IEEE Open Journal of Signal Processing

Received 6 September 2023; revised 27 November 2023; accepted 27 November 2023. Date of publication 14 December 2023; date of current version 4 January 2024. The review of this article was arranged by Associate Editor S. Watanabe.

Digital Object Identifier 10.1109/OJSP.2023.3343343

Masked Spectrogram Prediction for Unsupervised Domain Adaptation in Speech Enhancement

KATERINA ZMOLIKOVA [1,2], MICHAEL SYSKIND PEDERSEN [1], AND JESPER JENSEN [1,2]

[1]Demant A/S, 2765 Smorum, Denmark
[2]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark

CORRESPONDING AUTHOR: KATERINA ZMOLIKOVA (e-mail: k.zmolikova@gmail.com).

This work was supported by Innovation Fund Denmark.

ABSTRACT Supervised learning-based speech enhancement methods often work remarkably well in acoustic situations represented in the training corpus but generalize poorly to out-of-domain situations, i.e. situations not seen during training. This stands in the way of further improvement of these methods in realistic scenarios, as collecting paired noisy-clean recordings in the target application domain is typically not feasible. Recording noisy-only in-domain data is, though, much more practical. In this article, we tackle the problem of unsupervised domain adaptation in speech enhancement. Specifically, we propose a way to use in-domain noisy-only data in the training of a neural network to improve upon a model trained solely on out-of-domain paired data. For this, we make use of masked spectrogram prediction, a technique from self-supervised learning that aims to interpolate masked regions of a spectrogram. We hypothesize that masked spectrogram prediction encourages learning of features that represent well both speech and noise components of the noisy signals. These features can then be used to train a more robust speech enhancement system. We evaluate the proposed method on the VoiceBank-DEMAND and LibriFSD50k databases, with WSJ0-CHiME3 serving as the out-of-domain database. We show that the proposed method outperforms both the out-of-domain system and the baseline approaches, i.e. RemixIT and noisy-target training, and also combines well with the previously proposed RemixIT method.

INDEX TERMS Masked spectrogram prediction, speech enhancement, unsupervised domain adaptation.

I. INTRODUCTION

Speech enhancement (SE) is nowadays dominated by methods based on neural networks (NNs), which allow for significant improvements in intelligibility and quality of the enhanced speech when compared to traditional methods [1], [2], [3], [4]. This remarkable performance has motivated numerous practical applications and even commercial products (such as use in hearing aids [5], [6]). However, in contrast to controlled laboratory experiments, the results in real-world settings are often somewhat disappointing [7], [8]. This is largely due to domain mismatch i.e. the discrepancies between the data used to train the NN and the data encountered when the NN is applied [9], [10], [11]. Unseen noises and reverberation patterns, as well as microphone mismatch, are a few examples of such discrepancies.

A straightforward way to overcome this issue would be to train the NN using data collected in the target domain for the given application (e.g. data collected from hearing aids worn by hearing-aid users in everyday situations). However, in order to train the NN for speech enhancement using traditional supervised learning techniques, pairs of noisy and corresponding clean recordings are required. Such paired data are extremely difficult to record in the real target domain. As a result, the SE-NN training methods mostly turn towards simulating data to resemble the target domain, which seldom captures all of its properties, such as realistic reverberation patterns, particular noise types, speaker turns, microphone characteristics, or acoustic surroundings of the microphones (e.g., for body-worn devices). While paired data are difficult to acquire, in-domain data of noisy-only realistic mixtures

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

(without corresponding clean ones) are much easier to get. The goal of unsupervised domain adaptation[1] is to use such in-domain noisy mixtures alongside out-of-domain (OOD) paired data to improve the system performance [12]. For instance, when developing an SE system for a product, we might collect noisy-only recordings from the actual product usage and utilize them during the training instead of using a system trained solely on publicly available synthetic paired datasets.

Self-supervised learning approaches have recently gained popularity for training neural networks without supervision [14], [15]. Self-supervised approaches employ training procedures similar to those used for supervised tasks but with targets constructed from the input data itself. Typically, self-supervised approaches have been used to pre-train a system on a large amount of unlabeled data, before fine-tuning it using supervised techniques applied to smaller labeled target domain data. The concepts of self-supervision might however also be applied to the stated unsupervised domain adaptation setting (with unlabeled in-domain and labeled OOD data). Indeed, in computer vision, self-supervised objectives have been successfully used to leverage unlabeled in-domain data in the context of unsupervised domain adaptation [16], [17].

Among the plethora of available self-supervised techniques, masked spectrogram prediction (MSP) [18], [19] is particularly well-suited for the SE task. In MSP, part of the speech spectrogram is masked (i.e. removed from the NN input) and the objective is to predict it from its surroundings. Unlike other self-supervised techniques, which often form their objectives on higher-level representations, MSP aims to estimate the input features directly. This is a good fit for SE task, which likewise requires synthesis of the signal in its original form. Furthermore, masked modeling in general has been shown to learn features that well represent the underlying latent factors in the data [20]. In our case, we hypothesize that predicting masked parts of noisy spectrograms incentivizes learning features that well represent underlying speech and noise components.

Given the above reasoning, we propose to use MSP to leverage in-domain noisy mixtures during SE-NN training, as inspired by similar approaches in image classification [16], [17]. In particular, we propose a two-stage method depicted in Fig. 1; In the first stage, both in-domain noisy-only and OOD paired data are used with the MSP objective. For the noisy-only in-domain data, the objective is to predict the masked part of the noisy mixtures. The OOD paired data allow us to additionally predict the clean components of the masked parts. The MSP task should lead to learning features well representing both domains. In the second stage, we use the first part of the network trained during the first stage as a fixed feature extractor and use the resulting features to train an SE network on OOD paired data. The final model resulting from this two-stage process should generalize well toward the target domain thanks to the features learned in the first stage.

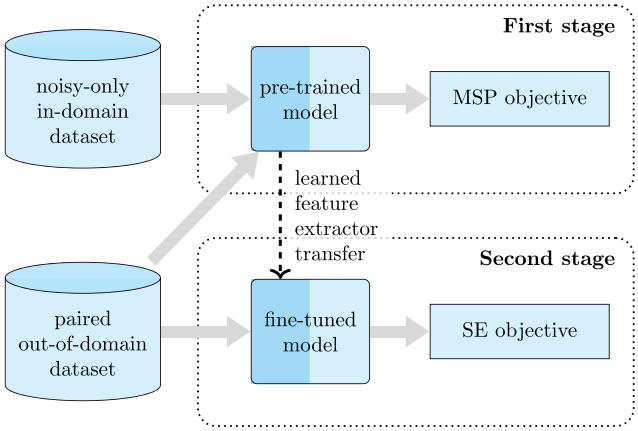

FIGURE 1. High-level scheme of the proposed two-stage method.

