Applied Soft Computing 120 (2022) 108568

Contents lists available at ScienceDirect

Applied Soft Computing

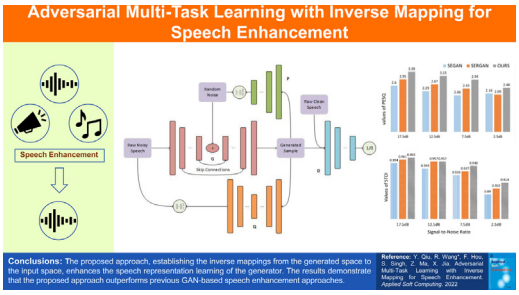
journal homepage: www.elsevier.com/locate/asoc

Adversarial multi-task learning with inverse mapping for speech enhancement

Yuanhang Qiu, Ruili Wang *, Feng Hou *, Satwinder Singh, Zhizhong Ma, Xiaoyun Jia

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 May 2021
Received in revised form 29 January 2022
Accepted 1 February 2022
Available online 8 February 2022

Keywords:
Speech enhancement
Adversarial multi-task learning
Inverse mapping learning
Deep neural networks

ABSTRACT

Adversarial Multi-Task Learning (AMTL) has demonstrated its promising capability of information capturing and representation learning, however, is hardly explored in speech enhancement. In this paper, we propose a novel adversarial multi-task learning with inverse mapping method for speech enhancement. Our method focuses on enhancing the generator's capability of speech information capturing and representation learning. To implement this method, two extra networks (namely P and Q) are developed to establish the inverse mapping from the generated distribution to the input data domains. Correspondingly, two new loss functions (i.e., latent loss and equilibrium loss) are proposed for the inverse mapping learning and the enhancement model training with the original adversarial loss. Our method obtains the state-of-the-art performance in terms of speech quality (PESQ=2.93, CVOL=3.55). For speech intelligibility, our method can also obtain competitive performance (STOI=0.947). The experimental results demonstrate that our method can effectively improve speech representation learning and speech enhancement performance.

1. Introduction

Speech enhancement, one of the most important topics in speech signal processing [1], aims at improving the intelligibility and overall perceptual quality of degraded speech signals. The intelligibility is a measurement of how comprehensible a speech is, while the perceptual quality measures how easy it is for a listener to perceive the content of a speech. Normally, a perceptual high-quality speech sounds more natural, rhythmic, yet less raspy, hoarse, or scratchy. In real life, there are various negative environmental interferences such as additive noise (e.g., fan noise) and convolutional noise (e.g., room reverberation), which can badly degrade speech intelligibility and overall perceptual quality. In practice, speech enhancement has been widely applied in many scenarios such as mobile communications [2], hearing aids [3], and noise-robust speech recognition [4] or speaker recognition [5].

There have been various speech enhancement methods proposed to eliminate the negative effect of the environmental noise and improve the intelligibility and overall perceptual quality of degraded speech signals. For example, Wiener filtering [6], a classic single-channel statistical estimation based approach, is

* Corresponding authors.
E-mail addresses: ruili.wang@massey.ac.nz (R. Wang), f.hou@massey.ac.nz (F. Hou).

https://doi.org/10.1016/j.asoc.2022.108568
1568-4946/

effective in reducing stationary additive noise. However, for reverberation or unknown noise interference, the statistical estimation based approaches perform unsatisfactorily in complex noisy environment generally. Another prominent approach is the microphone arrays based multi-channel speech enhancement [7]. For example, the acoustical beamforming algorithm [8] is conducted on the output signals of microphone arrays and converts them into a single-channel speech signal while amplifying the speech signals from the targeted direction and attenuating the noise signals coming from other directions. The microphone arrays based multi-channel approaches usually take the spatial position information into account and can effectively mitigate the reverberation problem [9].

With the rapid development of intelligent technologies and hardware resources, data-driven based deep neural network such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) has been a thriving technology in speech signal processing [10], computer vision [11], and natural language processing [12]. Long Short-Term Memory (LSTM) [13], a variant of RNN with multiple gates control operation, is suitable for sequences of data learning. CNN [14] is based on the shared-weight architecture of the convolution kernels that slide along input features and provide feature maps of translation responses. LSTM and CNN are basic neural network blocks, which can be accumulated to multiple layers for complex data learning. For speech enhancement, Zhao et al. [15] introduced convolutional-recurrent neural networks to exploit local structures in the frequency and temporal domains. The results showed that their method was more data-efficient and achieved better generalization on both seen and unseen noise. Deep neural network based approaches display a huge potential in dealing with complex speech signal processing and specific speech representation learning [16]. However, the performance of deep architecture degrades inversely if we exhaustively enlarge the network scale only [17], which would cause vanishing gradients or degradation problems. Benefiting from the normalized initialization [18], intermediate normalization layers [19], and skip connections of residual network [20], these aforementioned problems can be largely addressed.

Moreover, to further improve the generalization and performance of speech enhancement models, Meng et al. [21] proposed an Adversarial Multi-Task Learning (AMTL) based method. AMTL, which combines Generative Adversarial Network (GAN) [22] with Multi-Task Learning (MTL), is an effective method to improve the complex information learning of multiple domains and the generalization of models [23]. In [21], two discriminators were added on top of the basic cycle-consistent framework. The multiple loss functions (i.e., discrimination losses, reconstruction losses, and identity-mapping losses) were jointly optimized to distinguish the enhanced and noised features from the real samples. The experimental results showed that the AMTL-based speech enhancement methods effectively reduced the Word Error Ratio (WER) of noise-robust speech recognition.

The method in [21] focused on discriminability improvement of the discriminator, however, little attention was paid to improve the specific information capturing and speech representation learning of the generator. To address this issue, we propose a novel adversarial multi-task learning based method for speech enhancement in this paper. Based on the architecture of GAN, two extra networks (namely P and Q) are developed to establish the inverse mapping from the generated distribution to the input data domains. Correspondingly, two new loss functions (i.e., latent loss and equilibrium loss) are proposed for the inverse mapping learning and the enhancement model training based on the original adversarial loss. With the latent loss function, network P aims to explore relevant latent information from the latent space (i.e., random noise domain) and further facilitate the sample generation. The network Q is developed to balance the adversarial representation learning by mapping the generated distribution to the noisy speech domain with an equilibrium loss function.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 details the proposed method. Section 4 provides the details of our experiments. The results and discussions are presented in Section 5. Finally, the conclusions are shown in Section 6.

2. Related work

With promising adversarial training mechanism, GAN-based speech enhancement methods have successfully attracted much attention. Moreover, inverse mapping learning becomes hot research topic gradually for representation learning improvement. In this section, we introduce the related GAN-based speech enhancement and inverse mapping learning in detail.

2.1. GAN-based speech enhancement

The original GAN [22] consists of a Generator (G) and a Discriminator (D). G is set to learn an effective mapping between the given random noise (z), which is subject to the normal distribution (i.e., mean $= 0$, stddev $= 1$) generally, and the ground-truth (x). Differently, D is considered as an initialized binary classifier, which is trained to give a corresponding judgment of x (real) and $G(z)$ (fake). With the continuous iterative processing, the procedure is trained adversarially up to a Nash Equilibrium [22].

