

CONTINUAL SELF-TRAINING WITH BOOTSTRAPPED REMIXING FOR SPEECH ENHANCEMENT

Efthymios Tzinis^{1,*}, Yossi Adi², Vamsi K. Ithapu³, Buye Xu³, Anurag Kumar³

¹University of Illinois at Urbana-Champaign, ²Meta AI Research, ³Meta Reality Labs Research
etzinis2@illinois.edu, {adiyoss, ithapu, xub, anuragrkr}@fb.com

ABSTRACT

We propose *RemixIT*, a simple and novel self-supervised training method for speech enhancement. The proposed method is based on a continuously self-training scheme that overcomes limitations from previous studies including assumptions for the in-domain noise distribution and having access to clean target signals. Specifically, a separation teacher model is pre-trained on an out-of-domain dataset and is used to infer estimated target signals for a batch of in-domain mixtures. Next, we bootstrap the mixing process by generating artificial mixtures using permuted estimated clean and noise signals. Finally, the student model is trained using the permuted estimated sources as targets while we periodically update teacher’s weights using the latest student model. Our experiments show that *RemixIT* outperforms several previous state-of-the-art self-supervised methods under multiple speech enhancement tasks. Additionally, *RemixIT* provides a seamless alternative for semi-supervised and unsupervised domain adaptation for speech enhancement tasks, while being general enough to be applied to any separation task and paired with any separation model.

Index Terms— Self-supervised learning, speech enhancement, unsupervised denoising, zero-shot learning, domain adaptation

1. INTRODUCTION

Neural networks have been found to be highly effective and widely applicable to a large number of audio and speech problems, including speech enhancement, where the goal is to improve the quality and intelligibility of degraded speech signals. In recent years, several neural architectures have reported state-of-the-art results for supervised [1], real-time [2] and semi-supervised [3] speech enhancement tasks. Mostly driven by supervised learning, training such models requires large amounts of audio data which are expected to closely match the distribution of the test time input noisy recordings. While limited supervised data might often be available, supervised speech enhancement systems trained on them provides severely inferior performance due to the mismatch with the actual test distribution.

To address these problems and to reduce reliance on purely supervised data, several speech enhancement and audio source separation studies have shifted their focus towards self-supervised methods [4]. In [5], a model was trained to estimate the signal-to-noise ratio (SNR) of noisy mixture recordings and assign a confidence value on each noisy segment. Next, a separation model was trained on the noisy speech mixtures using a weighted reconstruction loss function to filter out the contribution of noisy ground-truth speech utterances. Lately, mixture invariant training (MixIT) has been proposed in [6]

which enables unsupervised training of separation models by generating artificial mixtures of mixtures (MoMs) and letting the separation model estimate and reassign the sources back to the ground-truth mixtures. MixIT has been experimentally shown to provide a robust unsupervised solution for speech enhancement under a variety of setups [6, 7, 8]. However, MixIT assumes access to in-domain noise samples and slightly alters the input SNR distribution by training on artificial MoMs with more than one noise sources [9].

On the other hand, teacher-student setups have shown significant improvements in several audio processing tasks [10]. In [11], a student model was trained at the outputs of a pre-trained MixIT model for solving the problem of the artificially created input SNR mismatch between the train and test mixture distributions. An energy threshold was used to reduce the number of sources appearing in the noisy mixtures. Moreover, a student model can be adapted to a given test set using regression over the pre-trained teacher’s estimates [12]. Closest to our work is the self-training framework, proposed in [13], for semi-supervised singing voice separation where the teacher was pre-trained on an out-of-domain (OOD) supervised data and used for predicting estimated sources on the larger in-domain noisy dataset. The new self-labeled version of the dataset was filtered for low-quality separated sources and stored for offline training of a new student model using artificially generated mixtures from the estimated self-labeled estimated sources. Although all the aforementioned works assumed a frozen teacher, other works in automatic speech recognition have shown significant benefits when updating the teacher model using a moving mean teacher [14].

