ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446056

REMIXED2REMIXED: DOMAIN ADAPTATION FOR SPEECH ENHANCEMENT BY NOISE2NOISE LEARNING WITH REMIXING

Li Li, Shogo Seki

CyberAgent, Inc.

ABSTRACT

This paper proposes a domain adaptation method for speech enhancement called Remixed2Remixed. The proposed method adopts Noise2Noise (N2N) learning to adapt models trained on artificially generated (out-of-domain: OOD) noisy-clean pairs of data to better separate real-world recorded (in-domain) noisy data. The proposed method employs a teacher model trained on OOD data to acquire pseudo-in-domain speech and noise signals, which are shuffled and remixed twice in each batch to generate two bootstrapped mixtures. The student model is then trained by optimizing an N2N-based cost function computed using these two bootstrapped mixtures. As the training strategy is similar to that of the recently proposed RemixIT, we also investigate the effectiveness of the N2N-based loss as a regularization of RemixIT. Experimental results on the CHiME-7 unsupervised domain adaptation for conversational speech enhancement (UDASE) task revealed that the proposed method outperformed the challenging baseline system, RemixIT, and reduced the performance blurring caused by the teacher models.

Index Terms— Speech enhancement, self-supervised learning, domain adaption, Noise2Noise learning, RemixIT

1. INTRODUCTION

Speech enhancement (SE) [1] is one of the fundamental problems in speech signal processing and has many applications, either as a hearing aid or as a frontend system for many other tasks. It aims to improve the speech quality recorded in the presence of noise, interference, and reverberation, which has been greatly improved by deep neural networks (DNNs).

Supervised learning is the most studied approach to SE [2], wherein the model is trained on noisy-clean paired data to predict clean signals either directly [3, 4] or via masking [5–7]. Since recording such parallel pair data is impossible owing to crosstalk [8], generally, artificially synthesized noisy data are used to train SE models. However, due to the distribution mismatch primarily caused by the different acoustic conditions between synthetic (out-of-domain: OOD) and real-world recorded (in-domain) data, trained models are prone to performance degradation in case of recorded data. Several methods have recently been proposed to address this issue, including unsupervised methods aimed at learning models using nonparallel data. For example, machine learning methods that learn from positive and unlabeled data [8], replacement of the ground truth of clean speech with evaluation metric scores [9,10], and use of observation consistency [11,12] have been proposed.

Another effective solution involves performing domain adaptation, which adjusts an SE model pre-trained on OOD data to formulate an accurate noisy-clean mapping that matches in-domain data. The existing methods include adaptive mechanisms such as adversarial learning, optimal transport [13, 14], and self-supervised learning. RemixIT [15] is a method that employs self-distillation and comprises two networks. A teacher model pre-trained with synthesized OOD pair data[1] is used to produce pseudo-paired data of noisy speech and target signals for student training by remixing the separated speech and noise signals in each batch. Subsequently, a student model is trained using the generated pseudo-paired data by minimizing the loss between the predicted signals and pseudo-targets. The teacher model is continually updated via a weighted moving average (WMA) using the weights of the student model. Although RemixIT loss has been theoretically shown to ideally approach the supervised loss when the teacher model accurately predicts signals or when the student model observes a large number of pseudo-mixtures containing the same teacher estimates, this is not feasible with limited training resources. Consequently, the performance of RemixIT depends to some extent on the performance of its teacher model.

On the other hand, approaches applying basic statistical reasoning have been proposed for DNN-based image denoising. Based on the principle that corrupting the training target of the network with zero-mean noise does not change what the denoising network learns from the clean signal, Noise2Noise (N2N) [16] has demonstrated that a denoising model can be trained on noisy-noisy paired data, which was later extended to SE [17]. However, the collection of paired data containing two independent noisy realizations of the same clean signal is challenging, particularly for audio signals. This has motivated the proposal of improved methods to further remove the demands on the data. The Noisier2Noise (Nr2N) [18] and recorrupted-to-recorrupted (R2R) [19] methods use noise sampled from a known prior distribution to generate noisy pair data for image denoising. Noisy-target training (NyTT) [20, 21] uses noisy speech with additional noise to obtain noisy pair data for SE. Further, NyTT has been demonstrated to reduce noise close to the additional noise used in training; however, its performance degrades in case of other noise [22].