Speech Enhancement GAN (SEGAN) [24] is one of the most famous GAN-based frameworks for time-domain speech enhancement as shown in Fig. 1. SEGAN combines the conditional GAN [25] with the Least-Squares GAN (LSGAN) [26] to further alleviate vanishing gradients. The Generator (G) usually takes original random noise z and extra noisy speech x_c as input and exports targeted data distribution (i.e., $G(z, x_c)$). The Discriminator (D) is considered as a binary classifier trained to distinguish generated samples $G(z, x_c)$ and clean speech x as fake or real (i.e., $D(x, x_c)$ and $D(G(z, x_c), x_c)$). This modification is proved to be effective for performance improvement. Below is the loss functions of its discriminator (i.e., L_D) and generator (i.e., L_G):

$$L_D = \frac{1}{2}E_{x\sim P_x, x_c\sim P_{x_c}}[(D(x, x_c) - 1)^2] + \frac{1}{2}E_{z\sim P_z, x_c\sim P_{x_c}}[(D(G(z, x_c), x_c))^2], \quad (1)$$

$$L_G = \frac{1}{2}E_{z\sim P_z, x_c\sim P_{x_c}}[(D(G(z, x_c), x_c) - 1)^2], \quad (2)$$

where x_c denotes noisy speech; x denotes clean speech; z denotes random noise subject to Gaussian distribution. The binary classifier D codes 1 for real sample and 0 for fake sample.

SEGAN operated directly on the raw speech waveform rather than on the processed spectral features with an end-to-end architecture. The fully convolutional architecture consisted of a downsampling and upsampling module (i.e., encoder and decoder). This enforced the network to focus on temporally close correlations of the input speech signal and throughout the entire processing of the network [24]. The random noise z (i.e., latent vectors) was added to the bottleneck layer for information compensation. However, the latent vectors may be used by the generator in a highly entangled way [27]. For inducing latent vectors, Chen et al. [27] proposed to adopt a mutual information strategy, which decomposed the input noise vectors into a set of semantically meaning factors of variation rather than using single

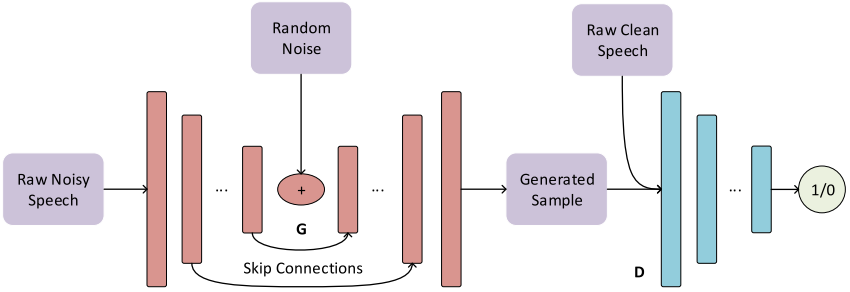


Fig. 1. The framework of SEGAN. The Generator (G) consumes raw noisy speech and latent vector (i.e., random noise) as input. The Discriminator (D) is a binary classifier aiming to judge the similarity between the generated sample and raw clean speech.

unstructured noise vectors. They discovered that these latent factors could target salient semantic features of data distribution. In our speech enhancement, the latent space information is explored further.

Speech Enhancement Relativistic GAN (SERGAN) [28] is another effective GAN-based framework for speech enhancement. As we know in the original GAN, the discriminator was trained to detect if a sample was an original one or a generated one, while the generator was trained to generate data to be more similar to original data to fool the discriminator. The relativistic GAN [29] argued that the probability of $D(x)$ should decrease as the probability of $D(G(z, x_c))$ increases. However, the original GAN cannot incorporate this situation described above, since G does not influence $D(x)$. To circumvent this problem, the relativistic loss function was proposed and used in speech enhancement [28]. Below is the loss functions of SERGAN:

$$L_D = -E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))], \tag{3}$$

$$L_G = -E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))], \tag{4}$$

where $C(x)$ denotes the discriminator without the final sigmoid layer; σ is the sigmoid non-linearity, and thus $D(x) = \sigma(C(x))$. With a similar architecture of SEGAN, SERGAN adopted a new loss function to boost information communication between the generator and discriminator.

Benefiting from adversarial training, more GAN-based methods have been proposed for speech enhancement. Michelsanti and Tan [30] explored the potential of the conditional GAN for speech enhancement; Soni et al. [31] exploited GAN with time–frequency mask based enhancement framework; Donahue et al. [32] conducted a detailed study to measure the effectiveness of GAN-based speech enhancement for robust speech recognition where the speech is contaminated by both additive and convolutional noise. With various model architectures and task requirements, GAN-based speech enhancement has demonstrated a promising capability of complex distribution modeling and speech representation learning.

2.2. Inverse mapping learning

For effective information capturing and data representation learning, the inverse mapping learning with GAN's architecture has shown its success in image processing. Donahue et al. [33] noticed that GAN models could capture semantic variation from latent space but with no means of projecting data back into the latent space. Thus, the GAN architecture ignored much of the useful information found in the structure of the data itself. Besides, interpolations in the latent space of the generator produced smooth and plausible semantic variations and made the model learn to associate particular latent directions with specific features. Thus, the Bidirectional Generative Adversarial Networks (BiGAN) was proposed to learn an inverse mapping from the projecting data back into the latent spaces [33] with a new encoder model. The learned feature representation was thus useful for auxiliary supervised discrimination tasks. Another similar work about latent space exploration with multiple models was proposed in [34]. As introduced in [34], the generation network mapped samples from stochastic latent variables to the data domain while the inference network mapped training examples in data domain to the space of latent variables inversely. The operation could effectively enhance representation learning and sample reconstruction with an adversarial process.

In addition, Huang et al. proposed the Stacked GAN (SGAN) [35] to invert the hierarchical representations of a traditional bottom-up encoder to a stack of top-down generators for high-quality image generation. In SGAN, each generator learned to generate lower-level data representations conditional on high-level representations. The bottom-up encoder, employing a fully connected network, was pre-trained to provide inherent information for the stacked layers of generators. The separated generators and discriminators were trained independently and then jointly to invert the hidden layers information of the encoders. SGAN improved the inherent information learning and enhanced high-resolution image generation by inverting the hidden layers information to the target data domain inversely.

3. Methodology

Based on adversarial multi-task learning and inverse mapping learning, we propose a novel method to further enhance speech representation learning and the performance of speech enhancement. As shown in Fig. 2, our method consists of four networks: a Generator G, a Discriminator D, and the proposed two new networks P and Q. G consumes raw noisy speech and random noise as input and output the generated sample; D aims to judge the similarity between the generated sample and the raw clean speech. The networks P and Q compensate the information learning of G by the inverse mappings with two new loss functions (i.e., latent loss and equilibrium loss).

3.1. Loss functions

The loss function of G consists of three parts: adversarial loss (L_{G_adv}), latent loss (L_{G_lat}), and equilibrium loss (L_{G_equ}). The weighted sum of these three parts is expected to capture real-data information and learn an effective representation of G. Below is the combined loss function of G:

$$L_G = L_{G\_adv} + \lambda_1 L_{G\_lat} + \lambda_2 L_{G\_equ}, \tag{5}$$

where P and Q can be activated or deactivated by setting λ_1 and λ_2 as 0 or 1. We can also try different weight groups to evaluate the effect of P and Q to the whole model.