In this paper, we propose a self-training method capable of performing self-supervised learning using large in-domain noisy datasets, while requiring only an OOD pre-trained teacher model (e.g. MixIT on an OOD dataset). In contrast to self-training methods in the literature which use ad-hoc filtering procedures to enhance the quality of the teacher model estimates, our method trains the student model by performing online remixing of the teacher’s estimated sources. Moreover, instead of freezing the teacher model, *RemixIT* treats the self-training process as a lifelong learning procedure by using sequential and moving averaging update schemes which enables faster convergence. Our experiments showcase the general applicability of our method towards *self-supervised speech enhancement*, *semi-supervised OOD generalization*, and *zero-shot domain adaptation*. We also provide an intuitive explanation of why our bootstrapped remixing process works under minimal assumptions.

2. REMIXIT METHOD

We present *RemixIT* for the general case of speech enhancement where the goal is to reconstruct the clean speech from a noisy speech signal. Formally, we train a separation model f_S which outputs M

*Work done during an internship at Meta Reality Labs Research.

Algorithm 1: REMIXIT for the noisy dataset \mathcal{D}_m .

```
 $\theta_{\mathcal{T}}^{(0)} \leftarrow \text{PRETRAINTEACHER}(f_{\mathcal{T}}, \mathcal{D}')$ 
 $\theta_{\mathcal{S}} \leftarrow \text{INITIALIZESTUDENT}(f_{\mathcal{S}})$ 
for  $k = 0; k++;$  while  $k \leq K$  do
  for SAMPLEBATCH  $\mathbf{m} \in \mathcal{D}_m, \mathbf{m} \in \mathbb{R}^{B \times T}$  do
     $\tilde{\mathbf{s}}, \tilde{\mathbf{n}} \leftarrow f_{\mathcal{T}}(\mathbf{m}; \theta_{\mathcal{T}}^{(k)})$  // Noisy estimates
     $\tilde{\mathbf{m}} = \tilde{\mathbf{s}} + \mathbf{\Pi} \tilde{\mathbf{n}}$  // Bootstrapped remixing
     $\hat{\mathbf{s}}, \hat{\mathbf{n}} \leftarrow f_{\mathcal{S}}(\tilde{\mathbf{m}}; \theta_{\mathcal{S}}^{(k)})$  // Student estimates
     $\mathcal{L}_{\text{RemixIT}} = \sum_{b=1}^B [\mathcal{L}(\hat{\mathbf{s}}_b, \tilde{\mathbf{s}}_b) + \mathcal{L}(\hat{\mathbf{n}}_b, [\mathbf{\Pi} \tilde{\mathbf{n}}]_b)]$ 
     $\theta_{\mathcal{S}} \leftarrow \text{UPDATESTUDENT}(\theta_{\mathcal{S}}, \nabla_{\theta_{\mathcal{S}}} \mathcal{L}_{\text{RemixIT}})$ 
  end
   $\theta_{\mathcal{T}}^{(k+1)} \leftarrow \text{UPDATETEACHER}(\theta_{\mathcal{T}}^{(k)}, \theta_{\mathcal{S}})$ 
end
```

sources for each noisy recording from an input batch $\mathbf{x} \in \mathbb{R}^{B \times T}$ containing B waveforms, each with T samples in the time-domain:

$$\hat{\mathbf{s}}, \hat{\mathbf{n}} = f_{\mathcal{S}}(\mathbf{x}; \theta_{\mathcal{S}}), \mathbf{x} = \mathbf{s} + \sum_{i=1}^{M-1} \mathbf{n}_i = \hat{\mathbf{s}} + \sum_{i=1}^{M-1} \hat{\mathbf{n}}_i, \quad (1)$$

where $\hat{\mathbf{s}}, \mathbf{s} \in \mathbb{R}^{B \times T}$, $\hat{\mathbf{n}}, \mathbf{n} \in \mathbb{R}^{(M-1) \times B \times T}$, $\theta_{\mathcal{S}}$ are: the estimated speech signals, the clean speech targets, the estimated noise signals, the noise targets and the parameters of the model, respectively. In this work, we force the estimated sources $\hat{\mathbf{s}}, \hat{\mathbf{n}}$ to add up to the initial input mixtures \mathbf{x} using a mixture consistency layer [15].

