INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Domain Adaptation for Speech Enhancement in a Large Domain Gap

*Lior Frenkel*[1,2], *Jacob Goldberger*[2] *and Shlomo E. Chazan*[1,2]

[1]OriginAI, Israel
[2]Faculty of Engineering, Bar-Ilan University, Israel
{liorf,shlomi}@originai.co, jacob.goldberger@biu.ac.il

## Abstract

Speech enhancement approaches based on neural networks, aim to learn a noisy-to-clean transformation using a supervised learning paradigm. However, networks trained in this way may not be effective at handling languages and types of noise that were not present in the training data. To address this issue, this study focuses on unsupervised domain adaptation, specifically for large-domain-gap cases. In this setup, we have noisy speech data from the new domain but the corresponding clean speech data are not available. We propose an adaptation method that is based on domain-adversarial training followed by iterative self-training where the quality of the estimated speech used as pseudo labels is monitored by the performance of the adapted network on labeled data from the source domain. Experimental results show that our method effectively mitigates the domain mismatch between training and test sets, and surpasses the current baseline.

**Index Terms**: speech enhancement, domain shift, domain adaptation, self-training

## 1. Introduction

Speech enhancement (SE) aims to improve the quality of speech signals in various noisy environments. Its goal is to remove or reduce the background noise and to enhance the speech signal for better intelligibility and an improved listening experience [1, 2]. It is a key element for immersive audio experiences in telecommunication as well as a crucial front-end processor for robust speech recognition, assistive hearing, and robust speaker recognition.

Classical approaches to speech enhancement rely on certain assumptions about the statistical properties of the signals being analyzed (see e.g. [3]). The goal is to use mathematical criteria to estimate the original speech signal that is being obscured by noise or other interference. In contrast, more recent approaches based on deep learning are moving away from this traditional modeling approach and instead embracing a data-driven approach. A plethora of deep-learning-based approaches have been presented [4]. The main idea is to train a DNN in a supervised manner to enhance the noisy input and estimate the clean speaker. Training is usually carried out on a synthetically built dataset constructed from clean and noise signals. These DNN-based approaches, which can be divided into frequency-domain [4] and time-domain methods [5, 6, 7], outperform classic model-based approaches.

A speech enhancement network can be exploited to obtain speech signals in the context of other languages, speakers, recording environments and noise types. Introducing unfamiliar scenarios to a well-trained SE system can cause severe performance degradation. This is caused by a mismatch between the speech characteristics used to train the network (source domain) and the speech characteristics it encounters in the target domain, which is commonly referred to as the domain shift problem. Collecting enough annotated data for each new domain is not always possible and training from scratch a separate SE system for each noisy speech type is impractical. In an Unsupervised Domain Adaptation (UDA) setup we assume the availability of noisy speech from the target domain but without the corresponding clean speech. Despite variations in languages, speakers, genders, and environments, human languages possess similar acoustic structures. Hence, it is still possible to adapt a well-trained SE network to different settings in an unsupervised manner. In recent years, several studies have addressed the UDA problem in speech enhancement. Liao et al. [8] applied Domain-Adversarial Neural Network (DANN) [9] for speech enhancement in a domain shift scenario. They assumed that the noise type of all the samples of both source and target domains is known and the adversarial classifier's goal is to predict the noise type (rather than the domain-identify as was done in [9]). Mixture Invariant Training (MixIT) is a self-supervised approach that enables unsupervised domain adaptation without the need for ground-truth source waveforms [10]. Although MixIT has been successfully used for various SE tasks, it requires access to the in-domain noise. To address this issue, Tzinis et al. proposed RemixIT [11], which adapts a teacher-student training framework to achieve state-of-the-art performance on various SE tasks. The flexibility of the framework allows for the use of any SE model as the teacher model. Although RemixIT scored high on the DNS-Challenge [12] test set, their approach depends on obtaining predicted speech with reasonable quality from the target domain noisy samples. Therefore, in cases where the source and target domains are far apart, this method tends to fail due to the low-quality pseudo-labels provided to the student. This situation might occur if the speech data in the two domains are in different languages with distinct phoneme sets for instance. In the case of classification tasks, examples can be selected based on the confidence of the teacher network in the prediction [13]. However, SE networks do not provide any indication as to the quality of the predicted speech.

Here we propose a two-stage UDA algorithm designed to handle large-domain-gap scenarios. First, a domain-adversarial model is applied to bring the two domains closer in the feature space. Then, a self-training framework uses this pre-trained model as initialization for a teacher network that produces qualitative pseudo-labels on the target domain for the student, even for distant domains. We also combine supervised training on source samples, in decreasing proportions during the self-training. Training to predict clean speech from the source domain monitors the quality of the predicted speech in the target domain where clean speech is not available. Experi-


10.21437/Interspeech.2023-705

mental results on standard publicly available datsets, show that our method effectively mitigates the domain mismatch between training and test sets, and surpasses the current baseline.

