JOINT LEARNING WITH SHARED LATENT SPACE FOR SELF-SUPERVISED MONAURAL SPEECH ENHANCEMENT

Yi Li[1]*, Yang Sun*[2]*, Wenwu Wang*[3]*, Syed Mohsen Naqvi*[1]

[1] Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2] Big Data Institute, University of Oxford, UK
[3] Centre for Vision Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Supervised learning has been used to solve monaural speech enhancement problem, offering state-of-the-art performance. However, clean training data is difficult or expensive to obtain in real room environments, which limits the training of supervised learning-based methods. In addition, mismatch conditions e.g., noises in the testing stages may be unseen in the training stage, present a common challenge. In this paper, we propose a self-supervised learning-based monaural speech enhancement method, using two autoencoders i.e., the speech autoencoder (SAE) and mixture autoencoder (MAE), with a *shared layer*, which help to mitigate mismatch conditions by learning a shared latent space between speech and mixture. To further improve the enhancement performance, we also propose phase-aware training and multi-resolution spectral losses. The latent representations of the amplitude and phase are independently learned in two decoders of the proposed SAE with only a very limited set of clean speech signals. Moreover, multi-resolution spectral losses help extract rich feature information. Experimental results on a benchmark dataset demonstrate that the proposed method outperforms the state-of-the-art self-supervised and supervised approaches. The source code is available at https://github.com/Yukino-3/Complex-SSL-SE.[1]

***Index Terms*—** monaural speech enhancement, self-supervised learning, multi-resolution spectral losses, phase-aware, joint training

1. INTRODUCTION

Monaural speech enhancement has attracted considerable research attention and deep learning techniques have significantly improved its performance with a supervised learning (SL) strategy [1, 2, 3, 4]. However, supervised training of the networks requires large sets of labelled paired data. Moreover, a trained model may suffer from performance degradation when deployed in previously unseen conditions e.g., a mismatch of room environments between the training and testing sets. To address the above limitations, self-supervised learning (SSL) techniques are applied as an effective alternative for monaural speech enhancement [5, 6, 7, 8].

The first SSL-based speech enhancement (SSE) method is proposed by Wang et al. [6], where an autoencoder is used to learn a latent representation of clean speech signals as the pretask, and another autoencoder is used to learn the shared representation between the clean speech and its mixtures. However, the SSE method only learns a shared latent space with unseen speakers [6], its generalization ability to unseen noises and room environments is still limited. Moreover, the phase information of speech signals is ignored in [6]. To address the limitations in [6], we propose a joint training algorithm to improve the speech enhancement performance by using two autoencoders, namely, the speech autoencoder (SAE) and the mixture autoencoder (MAE). The SAE is trained with clean speech signals to learn their latent representations with the amplitude and phase information processed with two individual decoders. The MAE is trained with noisy mixtures recorded in real room environments, where a shared layer from the SAE and MAE is used to obtain a joint latent space of the learned clean speech and noisy mixture representations. The last layer of the encoder in the MAE is replaced by the one in the shared layer after the training stage is completed. To improve the generalization ability of the network model, the training data used for the MAE is unseen in the training data (i.e. unseen room environments) used for the SAE, which helps to train the shared layer to address the mismatch conditions between the training and testing stages.

2. PROPOSED METHOD

2.1. Network Architecture

The block diagram of the proposed method is shown in Fig. 1. Initially, multi-resolution features are extracted from the spectra **S** i.e., the input of the SAE. In order to preserve the desired information in the signal, in the encoder named E_S, each convolutional layer generates the feature map of a specific resolution, which is then scaled to produce the latent representation