2.1. Mixture invariant training

MixIT [6] has proven its effectiveness under various self-supervised speech enhancement settings [7, 8]. Specifically, MixIT assumes that the training dataset consists of two portions ($\mathcal{D}_m, \mathcal{D}_n$), where \mathcal{D}_m is the part of the dataset which carries mixtures of speech and one noise source while \mathcal{D}_n contains isolated noise recordings. During training, a new batch of artificial mixtures of mixtures (MoMs) is generated, $\mathbf{x} = \mathbf{s} + \mathbf{n}_1 + \mathbf{n}_2$, by sampling a batch of noisy speech recordings $\mathbf{m} \sim \mathcal{D}_m$ and a batch of clean noise samples $\mathbf{n}_2 \sim \mathcal{D}_n$, where $\mathbf{m} = \mathbf{s} + \mathbf{n}_1$. The separation model always estimates $M = 3$ sources ($\hat{\mathbf{s}}, \hat{\mathbf{n}}_1, \hat{\mathbf{n}}_2 = f_{\mathcal{S}}(\mathbf{x}; \theta_{\mathcal{S}})$) and the following permutation invariant loss is minimized for the b -th input MoM in the batch:

$$\mathcal{L}_{\text{MixIT}}^{(b)} = \min_{\pi} [\mathcal{L}(\hat{\mathbf{s}}_b + \hat{\mathbf{n}}_{\pi_1, b}, \mathbf{m}_b) + \mathcal{L}(\hat{\mathbf{n}}_{\pi_2, b}, \mathbf{n}_{2, b})], \quad \forall b, \quad (2)$$

where $\pi \in \{(2, 3), (3, 2)\}$ is the set of permutations of the output noise slots. However, MixIT depends on in-domain isolated noise recordings which makes it impractical for real-world settings. In most of the cases, matching the real noise distribution with the available noise set \mathcal{D}_n is a strenuous task. The data augmentation proposed in [9] shows some improvements when an extra noise source from an OOD noise distribution is injected to the input MoM. Nevertheless, the performance of that method depends on the level of distribution shift between the actual noise distribution and \mathcal{D}_n .

2.2. Continual self-training with bootstrapped remixing

RemixIT does not assume access to in-domain information. Thus, we can only draw mixtures from the in-domain noisy dataset $\mathbf{m} = \mathbf{s} + \mathbf{n}$ ($\mathbf{m} \sim \mathcal{D}_m$) where the noisy speech recordings contain a single noise source each and thus, $\mathbf{m}, \mathbf{s}, \mathbf{n} \in \mathbb{R}^{B \times T}$. *RemixIT* leverages a student-teacher framework to bootstrap the remixing process by permuting the previous noisy estimates, remixing them and using them as targets for training. We summarize *RemixIT* in Algorithm 1.

2.2.1. *RemixIT's* continual self-training framework

We assume that we can pre-train in a supervised or a self-supervised way a teacher model $f_{\mathcal{T}}$ on an OOD dataset \mathcal{D}' which meets the specifications of Equation 1. Now, the first step is to use the teacher model to estimate some new noisy targets for a given mixture batch $\mathbf{m} = \mathbf{s} + \mathbf{n} \in \mathbb{R}^{B \times T}$, $\mathbf{m} \sim \mathcal{D}_m$ as follows:

$$\tilde{\mathbf{s}}, \tilde{\mathbf{n}} = f_{\mathcal{T}}(\mathbf{m}; \theta_{\mathcal{T}}^{(k)}), \mathbf{m} = \mathbf{s} + \mathbf{n} = \tilde{\mathbf{s}} + \sum_{i=1}^{M-1} \tilde{\mathbf{n}}_i, \quad (3)$$

where k denotes the optimization step. If the teacher network was obtained by supervised (unsupervised via MixIT) OOD pre-training, we would have $M = 2$ ($M = 3$) output slots. Next, we use these estimated sources to generate new noisy mixtures as shown below:

$$\tilde{\mathbf{m}} = \tilde{\mathbf{s}} + \tilde{\mathbf{n}}^{(\mathbf{\Pi})} \in \mathbb{R}^{B \times T}, \quad \tilde{\mathbf{n}}^{(\mathbf{\Pi})} = \mathbf{\Pi} \tilde{\mathbf{n}}, \quad \mathbf{\Pi} \sim \mathcal{P}_{B \times B}, \quad (4)$$