Considering the potential of learning models with less in-domain data than unsupervised learning that learns from scratch, this paper focuses on the domain adaptation approach and proposes a method called *Remixed2Remixed (Re2Re)*, which employs a teacher-student architecture similar to RemixIT and N2N learning. Specifically, the teacher model is used to generate pseudo-noisy pair data by performing the remix procedure twice, and the student model is trained using an N2N-based cost function. This facilitates

[1] RemixIT can be trained in a fully unsupervised manner, where the teacher model is trained solely using noisy speech by MixIT [11].

the obtaining of both in-domain speech and noise from only noisy speech. Moreover, through the explicit optimization of the cost function defined for denoising, the proposed method is expected to perform more consistently than RemixIT, regardless of the performance of the teacher model.

2. CONVENTIONAL METHOD: REMIXIT

2.1. Supervised learning

The speech and noise signals drawn from the corresponding distributions are denoted by $\boldsymbol{s} \sim \mathcal{D}_{\boldsymbol{s}}$ and $\boldsymbol{n} \sim \mathcal{D}_{\boldsymbol{n}}$, respectively. Synthetic noisy speech is obtained as $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$. With paired data $(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{n})$, a model predicting both speech and noise $\hat{\boldsymbol{s}}, \hat{\boldsymbol{n}} = \mathcal{F}(\boldsymbol{x}; \theta)$ parameterized by θ is trained under full supervision by optimizing the following cost function (i.e., minimizing the reconstruction error of both signals):

$$\mathcal{L}_{\text{supervised}} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s},\boldsymbol{n})}\big[\mathcal{L}(\hat{\boldsymbol{s}}, \boldsymbol{s}) + \mathcal{L}(\hat{\boldsymbol{n}}, \boldsymbol{n})\big]. \tag{1}$$

2.2. RemixIT

RemixIT [15] comprises a teacher model $\mathcal{F}_{\mathcal{T}}$ and student model $\mathcal{F}_{\mathcal{S}}$. Both models are initialized with a supervised pre-trained model using synthetic OOD pair data $(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{n})$ and further trained to enhance the real-world recorded data $\boldsymbol{x}' \sim \mathcal{D}_{\boldsymbol{x}'}$ with only in-domain data accessible. Given a mini-batch of in-domain noisy data $\mathbf{x}' = \mathbf{s}' + \mathbf{n}' \in \mathbb{R}^{B\times T}$, the teacher model estimates the speech and noise signals as follows:

$$\tilde{\mathbf{s}}', \tilde{\mathbf{n}}' = \mathcal{F}_{\mathcal{T}}(\mathbf{x}'; \theta_{\mathcal{T}}^{(k)}), \tag{2}$$

where the bold Roman font represents a batch $\mathbf{a} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_B]^{\mathsf{T}}$ including multiple signals $\boldsymbol{a}_b$ drawn from distribution $\mathcal{D}_{\boldsymbol{a}}$ and $\theta_{\mathcal{T}}^{(k)}$ denotes the parameters of teacher model at the k-th training epoch. Here, $^{\mathsf{T}}$ denotes the transpose operator and B and T denote the mini-batch size and signal length, respectively. The estimated signals are then shuffled and remixed to generate a bootstrapped mixture $\tilde{\mathbf{x}}'$, expressed as

$$\tilde{\mathbf{x}}' = \tilde{\mathbf{s}}' + \mathbf{P}\tilde{\mathbf{n}}'. \tag{3}$$