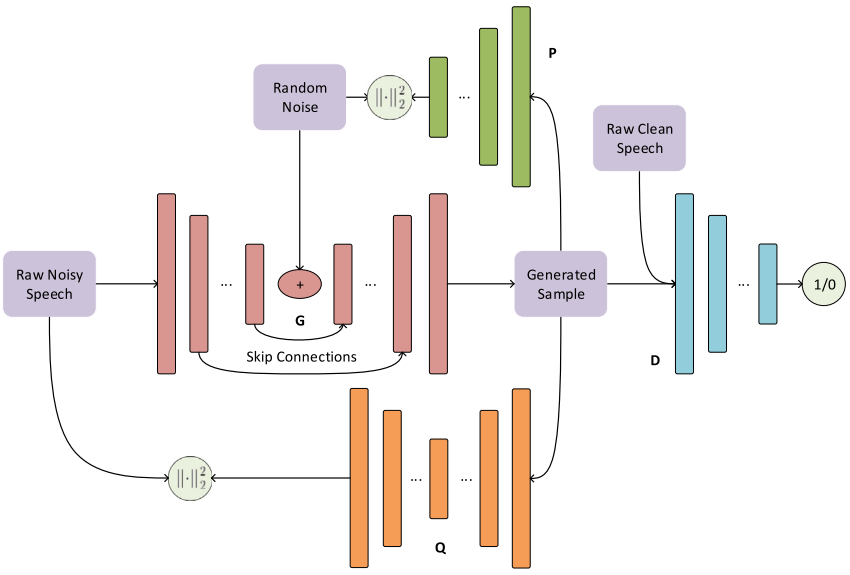


Fig. 2. The framework of our proposed method. The Generator (G) receives raw noisy speech and random noise as input. The Discriminator (D) gives the judgment (i.e., fake or true) of the generated sample and raw clean speech. The networks P and Q are proposed to establish the inverse mapping from the generated distribution to the input data domain for information capturing and representation learning.

The basic adversarial loss function can learn necessary information for G when D is frozen. Here, we adopt the adversarial loss function used in SERGAN [28]. Below is the adversarial loss function:

$$L_{G\_adv} = -E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))], \tag{6}$$

where x denotes the clean speech subject to data distribution P_x; x_c denotes the noisy speech subject to data distribution P_{x_c}; z denotes the random noise subject to distribution P_z (i.e., latent space); $D(x, x_c) = \sigma(C(x, x_c))$ as mentioned above.

The latent space plays an important role in GAN architecture-based representation learning and stable model training [33]. However, current models generally ignore thoroughly exploring latent space information. Thus, in our method, P is built to excavate latent space information by mapping the generated distribution to the latent space inversely. Below is the latent loss function:

$$L_{G\_lat} = -E_{z\sim P_z, x_c\sim P_{x_c}}[\|P(G(z, x_c)) - z\|_2^2], \tag{7}$$

where the squared Euclidean distance $\|\cdot\|_2^2$ is adopted to measure the similarity of random noise distribution z with the output distribution of P. Here, the distance measurement can be designed in other ways, but we choose $\|\cdot\|_2^2$ because it makes the hyper-parameter tuning easier [36].

We propose to establish the inverse mapping from generated data distribution to latent space with network P. The latent loss function works with the adversarial loss function together to enhance G to capture more effective information for information reconstruction. However, a potential unbalanced learning problem may appear and result in defective real-data distribution modeling. Also, unnecessary complexity and model instability may be introduced if just feeding more extra conditional information [35]. Thus, another network Q with the equilibrium loss function is developed as well to obtain a trade-off during model training. Below is the equilibrium loss function:

$$L_{G\_equ} = -E_{z\sim P_z, x_c\sim P_{x_c}}[\|Q(G(z, x_c)) - x_c\|_2^2], \tag{8}$$

where $\|\cdot\|_2^2$ is also adopted to measure the similarity of noisy speech distribution with the output of Q.

With the pre-set weights, the adversarial loss, latent loss, and equilibrium loss functions form the overall loss function of G. The multi-task learning based architecture achieves the related distribution mapping and representation learning with an adversarial training mechanism.

Following [28], the gradient penalty regularization is also applied in our work to avoid further vanishing gradients. Below is the loss function of D:

$$\begin{aligned} L_D = &- E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))] \\ &- \gamma E_{\tilde{x}, x\sim P_{(\tilde{x},x)}}[(\| \bigtriangledown_{\tilde{x},x} C(\tilde{x}, x)\|_2 - 1)^2] \end{aligned} \tag{9}$$

where x and x_c denote the clean and noisy speech pair subject to data distribution P_x and P_{x_c}, respectively; z denotes random noise subject to P_z; $P_{(\tilde{x},x)}$ is the joint probability of $\tilde{x} = \epsilon x + (1-\epsilon)G(z, x_c)$ and x; ϵ is sampled from a uniform distribution in $[0, 1]$, and $\gamma = 10$ is the hyper-parameter that controls the gradient penalty.

3.2. Model architecture setup

Algorithm 1 Training procedure of our speech enhancement method

Require:
Raw clean-noisy speech pairs (x, x_c) and random noise z
Initialized weights and biases of D, G, P, Q networks
Ensure:
Trained speech enhancement model
$\theta_D, \theta_G, \theta_P, \theta_Q \leftarrow$ initialize networks with Xavier-initialization [37]
for epoch (number of training iterations) **do**
  speech signal pre-processing
  $(z, x_c) \leftarrow$ batch input of G
  $G(z, x_c) \leftarrow$ enhanced output of G
  $(x, G(z, x_c)) \leftarrow$ batch input of D
  $L_D, L_G \leftarrow$ loss calculation
  $\theta_D, \theta_G, \theta_P, \theta_Q \leftarrow$ parameters update
end for
return trained model

In this subsection, we introduce the architecture of our model. As shown in Fig. 3, G is a standard downsampling and upsampling

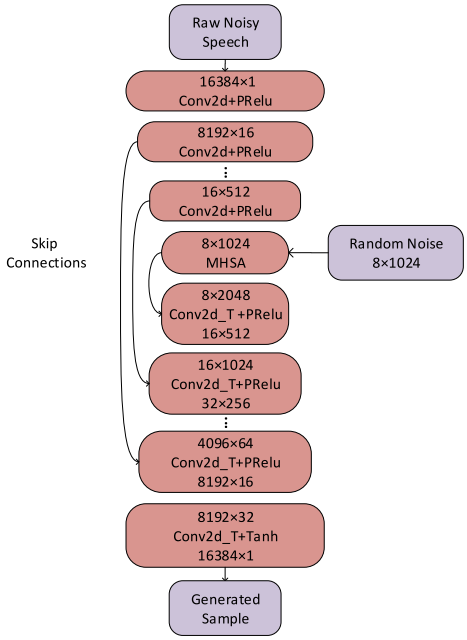


Fig. 3. The details of Generator (G). The downsampling adopts 2D convolutional kernels followed by PReLU for information capturing. The upsampling adopts 2D transposed convolutional kernels followed by PReLU for sample reconstruction. Latent vector z gets concatenated with the condensed representation of the bottleneck layer. The skip connections are used to boost the stability of model training.

