. Hu, H. Zou, L. Sun, and E. S. Chng, "Unsupervised noise adaptation using data simulation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] M. Lavechin, M. Métais, H. Titeux, A. Boissonnet, J. Copet, M. Rivière, E. Bergelson, A. Cristia, E. Dupoux, and H. Bredin, "Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation," *arXiv preprint arXiv:2210.13248*, 2022.

[12] R. Chao, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," *arXiv preprint arXiv:2203.17152*, 2022.

[13] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[14] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[15] C. Pavlovic, "Sii—speech intelligibility index standard: Ansi s3. 5 1997," *the Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1906–1906, 2018.

[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)*. IEEE, 2015, pp. 5206–5210.

[17] 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," *arXiv preprint arXiv:1907.01160*, 2019.

[18] E. Bezzam, R. Scheibler, C. Cadoux, and T. Gisselbrecht, "A study on more realistic room simulation for far-field keyword spotting," in *APSIPA ASC*. IEEE, 2020.

[19] 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.

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