Here, $\mathbf{P} \sim \Pi_{B\times B}$ is a permutation matrix. The bootstrapped mixture is then used to generate the in-domain pseudo-paired data $(\tilde{\mathbf{x}}', \tilde{\mathbf{s}}', \tilde{\mathbf{n}}')$. The student model $\mathcal{F}_{\mathcal{S}}$ with parameter $\theta_{\mathcal{S}}^{(k)}$ is then trained by minimizing the reconstructed error between the outputs of the model and the pseudo-targets $\tilde{\mathbf{s}}'$ and $\tilde{\mathbf{n}}'$ as follows:

$$\hat{\mathbf{s}}', \hat{\mathbf{n}}' = \mathcal{F}_{\mathcal{S}}(\tilde{\mathbf{x}}'; \theta_{\mathcal{S}}^{(k)}), \tag{4}$$

$$\mathcal{L}_{\text{RemixIT}} = \sum_{b=1}^{B} \big[\mathcal{L}(\hat{\mathbf{s}}'_b, \tilde{\mathbf{s}}'_b) + \mathcal{L}(\hat{\mathbf{n}}'_b, [\mathbf{P}\tilde{\mathbf{n}}']_b)\big]. \tag{5}$$

To generate more accurate pseudo-targets, the teacher model is continuously updated using the weighted moving average (WMA) with the weights of the student model at constant epoch, which is expressed as $\theta_{\mathcal{T}}^{(k+1)} = \gamma\theta_{\mathcal{S}}^{(k)} + (1-\gamma)\theta_{\mathcal{T}}^{(k)}$, where $0 \le \gamma \le 1$ is the weight parameter.

Notably, the cost function of RemixIT $\mathcal{L}_{\text{RemixIT}}$ exhibits convergence properties when the Euclidean norm-based metric is used to measure the reconstruction error:

$$\begin{aligned}
&\mathcal{L}_{\text{RemixIT}} \propto \mathbb{E}\big[||\hat{\mathbf{s}}' - \tilde{\mathbf{s}}'||_2^2\big] = \mathbb{E}\big[||(\hat{\mathbf{s}}' - \mathbf{s}') - (\tilde{\mathbf{s}}' - \mathbf{s}')||_2^2\big] \\
&= \mathbb{E}\big[||(\hat{\mathbf{s}}' - \mathbf{s}')||_2^2\big] + \mathbb{E}\big[||(\tilde{\mathbf{s}}' - \mathbf{s}')||_2^2\big] - 2\mathbb{E}\big[(\tilde{\mathbf{s}}' - \mathbf{s}')^{\mathsf{T}}(\hat{\mathbf{s}}' - \mathbf{s}')\big] \\
&\approx \underbrace{\mathbb{E}\Big[||\boldsymbol{\epsilon}'_{\mathcal{S}}||_2^2\Big]}_{\text{supervised loss}} - \underbrace{\mathbb{E}\Big[||\boldsymbol{\epsilon}'_{\mathcal{T}}||_2^2\Big]}_{\text{constant w.r.t } \theta_{\mathcal{S}}} - 2\mathbb{E}\Big[\underbrace{(\tilde{\mathbf{s}}' - \mathbf{s}')^{\mathsf{T}}}_{\text{teacher error}} \underbrace{\frac{1}{M}\sum_{m=1}^{M}(\hat{\mathbf{s}}'_m - \tilde{\mathbf{s}}')}_{\substack{\text{empirical mean student error} \\ \text{w.r.t student model input}}}\Big],
\end{aligned} \tag{6}$$

where $\boldsymbol{\epsilon}'_{\mathcal{S}}$ and $\boldsymbol{\epsilon}'_{\mathcal{T}}$ are the reconstruction errors between the target signal $\mathbf{s}'$ and the outputs of the student and teacher models, respectively, and $||\cdot||_2^2$ denotes the squared $L2$ norm. (6) shows that when the third term is zero, the RemixIT loss approaches the supervised loss. This could be achieved by reducing either the teacher error to zero with an accurately estimated signal in the teacher model or the empirical mean student error to zero by exposing the student to various bootstrapped mixtures $\tilde{\mathbf{x}}'_m = \tilde{\mathbf{s}}' + \tilde{\mathbf{n}}'_m,\ m = 1, \ldots, M$ involving the same teacher estimate $\tilde{\mathbf{s}}'$ such that $\mathbb{E}_m[\hat{\mathbf{s}}'_m|\tilde{\mathbf{x}}'_m]$ would approach $\tilde{\mathbf{s}}'$ when $M \to \infty$. This property is important for ensuring that RemixIT can learn models as supervised learning. However, reducing the third term to zero with limited training resources, for example, with $M = 1$ is not feasible. Thus, the performance of RemixIT inevitably depends, to a certain extent, on the performance of its teacher model. Furthermore, a gap may remain with supervised learning.