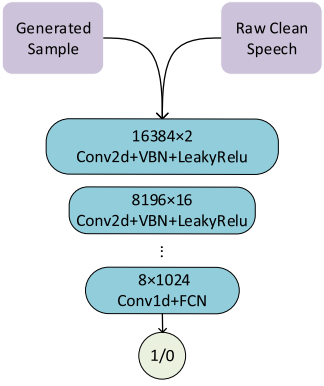


Fig. 4. The details of Discriminator (D). D is a binary classifier to give the judgment (fake or true) of the ground truth and the generated sample. The main components are the 2D convolutional kernels followed by Virtual Batch Normalization (VBN) and LeakyReLU for distinguishable information learning.

architecture developed for information learning and reconstruction. Before the intermediate bottleneck layer, the normal 2D convolutional kernels followed by Parametric Rectified Linear Units (PReLU) [18] are adopted for information capturing from real-data distributions. Then, the 2D transposed convolutional kernels with PReLU are applied for desirable sample reconstruction. Latent vector z gets concatenated with the condensed representation of the bottleneck layer. Additionally, the skip connections linking the downsampling and upsampling of G can transfer the fine-grained information of the speech waveform to the upsampling stage and boost the stability of model training [38].

As shown in Fig. 4, D is considered as a binary classifier to judge the similarity between the ground truth and the generated sample. The main components of D are also the 2D convolutional kernels but followed by Virtual Batch Normalization (VBN) and LeakyReLU [24]. This architecture is suitable for learning distinguishable information.

In this work, network P also employs 2D convolutional kernels with PReLU similar to D but removes the final fully connected layer to match the dimension of random noise. Network Q employs a downsampling and upsampling architecture similar to G but reduces the number of layers and discards the skip connections. The model's training procedure is presented in Algorithm 1 (the initialization refers to [37]). In particular, we also apply the multi-head self-attention to G and Q in the bottleneck layer for locating specific speech information and learning the contextual long-range dependencies.

3.3. Multi-head self-attention

In this subsection, we introduce multi-head self-attention carefully. The self-attention mechanism, also called intra-attention, has demonstrated a better balance of learning between modeling long-range dependencies and statistical efficiency [39]. For every input sequence, the query (q), key (k), and value (v) vectors will be created by applying learned linear projection or using feed-forward layers. Then, the attention will be applied to all other positions with the three vectors. The procedure can be described as below:

$$Attention(q, k, v) = softmax(\frac{qk^T}{\sqrt{d_k}}) \cdot v, \tag{10}$$

where d_k is the dimension of the key vectors. The purpose of this scaling is to improve numerical stability as the dimensions of keys, values, and queries grow. The obtained attention at each position will be used to times the value vector of all other positions including itself. This will produce multiple results called multi-head attention. The sum of all heads will be the final result of the first position input. The same operation will be applied to each subsequent position.

$$MultiHead(q, k, v) = concat(head_1, \ldots, head_h)W^o, \tag{11}$$

where $head_i = Attention(qW_i^q, kW_i^k, vW_i^v)$, and the matrices W_i^q, W_i^k, W_i^v, and W^o are the projection weight matrices, respectively.

The multi-head self-attention can calculate the response at a specific local position based on the resource collecting from all positions, where the attention vectors are calculated with a small computational cost. Moreover, as described in [40] that a self-attention module is complementary to the generator architecture in detail and the discriminator can also enforce complicated geometric constraints on the global structure. To further improve information capturing and the long-range dependencies learning, the multi-head self-attention is applied in this work.

4. Experiments

4.1. Database

The selected database [41] is an open and standard speech corpus for the evaluation of speech enhancement systems [41]. The database contains selected speech resources from multiple speech corpus. Some of the noise files were obtained from the DEMAND corpus.[1] Another two noise files[2] (i.e., the speech-shaped and babble noise) were also selected for noisy speech production. The original clean speech was selected from the Voice Bank corpus [42]. According to the number of speakers, two sub-databases were created: one includes 28 speakers (14 males and 14 females) with the same accent (England); another one includes

[1] http://parole.loria.fr/DEMAND/.

[2] http://homepages.inf.ed.ac.uk/cvbotinh/se/noises/.

56 speakers (28 males and 28 females) with different accents (Scotland and United States).

As mentioned above, the database added ten different noise types to the clean speech waveform using the ITU-T P.56 method [1], including eight real noise types and two artificially generated noises. In detail, the eight real noise types include a domestic kitchen room noise, a meeting room noise, three public space noises including cafeteria, restaurant, and subway station, two transportation noises including car and metro, a busy traffic intersection noise; the two artificially generated noises contains a speech-shaped noise by adding white noise and a babble noise by adding extra speech.

For training data, the Signal-to-Noise Ratio (SNR) values were set to 15 dB, 10 dB, 5 dB, and 0 dB. It signified that each clean sentence would produce 40 noisy sentences with different noise types. Each speaker contributed with 10 clean sentences. Thus, each speaker would contribute with 400 sentences to the database. Moreover, each clean speech waveform would be normalized, and the silence segments would be trimmed off when the silence segments were longer than 200 ms at the beginning and at the end.

Another two speakers (a male and a female), not included in the training data, were selected for the testing data with an accent from England. Five other noise types were selected from the DEMAND database, including a domestic living room noise, an office room noise, a transport noise of a bus, and two street noises including an open area and a public square. The SNR values were 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB, respectively.

4.2. Setup

Our model is trained using the RMSprop optimizer [43] with a learning rate of 0.0002. The number of epochs is 180 and the batch size is 100. As an end-to-end architecture, our model takes in the raw speech waveform and outputs the enhanced waveform directly, which is considered to preserve the original content of speech signals including phase information. About one-second speech chunks (16 384 samples) are segmented by a sliding window (500 ms overlap) during training, however there is no overlap during the test. Besides, a high-frequency pre-emphasis filter of coefficient 0.95 is applied to all input samples. We train the models with 2 NVIDIA GTX 1080 Ti GPUs. The model Ours_*I* and Ours_*II* spend about 2.5 and 3 days achieving convergence. The models Ours_*III* and Ours_*IV* spend about 4 days achieving convergence. The model sizes of Ours_*I*, Ours_*II*, Ours_*III*, and Ours_*IV* are 1.8G, 2.0G, 2.3G and 2.3G, respectively.

4.3. Evaluation metrics

Although subjective evaluation is more accurate and reliable, it is costly and time-consuming [48]. Many objective evaluation measures can evaluate enhanced speech with high correlation. The Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5) for wideband speech is an effective full-reference speech quality evaluation algorithm [48]. Moreover, we also implement the evaluation metrics of the enhanced speech including the Composite mean opinion score predictor of Signal distortion (CSIG: from 1 to 5), Background noise distortion (CBAK: from 1 to 5), and Overall quality (COVL: from 1 to 5). The Segmental Signal-to-Noise Ratio (SSNR: from 0 to ∞) is another crucial evaluation metric for speech quality.

The intelligibility of enhanced speech is also tested in this work. The Coherence-based Speech Intelligibility Index (CSII) measure is computed for the low-level high-level (CSII_{high}), medium-level (CSII_{mid}), and (CSII_{low}) segments of each speech sentence, which can predict the intelligibility of peak-clipping and centering-clipping distortions in a speech signal [49]. Besides, the Normalized Covariance Metric (NCM) [49] and the Short-Time Objective Intelligibility (STOI) [50] are also conducted for intelligibility evaluation of enhanced speech.

Table 1
The evaluation results of our method compared with previous methods including Wiener filtering [44], SEGAN [24], SERGAN [28], MMSE-GAN [31], BiLSTM [45], CRN-MSN [46], NAAGN [47]. All the presented methods were trained with the 28-speaker database. "†" denotes that we reproduced the results with the provided open resource. "-" denotes that the result is not reported or not available. The best scores are highlighted in bold.