We test the proposed technique in two settings which differ in how much discrepancy there is between the in-domain and out-of-domain data. Fundamentally, the extent to which the two domains differ limits the possible success of the unsupervised domain adaptation. While this limit is difficult to characterize theoretically [21], we study it experimentally, by comparing the achieved improvements in enhancement performance in the two settings. We further analyze the gains of the adaptation in different noise conditions and compare them to several baselines. In all experiments, we test the method in artificially created scenarios where we abstain from using clean data from the target domain, to be able to evaluate and study the performance in controlled settings. The experimental study is designed to demonstrate the ability of the proposed method to utilize the in-domain noisy-only data, rather than claiming the best performance on the test datasets. For instance, a general model trained on various types of speech and noise signals might perform strongly in the tested scenarios. However, the potential of the proposed method is to improve performance in target realistic domains, where even general models often degrade, especially in low-complexity, low-latency use cases where the model size is limited.

The rest of the paper begins with an overview of related works in Section II, followed by a description of the problem setting in Section III. The proposed method is explained in detail in Section IV and contrasted with baseline methods in Section V. Finally, we provide our experimental results in Section VI and discuss the limitations and future directions in Section VII.

II. RELATED WORKS

Various ways for exploiting noisy-only recordings during training have been investigated in SE research. Prior to the expansion of NNs for SE, methods for domain adaptation were investigated for hidden Markov models [22] or non-negative matrix factorization methods [23]. Nowadays, some works use noisy recordings by adding extra noise to them

[1] In many works the term "domain adaptation" is used for post-hoc modification of an already trained model towards a new domain. In our work, we adopt the term "unsupervised domain adaptation" used e.g. in [12], [13] in a broader sense for any methods using in-domain unlabeled and out-of-domain labeled data.

and constructing the loss function using the original noisy mixture as the target signal [24], [25], [26]. Other works employ domain discriminators with adversarial training to create domain-robust features [27], [28], [29], [30], [31]. Adversarial loss functions were also studied on the feature level together with cycle consistency techniques [32], [33], [34] or combined with optimal transport strategies [35]. In [36], a pre-trained model estimating speech quality was utilized to provide an objective to optimize on the noisy-only data. The problem has been also tackled from the perspective of generative modeling, with the use of variational auto-encoders [9], [37]. Finally, teacher-student training schemes have been employed, in which an OOD teacher model is used to provide targets for supervised training of a student model on target data [38], [39]. Although numerous approaches are available, no systematic comparison on a common data corpus has been done in the literature. In this article, we compare our method to noisy-target training (Nytt) [24], and RemixIT [38], a recent teacher-student training scheme.

In parallel with our work, a similar masking scheme has been used in [40] for self-supervised pre-training for the task of speech enhancement. In contrast with our work, [40] does not focus on unsupervised domain adaptation but rather a self-supervised scenario, where a model pre-trained on a huge unlabeled dataset is fine-tuned towards a labeled target one. Similarly to our work, findings in [40] also point to the suitability of masked spectrogram prediction for speech enhancement, and as such, it complements well the conclusions presented here.

III. PROBLEM SETTING

A. SPEECH ENHANCEMENT TASK

The input to a speech enhancement system is an observed mixture $\mathbf{y}$ composed of speech and noise:

$$\mathbf{y} = \mathbf{s} + \mathbf{v}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^T$, $\mathbf{s} \in \mathbb{R}^T$, $\mathbf{v} \in \mathbb{R}^T$ are the observed mixture, clean speech signal and noise signal, respectively, in time-domain, and T is the number of observed samples. In our work, we model the signals in the short-time Fourier transform (STFT) domain, where — due to the linearity of the STFT — the relationship is also additive:

$$\mathbf{Y} = \mathbf{S} + \mathbf{V}, \tag{2}$$

where $\mathbf{Y} \in \mathbb{C}^{N\times F}, \mathbf{S} \in \mathbb{C}^{N\times F}, \mathbf{V} \in \mathbb{C}^{N\times F}$ are the STFT counterparts of $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{v}$, respectively, N is the number of STFT time-frames and F the number of STFT frequency bins.

The basic goal of any SE system is to estimate the clean speech $\mathbf{S}$ given the mixture $\mathbf{Y}$. Some methods additionally estimate the noise component $\mathbf{V}$[2]

$$\hat{\mathbf{S}}, \hat{\mathbf{V}} = f_\Theta(\mathbf{Y}), \tag{3}$$

where $\hat{\mathbf{S}} \in \mathbb{C}^{N\times F}$ and $\hat{\mathbf{V}} \in \mathbb{C}^{N\times F}$ are the estimated speech and noise in STFT domain, respectively, with the time-domain counterparts $\hat{\mathbf{s}} \in \mathbb{R}^T$ and $\hat{\mathbf{v}} \in \mathbb{R}^T$. Furthermore, f is an enhancement function parameterized by a set of parameters Θ. In our case, f is modeled by an NN. For later convenience, we divide the parameters Θ of the NN into two disjoint parts: encoder Θ_{enc} and decoder Θ_{dec} parameters, $\Theta = \Theta_{\text{enc}} \cup \Theta_{\text{dec}}$. The NN function can then be composed as $f_\Theta = f_{\Theta_{\text{dec}}} \circ f_{\Theta_{\text{enc}}}$, with $\circ$ denoting the function composition operator.

B. PROBLEM WITH SUPERVISED SE SYSTEMS

Typically, an NN for SE is trained using a supervised learning paradigm on a set of examples, consisting of pairs of noisy and clean speech signals. Such paired examples are difficult to collect in the target domain; it is, however, possible to use an OOD paired dataset $\mathcal{D}^{(\text{ood})}_{\text{paired}} = \{(\mathbf{Y}^{(\text{ood})}_i, \mathbf{S}^{(\text{ood})}_i)\}^{D^{(\text{ood})}}_{i=1}$, constructed for example by adding separate OOD speech and OOD noise recordings to form synthetic noisy signals. Given such dataset, the training procedure of the NN can optimize — in a supervised manner — a loss function measuring the discrepancy between the estimated and target clean speech, and, optionally, the estimated and target noise. One such commonly used loss function is scale-invariant signal-to-noise ratio (SI-SNR) [3], [41]:

$$\Theta^{(\text{ood})} = \operatorname*{argmin}_\Theta \mathcal{L}^{(\text{ood})}, \tag{4}$$

$$\mathcal{L}^{(\text{ood})} = \sum_{\mathbf{Y},\mathbf{S}\in\mathcal{D}^{(\text{ood})}_{\text{paired}}} \text{SI-SNR}(\mathbf{s}, \hat{\mathbf{s}}) + \text{SI-SNR}(\mathbf{v}, \hat{\mathbf{v}}), \tag{5}$$

where the target noise signal can be obtained as $\mathbf{v} = \mathbf{y} - \mathbf{s}$, and where $\hat{\mathbf{s}}$ and $\hat{\mathbf{v}}$ are functions of Θ, as defined in (3).