where $\mathbf{\Pi}$ is drawn uniformly from the set of all $B \times B$ permutation matrices. Now, we simply use the permuted target pairs to train the student model $f_{\mathcal{S}}$ on the bootstrapped mixtures $\tilde{\mathbf{m}}$ as follows:

$$\hat{\mathbf{s}}, \hat{\mathbf{n}} = f_{\mathcal{S}}(\tilde{\mathbf{m}}; \theta_{\mathcal{S}}^{(k)}), \quad \hat{\mathbf{s}}, \hat{\mathbf{n}} \in \mathbb{R}^{B \times T} \quad (5)$$
$$\mathcal{L}_{\text{RemixIT}}^{(b)} = \mathcal{L}(\hat{\mathbf{s}}_b, \tilde{\mathbf{s}}_b) + \mathcal{L}(\hat{\mathbf{n}}_b, \tilde{\mathbf{n}}_b^{(\mathbf{\Pi})}), \quad b \in \{1, \dots, B\},$$

where the proposed loss function resembles a regular supervised setup with the specified signal-level loss function \mathcal{L} . Our method does not artificially alter the input SNR distributions similar to MixIT-like [6, 7, 9] training recipes. Instead, the student model is trained on mixtures with the same number of sources for the bootstrapped mixtures where the teacher model had performed adequately. Unlike previous teacher-student methods which use the same teacher-estimated source-pairs as the targets for the student network [11, 12], the proposed bootstrapped mixtures increase the input mixture diversity and allow faster model training. This is especially useful in settings with a large distribution shift between teacher's and student's training data. Moreover, in contrast to the self-training approach in [13], where the teacher model is frozen and the inference on the in-domain dataset \mathcal{D}_m is performed offline and the new self-labeled dataset is stored, *RemixIT* employs a lifelong learning process. Our method is general enough that could be paired with any online co-training method which continuously updates the teacher's weights in addition to the main student training.

2.2.2. Error analysis under the Euclidean norm

We can express the errors produced by the student $\hat{\mathbf{R}}_{\mathcal{S}}$ and the teacher $\tilde{\mathbf{R}}_{\mathcal{T}}$ w.r.t. the initial clean targets as random variables:

$$\hat{\mathbf{R}}_{\mathcal{S}} = \hat{\mathbf{S}} - \mathbf{S}, \quad \tilde{\mathbf{R}}_{\mathcal{T}} = \tilde{\mathbf{S}} - \mathbf{S}, \quad (\mathbf{S}, \mathbf{N}) \sim \mathcal{D} \quad (6)$$
$$\hat{\mathbf{R}}_{\mathcal{S}} \sim P(\hat{\mathbf{R}}_{\mathcal{S}} | \tilde{\mathbf{S}}, \tilde{\mathbf{N}}, \mathbf{\Pi}), \quad \tilde{\mathbf{R}}_{\mathcal{T}} \sim P(\tilde{\mathbf{R}}_{\mathcal{T}} | \mathbf{S}, \mathbf{N}).$$

Now, we focus on the part of the objective function which is minimized at every student optimization step w.r.t. the speech component. Assuming unit-norm vector signals and using a signal-level loss \mathcal{L} that minimizes the squared error between the estimated and target signals, the student's *RemixIT* loss function is equivalent to:

$$\mathcal{L}_{\text{RemixIT}} \propto \mathbb{E} \left[\|\hat{\mathbf{S}} - \tilde{\mathbf{S}}\|_2^2 \right] = \mathbb{E} \left[\|\hat{\mathbf{S}} - \mathbf{S}\|_2^2 - \|\tilde{\mathbf{S}} - \mathbf{S}\|_2^2 \right]$$
$$\propto \underbrace{\mathbb{E} \left[\|\hat{\mathbf{R}}_{\mathcal{S}}\|_2^2 \right]}_{\text{Supervised Loss}} + \underbrace{\mathbb{E} \left[\|\tilde{\mathbf{R}}_{\mathcal{T}}\|_2^2 \right]}_{\text{Constant w.r.t. } \theta_{\mathcal{S}}} - 2 \underbrace{\mathbb{E} \left[\langle \hat{\mathbf{R}}_{\mathcal{S}}, \tilde{\mathbf{R}}_{\mathcal{T}} \rangle \right]}_{\text{Errors correlation}}. \quad (7)$$

Ideally, this loss could lead to the same optimization objective with a supervised setup if the last inner-product term was zero.

