ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447573

LEVERAGING SELF-SUPERVISED SPEECH REPRESENTATIONS FOR DOMAIN ADAPTATION IN SPEECH ENHANCEMENT

Ching-Hua Lee, Chouchang Yang, Rakshith Sharma Srinivasa, Yashas Malur Saidutta, Jaejin Cho, Yilin Shen, Hongxia Jin

Samsung Research America

ABSTRACT

Deep learning based speech enhancement (SE) approaches could suffer from performance degradation due to mismatch between training and testing environments. A realistic situation is that an SE model trained on parallel noisy-clean utterances from one environment, the source domain, may fail to perform adequately in another environment, the target (new) domain of unseen acoustic or noise conditions. Even though we can improve the target domain performance by leveraging paired data in that domain, in reality, noisy data is more straightforward to collect. Therefore, it is worth studying unsupervised domain adaptation techniques for SE that utilize only noisy data from the target domain, together with exploiting the knowledge available from the source domain paired data, for improved SE in the new domain. In this paper, we present a novel adaptation framework for SE by leveraging self-supervised learning (SSL) based speech models. SSL models are pre-trained with large amount of raw speech data to extract representations rich in phonetic and acoustics information. We explore the potential of leveraging SSL representations for effective SE adaptation to new domains. To our knowledge, it is the first attempt to apply SSL models for domain adaptation in SE.

***Index Terms*—** Speech enhancement, unsupervised domain adaptation, self-supervised representation, wav2vec

1. INTRODUCTION

The performance of learning-based speech enhancement (SE) systems can degrade due to mismatch between training and testing conditions. Domain adaptation in SE [1, 2, 3] aims to mitigate the degradation issue by adapting an SE model trained under one condition, the source domain, towards another condition, the target domain. In the case where the target (new) domain data labels (i.e., clean utterances) are not available, it is referred to as the unsupervised domain adaptation problem [4] for SE [5, 6, 7, 8, 9]. The goal in general is to use the unlabeled noisy data from the target domain, along with the noisy-clean labeled data from the source domain, to learn an adequate model for the new domain. This problem has good applicability to real-world scenarios, as collecting just noisy recordings in new environments is more straightforward than collecting noisy-clean paired data. However, compared to the vast in-domain, supervised learning SE literature [10, 11, 12, 13, 14, 15, 16], unsupervised domain adaptation in SE has received relatively little attention.

On the other hand, self-supervised learning (SSL) based pre-trained speech models have attracted considerable attention [17, 18, 19, 20]. They can extract rich acoustic information from huge amount of raw speech data to benefit various downstream tasks [21], including a handful of works on supervised SE [22, 23, 24, 25, 26, 27, 28]. A nice property of SSL models for SE is that, most of the noisy and clean speech can be highly distinguishable in the SSL latent space, which is preferable for performing SE [25]. Fig. 1 (Left) presents this observation by using a popular SSL model, wav2vec [17], to encode noisy and clean utterances from the two public datasets, CHiME-3 [29] and VoiceBank+DEMAND [30]. The representations are visualized by the t-SNE analysis [31] of temporally-averaged wav2vec encoded features of the utterances.

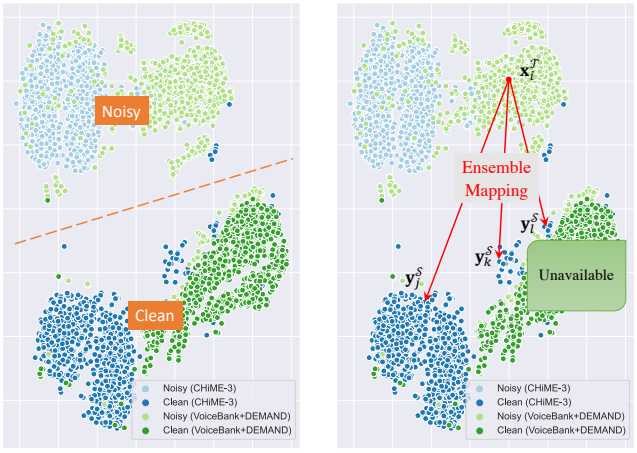

Fig. 1. Representations extracted by wav2vec encoder. (Left) It can be seen that noisy and clean data are well-separated in the SSL latent space. (Right) When the clean data of the target domain (VoiceBank+DEMAND) are not available, we approximate the exact noisy-clean mapping for SE through ensemble mapping, i.e., mapping the noisy sample into the clean speech neighborhood by using multiple source domain (CHiME-3) clean samples.