Models	PESQ	CSIG	CBAK	CVOL	SSNR	STOI
Noisy	1.97	3.35	2.44	2.63	1.68	0.921
Wiener filtering	2.22	3.23	2.68	2.67	5.07	–
SEGAN†	2.16	3.48	2.94	2.80	7.73	0.928
SERGAN†	2.52	3.66	3.18	3.06	9.40	0.937
MMSE-GAN	2.53	3.80	3.12	3.14	–	0.930
BiLSTM	2.70	3.99	2.95	3.34	–	0.925
MDPhD	2.70	3.85	3.39	3.27	10.2	–
CRN-MSE	2.74	3.86	3.14	3.30	–	0.934
NAAGN	**2.90**	**4.13**	**3.50**	3.51	10.3	**0.948**
Ours_$I_{(\lambda_1=1,\lambda_2=0)}$	2.57	3.78	3.23	3.16	9.32	0.937
Ours_$II_{(\lambda_1=0,\lambda_2=1)}$	2.52	3.73	3.22	3.11	9.06	0.935
Ours_$III_{(\lambda_1=1,\lambda_2=1)}$	2.79	3.90	3.34	**3.56**	9.67	0.941
Ours_$IV_{(\lambda_1=1,\lambda_2=1,MHSA)}$	2.88	4.01	**3.50**	3.51	9.72	0.945

5. Results and discussions

In this section, we report the experimental results and give the related discussion. We firstly conduct our experiments on 28-speaker database. The experimental results are demonstrated in terms of several main evaluation metrics including PESQ, CSIG, CBAK, CVOL, SSNR, and STOI as shown in Table 1.

In our ablation experiments, we activate the networks P and Q with corresponding loss functions by controlling the parameter λ for inverse mapping learning from output space to input space. We adopt SERGAN architecture in this paper. Thus, the experimental results are the same as SERGAN† if we set $\lambda_1 = 0, \lambda_2 = 0$. We just activate the network P and learn the inverse mapping from the generated space to the latent space when we set $\lambda_1 = 1, \lambda_2 = 0$. Compared with the original architecture (SERGAN), the experimental results show that our method achieves higher evaluation scores in terms of PESQ (2.52 to 2.57), CSIG (3.66 to 3.78), CBAK (3.18 to 3.23), CVOL (3.06 to 3.16), which relatively improves by 1.98%, 3.28%, 1.57%, and 5.23%, respectively. Moreover, we also find that the evaluation score decreases slightly in terms of SSNR (9.40 to 9.32) and remains the same in terms of STOI (0.937). We can infer that the proposed method can further improve the speech representation learning and speech enhancement performance.

Further, when we activate Q and inactivate P (i.e., $\lambda_1 = 0, \lambda_2 = 1$), the performance degrades slightly compared with the first ablation experiment (i.e., $\lambda_1 = 1, \lambda_2 = 0$) but still obtains a slight improvement compared with the original SERGAN architecture. We infer that re-excavating information from the input data domain by inverse mapping learning can improve representation learning and speech enhancement performance. However, the network P learning inverse mapping from the generated data domain to the latent domain is more effective in speech enhancement improvement than network Q.

Naturally, when we activate P and Q simultaneously (i.e., $\lambda_1 = 1, \lambda_2 = 1$), our method further improves the enhancement performance. Thus, we can infer that our method can improve speech representation learning and speech enhancement performance by inverse mapping learning. Moreover, when we add the multi-head self-attention layer in the bottleneck layer of network G

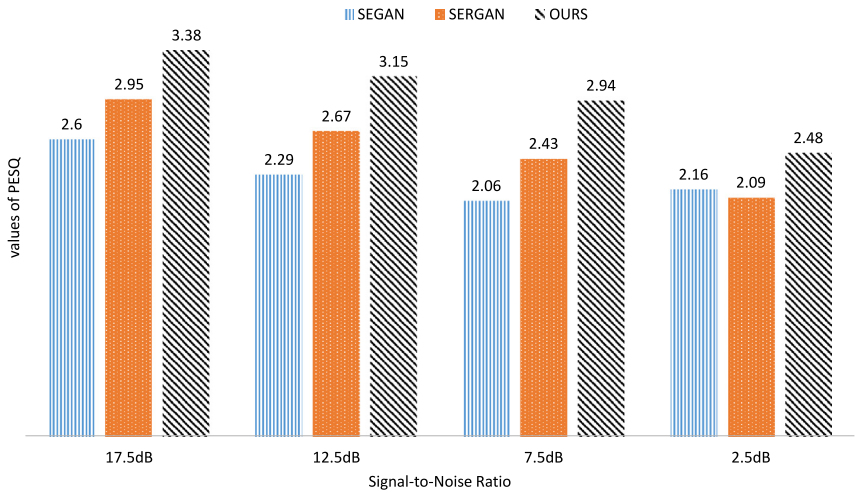


Fig. 5. The bar plot of PESQ on different SNR values (i.e., 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB). The compared methods are SEGAN, SERGAN, and Ours_*IV*.

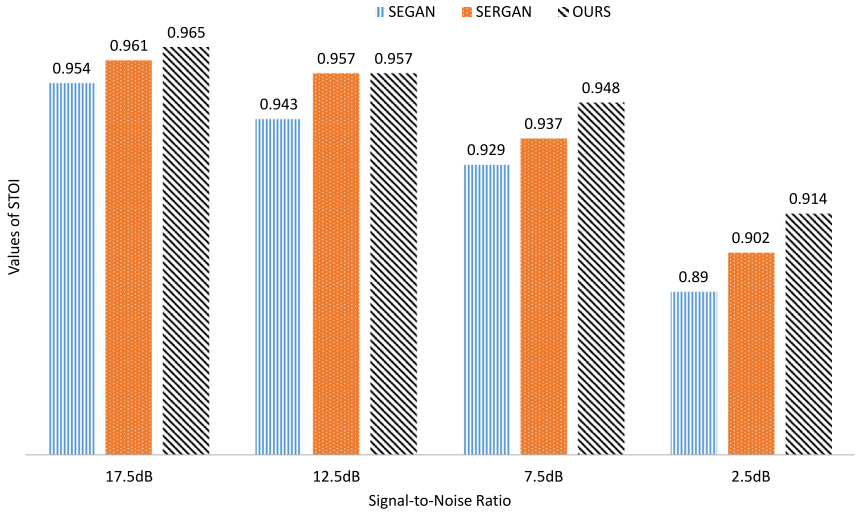


Fig. 6. The bar plot of STOI on different SNR values (i.e., 17.5 dB, 12.5 dB, 7.5 dB, 2.5 dB). The compared methods are SEGAN, SERGAN, and Ours_*IV*.

and Q, the evaluation results are further improved and obtain a competitive score compared with the state-of-the-art method (i.e., NAAGN [47]). Although more values of λ_1 and λ_2 have been tested, the most representative experimental results are presented in Table 1.