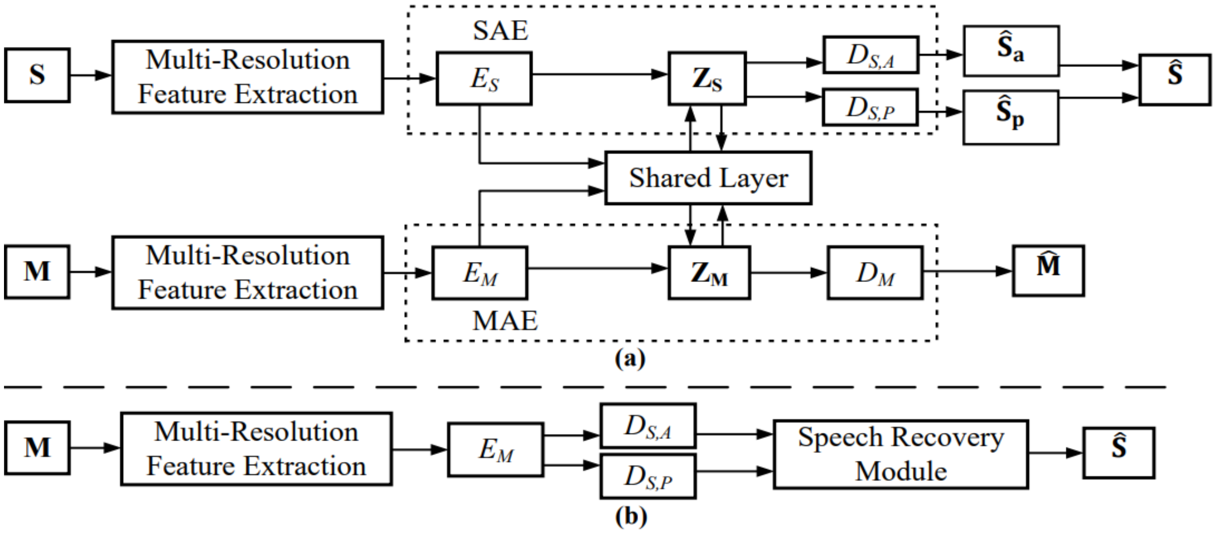

Fig. 1. The overall architecture of the proposed method. (a) Training: From speech spectra $\mathbf{S}$, the multi-resolution features are extracted with different window sizes as the input of E_S. Then, the latent representation of the speech feature $\mathbf{Z_S}$ is learned via E_S. Then, the reconstructed amplitude and phase of clean spectra are independently obtained as the output by using $D_{S,A}$ and $D_{S,P}$. Similarly, from unseen noisy mixture $\mathbf{M}$, the multi-resolution features are extracted as the input of the E_M and the mixture feature map $\mathbf{Z_M}$ is learned. Meanwhile, a shared latent space between $\mathbf{Z_S}$ and $\mathbf{Z_M}$ improves the generalization ability of the MAE. (b) Testing: Multi-resolution mixture spectra $\mathbf{M}$ which are unseen with $\mathbf{M}$ in (a) are fed into the trained E_M. The enhanced signal $\hat{\mathbf{S}}$ is obtained with the estimate spectrogram from the speech recovery module.

$\mathbf{Z}_S$ with multi-resolutions. The optimal weights for combining the spectra with each resolution are learned with the target i.e., the feature map of the clean speech, during the training of E_S.

In the proposed method, two decoders $D_{S,A}$ and $D_{S,P}$ are applied in SAE to learn the amplitude and phase of speech, respectively. In detail, the latent representations of both the amplitude and phase are learned by minimizing the discrepancy between the input representation and the corresponding reconstruction. The multi-resolution spectra of the estimated speech signals are obtained.

Different from the SAE, the MAE only requires access to unseen noisy mixtures $\mathbf{M}$. The multi-resolution features are extracted from the noisy mixture and fed to E_M. Consequently, the latent representation of the mixture is obtained as the output of E_M and exploited to modify the loss functions. Then, the speech feature representation $\mathbf{Z}_S$ and mixture representation $\mathbf{Z}_M$ are used to learn a cross-domain latent space. To achieve that, we concatenate the last layer from both E_S and E_M and create the shared layer between two autoencoders. The mixture representation is passed through the decoders of the SAE to get the enhanced version of the mixture representation. Benefiting from the learned speech representation, a mapping relationship from the mixture to the target speech is learned through $D_{S,A}$ and $D_{S,P}$. The shared latent space between the SAE and MAE is used to further learn the latent representation of the unseen mixture spectra. The last layer of E_M is replaced by the one in the shared layer after the training stage is completed.

In the testing stage, the feature of the noisy mixture is extracted and fed into the trained E_M to obtain the latent representation of the mixture feature. This representation is then used with the decoders $D_{S,A}$, and $D_{S,P}$ to decode the estimated amplitude and phase of the target speech spectra, respectively. Finally, in the speech recovery module, the phase is recovered by re-wrapping the estimated unwrapped phase of speech. Then, it is used with the recovered speech amplitude to reconstruct the estimated speech signal.

2.2. Loss Functions

Different from previous SSL methods [6, 7, 8, 9, 10], the proposed method exploits multi-resolution feature maps for the network training. Inspired by [11], we use the multi-resolution STFT loss as an auxiliary loss to improve the stability and efficiency for model training. The feature map is rescaled with the same frame shift (i.e. 32), but with different window sizes (1024, 512, 256, and 128). Each STFT loss term estimates the frame-level difference between the clean speech spectrogram and the corresponding reconstructed speech spectrogram.