In this paper, motivated by such observation, we explore the potential of leveraging SSL representations for unsupervised domain adaptation in SE. At a high level, as illustrated in Fig. 1 (Right), we assume VoiceBank+DEMAND data as the target domain whose clean utterances are not available. As there is no corresponding clean reference for a noisy utterance $\mathbf{x}_i^{\mathcal{T}}$ in that domain, the exact noisy-clean mapping is not available for training SE models. As such, we propose to leverage multiple clean utterances $\mathbf{y}_j^{\mathcal{S}}$, $\mathbf{y}_k^{\mathcal{S}}$, $\mathbf{y}_l^{\mathcal{S}}$, ... from the source domain (CHiME-3) to realize ensemble mapping of the noisy sample into the vicinity of clean speech. We show that by approximating the exact mapping with a tactically designed ensemble mapping process in the SSL feature space, improved SE can be achieved with the proposed adaptation strategy. Our framework is referred to as Self-Supervised Representation based Adaptation (SSRA), taking advantage of publicly available SSL models as a powerful tool for adaptation. To our knowledge, it is the first attempt to leverage SSL representations for domain adaptation in SE.

2. RELATED WORK

Unsupervised domain adaptation in SE: Works on unsupervised domain adaptation for SE are relatively few [5, 6, 7, 8, 9] compared to the in-domain, supervised learning SE schemes. Notably, [5, 6] utilize domain adversarial training (DAT) for learning domain-invariant features to reduce the mismatch between source and target domains. Despite effectiveness, DAT-based adaptation requires the SE model to have an encoder-decoder architecture for learning domain-invariant features. Moreover, it could be challenging to train a stable model with the domain classifier in adversarial training [32].

SSL models for SE: SSL pre-trained speech models have been applied to SE in various ways. For example, [22, 23, 24] utilize SSL representations as input features to downstream SE models. However, the entire denoising module (i.e., SSL + SE) becomes huge for inference (as SSL model $\gg$ SE model typically). On the other hand, [25, 26, 27, 28] leverage SSL models only for SE network training by using loss functions defined in the SSL feature space instead of the signal space. The SE performance improves by carefully engaging the SSL-based supervision. Despite demonstrating the benefits of using SSL models, most of the existing applications to SE are under in-domain, supervised SE settings. The potential of SSL models for unsupervised domain adaptation in SE remains unexplored.

3. PROBLEM FORMULATION

We consider the problem of finding an estimator $f(\cdot)$ that maps the noisy speech utterance $\mathbf{x} \in \mathcal{X}$ into its clean reference $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the spaces of noisy and clean speech, respectively. We further consider paired noisy-clean speech data $\{(\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$ of a source domain distribution $\mathcal{S}(\mathbf{x}, \mathbf{y})$ available for training, where i denotes the sample index and $N_{\mathcal{S}}$ is the number of samples. We then assume there is a new target domain following the distribution $\mathcal{T}(\mathbf{x}, \mathbf{y})$ with $N_{\mathcal{T}}$ samples of noisy, unlabeled data $\{\mathbf{x}_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$ available. Suppose that we now have an adequate estimator $f(\cdot; \boldsymbol{\theta}_{\mathcal{S}})$ parameterized by the SE model $\boldsymbol{\theta}_{\mathcal{S}}$ for the source domain obtained by training using the paired data $\{(\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$. Due to domain shift caused by unseen noise types and acoustic environments, the SE model $\boldsymbol{\theta}_{\mathcal{S}}$ could suffer from performance degradation in the target domain. Our goal is to seek an adapted version of the SE model $\boldsymbol{\theta}_{\mathcal{T}}$ that mitigates such degradation, by also leveraging target domain noisy data for learning the estimator. The search of an adapted SE model by utilizing both target domain unlabeled and source domain labeled data is called unsupervised domain adaptation for SE [6].