To further demonstrate the effectiveness of our proposed method, we also unfold the details of the evaluation results with two reproducible methods (i.e., SEGAN [24] and SERGAN [28]) with a more comprehensive evaluation metric. As shown in Table 2, our method improves speech enhancement performance in each SNR condition and evaluation metrics. Moreover, through careful comparison from high SNR to low SNR (17.5 dB to 2.5 dB), we find that our method performs better in lower SNR. In particular, the intelligibility improvement is dramatic. For example, in the 17.5 dB scene, the NCM evaluation score of our method (0.997) is similar to other methods (0.994 and 0.997). However, in 2.5 dB, the NCM score of our method (0.967) is much higher than other methods (0.928 and 0.959). With the same training resource, the models SEGAN, SERGAN, and Ours_*IV* spend about 2 days, 2.2 days, and 4 days achieving convergence, respectively. The trained model sizes of SEGAN, SERGAN and Ours_*IV* are about 1.1G, 1.4G, and 2.3G, respectively. Compared with single-task learning, multi-task learning generally increases the computational overhead. However, in consideration of performance gain, the increased computational load is acceptable.

To visualize the performance of compared methods on different SNR, we present the histogram of PESQ in Fig. 5 and STOI in Fig. 6. We can see the obvious improvements of our method in terms of both speech intelligibility and quality. We also present the spectrograms of four selected speech utterances in different SNR in Fig. 7 (2.5 dB and 7.5 dB) and Fig. 8 (12.5 dB and 17.5 dB). From the top to bottom, the figures are noisy, clean, and enhanced speech waveforms. We can observe that our method can enhance noisy speech. Compared to Figs. 7 and 8, we can see that the enhancement performance in low SNR (2.5 dB and 7.5 dB) is more effective than in high SNR (12.5 dB and 17.5 dB).

Moreover, we conduct experiments in different sizes of training data to further explore the effectiveness of our method. As we can see from Table 3, Ours_*IV* obtains further improvement in terms of speech quality and intelligibility with more training data. However, not all the evaluation metrics can achieve further improvement on this occasion. We speculate that the results may be caused by the different distribution of training data.

6. Conclusions

In this paper, we propose a novel adversarial multi-task learning with inverse mapping method for speech enhancement. The proposed networks P and Q establish the inverse mapping from

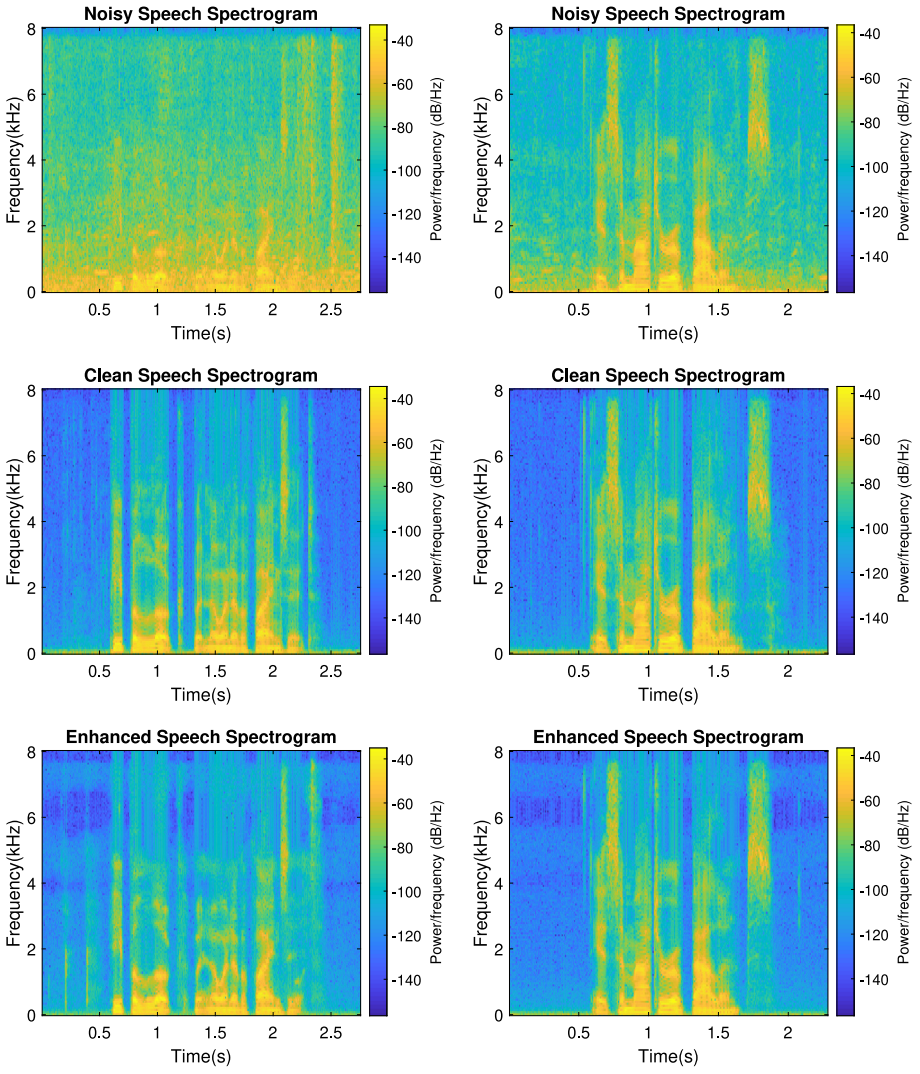


Fig. 7. Spectrograms of selected utterance in different SNR enhanced with our method. Top: noisy speech. Middle: clean speech. Bottom: enhanced speech (a) SNR = 2.5 dB, (b) SNR = 7.5 dB.

Table 2
The unfolded evaluation results on different SNR values (i.e., 17.5 dB, 12.5 dB, 7.5 dB, 2.5 dB, and overall). We evaluate SEGAN [24], SERGAN [28] and our method with more comprehensive speech quality and intelligibility metrics on the 28-speaker database. "†" denotes that we reproduced the results with the provided open resources.

Strategies		Quality					Intelligibility				
Methods	SNR	PESQ	CSIG	CBAK	CVOL	SSNR	$CSII_{high}$	$CSII_{mid}$	$CSII_{low}$	NCM	STOI
SEGAN†	17.5 dB	2.60	3.93	3.28	3.26	9.23	0.997	0.956	0.684	0.994	0.954
	12.5 dB	2.29	3.65	3.06	2.96	8.46	0.991	0.911	0.587	0.989	0.943
	7.5 dB	2.06	3.36	2.87	2.69	7.52	0.977	0.852	0.486	0.972	0.929
	2.5 dB	1.76	3.02	2.59	2.35	5.88	0.931	0.748	0.338	0.928	0.890
	Overall	2.16	3.48	2.94	2.80	7.73	0.973	0.864	0.518	0.970	0.928
SERGAN†	17.5 dB	2.95	4.10	3.56	3.51	11.7	0.998	0.969	0.738	0.997	0.961
	12.5 dB	2.67	3.81	3.31	3.21	10.2	0.994	0.934	0.644	0.994	0.957
	7.5 dB	2.43	3.57	3.09	2.97	8.83	0.982	0.879	0.561	0.985	0.937
	2.5 dB	2.09	3.22	2.79	2.61	7.13	0.946	0.784	0.425	0.959	0.902
	Overall	2.52	3.66	3.18	3.06	9.40	0.979	0.889	0.587	0.983	0.937
Ours_*IV*	17.5 dB	3.38	4.40	3.65	3.90	12.0	0.999	0.970	0.731	0.997	0.965
	12.5 dB	3.15	4.20	3.36	3.66	10.5	0.996	0.938	0.651	0.996	0.957
	7.5 dB	2.94	3.97	3.13	3.46	9.21	0.987	0.888	0.573	0.989	0.948
	2.5 dB	2.48	3.51	2.82	3.10	7.40	0.957	0.795	0.443	0.967	0.914
	Overall	2.88	4.01	3.50	3.51	9.72	0.984	0.895	0.595	0.987	0.945