3. PROPOSED METHOD: REMIXED2REMIXED

N2N [16] is an image denoising method utilizing basic statistical reasoning. It has demonstrated the feasibility of training a denoising model using noisy pair $(\boldsymbol{x}, \bar{\boldsymbol{x}})$ instead of $(\boldsymbol{x}, \boldsymbol{s})$ provided the noisy signal $\bar{\boldsymbol{x}} = \boldsymbol{s} + \bar{\boldsymbol{n}}$ satisfies $\mathbb{E}[\bar{\boldsymbol{x}}|\boldsymbol{x}] = \boldsymbol{s}$. Here, $\mathbb{E}[\bar{\boldsymbol{x}}|\boldsymbol{x}]$ represents the expected value of noisy signals when another noisy realization of clean signal is provided. This can be achieved when $\mathbb{E}[\bar{\boldsymbol{n}}] = 0$ and $\bar{\boldsymbol{n}}$ and $\boldsymbol{n}$ are independent of each other; that is, $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are two independent noisy realizations of $\boldsymbol{s}$. Inspired by the success of N2N, we extend it to SE with a motivation similar to that of [17]. In contrast to [17], where the paired data of two noisy realizations are obtained synthetically, we utilize the teacher-student architecture in RemixIT to generate paired noisy data by remixing in-domain speech and noise signals separated by a pre-trained OOD model. This renders it easy to obtain two in-domain noisy realizations that contain the same signals from only recorded noisy signals.

Fig. 1 presents a flowchart of the proposed method, *Remixed2Remixed (Re2Re)*. Re2Re has a teacher-student architecture similar to RemixIT, with the difference being that it generates in-domain paired data of two noisy realizations $(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')$ by performing the remixing process twice to generate two bootstrapped mixtures for every training iteration. In addition to the bootstrapped mixture $\tilde{\mathbf{x}}'$ generated by using (3), another bootstrapped mixture containing the teacher estimate $\tilde{\mathbf{s}}'$ is expressed as

$$\bar{\mathbf{x}}' = \tilde{\mathbf{s}}' + \mathbf{Q}\tilde{\mathbf{n}}', \tag{7}$$

where $\mathbf{Q}$ is uniformly sampled from a set of $B \times B$ permu-

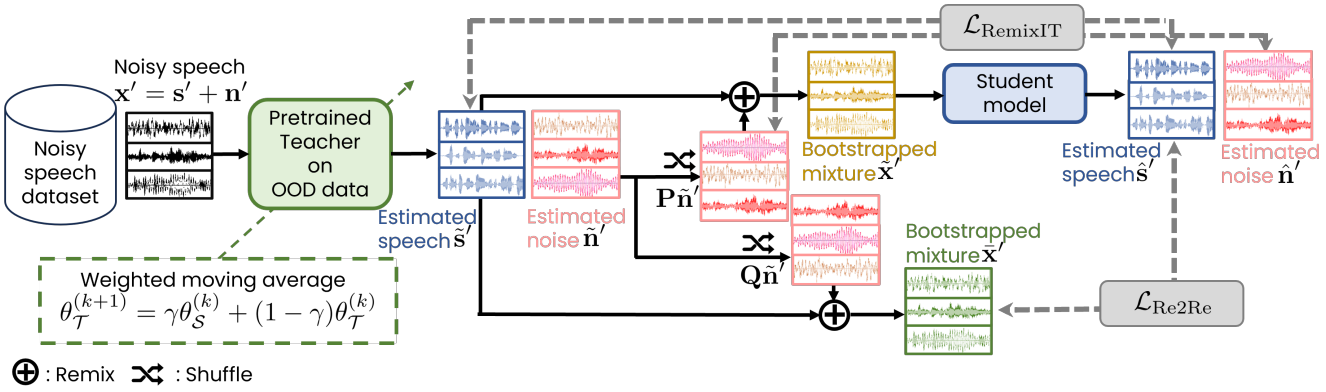

Fig. 1. Flowchart of proposed Remixed2Remixed.

