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

Contact Information:
Samsung Research America
665 Clyde Ave, Mountain View, CA 94043

Email: chinghua.l@samsung.com

Abstract

Objective:

- Develop a novel domain adaptation technique for speech enhancement (SE) to mitigate performance degradation due to mismatch between source and target domains, *where we only have access to noisy data in the target domain*
- Explore the idea of utilizing Self-Supervised Learning (SSL) speech representations for domain adaptation in SE

Method:

- Propose the **Self-Supervised Representation based Adaptation (SSRA)**:
  - leveraging SSL speech models (e.g., *wav2vec*) pre-trained with large amount of raw speech data which extract representations rich in phonetic & acoustics information
  - exploiting decent separability of clean and noisy speech in the SSL space
  - utilizing a novel similarity-based loss in the SSL space to handle unpaired data

Results:

- The proposed SSRA framework demonstrates the potential of exploiting SSL representations for adapting SE models to new domains

1 Overview

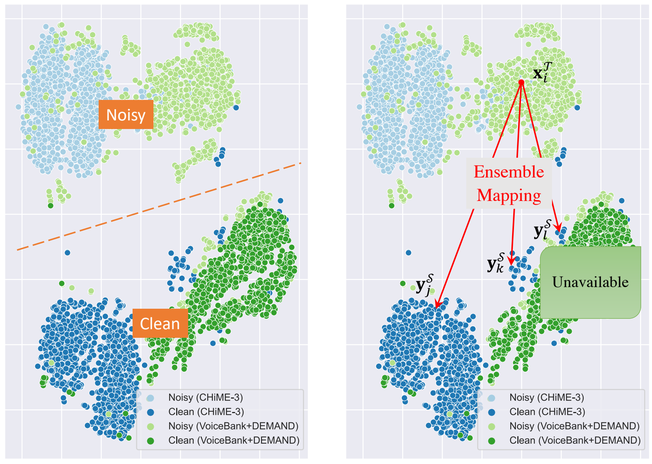

Figure 1: Representations extracted by SSL (*wav2vec*) encoder. (Left) It can be seen that noisy and clean data are well-separated in the SSL latent space. (Right) When the clean data of the target domain are not available, we approximate the exact noisy-clean mapping for SE through ensemble mapping.

2 Unsupervised Domain Adaptation for SE

- **General SE:** To find an estimator $f(\cdot;\boldsymbol{\theta})$ that maps the noisy 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
- **Scenario:**
  - *Source domain*: noisy-clean speech pairs $\{(\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$ of a source domain distribution $\mathcal{S}(\mathbf{x}, \mathbf{y})$ available for training
  - *Target domain*: a new domain following the distribution $\mathcal{T}(\mathbf{x}, \mathbf{y})$ with only noisy data $\{\mathbf{x}_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$ accessible for training
- **Issue:** Domain shift caused by unseen environments leads to an adequate SE model $\boldsymbol{\theta}_{\mathcal{S}}$ trained on source domain $\mathcal{S}$ suffering from performance degradation in target domain $\mathcal{T}$
- **Goal:** To seek an adapted version of the SE model $\boldsymbol{\theta}_{\mathcal{T}}$ that mitigates such degradation by leveraging target domain noisy data

3 Proposed Method

- **Main idea:** to take advantage of SSL representations for guiding SE model adaptation to the target domain, based on:
  - decent *separability* of clean noisy speech in the SSL space
  - rich acoustic and phonetic information in SSL representations
- Notably, the SSL encoder $h(\cdot)$ is utilized only during training and does not increase the complexity in inference time