For the SAE training, the loss $\mathcal{L}_\mathbf{S}$ is the sum of four multi-resolution losses defined on amplitude and phase between the clean speech feature and the reconstructed speech feature as:

$$\mathcal{L}_\mathbf{S} = \sum_{i=1}^{I}(\|\mathbf{S}_a^i - \hat{\mathbf{S}}_a^i\|_2^2 + \|\mathbf{S}_p^i - \hat{\mathbf{S}}_p^i\|_2^2) \qquad (1)$$

where i refers to the index of the multi-resolution feature maps, subscripts a and p denote the amplitude and phase, respectively. Once the loss function is minimized, we now use the trained SAE and noisy mixtures to train the MAE. The loss $\mathcal{L}_{\mathbf{M}}$ denotes the sum of the multi-resolution losses between the noisy mixture feature and the corresponding reconstruction as:

$$\mathcal{L}_{\mathrm{M}} = \sum_{i=1}^{I} (\|\mathbf{M}^i - \hat{\mathbf{M}}^i\|_2^2) \tag{2}$$

Then, the shared layer between the two autoencoders is used to learn a shared latent representation to mitigate the mismatch between the training and testing conditions. To achieve this, the amplitude and phase of $\mathbf{Z}^i_{\mathbf{M}}$ are enhanced by the trained $D_{S,A}$ and $D_{S,P}$, respectively. Then, the amplitude and phase of the enhanced spectra are mapped back by E_S to produce the estimated mixture representation $\hat{\mathbf{Z}}^i_{\mathbf{M}}$. The overall MAE loss with the hyper-parameter λ is given as:

$$\mathcal{L}_{\mathrm{MAE}} = \mathcal{L}_{\mathbf{M}} + \lambda \cdot \sum_{i=1}^{I} \left\|\mathbf{Z}^i_{\mathbf{M}} - \hat{\mathbf{Z}}^i_{\mathbf{M}}\right\|_2^2 \tag{3}$$

3. EXPERIMENTAL RESULTS

3.1. Datasets

The Device And Produced Speech (DAPS) dataset [12] is used in these experiments as [6]. The noisy data consists of 20 speakers (10 female and 10 male) each reading out 5 story excerpts in indoor environments with different real room impulse responses (RIRs). In addition, the clean raw data are collected in an acoustically treated low noise, low reflection vocal booth of a professional recording studio using a microphone with a flat frequency response [12]. Most non-speech sounds such as breaths and lip smacks were removed from the recordings by the sound engineer to create clean speech [12]. We cut 14 minutes of data from each speaker into 28 clips where each clip has 30 seconds long. To show the generalization ability of the proposed SSL method, we split utterances from different speakers in the data preprocessing stage. In the training stage, 420 clean utterances from 15 speakers are randomly selected. For each environment, we first randomly select 28 utterances from a speaker to generate the training data for the SAE. To train the MAE, 392 utterances from 14 speakers are used to generate the mixtures with three different background noises ($factory$, $babble$, and $cafe$) from the NOISEX dataset [13] with four SNR levels (-10, -5, 0, and 5 dB). Therefore, the data used for training MAE is unseen in the data used for training SAE. Moreover, in the testing stage, the remaining 140 utterances of 5 speakers, which are unseen from those in the training stage, are used to generate the mixtures with the same SNR levels but different background noise types and room environments as those in the training stage.

3.2. Experimental Setup and Performance Metrics

Similar to [6, 14, 15], the proposed autoencoders use variational autoencoders (VAEs) as the backbone. In the SAE, E_S, $D_{S,A}$, and $D_{S,P}$ all consist of four 1-D convolutional layers. In the MAE, E_M, $D_{M,A}$, and $D_{M,P}$ all consist of six 1-D convolutional layers. The proposed method is trained by using the Adam optimizer with a learning rate of 0.001 and batch size of 20. The coefficient λ is used in (3) to constraint loss terms and is set empirically with different experiments. For most of the experiments, it is set to 0.01 according to the grid search results by using 0.001, 0.01, 0.1, 1, and 10 as options for the parameter values. However, it is set to 0.1 because the latent representation loss plays a more important role in some specific experiments. The number of training epochs is 700 and 1500 for SAE and MAE, respectively.