4. PROPOSED FRAMEWORK

4.1. Self-supervised representation based adaptation (SSRA)

Given the training data of the source domain $\{(\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$ and target domain $\{\mathbf{x}_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$, our framework aims to seek an optimal parameter set $\boldsymbol{\theta}_*$ (in some sense) for the SE model $f(\cdot; \boldsymbol{\theta})$ by:

$$\min_{\boldsymbol{\theta}} \ \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{ssra}} = \min_{\boldsymbol{\theta}} \ \underbrace{\frac{1}{N_{\mathcal{S}}} \sum_{i=1}^{N_{\mathcal{S}}} D_1(f(\mathbf{x}_i^{\mathcal{S}}; \boldsymbol{\theta}), \mathbf{y}_i^{\mathcal{S}})}_{\mathcal{L}_{\text{rec}}\text{: Rec Loss}} + \underbrace{\frac{\lambda}{N_{\mathcal{S}} N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} \sum_{j=1}^{N_{\mathcal{S}}} w_{ij} D_2(h(f(\mathbf{x}_i^{\mathcal{T}}; \boldsymbol{\theta})), h(\mathbf{y}_j^{\mathcal{S}}))}_{\mathcal{L}_{\text{ssra}}\text{: SSRA Loss}}, \tag{1}$$

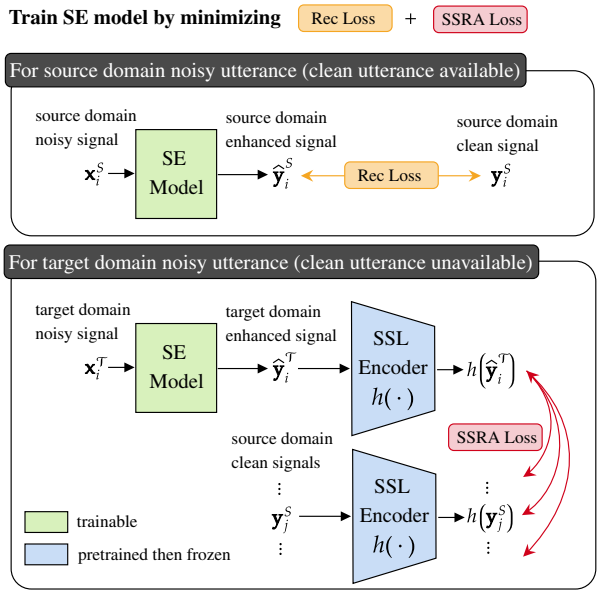

Fig. 2. Illustration of the proposed SSRA framework where the SE model is trained by jointly minimizing two loss terms. (Top) For a source domain noisy sample $\mathbf{x}_i^{\mathcal{S}}$ we perform typical SE training to estimate the corresponding clean reference $\mathbf{y}_i^{\mathcal{S}}$ via normal reconstruction loss (Rec Loss). (Bottom) For a target domain noisy sample $\mathbf{x}_i^{\mathcal{T}}$, since the corresponding clean speech $\mathbf{y}_i^{\mathcal{T}}$ is not available, we propose to leverage multiple clean utterances $\{\mathbf{y}_j^{\mathcal{S}}\}$ from the source domain to guide the SE model learning through the SSRA Loss.

where $D_1(\cdot, \cdot)$ and $D_2(\cdot, \cdot)$ are some distance measures and $\lambda > 0$ is a weighting term to control the relative strength between the two loss terms. The framework is graphically illustrated in Fig. 2. The first term in (1), $\mathcal{L}_{\text{rec}}$, is the typical signal reconstruction loss in normal supervised SE settings for the source domain labeled data. It measures the discrepancy between each enhanced speech $\hat{\mathbf{y}}_i^{\mathcal{S}} = f(\mathbf{x}_i^{\mathcal{S}}; \boldsymbol{\theta})$ and its corresponding clean reference $\mathbf{y}_i^{\mathcal{S}}$. This term helps the SE model transfer knowledge learned from the source domain.