The SE system trained in this way on the OOD dataset will, however, often generalize poorly [7], [8] and thus might fail when applied in the target domain.

C. UNSUPERVISED DOMAIN ADAPTATION

The goal of unsupervised domain adaptation is to leverage in-domain dataset of noisy mixtures $\mathcal{D}^{(\text{id})}_{\text{noisy}} = \{\mathbf{Y}^{(\text{id})}_i\}^{D^{(\text{id})}}_{i=1}$ in the training procedure alongside $\mathcal{D}^{(\text{ood})}_{\text{paired}}$ to obtain a set of NN parameters $\Theta^{(\text{uda})}$, for which the NN performs better in the target domain. The in-domain dataset cannot be simply included in the conventional optimization in (4) due to the lack of clean speech references. A novel training procedure thus needs to be designed.

IV. MASKED SPECTROGRAM PREDICTION FOR UNSUPERVISED DOMAIN ADAPTATION

We propose a two-stage process for unsupervised domain adaptation, where in the first stage, all available data are used (both noisy-only in-domain and paired OOD data) to train an encoder, and in the second stage, the decoder is trained on the paired OOD data. In this section, we first describe both stages and then discuss the masking strategy and NN architecture used in MSP.

[2]In our work, we mainly introduce the estimation of the noise component for consistency with the baseline method RemixIT. However, we also found that the models estimating both components seem to be more robust to domain change — making our mismatched baseline model stronger.

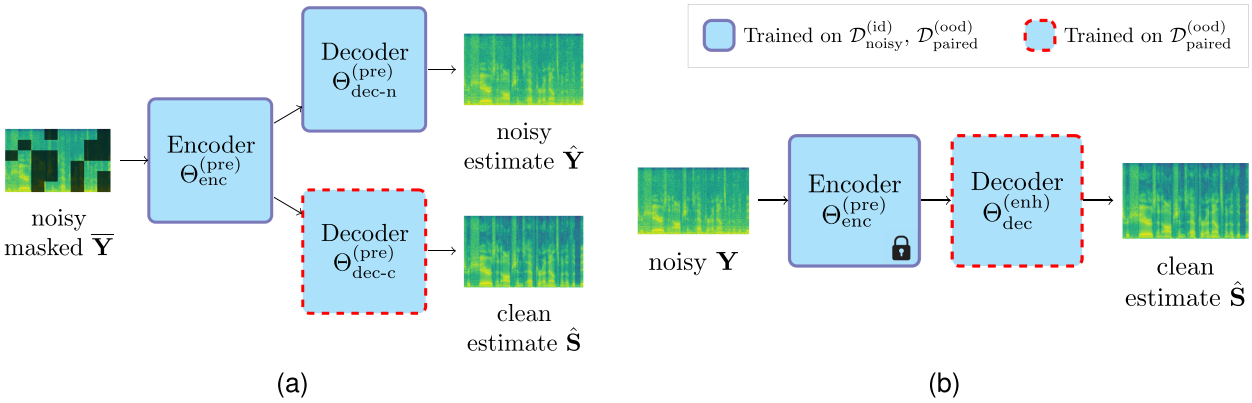

FIGURE 2. **Overall training scheme with the proposed MSP method. The left sub-figure (a) shows the first pre-training stage (described in Section IV-A) and the right sub-figure (b) shows the second supervised fine-tuning stage (described in Section IV-B). For simplicity, we omitted the noise estimation in the figure.**

A. FIRST STAGE: MASKED SPECTROGRAM PREDICTION PRE-TRAINING

With the encoder-decoder structure described in Section II-I-A, the goal of the first stage is to train an encoder Θ_{enc} to extract useful features, using all available training data $\mathcal{D}^{(\text{id})}_{\text{noisy}}$, $\mathcal{D}^{(\text{ood})}_{\text{paired}}$. For this, we make use of the MSP technique. The overall model used for the first stage is depicted in Fig. 2(a). The model consists of one encoder Θ_{enc} and two decoders $\Theta_{\text{dec-n}}$, $\Theta_{\text{dec-c}}$. The input of the encoder is a masked noisy spectrogram, $\overline{\mathbf{Y}} = \mathbf{Y} \odot \mathbf{M}$, created using a binary mask $\mathbf{M} \in \{0, 1\}^{N \times F}$, with $\odot$ denoting element-wise multiplication. The decoder $\Theta_{\text{dec-n}}$ is used to recover the original noisy spectrogram $\mathbf{Y}$, while the decoder $\Theta_{\text{dec-c}}$ estimates the clean spectrogram $\mathbf{S}$:

$$\hat{\mathbf{Y}} = f_{\Theta_{\text{dec-n}}} \circ f_{\Theta_{\text{enc}}}(\overline{\mathbf{Y}}), \tag{6}$$

$$\hat{\mathbf{S}} = f_{\Theta_{\text{dec-c}}} \circ f_{\Theta_{\text{enc}}}(\overline{\mathbf{Y}}). \tag{7}$$

Here, the parameters $\Theta_{\text{dec-n}}$ can be learned on all available data $\mathcal{D}^{(\text{id})}_{\text{noisy}}$, $\mathcal{D}^{(\text{ood})}_{\text{paired}}$ by a loss function measuring discrepancy between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. The parameters $\Theta_{\text{dec-c}}$ can be trained only on a paired (hence OOD) dataset $\mathcal{D}^{(\text{ood})}_{\text{paired}}$ using a loss function measuring discrepancy between $\hat{\mathbf{S}}$ and $\mathbf{S}$. Finally, the encoder parameters Θ_{enc} can be learned jointly with both decoders, thus using all available data $\mathcal{D}^{(\text{id})}_{\text{noisy}}$, $\mathcal{D}^{(\text{ood})}_{\text{paired}}$.