Similar to [6], we use composite metrics that approximate the Mean Opinion Score (MOS) including COVL, i.e. the MOS predictor of overall signal quality, CBAK, i.e. the MOS predictor of background-noise intrusiveness, CSIG, i.e. the MOS predictor of signal distortion [16], and Perceptual Evaluation of Speech Quality (PESQ). Higher values of these performance metrics imply better enhancement performance.

3.3. Comparisons with SSL Methods

In this section, we compare the proposed method with three state-of-the-art SSL speech enhancement approaches [6, 7, 8]. The first method is SSE [6] which exploits two autoencoders to estimate speech and mixture, respectively. The second method is the pre-training vector quantization method (PT-VQ) [7], which combines WavLM [17] and Transformer encoder. The third method applies a cross-domain feature (CF) which integrates the SSL representation and spectrogram [8]. This baseline consists of 2 linear layers, two-layered bidirectional long short-term memory (BLSTM) of 256 hidden units and a sigmoid activation to generate the prediction mask. Table 1 shows the speech enhancement performance with PESQ, CSIG, CBAK, and COVL at different SNR levels.

It can be seen from Table 1 that the proposed method outperforms the state-of-the-art SSL methods in terms of all four performance measures. The proposed method and baselines are also compared at different SNR levels. From the experimental results, it can be seen that the proposed method outperforms the baselines even at a relatively low SNR level i.e., -5 dB. The proposed method has 7.6%, 7.0%, 7.8%, and 7.4% improvements compared with the CF method in terms of four performance measures at -5 dB SNR level.

3.4. Comparisons with SL Methods

In this section, we further compare the proposed method with state-of-the-art SL approaches [1, 2, 4]. The DBT-Net aims to recover the coarse- and fine-grained regions of the overall spectrogram in parallel [2]. An attention-in-attention

Table 1. Comparison with SSL methods. Each result is the average value of 1,260 (140 signals×3noise types×3 room environments) experiments. *Italic* shows the proposed methods. **Bold** indicates the best results.

	PESQ			CSIG			CBAK			COVL		
SNR (dB)	-5	0	5	-5	0	5	-5	0	5	-5	0	5
SSE [6]	1.32	1.33	1.34	1.97	2.04	2.09	1.74	1.76	1.77	1.59	1.65	1.68
PT-VQ [7]	1.68	1.70	1.71	2.24	2.27	2.29	1.76	1.79	1.80	1.72	1.77	1.81
CF [8]	1.71	1.74	1.77	2.29	2.30	2.35	1.80	1.89	1.96	1.76	1.80	1.86
Proposed	**1.84**	**1.89**	**1.91**	**2.45**	**2.47**	**2.49**	**1.94**	**2.10**	**2.23**	**1.89**	**1.96**	**2.03**

transformer-based network is adopted for better feature learning. The second method is frequency recurrent convolutional recurrent network (FRCRN) which boosts feature map along the frequency axis [4]. Moreover, in the spectrogram decomposition (SD) method, feature maps are composed of spectra containing evident speech components according to the mask value [1]. These feature maps make the boundary information of speech components clear by ignoring others, thus boosting the sensitivity of the model to input features. Table 2 shows the speech enhancement performance with PESQ, CSIG, CBAK, and COVL.

Table 2. Comparison with SL methods. Each result is the average value of 3,780 experiments (140 signals×3 noise types×3 room environments×3 SNR levels). *Italic* shows the proposed methods. **Bold** indicates the best results.

	PESQ	CSIG	CBAK	COVL
SD [1]	1.68	2.21	1.72	1.66
DBT-Net [2]	1.69	2.21	1.76	1.68
FRCRN [4]	1.72	2.25	1.83	1.74
Proposed	**1.88**	**2.47**	**2.09**	**1.96**

These SL methods are originally trained with large datasets e.g., VoiceBank [18] and DEMAND datasets [19] which contain 11,572 utterances in [2]. However, in these comparison experiments, we use only 420 utterances in the DAPS dataset to train all the methods because the clean speech data is difficult or expensive to obtain in real-world scenarios, e.g., talking in an office. The training of the supervised methods strongly relies on the large-scale data to facilitate the model to learn structural information [20]. Therefore, the speech enhancement performance of supervised methods suffers from significant degradation compared with its original implementation. In addition, different from the original implementation [2, 4, 1], unseen speakers, noises, and room environments are also used to generate noisy mixtures in the testing stage, which leads to a further drop in the reproduced performance results. In this work, the proposed method uses the shared layer to learn a joint latent space between the SAE and MAE in unseen cases. Thus, the speech enhancement performance is improved although the model is tested in unseen cases.