The second term in (1), $\mathcal{L}_{\text{ssra}}$, is the proposed SSRA loss responsible for adapting the SE model to the new domain. Here, the target domain enhanced speech $\hat{\mathbf{y}}_i^{\mathcal{T}} = f(\mathbf{x}_i^{\mathcal{T}}; \boldsymbol{\theta})$ and source domain clean speech $\mathbf{y}_j^{\mathcal{S}}$ are first transformed to SSL representations via the pre-trained SSL encoder $h(\cdot)$. Then, the distance between each i, j pair is computed for the two SSL representations, $h(\hat{\mathbf{y}}_i^{\mathcal{T}})$ and $h(\mathbf{y}_j^{\mathcal{S}})$. Summation of the distance over all the formed i, j pairs is then carried out, aiming to leverage multiple source domain clean samples to guide the SE model learning through mapping each target domain noisy sample into the clean speech neighborhood. This way, we are approximating the exact noise-clean mapping by an ensemble mapping process. Here in $\mathcal{L}_{\text{ssra}}$, we heuristically adopt a weighting term $w_{ij} \in [0, 1]$ that is supposed to reflect the similarity between the target domain and source domain noisy samples, $\mathbf{x}_i^{\mathcal{T}}$ and $\mathbf{x}_j^{\mathcal{S}}$. If the two noisy samples are similar, then a larger weight is assigned to the corresponding distance, and vice versa. Intuitively, in the extreme case that for a given target domain $\mathbf{x}_i^{\mathcal{T}}$, if an identical $\mathbf{x}_j^{\mathcal{S}}$ from the source domain is found, then we have the largest $w_{ij} = 1$, meaning that we are confident the unknown target domain label can be constructed by the corresponding source domain label, i.e., $h(\mathbf{y}_i^{\mathcal{T}}) = h(\mathbf{y}_j^{\mathcal{S}})$.

In sum, the main idea of the SSRA framework is to take advantage of SSL representations for guiding the SE model adaptation to

the target domain. This is motivated by the decent separability of clean and noisy speech in the SSL latent space and that SSL representations are rich in acoustic and phonetic information of speech. Notably, the SSL model is utilized only during training and does not increase the complexity in inference time. The framework is general and can potentially adopt various SE networks and SSL models.

4.2. Practical implementations

The reconstruction loss: There are various choices for the distance metric $D_1(\cdot,\cdot)$ in the reconstruction loss $\mathcal{L}_{\text{rec}}$ and our framework is not restricted to any specific one. For example, the time-domain mean-squared-error loss (MSE), mean-absolute-error loss (MAE), or scale-invariant signal-to-distortion ratio (SI-SDR) loss [33], etc. It can also be computed in the time-frequency domain based on the short-time Fourier transform (STFT), e.g., the power-law compressed combined MSE loss and others [34].

The SSRA loss: Though generally speaking the $D_2(\cdot,\cdot)$ in $\mathcal{L}_{\text{ssra}}$ can utilize various distance measures, we present an effective choice that is usually used for contrasting embeddings – the negative cosine similarity, imposed on temporally averaged SSL representations:

$$D_2(h(f(\mathbf{x}_i^{\mathcal{T}};\boldsymbol{\theta})), h(\mathbf{y}_j^{\mathcal{S}})) = -\text{cossim}(\bar{h}(f(\mathbf{x}_i^{\mathcal{T}};\boldsymbol{\theta})), \bar{h}(\mathbf{y}_j^{\mathcal{S}})), \quad (2)$$

where $\text{cossim}(\mathbf{a},\mathbf{b}) = \mathbf{a}^T\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\|$ is the cosine similarity of two vectors $\mathbf{a}$ and $\mathbf{b}$, and $\bar{h}(\cdot)$ stands for the averaged SSL representation over time frames. The intuition behind using (2) is to align the two representations in a softer manner rather than strictly forcing them to be frame-wise identical, as it is unlikely to be an exact noisy-clean mapping from two different domains. Our findings suggest that such soft alignment is more beneficial for adapting the SE model.