We found that with MSP, it is especially difficult to predict the phase information correctly, which makes the SI-SNR loss as in (5) unsuitable.[3] For MSP, we thus opt for a loss function where the magnitude and phase estimation are decoupled. More specifically, for ground-truth STFT-domain values $\mathbf{X}$ and estimated STFT-domain values $\hat{\mathbf{X}}$, we propose to use:

$$\mathcal{L}^{(\text{mag})}(\mathbf{X}, \hat{\mathbf{X}}) = \log \sum_{n,f} (|\mathbf{X}_{n,f}| - |\hat{\mathbf{X}}_{n,f}|)^2, \tag{8}$$

$$\mathcal{L}^{(\text{phase})}(\mathbf{X}, \hat{\mathbf{X}}) = \log \sum_{n,f} |\mathbf{X}_{n,f}|^2 \left| \frac{\mathbf{X}_{n,f}}{|\mathbf{X}_{n,f}|} - \frac{\hat{\mathbf{X}}_{n,f}}{|\hat{\mathbf{X}}_{n,f}|} \right|^2, \tag{9}$$

$$\mathcal{L}^{(\text{magphase})}(\mathbf{X}, \hat{\mathbf{X}}) = \mathcal{L}^{(\text{mag})}(\mathbf{X}, \hat{\mathbf{X}}) + \lambda \mathcal{L}^{(\text{phase})}(\mathbf{X}, \hat{\mathbf{X}}), \tag{10}$$

where λ is a weighting factor. Overall, the MSP loss function for the noisy estimate is then:

$$\mathcal{L}^{(\text{msp-n})} = \sum_{\mathbf{Y} \in \mathcal{D}^{(\text{id})}_{\text{noisy}} \cup \mathcal{D}^{(\text{ood})}_{\text{noisy}}} \mathcal{L}^{(\text{magphase})}(\mathbf{Y}, \hat{\mathbf{Y}}), \tag{11}$$

where $\mathcal{D}^{(\text{ood})}_{\text{noisy}} = \{\mathbf{Y}^{(\text{ood})}_i\}_{i=1}^{D^{(\text{ood})}}$ consists of the noisy signals only from $\mathcal{D}^{(\text{ood})}_{\text{paired}}$. Analogously, the MSP loss function for the clean speech estimate is:

$$\mathcal{L}^{(\text{msp-c})} = \sum_{\mathbf{Y}, \mathbf{S} \in \mathcal{D}^{(\text{ood})}_{\text{paired}}} \mathcal{L}^{(\text{magphase})}(\mathbf{S}, \hat{\mathbf{S}}). \tag{12}$$

The first stage thus overall solves the following optimization problem:

$$\Theta^{(\text{pre})} = \underset{\Theta}{\text{argmin}}\, \mathcal{L}^{(\text{msp-n})} + \mathcal{L}^{(\text{msp-c})}, \tag{13}$$

where $\Theta^{(\text{pre})} = \Theta^{(\text{pre})}_{\text{enc}} \cup \Theta^{(\text{pre})}_{\text{dec-n}} \cup \Theta^{(\text{pre})}_{\text{dec-c}}$.

B. SECOND STAGE: SUPERVISED FINE-TUNING

The second stage is depicted in Fig. 2(b) and follows the conventional supervised SE training paradigm as described in Section III-B. That is, we employ only the paired data $\mathcal{D}^{(\text{ood})}_{\text{paired}}$ with the loss function $\mathcal{L}^{(\text{ood})}$ (5). The crucial difference is that in the second stage, we initialize the encoder Θ_{enc} to parameters learned in the first stage $\Theta^{(\text{pre})}_{\text{enc}}$ and keep them fixed. Only decoder parameters Θ_{dec} are thus trained

[3] In particular, when using SI-SNR loss, uncertainty in the phase impacts also prediction of the magnitude. As predicting phase information from masked patches is difficult, using SI-SNR loss for MSP hinders the overall learning.

at this stage:

$$\Theta_{\text{dec}}^{(\text{enh})} = \underset{\Theta_{\text{dec}}}{\text{argmin}}\, \mathcal{L}^{(\text{ood})}. \tag{14}$$

C. MASKING STRATEGY AND NN ARCHITECTURE

For MSP, a proper masking strategy is of great importance as it can change the difficulty of the MSP task and the nature of the features that are learned. In our work, we follow the previous literature on MSP and masked image modeling in general and mask rectangular portions (patches) of the spectrogram [18], [19], [42]. The spectrogram is thus divided into a grid of regularly sized patches of size $T_p \times F_p$ and each patch is masked with a probability p_{mask}. An example of a masked spectrogram is given in the left part of Fig. 2(a).

Previous works on MSP and masked image modeling used the Vision Transformer (ViT) architecture [43]. As this architecture directly works on the level of patches, it is very well-suited for masked pre-training, where masked patches can be simply omitted from the input of the attention layers of the ViT. However, in our preliminary experiments, we found that the ViT architecture is difficult to apply to the given speech enhancement task with satisfying results. We suspect that the successful application of ViT necessitates large datasets of millions of recordings as used in previous works [18], [19], [42]. We instead based our model on an off-the-shelf TF-GridNet architecture, which has proven to be efficient in speech separation and enhancement tasks [44].

Both the encoder and decoder of our architecture thus consist of TF-GridNet blocks. Each TF-GridNet block consists of sub-band temporal and intra-frame spectral modules, both implemented by BLSTMs. The blocks operate on embeddings of each time-frequency point. Overall, the input of the encoder is a spectrogram with masked patches replaced by zeros. At the input of the decoder, the embeddings corresponding to masked time-frequency points are replaced with a learnable masking token, as done in previous works [18], [42]. Note that the usage of the TF-GridNet instead of ViT architecture introduces a zero-mismatch problem, i.e., during pre-training, the encoder processes zeros at masked portions of the spectrogram, which are not present during test time. Although this might have a negative effect, we found the performance to exceed that of ViT for our task. While in other settings, ViT could lead to further improvements, we find it encouraging that the MSP strategy can work also with an off-the-shelf SE architecture.

V. BASELINES AND EXTENSIONS

A. NOISY TARGET TRAINING BASELINE

Nytt [24] is a training strategy for SE that does not require clean speech signals. Instead, it uses a database of noisy speech together with a database of noise-only signals. It trains the NN using the noisy signals $\mathbf{Y} \in \mathcal{D}_{\text{noisy}}^{(\text{id})}$ as targets, where the input signal is created by adding additional noise to the same noisy signal $\mathbf{Y}' = \mathbf{Y} + \mathbf{V}^{(\text{extra})}$.

As in our work, we assume no access to in-domain noises, we compare with Nytt in the same setting, where the extra noise signals $\mathbf{V}^{(\text{extra})}$ are obtained from the out-of-domain database. In this case, if the type of noise in the original noisy signal $\mathbf{Y}$ and the type of extra noise $\mathbf{V}^{(\text{extra})}$ differ significantly, Nytt might learn to remove only the extra noise and thus fail in denoising the in-domain recordings [24]. To add more context, we also provide results of Nytt in an extended setting with in-domain noises used in the training.