3.5. Ablation Study

In this section, we investigate the effectiveness of each contribution. Table 3 shows the speech enhancement performance with PESQ, CSIG, CBAK, and COVL.

Table 3. Ablation study of the three contributions in the proposed method. Each result is the average value of 3,780 experiments (140 signals×3 noise types×3 room environments×3 SNR levels). The shared layer and multi-resolution are abbreviated as S-L and M-R, respectively.

Ablation Settings			PESQ	CSIG	CBAK	COVL
Phase	S-L	M-R				
✗	✗	✗	1.33	2.03	1.76	1.64
✓	✗	✗	1.43	2.15	1.83	1.68
✗	✓	✗	1.59	2.26	1.98	1.81
✗	✗	✓	1.39	2.10	1.79	1.65

From Table 3, it can be observed that the performance is improved by each contribution among all four performance metrics. The improvement of the proposed shared layer is more significant than the use of the phase-aware and multi-resolution spectral losses. Because the shared latent space between the two autoencoders is learned at the last layers of E_S and E_M, the speech signal can be estimated from unseen noisy mixtures using a network that is trainable without labelled training data.

4. CONCLUSION

In this paper, we have presented a self-supervised learning based method with complex spectra and limited training data to address the monaural speech enhancement problem. The cross-domain latent representation for unseen noisy mixtures was learned by using the proposed shared layer. To further improve the generalization ability, we proposed phase-aware decoders and multi-resolution spectral losses based on the multi-resolution feature maps. The experimental results showed that the proposed method outperformed the state-of-the-art approaches in a challenging case where the speakers, background noises, and room environments are unseen in the testing stage. Furthermore, the relationship between the amplitude and phase may be relevant to future studies.

5. REFERENCES

[1] H. Shi, L. B. Wang, S. Li, J. W. Dang, and T. Kawahara, "Monaural speech enhancement based on spectrogram decomposition for convolutional neural network-sensitive feature extraction," *Interspeech*, 2022.

[2] G. C. Yu, A. D. Li, H. Wang, Y. T. Wang, Y. X. Ke, and C. S. Zheng, "DBT-Net: dual-branch federative magnitude and phase estimation With attention-in-attention Transformer for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[3] Y. Xian, Y. Sun, J. A. Chambers, and S. M. Naqvi, "Geometric information based monaural speech separation using deep neural network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[4] S. K. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: boosting feature representation using frequency recurrence for monaural speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[5] Y. Li, Y. Sun, and S. M. Naqvi, "Self-supervised learning and multi-task pre-training based single-channel acoustic denoising," *IEEE International Conference on Multisensor Fusion and Integration*, 2022.

[6] Y.-C. Wang, S. Venkataramani, and P. Smaragdis, "Self-supervised learning for speech enhancement," *International Conference on Machine Learning (ICML)*, 2020.

[7] X.-Y. Zhao, Q.-S. Zhu, and J. Zhang, "Boosting self-supervised embeddings for speech enhancement," *arXiv preprint arXiv:2209.14150*, 2022.

[8] 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," *Interspeech*, 2022.

[9] L. Jing, P. Vincent, Y. LeCun, and Y. D. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021.

[10] Z. H. Du, M. Lei, J. Q. Han, and S. L. Zhang, "Self-supervised adversarial multi-task learning for vocoder-based monaural speech enhancement," *Interspeech*, 2020.

[11] H. Y. Kim, J. Yoon, S. J. Cheon, W. H. Kang, and N. Kim, "A multi-resolution approach to GAN-based speech enhancement," *Applied Sciences*, vol. 11, no. 2, pp. 721, 2021.

[12] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006 – 1010, 2014.

[13] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 12, no. 3, pp. 247 – 251, 1993.

[14] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, "U-shaped transformer with frequency-band aware attention for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1511 – 1521, 2023.

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

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 – 238, 2008.

[17] S. Y. Chen, C. Y. Wang, Z. Y. Chen, Y. Wu, S. J. Liu, Z. Chen, J. Y. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Z. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *submitted to Journal of Selected Topics in Signal Processing*, 2022.

[18] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: design, collection and data analysis of a large regional accent speech database," *IEEE Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.

[19] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591 – 3591, 2013.

[20] S.-F. Huang, S.-P. Chuang, D.-R. Liu, Y.-C. Chen, G.-P. Yang, and H.-Y. Lee, "Stabilizing label assignment for speech separation by self-supervised pre-training," *Interspeech*, 2021.