For the weighting term w_{ij}, we propose to use:

$$w_{ij} = 0.5 * (\text{cossim}(\bar{h}(\mathbf{x}_i^{\mathcal{T}}), \bar{h}(\mathbf{x}_j^{\mathcal{S}})) + 1). \quad (3)$$

The value of w_{ij} given by (3) lies in $[0,1]$ and is proportional to the similarity between the time-averaged SSL representations of the i-th target domain and j-th source domain noisy utterances.

Mini-batch optimization: Instead of the full batch problem (1), practically we perform stochastic optimization based on using mini-batches of data samples. More specifically, we draw $B_{\mathcal{S}} \ll N_{\mathcal{S}}$ samples each time from the source domain data and $B_{\mathcal{T}} \ll N_{\mathcal{T}}$ samples from the target domain data. Gradient is computed on (1) with $N_{\mathcal{S}}$ and $N_{\mathcal{T}}$ replaced by $B_{\mathcal{S}}$ and $B_{\mathcal{T}}$, respectively, for back-propagation. In every new epoch, the two datasets can be reshuffled to increase data pairing diversity of the target domain and source domain samples for the SSRA loss.

5. EXPERIMENTS

Datasets: We validate the proposed SSRA approach for domain adaptation of SE models from the source domain (CHiME-3) to the target domain (VoiceBank+DEMAND), following the setup of [6].

- *Source domain – CHiME-3 dataset [29]:* The publicly available CHiME-3 dataset is a 6-channel microphone recording of talkers speaking in a noisy environment, sampled at 16 kHz. It consists of 7138 and 1320 simulated utterances for training and testing, respectively. In this paper, we only take the 5-th channel recordings for the experiments. We use all noisy-clean data pairs of the training set as the source domain labeled data. The noises and speakers are all different from the target domain data that we consider, i.e. the VoiceBank–DEMAND dataset.
- *Target domain – VoiceBank+DEMAND dataset [30]:* In this dataset, clean speech clips are collected from the VoiceBank corpus [35] with 28 speakers for training and another 2 unseen speakers for testing, mixed with noise profiles from the DEMAND database [36]. There are totally 11527 utterances for training and 824 for testing. In our experiments, we take the noisy utterances from the training set as the target domain unlabeled data and do not include their clean references for training the SE model. For testing, we evaluate the SE model performance on the test set data.

Model settings:

- *The SSL model:* we adopt the pre-trained *wav2vec large*[1]. It consists of a convolutional encoder network and a context network. We utilize the encoder network $h(\cdot)$ for extracting representations.
- *SE Network 1:* We use the gated recurrent units (GRUs) based SE network from [34], which was chosen to maintain real-time constraints without delay and moderate complexity. The SE model operates in the time-frequency domain, where an STFT size of 512 with 32 ms Hann window and 16 ms frame shift are used. The network takes the logarithmic power spectrum (LPS) of the noisy speech as input, and outputs a time-frequency mask to multiply with the noisy STFT for denoising. It consists of a feed-forward (FF) embedding layer followed by two GRUs and then three FF layers with ReLU activations, and finally an output layer with sigmoid activation. The architecture is depicted in Fig. 2 of [34].
- *SE Network 2:* We have another SE network for fair comparison with the DAT-based approach of [6], which consists of a bidirectional long short-term memory (BLSTM) layer with 512 units, followed by one FF layer of 257 units with sigmoid activation. It takes the stack of static, delta, and acceleration features of the noisy LPS as input and outputs a time-frequency mask, using 512-point STFT with 32 ms Hann window and 16 ms frame shift.