tation matrices, such that $\mathbf{Q} \perp \mathbf{P}$. Using noisy pair data $(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')$, the student model is trained by minimizing the N2N-based loss

$$\mathcal{L}_{\text{Re2Re}} = \mathbb{E}_{(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')}\left[\mathcal{L}(\mathcal{F}_{\mathcal{S}}(\tilde{\mathbf{x}}'; \theta_{\mathcal{S}}), \bar{\mathbf{x}}')\right] = \mathbb{E}_{(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')}\left[\mathcal{L}(\hat{\mathbf{s}}', \bar{\mathbf{x}}')\right], \tag{8}$$

that satisfies $\mathbb{E}[\bar{\mathbf{x}}'|\tilde{\mathbf{x}}'] = \mathbf{s}'$ when sufficient paired data $(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')$ the student model could obtain. To generate sufficient pair data, we update the teacher model at every epoch such that $\tilde{\mathbf{x}}'$ and $\bar{\mathbf{x}}'$ could be considered as two noisy realizations of signal $\mathbf{s}'$ generated in an on-the-fly manner by corrupting speech signal $\mathbf{s}'$ with $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)} + \mathbf{P}\tilde{\mathbf{n}}'$ and $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)} + \mathbf{Q}\tilde{\mathbf{n}}'$. $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)}$ is the estimated error of the teacher model in kth epoch. It is generally assumed that noise signals and estimated errors have zero means [23, 24]. Therefore, $(\bar{\mathbf{x}}', \tilde{\mathbf{x}}')$ satisfies the zero-mean condition. Although $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)} + \mathbf{P}\tilde{\mathbf{n}}'$ and $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)} + \mathbf{Q}\tilde{\mathbf{n}}'$ are not exactly independent due to the presence of $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)}$, the impact of $\boldsymbol{\epsilon}_{\mathcal{T}}^{(k)}$ could be reduced by increasing the power of $\mathbf{P}\tilde{\mathbf{n}}'$ and $\mathbf{Q}\tilde{\mathbf{n}}'$. We also consider applying the N2N loss as a regularization for RemixIT, referred to as Re2Re_reg, whose cost function is given by

$$\mathcal{L}_{\text{Re2Re\_reg}} = \mathcal{L}_{\text{RemixIT}} + \beta\mathcal{L}_{\text{Re2Re}}, \tag{9}$$

where $\beta \geq 0$ is a parameter balancing the importance of each term. By explicitly optimizing the cost defined for denoising (8) instead of the reconstruction error of (5) between the outputs of the teacher and student models, methods using N2N loss are expected to perform more consistently than RemixIT, regardless of the performance of the teacher model.

4. EXPERIMENTAL EVALUATION

4.1. Datasets and experimental conditions

To evaluate the performance of the proposed Re2Re for domain adaptation, we conducted speech enhancement experiments on the CHiME-7 unsupervised domain adaptation for conversational speech enhancement (UDASE) task [25, 26], which comprises three datasets: (1) the LibriMix paired dataset for training OOD supervised SE model and development; (2) the CHiME-5 in-domain unlabeled dataset for adopting domain adaptation, development, and evaluation; and (3) the reverberant LibriCHiME-5 close-to-in-

Table 1. SI-SDR [dB] in reverberant CHiME-5 dataset and DNS-MOS in 1-spk subset of CHiME-5 dataset. * denotes model checkpoints provided by CHiME-7. Other models were trained from Sudo rm-rf* checkpoints. Bold fonts indicate the best scores.