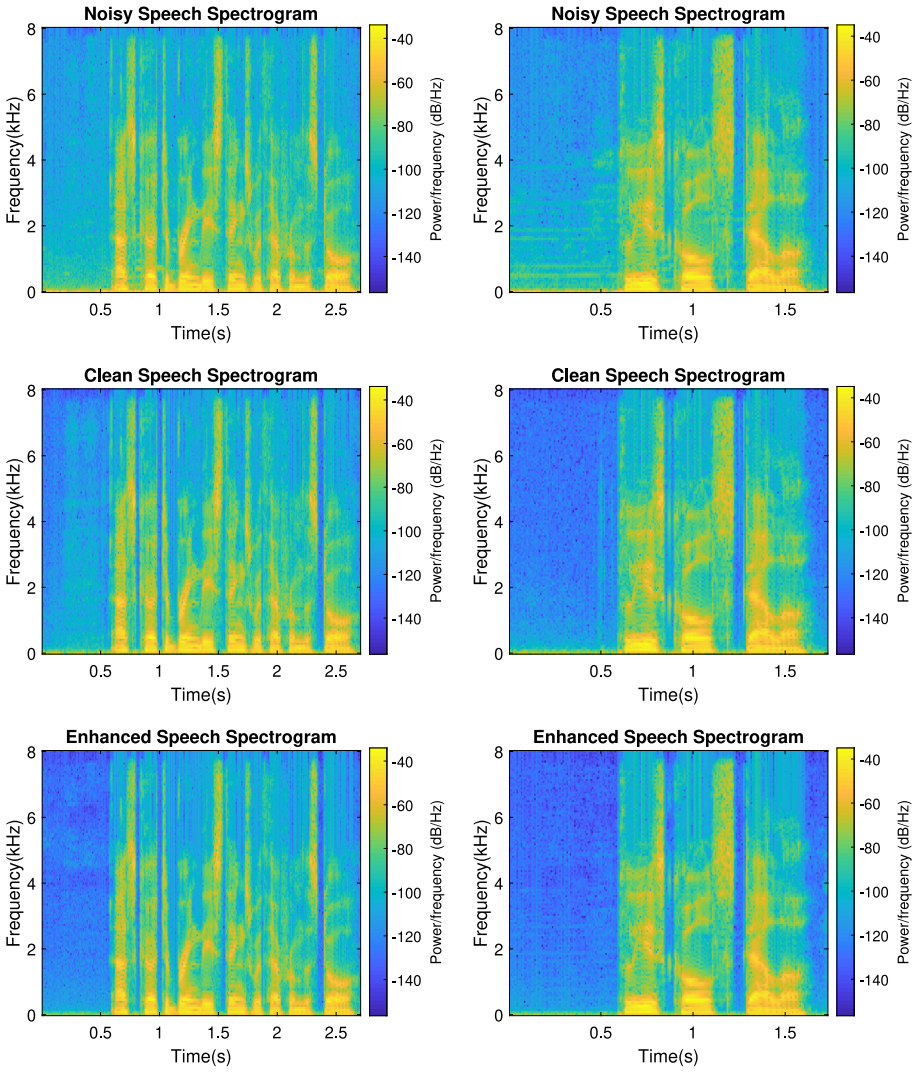


Fig. 8. Spectrograms of selected utterance in different SNR enhanced with our method. Top: noisy speech. Middle: clean speech. Bottom: enhanced speech (a) SNR = 12.5 dB, (b) SNR = 17.5 dB.

Table 3
The evaluation results of SEGAN [24], SERGAN [28] and our method (i.e., Ours_*IV*) with different scales of training database in terms of speech quality and intelligibility. "†" denotes that we reproduced the results with the provided open resources.

Strategies		Quality					Intelligibility				
Methods	Training data	PESQ	CSIG	CBAK	CVOL	SSNR	$CSII_{high}$	$CSII_{mid}$	$CSII_{low}$	NCM	STOI
SEGAN†	28spks	2.16	3.48	2.94	2.80	7.73	0.973	0.864	0.518	0.970	0.928
	56spks	2.46	3.63	3.10	3.03	7.83	0.977	0.879	0.554	0.973	0.934
	28+56spks	2.51	3.46	3.15	2.95	8.99	0.981	0.890	0.554	0.976	0.937
SERGAN†	28spks	2.52	3.66	3.18	3.06	9.40	0.979	0.889	0.587	0.983	0.937
	56spks	2.61	3.89	3.25	3.24	9.03	0.980	0.890	0.587	0.984	0.938
	28+56spks	2.60	3.79	3.28	3.18	9.64	0.981	0.893	0.593	0.983	0.938
Ours_*IV*	28spks	2.88	4.01	3.50	3.51	**9.72**	**0.984**	0.895	0.595	0.987	0.945
	56spks	2.90	4.03	**3.55**	3.52	9.70	0.983	**0.898**	0.593	0.987	**0.947**
	28+56spks	**2.93**	**4.08**	3.48	**3.55**	9.46	0.983	0.895	**0.596**	**0.988**	0.946

the generated distribution to the input data space. Based on the adversarial multi-task learning, our method effectively enhances the generator's capability of speech information capturing and representation learning. Our method obtains the state-of-the-art performance in terms of speech quality (i.e., PESQ = 2.93, CVOL = 3.55). For speech intelligibility, the proposed method obtains competitive performance (e.g., STOI = 0.947). The experimental results demonstrate that our proposed method can greatly improve speech enhancement performance in terms of speech quality and intelligibility, especially in a low SNR scene. Moreover, the multi-head self-attention is also effective to locate specific information and learn long-range dependencies of the speech signals, and improve speech enhancement further.

In future work, we will investigate our proposed method on different evaluation resources, reduce the computational load and

time complexity. In addition, we will try to apply our speech enhancement method to more applications, such as robust speaker identification and speech recognition.

CRediT authorship contribution statement

Yuanhang Qiu: Conceptualization, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft preparation, Writing– reviewing and editing. **Ruili Wang:** Conceptualization, Investigation, Project administration, Resources, Supervision, Writing– reviewing and editing. **Feng Hou:** Resources, Writing– reviewing and editing. **Satwinder Singh:** Resources, Writing– reviewing and editing. **Zhizhong Ma:** Resources, Writing– reviewing and editing. **Xiaoyun Jia:** Writing– reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the China Scholarship Council (CSC), and 2020 MBIE Catalyst: Strategic – New Zealand-Singapore Data Science Research Programme, New Zealand.