All waveforms are resampled to 16 kHz for processing. We use $B_{\mathcal{S}} = B_{\mathcal{T}} = 32$ for the batch size. During training, we randomly crop 4-sec long audio clips while during testing the entire sequence is processed and evaluated. For training SE Network 1, the combined power-law compressed MSE loss of [34] is used for $\mathcal{L}_{\text{rec}}$ with the suggested $c = 0.3$ and $\beta = 0.3$ in their paper. For SE Network 2 of [6], the squared loss on the static, delta, and acceleration features of the enhanced speech is utilized for $\mathcal{L}_{\text{rec}}$, following [6]. Adam optimizer is used with a step size of 0.0001 for SE Network 1 (80 epochs) and 0.001 for SE Network 2 (40 epochs), where an epoch means going through every sample of the source and target domain data at least once. The SSRA loss weight is $\lambda = 10^{-4}$ for SE Network 1 and $\lambda = 10^{-2}$ for SE Network 2 based on grid search.

Evaluation metrics: Here we use: **PESQ:** Perceptual Evaluation of Speech Quality [37] (value: -0.5 to 4.5). **SI-SNR:** Scale-Invariant SNR [38]. **CSIG, CBAK, COVL:** Mean opinion score (MOS) predictors of signal distortion, background-noise intrusiveness, and overall signal quality [39] (value: 0 to 5). **SSNR:** Segmental SNR [39]. In all metrics a higher score indicates better SE performance.

Results: Table 1 presents the results for the SE model using SE Network 1 and tested on the target domain test set. We compare the SE models trained with i) only using the source domain labeled (noisy-clean paired) data, referred to as "SE-unadapted" and ii) using both source domain labeled data and target domain unlabeled (noisy) data with our SSRA adaptation, referred to as "SE-SSRA". It can be seen

[1]We use the pre-trained wav2vec large model publicly provided at: https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec

that SSRA improves the denoising results over the unadapted baseline model for all metrics, validating its effectiveness. We further confirm its superiority in Fig. 3 by plotting the PESQ and COVL scores vs. input SNR. We see that SE-SSRA consistently outperforms the baseline unadapted model across difference SNR settings.

Being an adaptation method for SE, it would be interesting to validate SSRA's ability to take advantage of more target domain unlabeled data. To this end, we include extra 23075 noisy utterances taken from the larger 56-speaker training set of the VoiceBank+DEMAND dataset (note: there is no overlapping with the original 11527 utterances from the 28-speaker training set normally used) as additional target domain noisy samples for adaptation. In Table 1 we also present the result of using the extra noisy data from the target domain for SSRA-based adaptation (trained with $\lambda = 5 \times 10^{-5}$ for 55 epochs). It can be seen that with additional noisy samples leveraged, the SSRA performance can be furthered improved. Such a property of SSRA is desirable for domain adaptation in SE, as in reality, collecting just noisy samples in a new environment is more straightforward than collecting noisy-clean paired data.

Table 1. Performance on target domain (VoiceBank+DEMAND).

Methods	PESQ	SI-SNR	CSIG	CBAK	COVL
Noisy	1.97	8.46	3.35	2.44	2.63
SE-unadapted	2.43	17.22	3.09	3.15	2.75
SE-SSRA	**2.56**	**17.31**	**3.37**	**3.15**	**2.94**
SE-SSRA + extra noisy data	**2.61**	**17.43**	**3.51**	**3.17**	**3.02**

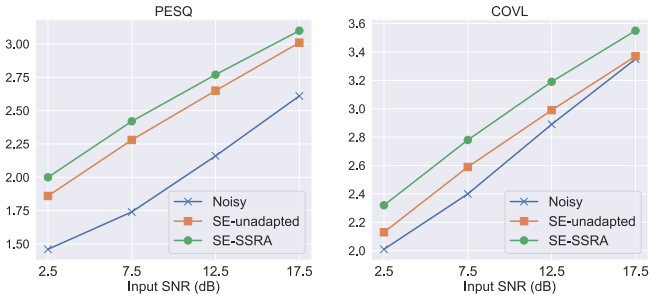

Fig. 3. Performance vs. input SNR for SSRA and baseline.