B. REMIXIT BASELINE

RemixIT [38] is a teacher-student training scheme, i.e. we assume having a teacher model $f_{\Theta}^{(\text{te})}$ and we use it to train a student model $f_{\Theta}^{(\text{st})}$. The student model $f_{\Theta}^{(\text{st})}$ is trained on in-domain noisy recordings $\mathcal{D}_{\text{noisy}}^{(\text{id})}$ while using targets provided by the teacher in place of ground-truth clean recordings. In particular, the teacher first processes a batch of noisy in-domain signals $\mathbf{Y}_i \in \mathcal{D}_{\text{noisy}}^{(\text{id})}$,

$$[(\hat{\mathbf{S}}_i^{(\text{te})}, \hat{\mathbf{V}}_i^{(\text{te})})]_{i=1..N} = [f_{\Theta}^{(\text{te})}(\mathbf{Y}_i)]_{i=1..N}. \tag{15}$$

The noise estimates $\hat{\mathbf{V}}_i^{(\text{te})}$ are then permuted and mixed with the clean estimates $\hat{\mathbf{S}}_i^{(\text{te})}$ to create new synthetic noisy signals $\mathbf{Y}_i^{(\text{te})} = \hat{\mathbf{S}}_i^{(\text{te})} + \hat{\mathbf{V}}_{P(i)}^{(\text{te})}$, where P is a permutation function. These synthetic noisy signals are then used to train the student model according to

$$\Theta^{(\text{remixit})} = \underset{\Theta}{\text{argmin}}\, \mathcal{L}^{(\text{remixit})}, \tag{16}$$

where

$$\mathcal{L}^{(\text{remixit})} = \sum_{\mathbf{Y} \in \mathcal{D}_{\text{noisy}}^{(\text{id})}} \text{SI-SNR}(\hat{\mathbf{s}}^{(\text{st})}, \hat{\mathbf{s}}^{(\text{te})}) + \text{SI-SNR}(\hat{\mathbf{v}}^{(\text{st})}, \hat{\mathbf{v}}^{(\text{te})}), \tag{17}$$

and

$$\hat{\mathbf{S}}^{(\text{st})}, \hat{\mathbf{V}}^{(\text{st})} = f_{\Theta}^{(\text{st})}(\mathbf{Y}^{(\text{te})}). \tag{18}$$

The RemixIT work [38] introduced several setups of the scheme with different types of data available. Here, we compare with the setup closest to ours, that is when the teacher is a mismatched supervised model (trained on OOD data).

C. MSP + REMIXIT EXTENSION

MSP and RemixIT have some complementary characteristics. RemixIT has the advantage that the final stage of training is performed directly on the target in-domain data, while in MSP the final stage of training employs the OOD paired dataset. However, since RemixIT starts with an out-of-domain teacher, the initial errors of the teacher might get propagated to the student model during the training. MSP, on the other hand, initiates the training by learning good representations on the target in-domain data, it thus does not suffer from similar error propagation. Due to this complementarity, combining both MSP and RemixIT might be beneficial.

We combine the methods by using the MSP model as the initial teacher for RemixIT. The training thus starts with the

TABLE 1. Description of the Used Datasets

	WSJ0-CHiME3 [10]	VoiceBank-DEMAND [45]	LibriFSD50k [46]
clean speech source	WSJ0 [47]	VoiceBank [48]	LibriSpeech [49]
noise source	CHiME3 [50]	DEMAND [51]	FSD50k [52]
SNRs	0-20 dB	0-20 dB	-20-30 dB*
noise types	street, transport, cafes, pedestrian	domestic, office, street, transport	sound events (music, sound of things, human sounds, natural sounds)
number of recordings in train / development / test	12776 / 1206 / 651	10802 / 770 / 824	45608 / 3044 / 3196
duration [s] of recordings (mean ± std)	7.0 ± 2.8	2.9 ± 1.1	4.0 ± 0.0

* There is no hard limit on signal-to-noise ratios (SNRs) in LibriFSD50k database. However, 95% of the recordings fall into the -20-30 dB range.

TABLE 2. Two Used Experimental Settings

Exp. setting	$\mathcal{D}^{(\text{ood})}_{\text{paired}}$	$\mathcal{D}^{(\text{id})}_{\text{noisy}}$
WSJCH → VBD	WSJ0-CHiME3	VoiceBank-DEMAND
WSJCH → LFSD	WSJ0-CHiME3	LibriFSD50k

two stages described in Section IV and finishes with RemixIT training on the target in-domain data.

VI. EXPERIMENTS

A. DATASETS

To study the proposed technique, we employ three different datasets: WSJ0-CHiME3 [10] as $\mathcal{D}^{(\text{ood})}_{\text{paired}}$, VoiceBank-DEMAND [45] as $\mathcal{D}^{(\text{id})}_{\text{noisy}}$ and LibriFSD50k [46] as $\mathcal{D}^{(\text{id})}_{\text{noisy}}$. The properties of the datasets are summarized in Table 1. Note that in the case of VoiceBank-DEMAND and LibriFSD50k, we use the clean speech and noise-only recordings only for evaluation, or when explicitly stated.

Overall, we have two experimental settings summarized in Table 2. Both are using WSJ0-CHiME3 as the out-of-domain database and VoiceBank-DEMAND or LibriFSD50 k as the target domains. In the first setting (WSJCH → VBD), the two datasets are more similar to each other, however still with some mismatch in the recording conditions and types of noises used. The second setting (WSJCH → LFSD) is used to test whether the presented methods are beneficial even in a scenario where the two databases are significantly different in terms of both signal-to-noise ratios and types of noise (see Table 1 for details).

B. CONFIGURATION

1) NN ARCHITECTURE

We use a scaled-down version of TF-GridNet [44] as neural network architecture for all experiments. The scaling down involves the removal of the cross-frame self-attention module and overall smaller size, as described below.[4] Such scale-down was necessary to enable us to run experiments on available hardware, i.e. each experiment was run on a single GPU with 11G memory. The notation below (B, D, H, I, J) refers to the original paper [44]. The scaled-down TF-GridNet consists of $B = 4$ blocks (in MSP experiments, two for the

[4] We compared the performance of the scaled-down TF-GridNet with the widely used ConvTasNet [3] with 5 M parameters. In matched settings, the performance of both is comparable, but TF-GridNet generalizes better in the case of different training and test databases (our baseline *Mismatched* scenario).