$\langle \widehat{\mathbf{R}}_S, \widehat{\mathbf{R}}_T \rangle = 0$ could be achieved if the teacher produced outputs indistinguishable from the clean target signals or the conditional error distributions in Equation 6 were independent. Intuitively, as we continually update the teacher model and refine its estimates, we minimize the norm of the teacher error. Additionally, the bootstrapped remixing process forces the errors to be more uncorrelated since the student tries to reconstruct the same clean speech signals \mathbf{s} , similar to its teacher, but under a different mixture distribution. Formally, the student tries to reconstruct \mathbf{s} when observing the bootstrapped mixtures $\widehat{\mathbf{m}} = \widehat{\mathbf{s}} + \widehat{\mathbf{n}}^{(\Pi)}$ while the teacher tries to reconstruct \mathbf{s} from the initial input mixtures $\mathbf{m} = \mathbf{s} + \mathbf{n}$.

3. EXPERIMENTAL FRAMEWORK

3.1. Datasets

DNS-Challenge (DNS): We use the DNSChallenge 2020 benchmark dataset [16] which covers a wide variety of noisy speech conditions. This dataset consists of 64,649 and 150 pairs of clean speech and noise recordings for training and testing, respectively. DNS is mainly used for showing the effectiveness of *RemixIT* at leveraging vast amounts of noisy mixture recordings.

LibriFSD50K (LFS): This dataset consists of a diverse set of speakers drawn from the LibriSpeech [17] corpus and a wide variety of background noises from FSD50K [18]. Specifically, 45,602 and 3,081 for training and validation, correspondingly. We follow the same data generation procedure as indicated in [8]. In this study, this dataset is mainly used for the OOD unsupervised or semi-supervised pre-training of speech enhancement models.

WHAM!: We follow the same procedure as in [8] in order to generate noisy mixtures using speakers and noise sources from the WHAM! [19] dataset. We use this dataset as a medium-sized dataset with 20,000 training noisy-speech pairs and 3,000 test mixtures.

VCTK: We use the test partition of the VCTK dataset proposed in [20] which includes 586 randomly generated noisy speech samples by mixing recordings from the VCTK speech corpus [21] and the DEMAND [22] noisy data collection.

3.2. Separation model

We want to underline that the proposed method can be applied alongside any separation model. In this work, we chose the Sudo `rm -rf` [23] architecture since it achieves a good trade-off between speech enhancement quality and time-memory computational requirements. Specifically, we use the Sudo `rm -rf` variation with group communication [24] and the default parameters in [25], which has shown promising results in speech enhancement tasks [8]. All models have $U = 8$ U-ConvBlocks except for the experiments where we increase the depth of the new student networks to 16 and 32. We fix the number of output slots to $M = 3$ for MixIT models or $M = 2$ otherwise.

3.3. *RemixIT* configurations

For the unsupervised *RemixIT*, we assume that the initial teacher model was pre-trained using MixIT at a specified OOD dataset. For semi-supervised *RemixIT*, we pre-train the teacher using conventional supervised training. We also experiment with various online teacher updating protocols such as:

$$\widehat{\theta}_T^{(K)} := \theta_S^{(K)}, \quad \widehat{\theta}_T^{(k+1)} := \gamma \theta_S^{(k)} + (1 - \gamma) \widehat{\theta}_T^{(k)}, \quad (8)$$

where k denotes the training epoch index. For the sequentially updated teacher, we replace the old teacher with the latest student every

$K = 20$ epochs. For the zero-shot domain adaptation experiments, we first set the student to be the same as the teacher $\theta_S^{(0)} := \widehat{\theta}_T^{(0)}$ and then use the moving average teacher update with $\gamma = 0.01$.

3.4. Training and evaluation details

Although we could use any valid signal-level loss function (see Equations 2, 5), we choose the negative scale-invariant signal to distortion ratio (SI-SDR) [26]:

$$\mathcal{L}(\widehat{y}, y) = -\text{SI-SDR}(\widehat{y}, y) = -10 \log_{10} \left(\frac{\|\alpha y\|^2}{\|\alpha y - \widehat{y}\|^2} \right), \quad (9)$$

where $\alpha = \widehat{y}^\top y / \|y\|^2$ makes the loss invariant to the scale of the estimated source \widehat{y} and the target signal y . We train all models using the Adam optimizer [27] with a batch size of $B = 2$ and an initial learning rate of 10^{-3} which is divided by 2 every 6 epochs.