After validating the performance of SSRA on the target domain, we come back to see how the adapted model performs on the original source domain of CHiME-3. Table 2 presents the results of the SE models evaluated on the CHiME-3 test set. We can see that the denoising performance of the SE-SSRA is not compromised for the source domain while adapting to the target domain. In fact, the results slightly improve with SSRA, and with extra target domain noisy data used. This may be attributed to that the SE model has also become more generalizable and robust to noise conditions while adapting to the new domain, by also exploiting the additional noise information inherent in the target domain data. Indeed, VoiceBank+DEMAND encompasses ten noise types as compared to only four types in CHiME-3, thus potentially richer in noise information.

With an aim to provide further insight on how the SSRA actually affects the denoising outcomes, in Fig. 4 we visualize the wav2vec features of the noisy, enhanced by SE-unadapted, enhanced by SE-SSRA, and the clean speech utterances via t-SNE analysis [31] for both target and source domains. Fig. 4 (Left) shows that for the target domain, without SSRA the enhanced signals (orange) are still close to the noisy ones (blue). The SSRA adaptation effectively pulls the enhanced signals (green) towards the clean utterances (red). This

Table 2. Performance on source domain (CHiME-3).

Methods	PESQ	SI-SNR	CSIG	CBAK	COVL
Noisy	1.27	7.51	2.61	1.92	1.88
SE-unadapted	1.70	12.58	3.04	2.53	2.34
SE-SSRA	**1.73**	**12.69**	**3.09**	**2.54**	**2.38**
SE-SSRA + extra noisy data	**1.74**	**12.93**	**3.11**	**2.57**	**2.40**

is in accordance to the improvements observed in Table 1 and Fig. 3. On the other hand, the feature maps shown in Fig. 4 (Right) for the source domain show that the enhanced signals of both with and without adaptation fall in the vicinity of the clean utterances. This corresponds to the similar SE performance observed in Table 2.

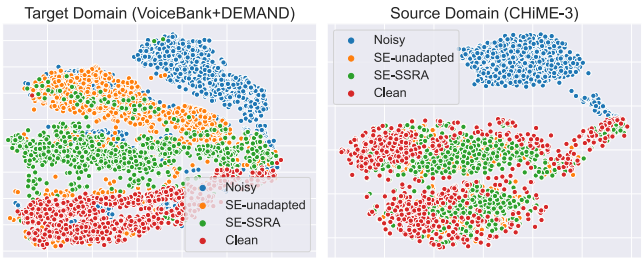

Fig. 4. t-SNE analysis on wav2vec encoded feature maps.

Finally, we compare our SSRA method to the DAT-based approach of [6], i.e., SE-DAT, on the target domain test set. For fairness, here we use SE Network 2 for the SE model, i.e., the same BLSTM-based SE model as in [6], following their routines. Table 3 presents the results (since [6] did not report SI-SNR, we report SSNR instead). For reference, we also show the conventional Wiener filter [40] results taken from [11]. From the table we can see that without adaptation (i.e., "SE-unadapted"), the model performs even worse than the Wiener filter. This indicates that the domain mismatch issue can lead to unsatisfactory performance for the learning-based SE approaches if careful consideration is not taken. With adaptation, the SE performance is clearly improved, and our SSRA demonstrates more improvements compared to DAT in most metrics.

Table 3. SE adaptation comparison to DAT-based approach on target domain (VoiceBank+DEMAND).

Methods	Training data	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	-	1.97	3.35	2.44	2.63	1.68
Wiener	none	2.22	3.23	2.68	2.67	5.07
SE-unadapted	source dom. labeled	2.12	3.38	2.46	2.66	1.76
SE-DAT [6]	source dom. labeled +	2.26	**3.72**	2.77	**2.98**	4.11
SE-SSRA (ours)	target dom. unlabeled	**2.46**	3.53	**3.10**	**2.98**	**7.76**

6. CONCLUSION

We presented SSRA, a domain adaptation framework for SE by leveraging speech representations extracted via SSL pre-trained models. We explored the possibility of exploiting the nice properties of SSL encoded features for adapting the SE model to perform adequately in a new domain with only noisy data available. Different from existing DAT-based methods, our framework provides yet another effective solution to unsupervised domain adaptation in SE.