Methods	CHiME-5 w/o VAD				CHiME-5 w/ VAD			
	SI-SDR	DNS-MOS			SI-SDR	DNS-MOS		
	[dB]	OVR	BAK	SIG	[dB]	OVR	BAK	SIG
Sudo rm-rf*	7.80	**2.88**	3.59	3.33	7.80	**2.88**	3.59	**3.33**
RemixIT*	9.44	2.83	**3.65**	3.25	10.05	2.84	**3.63**	3.27
RemixIT	10.94	2.84	3.63	3.29	10.68	2.85	3.51	**3.33**
Re2Re_reg	11.26	2.82	3.54	3.31	11.64	2.82	3.51	3.32
Re2Re	**11.65**	2.84	3.42	**3.37**	**11.76**	2.80	3.47	3.29

domain paired dataset for development and evaluation. All the datasets contain three subsets labeled with the maximum number of speakers: 1-spk, 2-spk, and 3-spk. **LibriMix [27]:** A noisy speech separation benchmark comprising clean speech and noise signals from LibriSpeech [28] and WHAM! [29], respectively. Libri2Mix and Libri3Mix, with two or three overlapping speakers in each mixture, were used as subsets of 2-spk and 3-spk, respectively, and a subset of 1-spk (Libri1Mix) was obtained by discarding one of the two speakers in the Libri2Mix mixtures. The proportions of the 1-spk, 2-spk, and 3-spk mixtures were 0.5, 0.25, and 0.25, respectively. **CHiME-5 [30]:** A dataset originally comprising noisy multi-speaker utterances of 20 conversation sessions recorded at 4-people dinner parties. CHiME-7 UDASE excerpted the recording channel where participants wearing microphones did not speak (i.e., the maximum number of simultaneously active speakers was three) and divided the signals into four subsets, including short segments of at least 3s length labeled by the maximum number of speakers according to the transcript. A subset containing noise-only segments was used to create the reverberant LibriCHiME-5 dataset for objective evaluation. Other subsets were further divided for train (≈83h), development (≈15.5h), and evaluation (≈7h), respectively. Segments for training were cut into chunks of up to 10s, and a voice activity detector (VAD) was applied for post-processing to obtain two versions of the training dataset: CHiME-5 w/o VAD and CHiME-5 w/ VAD. **Reverberant LibriCHiME-5:** A synthetic dataset that comprised reverberant noisy speech labeled with clean speech, where clean speech and noise signals were excerpted

Table 2. Average SI-SDRs and standard deviations [dB] over ten trials in reverberant LibriCHiME-5 dataset. All models were initialized by the same teacher models. Bold fonts indicate the best scores, and underlines indicate standard deviations smaller or equivalent to those achieved by RemixIT.

Methods	CHiME-5 w/o VAD				CHiME-5 w/ VAD			
	1-spk	2-spk	3-spk	Avg.	1-spk	2-spk	3-spk	Avg.
Sudo rm-rf	8.68 ± 0.63	8.76 ± 1.02	7.50 ± 1.55	8.67 ± 0.75	8.36 ± 0.86	8.46 ± 1.15	7.84 ± 1.43	8.37 ± 0.95
RemixIT	10.95 ± 0.94	10.76 ± 1.51	9.91 ± 2.13	10.87 ± 1.10	11.21 ± 0.56	11.25 ± 0.81	10.76 ± 1.05	11.20 ± 0.59
Re2Re_reg	$\mathbf{11.34 \pm 0.48}$	11.20 ± 0.92	10.53 ± 1.32	11.28 ± 0.57	11.35 ± 0.46	11.42 ± 0.61	10.84 ± 0.66	11.35 ± 0.48
Re2Re	11.24 ± 0.39	$\mathbf{11.75 \pm 0.77}$	$\mathbf{11.53 \pm 1.19}$	$\mathbf{11.38 \pm 0.45}$	$\mathbf{11.44 \pm 0.49}$	$\mathbf{11.83 \pm 0.73}$	$\mathbf{11.61 \pm 0.82}$	$\mathbf{11.55 \pm 0.53}$

from LibriSpeech [28] and the above-mentioned noise-only subset, respectively. Room impulse responses (RIRs) excerpted from the VoiceHome corpus were recorded in the living room, kitchen, and bedroom of three real homes with 18 different microphone arrays and loudspeaker settings. The mixtures were generated via the addition of noise segments to randomly sampled speech utterances convolved with randomly sampled RIRs, where the signal-to-noise ratio (SNR) for each speaker was distributed as a Gaussian distribution with a mean of 5 dB and a standard deviation (std) of 7 dB to match the CHiME-5 dataset. The proportions of the 1-spk, 2-spk, and 3-spk subsets were 0.6, 0.35, and 0.05, respectively. The data durations for development and evaluation were approximately 3h each.