TABLE 3. Results on VoiceBank-DEMAND in WSJCH → VBD Setting. All Baseline Methods are Our Re-Implementations. Bolded Numbers Denote the Best Achieved Performance Among Methods Not Using Any In-Domain Clean or Noise Data for Training (With Significance Tested Using Paired T-Test With P-Value < 0.05). Results are Further Discussed in Section VI-D

	Section	SI-SNR [dB]	PESQ	eSTOI
Unprocessed	-	8.4	1.97	0.79
Mismatched	III	12.6	2.33	**0.85**
Nytt (OOD noise)	V-A	9.4	2.31	0.82
RemixIT	V-B	12.3	**2.37**	0.84
MSP	IV	**15.4**	2.31	0.84
MSP + RemixIT	V-C	15.1	**2.37**	0.84
Nytt (ID noise)	V-A	15.9	2.26	0.81
Matched	-	18.9	2.55	0.85

* Note that other works in the literature report PESQ 2.45 [36] or PESQ 2.41 and SI-SNR 17.3 [9] in a setting similar to WSJCH → VBD. As these results were obtained with different NN architectures and training configurations, a direct comparison with our method is currently difficult.

TABLE 4. Results on LibriFSD50 k in WSJCH → LFSD Setting. All Baseline Methods are Our Re-Implementations. Bolded Numbers Denote the Best Achieved Performance Among Methods Not Using Any In-Domain Clean or Noise Data for Training (Significance Tested Using Paired T-Test With p-Value < 0.05). Results are Further Discussed in Section VI-D

	Section	SI-SNR [dB]	PESQ	eSTOI
Unprocessed	-	6.5	1.90	0.75
Mismatched	III	10.5	2.12	0.80
Nytt (OOD noise)	V-A	6.2	1.85	0.75
RemixIT	V-B	10.8	2.15	0.80
MSP	IV	11.2	2.12	0.80
MSP + RemixIT	V-C	**11.9**	**2.24**	**0.82**
Nytt (ID noise)	V-A	9.5	2.03	0.77
Matched	-	18.4	2.84	0.86

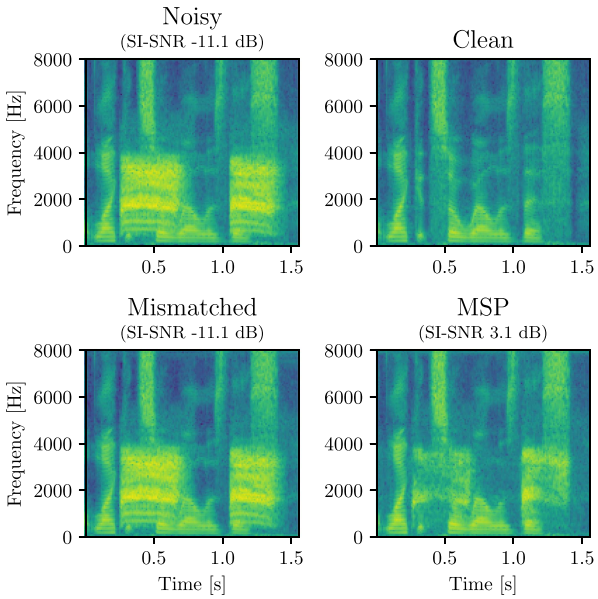

FIGURE 3. Example from LibriFSD50k dataset involving animal sound (bird croaking).

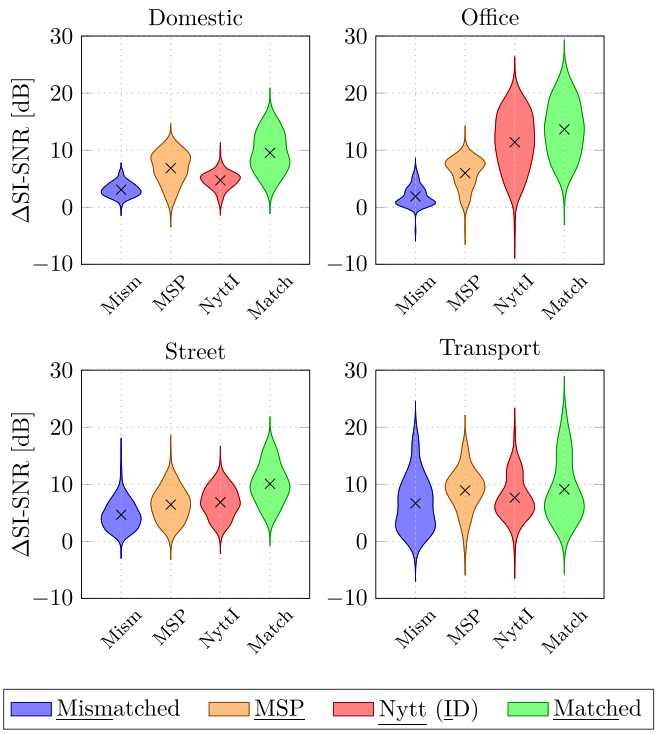

FIGURE 4. Results on VoiceBank-DEMAND in WSJCH → VBD setting, broken down into different types of noises. ΔSI-SNR denotes the improvement in SI-SNR over the observed noisy recordings. Results are further discussed in Section VI-E.

encoder and two for the decoder). The embedding dimension is $D = 16$, the number of units in BLSTMs is $H = 16$, and the kernel and stride size for Unfold and Deconv1D is $I = 4$, $J = 1$, respectively. The resulting model has 101 K trainable parameters. We used the original implementation.[5]

For all experiments except Nytt, we employ mixture consistency layer [38], [53] for ensuring that the speech and noise output sum to the original noisy signal. For Nytt, we observed significant degradation with mixture consistency, we thus refrained from using it.

2) OPTIMIZATION

We use the Adam optimizer [54] with a learning rate of 0.001 and batch size of 12 for all experiments. We train the models for 100 epochs. The learning rate is reduced by half after 5 epochs without improvement on the validation set.

3) INPUT/OUTPUT

The recordings are transformed to STFT domain using Hann windows of 400 samples (25 ms) with 160 sample shift (10 ms). The inputs and outputs of the neural network are concatenated real and imaginary components of the STFT-domain signal. The standard deviation of the inputs is further globally normalized for each dataset (with statistics estimated from training data) and de-normalized before computing the loss function. During training, we use random segments of 128 frames as inputs in the case of VoiceBank-DEMAND and 256 frames for LibriFSD50 k and WSJ0-CHiME3. The spectrogram masking uses patches of $T_p = 32$ time-frames and $F_p = 32$ frequency bins with probability of a patch removed (replaced by zeros) $p_{\text{mask}} = 0.6$.