We evaluate the robustness of our speech enhancement models on the each test set after 100 epochs for the supervised trained and pre-trained teachers as well as 60 epochs for all the other configurations. Specifically, we report the SI-SDR [26], the Short-Time Objective Intelligibility (STOI) [28] and the Perceptual Evaluation of Speech Quality (PESQ) [29] for 16kHz target signals.

4. RESULTS

4.1. Continuous refinement of teacher’s estimates

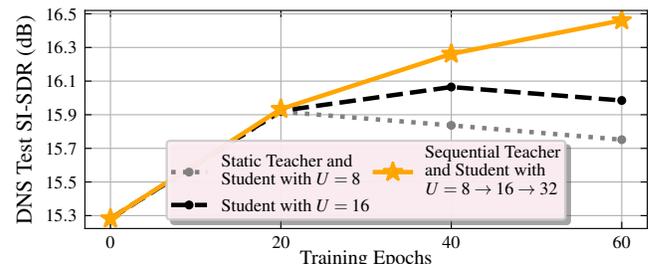


Fig. 1: Speech enhancement performance on the DNS test set when using *RemixIT* with different teacher update protocols. All approaches use the same teacher architecture with $U = 8$ which was pre-trained in a supervised way using the training split of WHAM!. Notice that by sequentially updating the teacher (orange solid line) every 20 epochs and replacing it with the previous student, we are able to obtain significant improvements over the methods which use the same static teacher (gray and black dashed lines). In this semi-supervised *RemixIT* setup, considering the large mismatch between WHAM! and DNS datasets, the student model significantly outperforms the initial OOD pre-trained teacher by showing an improvement of more than 1 dB in terms of SI-SDR performance.

Our experimental results validate our hypothesis that a speech enhancement model could be trained faster and more effectively under a lifelong process where the teacher would be continuously updated in parallel to the student. The speech enhancement performance obtained by the sequentially updated and the frozen teacher protocols are shown in Figure 1. We notice that all protocols perform similarly until the 20th epoch where the teacher is still static for all strategies. However, after the 20th epoch, the teacher model is replaced with the latest student while the depth of the next student is increased $8 \rightarrow 16$. As a result, the student trained with the

Training Method and Model Details		#Model Params (10^6)	Available Training Data (%)				Evaluation Metrics				
			Clean Speech \mathcal{D}_s		Clean Noise \mathcal{D}_n		Mixture \mathcal{D}_m		SISDR (dB)	PESQ	STOI
			DNS	LFSD	DNS	LFSD	DNS	LFSD			
Input Noisy Mixture		-							9.2	1.58	0.915
Unsupervised	In-domain	0.79			20%		80%		14.4	2.13	0.933
MixIT with	OOD noise	0.79				20%	100%		14.3	2.02	0.933
Student ($U = 8$)	Extra noise [9]	0.79				50%	100%		14.5	2.03	0.930
Unsupervised	Teacher ($U = 8$)	0.79				20%		80%	14.8	2.15	0.940
RemixIT	Student ($U = 32$)	0.97					100%		16.0	2.34	0.952
Semi-supervised	Teacher ($U = 8$)	0.56		100%		100%			17.6	2.61	0.958
RemixIT	Student ($U = 32$)	0.73					100%		18.0	2.60	0.959
Supervised	Student ($U = 8$)	0.56	100%		100%				18.6	2.69	0.962
In-domain	FullSubNet [2]	5.6	100%		100%				17.3	2.78	0.961

Table 1: Speech enhancement performance on the DNS test set using the proposed *RemixIT* methods, unsupervised MixIT approaches [6, 9] and supervised in-domain training with the same Sudo rm -rf model ($U = 8$) and the state-of-the-art *FullSubNet* model in the literature [2].

sequential method scales better for the same number of epochs (e.g. 40th epoch) compared to the same size student ($U = 16$) with a frozen teacher. We have experimentally seen that the sequentially updated teacher scales better than other protocols and this is the default strategy which we use across all other experiments except for the zero-shot adaptation where we also show that the running mean teacher updating scheme is also an effective option.