References

[1] P.C. Loizou, Speech Enhancement: Theory and Practice, second ed., CRC Press, Inc., USA, 2013.
[2] J. Benesty, J.R. Jensen, M.G. Christensen, J. Chen, Speech Enhancement: A Signal Subspace Perspective, Elsevier, 2014.
[3] M.S. Kavalekalam, J.K. Nielsen, J.B. Boldt, M.G. Christensen, Model-based speech enhancement for intelligibility improvement in binaural hearing aids, IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 27 (1) (2019) 99–113, http://dx.doi.org/10.1109/TASLP.2018.2872128.
[4] Y.-H. Tu, J. Du, C.-H. Lee, Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 27 (12) (2019) 2080–2091, http://dx.doi.org/10.1109/TASLP.2019.2940662.
[5] H. Taherian, Z.-Q. Wang, J. Chang, D. Wang, Robust speaker recognition based on single-channel and multi-channel speech enhancement, IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 28 (2020) 1293–1302, http://dx.doi.org/10.1109/TASLP.2020.2986896.
[6] M. Abd El-Fattah, M.I. Dessouky, S.M. Diab, F.E.-S. Abd El-Samie, Speech enhancement using an adaptive wiener filtering approach, Prog. Electromagn. Res. 4 (2008) 167–184, http://dx.doi.org/10.2528/PIERM08061206.
[7] X. Cui, Z. Chen, F. Yin, Multi-objective based multi-channel speech enhancement with BiLSTM network, Appl. Acoust. 177 (2021) 107927.
[8] Z. Zhang, J. Geiger, J. Pohjalainen, A.E.-D. Mousa, W. Jin, B. Schuller, Deep learning for environmentally robust speech recognition: An overview of recent developments, ACM Trans. Intell. Syst. Technol. (TIST) 9 (5) (2018) 1–28, http://dx.doi.org/10.1145/3178115.
[9] X. Li, L. Girin, S. Gannot, R. Horaud, Multichannel online dereverberation based on spectral magnitude inverse filtering, IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 27 (9) (2019) 1365–1377, http://dx.doi.org/10.1109/TASLP.2019.2919183.
[10] G. Saon, Z. Tüske, D. Bolanos, B. Kingsbury, Advancing RNN transducer technology for speech recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5654–5658.
[11] Z. Liu, Z. Li, R. Wang, M. Zong, W. Ji, Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition, Neural Comput. Appl. 32 (18) (2020) 14593–14602.
[12] F. Hou, R. Wang, J. He, Y. Zhou, Improving entity linking through semantic reinforced entity embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2020, pp. 6843–6848, http://dx.doi.org/10.18653/v1/2020.acl-main.612.
[13] T. Gao, J. Du, L.-R. Dai, C.-H. Lee, Densely connected progressive learning for LSTM-based speech enhancement, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5054–5058, http://dx.doi.org/10.1109/ICASSP.2018.8461861.
[14] Z. Ouyang, H. Yu, W.-P. Zhu, B. Champagne, A fully convolutional neural network for complex spectrogram processing in speech enhancement, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5756–5760, http://dx.doi.org/10.1109/ICASSP.2019.8683423.
[15] H. Zhao, S. Zarar, I. Tashev, C.-H. Lee, Convolutional-recurrent neural networks for speech enhancement, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 2401–2405, http://dx.doi.org/10.1109/ICASSP.2018.8462155.
[16] A. Nicolson, K.K. Paliwal, Deep learning for minimum mean-square error approaches to speech enhancement, Speech Commun. 111 (2019) 44–55, http://dx.doi.org/10.1016/j.specom.2019.06.002.
[17] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 31, 2018, pp. 1–17.
[18] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034, http://dx.doi.org/10.1109/ICCV.2015.123.
[19] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 448–456.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[21] Z. Meng, J. Li, Y. Gong, B.-H.F. Juang, Cycle-consistent speech enhancement, in: Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 1165–1169, http://dx.doi.org/10.21437/Interspeech.2018-2409.
[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.
[23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 1–35, http://dx.doi.org/10.1007/978-3-319-58347-1_10.
[24] S. Pascual, A. Bonafonte, J. Serrà, SEGAN: speech enhancement generative adversarial network, in: Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017, pp. 3642–3646, http://dx.doi.org/10.21437/Interspeech.2017-1428.
[25] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134, http://dx.doi.org/10.1109/CVPR.2017.632.
[26] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2794–2802.
[27] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN Interpretable representation learning by information maximizing generative adversarial nets, in: Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS), 2016, pp. 2172–2180.
[28] D. Baby, S. Verhulst, Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 106–110, http://dx.doi.org/10.1109/ICASSP.2019.8683799.
[29] A. Jolicoeur-Martineau, The relativistic discriminator: A key element missing from standard GAN, in: Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019, pp. 1–26.
[30] D. Michelsanti, Z.-H. Tan, Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification, in: Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017, pp. 2008–2012, http://dx.doi.org/10.21437/Interspeech.2017-1620.
[31] M.H. Soni, N. Shah, H.A. Patil, Time-frequency masking-based speech enhancement using generative adversarial network, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5039–5043, http://dx.doi.org/10.13140/RG.2.2.19312.15365.
[32] C. Donahue, B. Li, R. Prabhavalkar, Exploring speech enhancement with generative adversarial networks for robust speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5024–5028, http://dx.doi.org/10.1109/ICASSP.2018.8462581.

[33] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, in: Proceedings of the 5th International Conference on Learning Representations (ICLR), 2017, pp. 1–18.
[34] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, A. Courville, Adversarially learned inference, in: Proceedings of the 5th International Conference on Learning Representations (ICLR), 2017, pp. 1–18.
[35] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, S. Belongie, Stacked generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5077–5086, http://dx.doi.org/10.1109/CVPR.2017.202.
[36] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2016, pp. 658–666.
[37] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[38] S. Pascual, J. Serrà, A. Bonafonte, Towards generalized speech enhancement with generative adversarial networks, in: Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2019, pp. 1791–1795, http://dx.doi.org/10.21437/Interspeech.2019-2688.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[40] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 7354–7363.
[41] C. Valentini-Botinhao, X. Wang, S. Takaki, J. Yamagishi, Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks, in: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016, pp. 352–356, http://dx.doi.org/10.21437/Interspeech.2016-159.
[42] C. Veaux, J. Yamagishi, S. King, The voice bank corpus: Design, collection and data analysis of a large regional accent speech database, in: Proceedings of Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), IEEE, 2013, pp. 1–4, http://dx.doi.org/10.1109/ICSDA.2013.6709856.
[43] F. Zou, L. Shen, Z. Jie, W. Zhang, W. Liu, A sufficient condition for convergences of adam and rmsprop, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11127–11135, http://dx.doi.org/10.1109/CVPR.2019.01138.
[44] J. Lim, A. Oppenheim, All-pole modeling of degraded speech, IEEE Trans. Audio Speech Lang. Process. (TASLP) 26 (3) (1978) 197–210, http://dx.doi.org/10.1109/TASSP.1978.1163086.
[45] H. Erdogan, J.R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 708–712, http://dx.doi.org/10.1109/ICASSP.2015.7178061.
[46] K. Tan, D. Wang, A convolutional recurrent neural network for real-time speech enhancement, in: Proceedings of the 19th Conference of the International Speech Communication Association (INTERSPEECH), Vol. 2018, 2018, pp. 3229–3233, http://dx.doi.org/10.21437/Interspeech.2018-1405.
[47] F. Deng, T. Jiang, X. Wang, C. Zhang, Y. Li, NAAGN: Noise-aware attention-gated network for speech enhancement, in: Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2020, pp. 2457–2461.
[48] Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Audio Speech Lang. Process. (TASLP) 16 (1) (2007) 229–238, http://dx.doi.org/10.1109/TASL.2007.911054.
[49] J. Ma, Y. Hu, P.C. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, J. Acoust. Soc. Am. 125 (5) (2009) 3387–3405, http://dx.doi.org/10.1121/1.3097493.
[50] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Trans. Audio Speech Lang. Process. (TASLP) 19 (7) (2011) 2125–2136, http://dx.doi.org/10.1109/TASL.2011.2114881.