## 2. The domain adaptation method

The noisy speech signal can be represented as follows:

$$x(t) = s(t) + n(t), \tag{1}$$

where $x(t)$ represents the noisy speech, $s(t)$ and $n(t)$ are the speech and additive noise signals, respectively. The observed noisy signal (1) can be rephrased in the short-time Fourier transform (STFT) domain with

$$x(k,m) = s(k,m) + n(k,m),$$

where $s(k,m)$ is the speech, and $n(k,m)$ is the noise. The terms $k \in \{0,\dots,K-1\}$ and $m \in \{0,\dots,M-1\}$ represent the frequency and time-frame indices, respectively.

In deep learning-based speech enhancement, the noisy speech is fed to a network that performs a non-linear procedure to generate an estimate of the clean speech $\hat{\mathbf{s}}$ given the noisy input $\mathbf{x}$. The network is trained in a supervised way using pairs of clean and noisy speech.

We address a domain shift problem where the network was trained on noisy speech and we want to apply the network to enhance speech signals from another domain that consists of languages, recording environments and noise types that do not appear in the data used to train the model. We assume here the availability of noisy speech from the target domain but no corresponding clean speech. Our method is composed of two-stages. First, we apply adversarial training that aligns the distributions of the source and the target domains. The second step is a self-training algorithm based on computing pseudo labels for the target domain data. These two steps are described bellow.

**Domain-Adversarial Training (DAT).** Domain adversarial training is designed to learn a domain-invariant representation by reducing the bias presented in the data from different domains [9]. The Domain-Adversarial Neural Network (DANN) is composed of three main elements: an encoder network $\mathbf{f} = E(\mathbf{x};\theta_{enc})$, which takes in the noisy speech $\mathbf{x}$ and generates the feature vector $\mathbf{f}$, the decoder network $D(\mathbf{f};\theta_{dec})$ that utilizes $\mathbf{f}$ to produce the estimated clean acoustic feature $\hat{\mathbf{s}}$, and a binary classification network $Disc(\mathbf{f};\theta_{disc})$ that distinguishes between the source and target domains. In the learning phase we jointly minimize the enhancement loss and maximize the domain classifier loss:

$$L_{\text{DAT}}(\theta_{enc},\theta_{dec},\theta_{disc}) = L_{\text{enhancement}} - \beta \cdot L_{\text{domain-classifier}}. \tag{2}$$

where $\beta$ is a constant (see details in the experiments section). Labeled samples from the source domain contribute to both losses whereas samples from the target domain contribute only to the second loss. The optimization finds the best trade-off between producing features that are domain invariant and are also useful for the main task of speech enhancement. The model is trained by using a gradient reversal layer (GRL) which is placed between the encoder and discriminator in the initial architecture. This layer ensures that the gradient computed during back-propagation is negated in sign, which means that the encoder is trained to maximize the discriminator's loss instead of minimizing it. In this case, the discriminator can be pre-trained to classify the domains and is kept frozen during the UDA procedure. The DANN architecture is illustrated in Fig 1. We employed DAT as a pre-processing step for the second stage training.

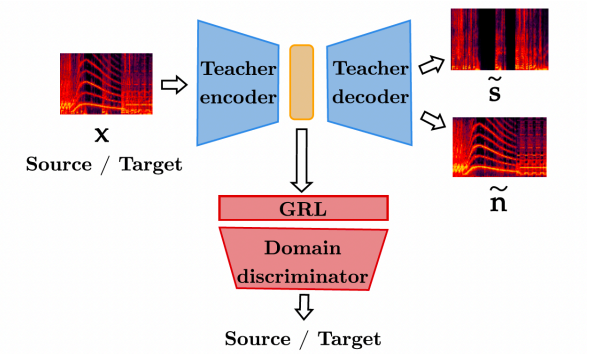


Figure 1: *The domain-adversarial training framework.*

**Self-training with pseudo labels.** Self-training [14] converts model predictions of unlabeled noisy samples from the target domain into pseudo labels. This method involves two key steps. The teacher network (obtained by the DAT step) generates a set of pseudo-labels in the target domain (predicted clean speech and noise). Then, a student network is trained using the noisy samples along with the predicted clean speech. Here we followed the RemixIT approach [11], where the teacher provides estimated speech and noise signals, $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{n}}$, as pseudo-labels. Then, the estimated noises are randomly shuffled within a mini-batch and combined with the teacher's speech estimates to generate a set of bootstrapped mixtures $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = \tilde{\mathbf{s}} + \mathbf{P}\tilde{\mathbf{n}} \in \mathbb{R}^{B\times T}, \tag{3}$$