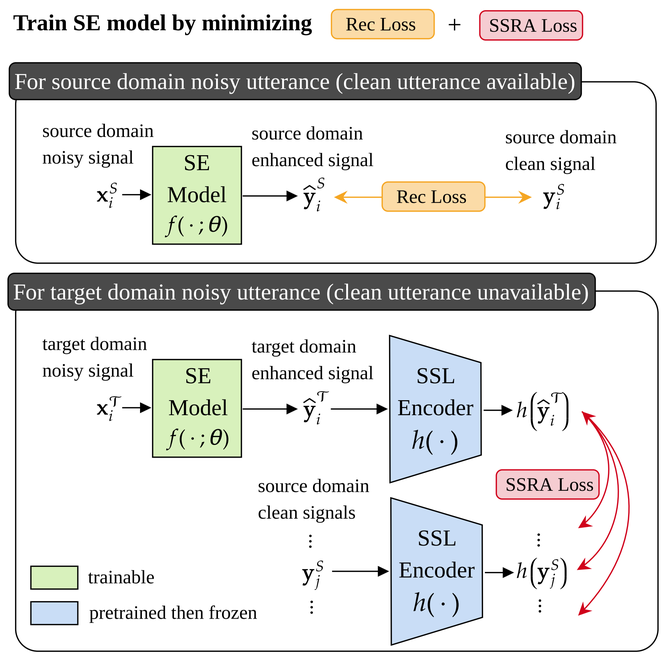

Figure 2: Illustration of the proposed SSRA framework where the SE model is trained by jointly minimizing two loss terms.

3.1 The SSRA framework

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:

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

where $D_1(\cdot,\cdot)$ and $D_2(\cdot,\cdot)$ are some distance measures and $\lambda > 0$ for weighting the two loss terms

- **The reconstruction loss:** various choices for the distance metric $D_1(\cdot,\cdot)$ in $\mathcal{L}_{\text{rec}}$ (e.g., MSE, SI-SNR)
- **The SSRA loss:** we present an effective choice for $D_2(\cdot,\cdot)$ with the negative cosine similarity imposed on temporally averaged SSL representations:

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

where $\bar{h}(\cdot)$ stands for the averaged SSL representation over time frames. For the weighting term w_{ij}, we propose to use:

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

which ranges in $[0, 1]$ and is proportional to the similarity of the comparing target and source domain noisy utterances

- **Mini-batch optimization** is performed for (1) in practice. In every epoch, datasets are reshuffled to increase data pairing diversity of the target and source domains

4 Experiments

Datasets: We validate the proposed SSRA for domain adaptation of single-channel SE models following the setup in [1]:

- *Source domain:* CHiME-3 dataset [2] (on the 5-th channel)
- *Target domain:* VoiceBank+DEMAND dataset [3]

Models:

- *SSL model:* pre-trained *wav2vec large* from [4]
- *SE Network 1:* a GRU based SE network from [5] (for results in Table 1, Table 2 and Figure 3)
- *SE Network 2:* a BLSTM based SE network from [1] (for Table 3 results)

Results:

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**

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**

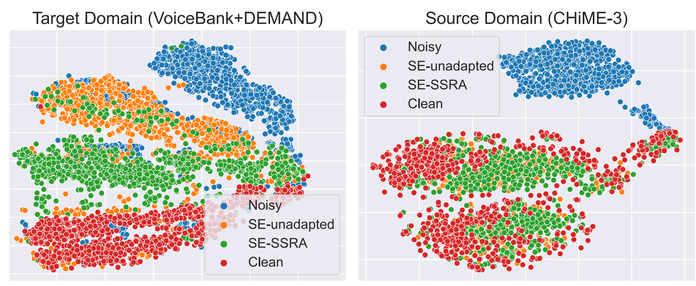

Figure 3: t-SNE analysis on *wav2vec* encoded feature maps.

Table 3: SE adaptation comparison to domain adversarial training (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 [1]	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**

5 Conclusion

We presented SSRA, a novel domain adaptation framework for SE based on using SSL representations. We explored the possibility of exploiting the nice properties of SSL features for adapting the SE model to new domains and demonstrated its effectiveness.

References

[1] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *Proc. of APSIPA ASC*, 2019.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME'speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, 2017.

[3] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *ISCA Speech Synthesis Workshop*, 2016.

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of Interspeech*, 2019.

[5] 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.*, 2021.