4.2. Self-supervised and Self-supervised speech enhancement

The speech enhancement results of the proposed method alongside supervised and unsupervised baselines are summarized in Table 1. The percentage of the available data denotes the portion of each disjoint splits from the DNS or the LFSD paired data collections. For instance, the unsupervised *RemixIT* teacher pre-training requires unsupervised MixIT using 80% of the LFSD data pairs to simulate the noisy recordings \mathcal{D}_m and the other 20% for the clean OOD noise recordings \mathcal{D}'_n , whilst the regular student training leverages the whole noisy DNS dataset.

Notice that unsupervised and semi-supervised *RemixIT* does not depend on clean in-domain noise samples. Despite that, the unsupervised student model significantly outperforms all MixIT-like approaches including the in-domain training and the recently proposed extra noise augmentation where MoMs contain 3 noise sources [9] (14.5dB \rightarrow 16.0dB in terms of SI-SDR). Moreover, the largest unsupervised student ($U = 32$) outperforms its OOD MixIT unsupervised teacher by a large margin across all evaluation metrics which shows the efficacy of *RemixIT* for self-supervised settings. The proposed method also yields noticeable gains for the semi-supervised case where the student model performs comparably with in-domain supervised training using the default Sudo rm -rf model with $U = 8$ and a recent state-of-the-art model. We want to underline that our method could be used with more complex models as teachers, rather than the efficient Sudo rm -rf architecture, and provide even higher quality speech enhancement performance.

4.3. Zero-shot domain adaptation

We show that *RemixIT* can also be used with low-resource datasets, where training data are limited but one has access to a test dataset for adapting a pre-trained model. In Figure 2, the performance improvement for the zero-shot speech enhancement task is depicted with a variety of supervised and unsupervised pre-trained networks on larger OOD datasets. Notably, *RemixIT* yields improvements of

up to 0.8dB in terms of SI-SDR compared to the uncalibrated pre-trained models while using a limited amount of in-domain mixtures. The performance of our model is correlated with the amount of available noisy mixtures and this is the reason we see the largest (smallest) gains for the WHAM! (DNS) test partition which has 3,000 (only 150) mixtures. Moreover, we also notice a large improvement in cases where there is a large distribution shift between the training data and the mixtures in the adaptation set (e.g. supervised training on WHAM! and adapting on the relatively small DNS test set).

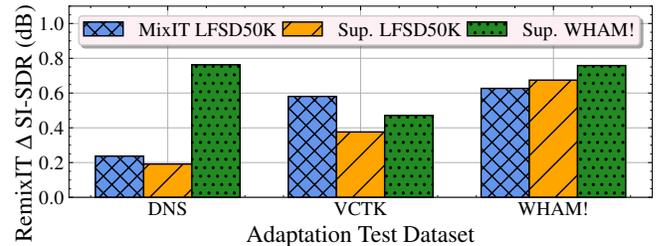


Fig. 2: Absolute gain in terms of SI-SDR when using *RemixIT* with an OOD pre-trained model and a given test set to adapt (e.g. DNS, LFSD and WHAM!, from left to right). We show our results on three different pre-trained Sudo rm -rf models with $U = 8$ and following MixIT unsupervised pre-training on LFSD (blue/leftmost), supervised training on LFSD (yellow) and supervised pre-training on WHAM! (green/rightmost).

5. CONCLUSION

We have proposed a novel continual self-training method for denoising and have experimentally showed its benefits on several realistic speech enhancement tasks. Our method depends only on the existence of in-domain noisy data and a pre-trained model using purely out-of-domain data which might not necessarily capture the in-domain distribution. The coupling of the bootstrap remixing process with the continuously bi-directional teacher-student self-training framework leads to significant improvements for zero-shot and self-supervised speech enhancement as well as semi-supervised domain adaptation. In the future, we aim to apply our method to other domains and denoising tasks as well as provide stronger theoretical guarantees for the convergence of our algorithm.