where $B$ is the mini-batch size, $T$ is the size of the speech samples and $\mathbf{P}$ is a permutation matrix that rearranges the predicted noise in the current mini-batch. The student model is then trained using these bootstrapped mixtures as inputs and by predicting the teacher's pseudo-target signals $\tilde{\mathbf{s}}$ and $\mathbf{P}\tilde{\mathbf{n}}$, by applying a standard supervised training procedure. Another property of the RemixIT method is that the teacher network is updated multiple times to learn from higher-quality source estimates, thus enabling the model to continually improve its performance.

**Reliable pseudo labels.** In the RemixIT implementation of self-training to speech enhancement, [11], the teacher network is pre-trained in a supervised manner on the source domain. Therefore, if the gap between the source and the target is relatively large, the quality of the teacher's enhancement results on the noisy target speech samples is poor which yields misleading pseudo-labels. The domain-adversarial training step produces a more suitable teacher network and thus enables a warm start for the noisy-labels steps. However, even with this adapted network, the resulting network may perform better on some target samples than others, and the quality of the pseudo-labels is uncertain.

The framework of adversarial training followed by iterative self-training with pseudo labels has been shown to effectively address domain shift issues in image classification and segmentation tasks. This approach, as demonstrated in studies such as [15, 13], involves selecting target samples with more reliable pseudo-labels, determined based on the confidence of the teacher network's class prediction. However, the major challenge in applying this approach to speech enhancement is that the output of the network is an estimated speech signal without any indication of the network's confidence in the estimated speech's quality. Without feedback on the quality of the pseudo-

**Algorithm 1** Source Regularization Self Training (SRST)

Input: labeled data from the source domain and unlabeled data from the target domain.
- Apply DAT domain adaptation algorithm.
- Train a source-target domain binary classifier.
- Select the source samples that are classified as targets.
- Generate speech and noise predictions using the teacher network.

**for** $k$ in 1 until convergence **do**
Train a student network by minimizing the loss:
$L = L_{\text{RemixIT}} + \lambda L_{\text{source}}$, s.t. $\lambda$ is decreased at each epoch $k$.
**end for**

labels, there is no explicit guidance for improving the quality of the predicted speech.

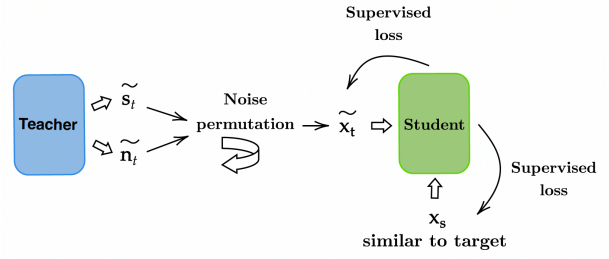


Figure 2: *A Scheme of the SRST loss function.*

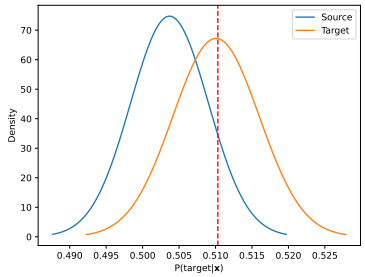


Figure 3: *Histogram of source and target domain soft decisions after the domain-adversarial algorithm. The dashed line indicates the 30% of the source samples that are classified as targets with the highest probability.*

Applying pseudo-labels to speech enhancement suffers from the problem of a lack of information on their quality. In the UDA setup there are accurate labels from the source domain data, and adding labeled data to the pseudo-labeled data from the target domain can encourage the student network to produce a meaningful estimation of the clean data. The problem, of course, is that the labeled dataset is from the source domain and hence is not suitable for our goal of learning an enhancement network for the target domain. Domain adversarial training is designed to reduce the distance between the features of samples from the two domains but we cannot remove it completely. The crux of our approach is adding source domain samples that are similar to the target domain samples to the RemixIT self-training scheme. As part of the domain-adversarial training step, we have already built a binary classifier that can distinguish between samples from the source and target domains. Our approach involves selecting the subset of samples from the source domain with the highest probability of being classified as target examples. Fig. 3 shows the densities of the soft decision outcomes of the binary domain classifier for the source and target domains after the domain adversarial adaptation step (see next section for details on the source and target domains data). It shows that the feature alignment DANN algorithm makes the two plots closer buts the classifier can still distinguish between the source and target domain. In addition, there are source domain examples that are more similar to the target domain than others. These source samples that look like targets, serve as intermediate training data to adapt the network to the target domain. The minibatch-level loss for training the student network is thus:

$$L = L_{\text{RemixIT}} + \lambda L_{\text{source}} \tag{4}$$

such that the first component is a loss function used for target domain data to penalize the reconstruction error between the estimates and their corresponding pseudo-labels. The second component is the same loss function which is this time applied on the source domain speech estimates and their corresponding true speech signal. The scalar $\lambda$ is monotonically decreased at each epoch. below we show that, unlike RemixtIT, there is no need for iterations of updating the teacher network and recomputing the pseudo-labels. We dub the proposed method Source Regularization Self Training (SRST). Fig. 2 shows a scheme of the SRST training procedure and the algorithm is summarized in Algorithm Box 1.

## 3. Experiments

We implemented SRST on various domain shift scenarios to evaluate its performance.

**Datasets.** The experimental setup included the following three standard speech datasets. **LibriSpeech** [16]: This corpus of read English speech is divided into a training set comprised of 960 hours of audio, a validation set that contains 5 hours and a test set with about 5 hours of recordings. **WHAM!** [17]: This dataset is a collection of speech recordings mixed with babble background noise from various urban locations. The speech recordings come from the WSJ0-2mix dataset and the background noise was collected at different places such as restaurants, cafes, bars, and parks. **DNS-Challenge (DNS)** [12]: This dataset is a collection of clean speech recordings mixed with different types of noise. The dataset includes 64,649 pairs of clean speech and noise recordings for training and 150 pairs for testing. We used the LibriSpeech and WHAM! datasets to create source domain speech and noise and the DNS-Challenge dataset was used as the target domain. In our experiments, the target domain was different from the source in terms of both language and noise types. The language of the source was English and the languages that were used in the target data were French, German, Italian, Mandarin, Russian and Spanish.

**The domain gap problem**: Most previous speech enhancement UDA works focused on the gap between the domain distributions caused by different *noise* types [8], [11]. Table 2 presents the gaps between enhancement measures when evaluating the target data with a model trained on the source domain and with a model trained on the target domain, using English and non-English speech in the target. The results show that language change had a major impact on network performance.

**Implementation details**: The noisy waveforms were extracted into time-frequency (T-F) complex features using a 512-

Table 1: *Speech enhancement results on the DNS non-English (**target** domain) test set (16kHz) with SNR values of -5, 0 and 5.*

| Method | SNR=-5 | | | SNR=0 | | | SNR=5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | PESQ | STOI | SI-SDR (dB) | PESQ | STOI | SI-SDR (dB) | PESQ | STOI | SI-SDR (dB) |
| Noisy | 1.126 | 0.648 | -4.867 | 1.177 | 0.714 | 0.151 | 1.299 | 0.823 | 4.969 |
| Source model | 1.325 | 0.707 | 3.315 | 1.443 | 0.777 | 8.234 | 1.872 | 0.888 | 12.556 |
| RemixIT [11] | 1.364 | 0.723 | 5.598 | 1.496 | 0.790 | 9.924 | 1.919 | 0.896 | 14.000 |
| DANN [9] | 1.597 | 0.772 | 8.762 | 1.859 | 0.834 | 12.857 | 2.301 | 0.915 | 14.951 |
| DANN+ RemixIT | 1.557 | 0.769 | 9.241 | 1.750 | 0.829 | 12.716 | 2.221 | 0.913 | 15.228 |
| + SRST (random source) | 1.603 | 0.767 | 8.999 | 1.855 | 0.833 | 13.006 | 2.290 | 0.915 | 15.237 |
| + SRST (similar source) | **1.616** | **0.775** | **9.248** | **1.861** | **0.837** | **13.260** | **2.317** | **0.917** | **15.336** |
| Target model | 1.602 | 0.770 | 8.972 | 1.900 | 0.842 | 13.170 | 2.375 | 0.916 | 15.118 |

point Short Time Fourier Transform (STFT) with a Hamming window and an overlap of 256. The input to the network was the real and imaginary parts of the T-F maps. The output real and imaginary maps for both the speech and noise estimates were reconstructed into the time domain using iSTFT with the same parameters. We developed and evaluated models that function with a sampling rate of 16kHz. Throughout all of our experiments, we used U-Net [18] as our encoder-decoder architecture. The classifier received the U-Net bottleneck as input and contained two convolution layers with a PReLU activation function followed by the mean over the time domain. Before using it for domain adaptation, the encoder-decoder was trained in a supervised manner on a set of 50,000 samples from the source domain, including mixtures generated by speech signals from LibriSpeech and noises from the WHAM! dataset. The supervised loss for the enhancement task was chosen to be the negative scale-invariant signal to distortion ratio (SI-SDR) [19] for both the estimated speech and noise:

$$L(\hat{\mathbf{s}}, \mathbf{s}) = -\text{SI-SDR}(\hat{\mathbf{s}}, \mathbf{s}) = -20\log_{10}(\frac{\|\alpha\mathbf{s}\|}{\|\alpha\mathbf{s} - \hat{\mathbf{s}}\|}), \quad (5)$$

where $\alpha = \hat{\mathbf{s}}^T\mathbf{s}/\|\mathbf{s}\|^2$. In the domain-adversarial training phase, we used a pre-trained domain classifier as initialization to the adversarial classifier. For the pre-trained classifier training, we used BCE loss and the Adam optimizer with a learning rate of $10^{-4}$, whereas the feature extractor was the frozen pre-trained encoder from the source supervised training. The loss for the adversarial training was the weighted sum of the SI-SDR loss for enhancement and the BCE loss for the domain classification:

$$L(\hat{\mathbf{s}}, \mathbf{s}, \hat{d}, d) = -\text{SI-SDR}(\hat{\mathbf{s}}, \mathbf{s}) + \beta \cdot \text{BCE}(\hat{d}, d) \quad (6)$$

where $d$ an $\hat{d}$ were the ground truth and estimated domain, respectively and $\beta$ was set to 0.05. Here we also used Adam with a learning rate of $10^{-4}$. In the self-training stage, we used teacher inference to create labels for the target domain. For each mini-batch we selected noisy data from the source and the target domain. We started with 30% most similar source samples and then decreased the ratio of source samples in each epoch by 3%.

**Compared methods**: We compared the quality of the predicted speech produced by the following methods: RemixIT [11], DANN [11] and DANN followed by RemixIT (DANN+RemixIT). We implemented two variants of our SRST method. In the first, we used 30% of the source samples that were found to be the most similar to the target domain and in the second we randomly selected 30% of the source samples. We denote these variants as SRST(similar source) and SRST(random source). We also report the speech quality of the input noisy signal, the results of a model trained on the source data which serves as a lower bound and a model trained on the target domain using noisy and clean speech which serves as an oracle upper bound.

Table 2: *Performance of English and non-English DNS test data. The source model was trained on English LibriSpeech + WHAM!. The target model was trained on English and non-English DNS train dataset.*

| Test speech | Method | PESQ | STOI | SI-SDR (dB) |
|---|---|---|---|---|
| DNS English | Noisy | 1.582 | 0.915 | 9.229 |
| | Source model | 2.320 | 0.952 | 15.696 |
| | Target model | 2.387 | 0.953 | 17.207 |
| | Gap | 0.067 | 0.010 | 1.511 |
| DNS non-English | Noisy | 1.177 | 0.714 | 0.151 |
| | Source model | 1.443 | 0.777 | 8.234 |
| | Target model | 1.900 | 0.842 | 13.170 |
| | Gap | **0.457** | **0.065** | **4.936** |

**Results**: Table 1 presents the SI-SDR [19], the short-time objective intelligibility (STOI) [20] and the perceptual evaluation of speech quality (PESQ) [21] on a separate test set from the target domain built from non-English languages speech and DNS noises. The results demonstrate that RemixIT itself was hardly useful at all in the scenario of a large domain gap, since it lacks high-quality pseudo-labels for training the student. The DANN strategy worked well here and yielded improved results. Adding to the DANN a post-processing step of RemixIT resulted in a slight improvement. The predicted speech of our SRST method was found to be the best for all the SNR values on all the speech quality measures. SRST(similar source) was better than SRST(random source), which indicates that it indeed helps to exclusively use source samples that resemble the target domain samples.

To conclude, in this paper, we studied the problem of language and noise mismatch in SE systems. We proposed a self-supervised speech enhancement method, that can successfully handle large gaps between the source and the target domains. The experiments showed the superiority of the proposed method in terms of performance, in an unsupervised speech enhancement domain adaptation. In the future, we aim to generalize and evaluate our method in a more general setup of noisy speech separation.

# 4. References


[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[2] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

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

[4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] A. Defosses, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[6] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[7] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[8] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

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

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

[12] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[13] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[14] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning*, 2013.

[15] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European Conference on computer vision (ECCV)*, 2018.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[17] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

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

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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.

[21] 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 *IEEE International Conference on Acoustics, speech, and Signal Processing (ICASSP)*, 2001.