[5] Available in ESPnet github.com/espnet

4) BASELINES

For RemixIT, we update the teacher every 30 epochs with the current student model. In Nytt, we mix the additional noise with SNR uniformly sampled from -5–5 dB, following the original work [24].

C. EVALUATION METRICS

We evaluate the performance of the proposed enhancement models and baselines using SI-SNR [41] and Perceptual Evaluation of Speech Quality (PESQ) [55] for estimated speech quality and extended Short-Time Objective Intelligibility (eSTOI) [56] for estimated speech intelligibility.

D. RESULTS: COMPARISON WITH BASELINES

In this section, we compare the overall results of the proposed method with several baselines.

First, we present the results for the WSJCH → VBD setting in Table 3. The *Mismatched* and *Matched* systems are fully supervised systems trained on WSJ0-CHiME3 and VoiceBank-DEMAND, respectively, and represent our baseline and topline. We compare the proposed methods (*MSP*, *MSP + RemixIT*) with further baselines introduced in Section V, namely *Nytt (OOD)* and *RemixIT*. Additionally, we show results of the *Nytt (ID)* system, which, in contrast with the other systems, has the advantage of using an in-domain noise database during training.

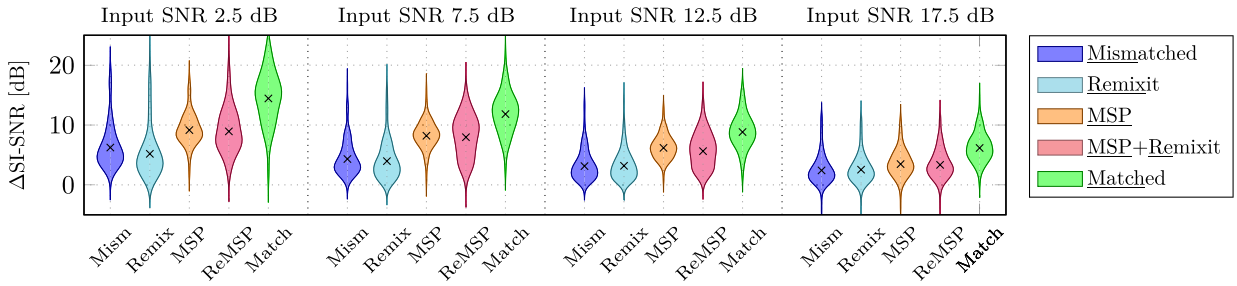

FIGURE 5. Results on VoiceBank-DEMAND in WSJCH → VBD setting, broken down into different input SNRs. ΔSI-SNR denotes the improvement in SI-SNR over the observed noisy recordings. Results are further discussed in Section VI-E.

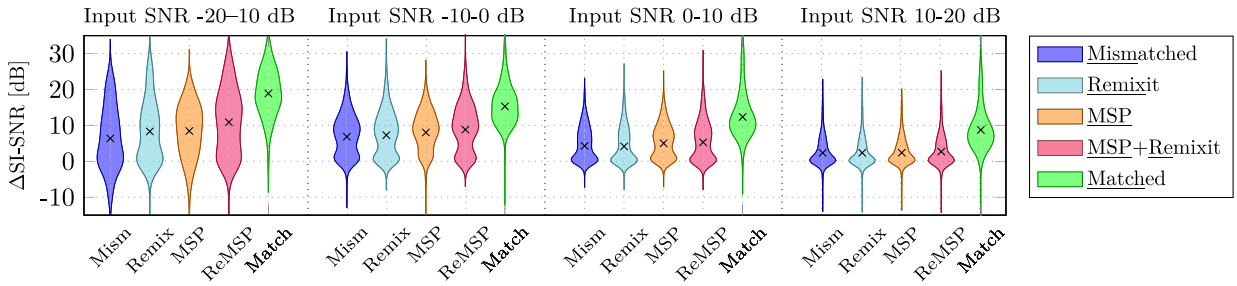

FIGURE 6. Results on LibriFSD50k in WSJCH → LFSD setting, broken down into different input SNRs. ΔSI-SNR denotes the improvement in SI-SNR over the observed noisy recordings. Results are further discussed in Section VI-E.

In terms of the estimated speech quality (SI-SNR, PESQ), the best results are achieved with a combination of *MSP* + *RemixIT*, with the SI-SNR metric being more improved by *MSP*, while PESQ by *RemixIT* method. The intelligibility results in terms of eSTOI generally do not show substantial differences, even between *Mismatched* and *Matched* systems. The *Nytt* method performs well when in-domain (DEMAND) noises are used during the training but stays far behind other methods when using OOD noise (CHiME3).

Analogous results for the WSJCH → LFSD setting are presented in Table 4. Several differences from previous results (in Table 3) can be observed. First, the overall improvements with all methods are smaller, with a bigger gap to the *Matched* results. This is caused by a larger difference between the in-domain and out-of-domain databases, as described in Section VI-A. Second, RemixIT compares better with the other methods in this condition, with the combination of *MSP* + *RemixIT* performing overall best. *Nytt* here does not lead to big improvements. We hypothesize that this is due to the database containing lower SNRs than VoiceBank-DEMAND, causing the targets used in the training to be more noisy and further from the optimal clean targets.

Fig. 3 shows an example from LibriFSD50k where the *MSP* adaptation significantly helps to remove noise. This example contains an animal noise (bird croaking). The *Mismatched* system trained on WSJ0-CHiME3 fails to remove this noise as it did not encounter a similar type of noise during training. In contrast, the *MSP* system encountered this type of noise in the pre-training stage and, as a consequence, manages to reduce the noise.

E. RESULTS: FURTHER ANALYSIS

The VoiceBank-DEMAND database contains conditions with different classes of noise (domestic, office, street, transport). For some of these, similar (but not identical) classes of noise are also present in the OOD CHiME3 data (transport, street), while some are not present (domestic, office). Fig. 4 shows the achieved improvements separately for the different noise conditions. Comparing *Mismatched*, *MSP* and *Matched* results, the advantage of *MSP* is more significant for domestic and office conditions, which shows that the method adapted to these conditions although they were unseen in the paired data. An interesting trend is also revealed for *Nytt (ID)*, where we can observe that it works greatly in some conditions (e.g. office) while in others it is outperformed by *MSP* (e.g. domestic or transport).

Fig. 5 shows the breakdown of the results on VoiceBank-DEMAND into different input SNR conditions. The improvements brought by MSP are fairly uniformly distributed for different input SNRs. Note that the WSJ-CHiME3 database contains a similar SNR range as the target VoiceBank-DEMAND; this setup thus does not introduce a mismatch in the levels of noise. For LibriFSD50k, a similar breakdown of the results is shown in Fig. 6. Here, modest improvements with *MSP+RemixIT* can be observed mostly on lower SNRs, which are mismatched with the SNRs contained in the in-domain WSJ0-CHiME3 database.