7. REFERENCES

[1] S. Pascual, M. Park, J. Serrà, A. Bonafonte, and K.-H. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5019–5023.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *Proc. Int. Symp. Chinese Spoken Lang. Process. (ISCSLP)*, 2014, pp. 336–340.

[3] C.-C. Lee, Y.-C. Lin, H.-T. Lin, H.-M. Wang, and Y. Tsao, "SERIL: Noise adaptive speech enhancement using regularization-based incremental learning," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2020, pp. 2432–2436.

[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Machine Learn. (ICML)*, 2015, pp. 1180–1189.

[5] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2019, pp. 3148–3152.

[6] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *Proc. Asia-Pacific Signal Inform. Process. Assoc. Annual Summit and Conf. (APSIPA ASC)*, 2019, pp. 667–672.

[7] S. Wang, W. Li, S. M. Siniscalchi, and C.-H. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6219–6223.

[8] H.-Y. Lin, H.-H. Tseng, X. Lu, and Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021.

[9] Y. Li, Y. Sun, K. Horoshenkov, and S. M. Naqvi, "Domain adaptation and autoencoder-based unsupervised speech enhancement," *IEEE Trans. Artificial Intelligence*, vol. 3, no. 1, pp. 43–52, 2021.

[10] 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, 2018.

[11] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2017, pp. 3642–3646.

[12] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6649–6653.

[13] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2021, pp. 14549–14557.

[14] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2020, pp. 2742–2746.

[15] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 6648–6652.

[16] C.-H. Lee, C. Yang, Y. Shen, and H. Jin, "Improved mask-based neural beamforming for multichannel speech enhancement by snapshot matching masking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2019, pp. 3465–3469.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020, pp. 12449–12460.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[20] S. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[21] S. w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[22] Z. Huang, S. Watanabe, S.-w. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 6837–6841.

[23] K.-H. Hung, S.-W. Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting self-supervised embeddings for speech enhancement," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2022, pp. 186–190.

[24] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

[25] T.-A. Hsieh, S.-W. Yu, C and. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using Wasserstein distance for speech enhancement," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2021, pp. 196–200.

[26] T. Sun *et al.*, "Boosting the intelligibility of waveform speech enhancement networks through self-supervised representations," in *Proc. IEEE Int. Conf. Machine Learn. Appl. (ICMLA)*, 2021, pp. 992–997.

[27] X. Xu, W. Tu, C. Han, and Y. Yang, "All information is necessary: Integrating speech positive and negative information by contrastive learning for speech enhancement," *arXiv preprint arXiv:2304.13439*, 2023.

[28] H. Sato *et al.*, "Downstream task agnostic speech enhancement with self-supervised representation loss," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2023, pp. 854–858.

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME'speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, vol. 46, pp. 605–626, 2017.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Int. Speech Comm. Assoc. Speech Synthesis Workshop (SSW)*, 2016, pp. 146–152.

[31] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *J. Machine Learn. Research*, vol. 9, no. 11, 2008.

[32] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Towards stable and comprehensive domain alignment: Max-margin domain-adversarial training," *arXiv preprint arXiv:2003.13249*, 2020.

[33] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 825–838, 2020.

[34] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. Int. Conf. Telecomm. Signal Process. (TSP)*, 2021, pp. 72–76.

[35] 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. Research Eval. (O-COCOSDA/CASLRE)*, 2013.

[36] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Mtgs. Acoust.*, 2013.

[37] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *Int. Telecomm. Union*, 2005.

[38] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 626–630.

[39] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2007.

[40] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1996, pp. 629–632.