To demonstrate the effectiveness of the cost function, we used the recipe provided by CHiME-7 without modifications except for the cost function. We used the Sudo rm-rf [6] architecture for both the teacher and student models, whose encoder and decoder comprised one-dimensional convolution and transpose convolution, respectively, with 512 filters of 41 taps and a hop size of 20 samples; the separator comprised 8 U-Conv blocks. The pre-trained teacher model initialized the student model and was continually updated by WMA with a weight of $\gamma = 0.01$ every epoch. The batch size was 24. The negative scale-invariant signal-to-distortion ratio (SI-SDR) [31] was used as the cost function for training the teacher and student models in RemixIT. We used the mean squared error between the estimated speech signal and bootstrapped mixture as $\mathcal{L}_{\mathrm{Re2Re}}$. For Re2Re_reg, we set $\beta = 100$ according to the development set. We calculated the DNS-MOS [32] scores on the 1-spk subset of the CHiME-5 dataset and SI-SDR [dB] on the reverberant LibriCHiME-5 dataset. Further details regarding the datasets and the baseline system can be found in [25, 26].

4.2. Experimental results

First, the proposed Re2Re and Re2Re_reg were compared with the CHiME-7 baseline system. Table 1 lists the SI-SDRs [dB] on the reverberant LibriCHiME-5 dataset and the DNS-MOS scores in the 1-spk subset of the CHiME-5 dataset. All models were trained using the Sudo rm-rf checkpoint provided by CHiME-7. The two proposed methods outperformed the baseline method in terms of SI-SDR, regardless of the application of VAD to the training data. Re2Re, using only the N2N loss, achieved SI-SDR that was approximately 0.71 dB and 1.08 dB higher than that achieved by using RemixIT. However, no improvement was observed for the DNS-MOS. This may be attributed to Re2Re only considering the reconstruction error of the speech signal, resulting in a less accurate estimation of background noise. Table 2 summarizes the SI-SDR[dB] and its std for each subset averaged over 10 teacher models. The two proposed methods achieved better and relatively stable performances in all cases. The models trained on data without and with VAD achieved SI-SDR improvements of 0.99 and 1.62 dB on the 2-spk and 0.58 and 0.85 dB on the 3-spk subsets, respectively, whereas the improvements on the 1-spk subset were limited to 0.29 and 0.23 dB. This could be another reason for the lack of improvement in the DNS-MOS. The standard deviations were approximately halved when trained on data without VAD and slightly reduced when trained on data with VAD, indicating that the performance of the student model relative to the teacher could be stabilized by N2N loss, even as a regularization. Subsequently, we compared our best systems to those submitted to the challenge, whose results are summarized in Table 3. The proposed methods achieved performance comparable to that of the system ranked second in the challenge regarding SI-SDR and the baseline RemixIT regarding DNS-MOS.

Table 3. SI-SDR [dB] in reverberant CHiME-5 dataset and DNS-MOS in 1-spk subset of CHiME-5 dataset achieved by our best systems and systems submitted to CHiME-7 challenge, ranked based on SI-SDR scores. Scores of other systems are obtained from [26]. The presence of "VAD" indicates the version of the CHiME-5 dataset used for training.