VII. LIMITATIONS AND FUTURE DIRECTIONS

In this article, we introduced MSP as a way to utilize in-domain noisy-only recordings to improve the performance of deep learning-based speech enhancement systems. The results show the potential of the proposed approach in comparison with previously proposed methods. This suggests that the proposed approach could help bridge a gap between the often-reported impressive performance of SE systems on artificial data and the more modest performance achieved in realistic settings. In the experiments presented above, we tested the method in artificially created scenarios where we abstained from using clean data from the target domain. This enabled us to develop the method and evaluate it thoroughly in a controlled scenario. To fully establish the usefulness of the method in real scenarios, future work includes experiments using realistic, more diverse datasets [57]. Furthermore, the current version of the method assumes the prior availability of in-domain noisy-only data at training time. While this is a valid assumption in many scenarios, it would be of even wider practical use if the system could be adapted to the target domain dynamically during test time. In the field of computer vision, similar methods [17] have been shown to work well for test-time adaptation as well, and we hope to extend our work in this direction in the future. Another interesting direction for extending the work is pre-training on unsegmented recordings, containing noise-only portions or multiple speakers. While the proposed method readily allows for such data, experimental validation is needed to demonstrate this potential.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[3] 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, Aug. 2019.

[4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[5] A. H. Andersen et al., "Creating clarity in noisy environments by using deep learning in hearing aids," *Seminars Hear.*, vol. 42, pp. 260–281, 2021.

[6] I. Fedorov et al., "TinyLSTMs: Efficient neural speech enhancement for hearing aids," in *Proc. Interspeech*, 2020, pp. 4054–4058.

[7] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2489–2499, 2020.

[8] C. Subakan, M. Ravanelli, S. Cornell, and F. Grondin, "REAL-M: Towards speech separation on real mixtures," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6862–6866.

[9] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2993–3007, 2022.

[10] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[11] P. Gonzalez, T. S. Alstrøm, and T. May, "Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3390–3403, 2023.

[12] X. Liu et al., "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 1, p. e25, 2022. [Online]. Available: https://www.nowpublishers.com/article/Details/SIP-2022-0019

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[14] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[15] A. Mohamed et al., "Self-supervised speech representation learning: A review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[16] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019, *arXiv:1909.11825*.

[17] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9229–9248.

[18] P.-Y. Huang et al., "Masked autoencoders that listen," in *Advances in NeurIPS*, vol. 35, Red Hook, NY, USA: Curran Associates, Inc, LA, New Orleans, USA, 2022, pp. 28708–28720.

[19] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," in *Proc. HEAR Holistic Eval. Audio Representations*2022, pp. 1–24.

[20] J. D. Lee, Q. Lei, N. Saunshi, and J. ZHUO, "Predicting what you already know helps: Provable self-supervised learning," in *Advances in NeurIPS*. vol. 34, Red Hook, NY, USA:Curran Associates, Inc., 2021, pp. 309–323.

[21] A. Mehra, B. Kailkhura, P.-Y. Chen, and J. Hamm, "Understanding the limits of unsupervised domain adaptation via data poisoning," in *Advances in NeurIPS*. vol. 34, Red Hook, NY, USA: Curran Associates, Inc, 2021, pp. 17347–17359.

[22] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[23] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[24] 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. 29th Eur. Signal Process. Conf. EUSIPCO*, 2021, pp. 436–440.

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

[26] V. A. Trinh and S. Braun, "Unsupervised speech enhancement with speech recognition embedding and disentanglement losses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 391–395.

[27] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. Interspeech*, 2019, pp. 3148–3152.

[28] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 667–672.

[29] J. Cheng, R. Liang, Z. Liang, L. Zhao, C. Huang, and B. Schuller, "A deep adaptation network for speech enhancement: Combining a relativistic discriminator with multi-kernel maximum mean discrepancy," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 41–53, 2021.

[30] N. Hou, C. Xu, E. S. Chng, and H. Li, "Learning disentangled feature representations for speech enhancement via adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 666–670.

[31] L. Frenkel, J. Goldberger, and S. E. Chazan, "Domain adaptation for speech enhancement in a large domain gap," in *Proc. Interspeech*, 2023, pp. 2458–2462.

[32] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Cycle-consistent speech enhancement," in *Proc. Interspeech*, 2018, pp. 1165–1169.
[33] W.-Y. Ting, S.-S. Wang, H.-L. Chang, B. Su, and Y. Tsao, "Speech enhancement based on cyclegan with noise-informed training," in *Proc. IEEE 13th Int. Symp. Chin. Spoken Lang. Process.*, 2022, pp. 155–159.
[34] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1826–1838, 2020.
[35] H.-Y. Lin, H.-H. Tseng, X. Lu, and Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*2021, pp. 19935–19946.
[36] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*., 2022, pp. 7412–7416.
[37] X. Lin, S. Leglaive, L. Girin, and X. Alameda-Pineda, "Unsupervised speech enhancement with deep dynamical generative speech and noise models," in *Proc. Interspeech*, 2023, pp. 5102–5106.
[38] 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. Topics Signal Process.*, vol. 16, no. 6, pp. 1329–1341, Oct. 2022.
[39] Y. Xia, B. Xu, and A. Kumar, "Incorporating real-world noisy speech in neural-network-based speech enhancement systems," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2021, pp. 564–570.
[40] Z. Zhong et al., "Extending audio masked autoencoders toward audio restoration," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/WASPAA58266.2023.10248171
[41] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?in," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.
[42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy
[44] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.
[45] C. V. Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 159–165.
[46] E. Tzinis, J. Casebeer, Z. Wang, and P. Smaragdis, "Separate but together: Unsupervised federated learning for speech enhancement from non-iid data," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 46–50.
[47] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Speech Natural Lang.: Proc. Workshop, New York, USA*, 1992, 1992, pp. 357–362.
[48] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly With Conf. Asian Spoken Lang. Res. Eval.*2013, pp. 1–4.
[49] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[50] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
[51] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoustical Soc. Amer.*, vol. 133, no. 5_Supplement, pp. 3591–3591, 2013.
[52] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50k: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.
[53] S. Wisdom et al., "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*., 2019, pp. 900–904.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[55] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.
[56] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[57] S. Leglaive et al., "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," *Proc. 7th Int. Workshop Speech Process. Everyday Environments (CHiME)*, 2023, pp. 7–12, doi: 10.21437/CHiME.2023-2.