6. REFERENCES

- [1] Ashutosh Pandey and DeLiang Wang, “Dense cnn with self-attention for time-domain speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.
- [2] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement,” in *Proc. ICASSP*, 2021, pp. 6633–6637.
- [3] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, “Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss,” in *Proc. Interspeech*, 2020, pp. 2487–2491.
- [4] Yangyang Xia, Buye Xu, and Anurag Kumar, “Incorporating real-world noisy speech in neural-network-based speech enhancement systems,” *arXiv preprint arXiv:2109.05172*, 2021.
- [5] Aswin Sivaraman, Sunwoo Kim, and Minje Kim, “Personalized speech enhancement through self-supervised data augmentation and purification,” in *Proc. Interspeech*, 2021, pp. 2676–2680.
- [6] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson, and John R Hershey, “Unsupervised sound separation using mixture invariant training,” in *Proc. NeurIPS*, 2020, vol. 33, pp. 3846–3857.
- [7] Takuya Fujimura, Yuma Koizumi, Kohei Yatabe, and Ryoichi Miyazaki, “Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech,” *arXiv preprint arXiv:2101.08625*, 2021.
- [8] Efthymios Tzinis, Jonah Casebeer, Zhepei Wang, and Paris Smaragdis, “Separate but together: Unsupervised federated learning for speech enhancement from non-iid data,” in *Proc. WASPAA*, 2021, pp. 46–50.
- [9] Koichi Saito, Stefan Uhlich, Giorgio Fabbro, and Yuki Mitsufuji, “Training speech enhancement systems with noisy speech datasets,” *arXiv preprint arXiv:2105.12315*, 2021.
- [10] Ryo Aihara, Toshiyuki Hanazawa, Yohei Okato, Gordon Wichern, and Jonathan Le Roux, “Teacher-student deep clustering for low-delay single channel speech separation,” in *Proc. ICASSP*, 2019, pp. 690–694.
- [11] Jisi Zhang, Cătălin Zorilă, Rama Doddipatla, and Jon Barker, “Teacher-Student MixIT for Unsupervised and Semi-Supervised Speech Separation,” in *Proc. Interspeech*, 2021, pp. 3495–3499.
- [12] Sunwoo Kim and Minje Kim, “Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation,” in *Proc. WASPAA*, 2021, pp. 176–180.
- [13] Zhepei Wang, Ritwik Giri, Umut Isik, Jean-Marc Valin, and Arvindh Krishnaswamy, “Semi-supervised singing voice separation with noisy self-training,” in *Proc. ICASSP*, 2021, pp. 31–35.
- [14] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori, “Momentum pseudo-labeling for semi-supervised speech recognition,” in *Proc. Interspeech*, 2021, pp. 726–730.
- [15] Scott Wisdom, John R Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A Saurous, “Differentiable consistency constraints for improved deep speech enhancement,” in *Proc. ICASSP*, 2019, pp. 900–904.
- [16] Chandan KA Reddy et al., “The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in *Proc. Interspeech*, 2020, pp. 2492–2496.
- [17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [18] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, “Fsd50k: an open dataset of human-labeled sound events,” *arXiv preprint arXiv:2010.00475*, 2020.
- [19] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, “WHAM!: Extending Speech Separation to Noisy Environments,” in *Proc. Interspeech*, 2019, pp. 1368–1372.
- [20] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy, “Attention wave-u-net for speech enhancement,” in *Proc. WASPAA*, 2019, pp. 249–253.
- [21] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonal, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” 2019.
- [22] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, “The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings,” in *Proc. ICA*, 2013.
- [23] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis, “Sudo rm-rf: Efficient networks for universal audio source separation,” in *Proc. MLSP*, 2020, pp. 1–6.
- [24] Yi Luo, Cong Han, and Nima Mesgarani, “Ultra-lightweight speech separation via group communication,” in *Proc. ICASSP*, 2021, pp. 16–20.
- [25] Efthymios Tzinis, Zhepei Wang, Xilin Jiang, and Paris Smaragdis, “Compute and memory efficient universal sound source separation,” *Journal of Signal Processing Systems*, 2021.
- [26] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, “Sdr-half-baked or well done?,” in *Proc. ICASSP*, 2019, pp. 626–630.
- [27] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [28] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [29] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP*, 2001, pp. 749–752.