Systems	SI-SDR [dB]	DNS-MOS		
		OVRL	BAK	SIG
NWPU and ByteAudio	13.0	3.07	3.93	3.39
Sogang ISDS1	12.4	2.90	3.60	3.39
RemixIT-VAD	10.1	2.84	3.62	3.28
Conformer Metric GAN	7.8	3.40	3.97	3.76
Sudo rm-rf	7.8	2.88	3.59	3.33
Input	6.6	2.84	2.92	3.48
Re2Re	12.41	2.85	3.42	3.35
Re2Re-VAD	12.41	2.79	3.39	3.32

5. CONCLUSIONS

The paper proposed applying N2N learning to SE domain adaptation. The proposed method, called Remixed2Remixed, used a teacher-student architecture, wherein a teacher model was pre-trained with OOD data and then used to generate pseudo-noisy pair data, and a student model was trained by minimizing an N2N-based loss function. Experimental results on the CHiME-7 UDASE task revealed that Re2Re outperformed RemixIT w.r.t SI-SDR with a more stable performance.

6. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice, CRC press*, 2007.

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

[3] C. Macartney, and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[4] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in Proc. *Interspeech*, pp. 3291–3295, 2020.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm -rf: Efficient networks for universal audio source separation," in Proc. *MLSP*, pp. 1–6, 2020.

[7] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in Proc. *ICASSP*, pp. 6648–6652, 2021.

[8] N. Ito and M. Sugiyama, "Audio Signal Enhancement with Learning from Positive and Unlabeled Data," in Proc. *ICASSP*, pp. 1–5, 2023.

[9] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "Speech enhancement using end-to-end speech recognition objectives," in Proc. *WASPAA*, pp. 234–238, 2019.

[10] 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 Proc. *ICASSP*, pp. 7412–7416, 2022.

[11] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in Proc. *Adv. NIPS*, 33, pp. 3846–3857, 2020.

[12] K. Saijo, and T. Ogawa, "Self-Remixing: Unsupervised Speech Separation VIA Separation and Remixing," in Proc. *ICASSP*, pp. 1–5, 2023.

[13] C. F. Liao, Y. Tsao, H. Y. Lee, and H. M. Wang, "Noise Adaptive Speech Enhancement Using Domain Adversarial Training," in Proc. *Interspeech*, pp. 3148–3152, 2019.

[14] H. Y. Lin, H. H. Tseng, X. Lu, and Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in Proc. *Adv. NIPS*, 34, pp. 19935–19946, 2021.

[15] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, A, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1329–1341, 2022.

[16] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning Image Restoration without Clean Data," in Proc. *PMLR* pp. 2965–2974, 2018.

[17] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, "Speech Denoising Without Clean Training Data: A Noise2Noise Approach," in Proc. *Interspeech*, pp. 2716–2720, 2021.

[18] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, "Noisier2noise: Learning to denoise from unpaired noisy data," in Proc. *CVPR*, pp. 12064–12072, 2020.

[19] T. Pang, H. Zheng, Y, Quan, and H. Ji, "Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising," in Proc. *CVPR*, pp. 2043–2052, 2021.

[20] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in Proc. *EUSIPCO*, pp. 436–440, 2021.

[21] A. Sivaraman, S. Kim, and M. Kim, "Personalized speech enhancement through self-supervised data augmentation and purification," in Proc. *Interspeech*, pp. 2676–2680, 2021.

[22] T. Fujimura and T. Toda, "Analysis Of Noisy-Target Training For Dnn-Based Speech Enhancement," in Proc. *ICASSP*, pp. 1–5, 2023.

[23] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[24] Y. Lu and P. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty," *IEEE TASLP*, vol. 19, no. 5, pp. 1123–1137, 2010.

[25] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," *arXiv preprint arXiv:2307.03533*, 2023.

[26] Website of CHiME-7 Task 2 UDASE: `https://www.chimechallenge.org/current/task2/index` (last access: Sep. 4, 2023)

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

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in Proc. *ICASSP*, pp. 5206–5210, 2015.

[29] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. LeRoux, "WHAM!: Extending speech separation to noisy environments," in Proc. *Interspeech*, pp. 1368–1372, 2019.

[30] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in Proc. *Interspeech*, pp. 1561–1565, 2018.

[31] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in Proc. *ICASSP*, pp. 626–630, 2019.

[32] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in Proc. *ICASSP*, pp. 886–890, 2022.