.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Method of Underwater Acoustic Signal Denoising Based on Dual-Path Transformer Network

YONGQIANG SONG[1] FENG LIUR[2], AND TONGSHENG SHEN.[2], (Member, IEEE)
[1]National Innovation Institute of Defense Technology, Beijing, 100071, China
[2]Chinese Academy of Military Science, Beijing, 100071, China

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This research was funded by the National Natural Science Foundation of China, grant number 41906169.

ABSTRACT The presence of natural ambient noise interferes with the system for locating and identifying underwater targets. This paper suggests that a Dual-Path Transformation Network(DPTN) reduces ambient noise in underwater acoustic signals. First, the input acoustic signals' higher-order non-linear features are extracted using a multi-scale convolutional encoder neural network. Second, sub-vectors with the same length are created according to the time dimension from the higher-order non-linear features. The sub-vectors are stitched together to form a three-dimensional tensor. Third, a neural network transformer based on the feed-forward network is constructed. Further, to capture long-term series features and separate the target signal from the noisy signals, the three-dimensional tensor is used as the input of the transformer-based masking network. Finally, overlap-add and transpose are used to obtain discernible target signals. The experimental results verify the effectiveness of the proposed underwater acoustic signal denoising algorithm and demonstrate that the proposed DPRN method can obtain higher output signal-to-noise ratio (SNR) and the scale-invariant signal-to-noise ratio (SI-SNR) compared with the other classical algorithms.

INDEX TERMS Deep learning; Underwater acoustics; Dual-Path transformer network

I. INTRODUCTION

The study of underwater acoustics is a crucial foundation for the passive sonar systems used by underwater vehicles to detect, track, localize, and identify targets in the marine environment [1]. However, the feature extraction of underwater acoustic signals by modern sonar systems is fraught with challenges due to the complex natural environment, variable acoustic sources, and high noise intensity. Multi-target signal mixing, complex spectral components, and interspersed noise signals are the leading causes of these issues because they lower the signal-to-noise ratio of the sonar acceptance signal. Therefore, it is necessary to implement noise reduction processing for the signals. At this stage, conventional denoising methods usually use transformation analysis methods and time-to-frequency conversion methods to achieve signal enhancement [2] [3]. For example, Amplitude-Aware Permutation Entropy [4] Wavelet and Block Thresholding [5] [6], multi-directional filters [7], Minimum Variance Distortionless Response [8], Variational Approach [9], Wavelet Transform [10], Empirical Mode Decomposition [11] [12] [13], Linear Spectrum [14], Singular Value Decomposition [15].

Conventional methods for improving underwater acoustic signals have apparent drawbacks in terms of signal-to-noise ratio. Large-scale data and ocean environment parameters cannot be modeled using the constrained conditions of conventional methods. It is not easy to iteratively update the data after system deployment, making the underwater acoustic signals detection for a long time dependent on manual interpretation. The background noise characteristics vary with the ocean scene and can produce considerable dynamic deviations at different times. Underwater acoustic signals are subject to ambient and machine interference, which can cause propagation attenuation and distortion, resulting in a relatively low signal-to-noise ratio of the signal received by the detection system. conventional methods in fitting the model need to know the frequency range of the signal affected by propagation characteristics, target interference, transient signals, and many other factors. There is a significant gap between the frequency range of different noises, which often leads to the model and noise type being unsuitable. The emergence of these problems makes conventional algorithms face many challenges.

Deep learning-based models have great potential for denoising applications compared with conventional methods. Deep learning can learn correlations before and after time series through neural networks. In addition, the introduction of filters allows the network to automatically capture the underlying features of the signal. This process eliminates manual extraction and automatically compiles the sequence in a coded fashion. This process eliminates manual extraction and automatically compiles the sequence in a coded fashion. The model's noise reduction effect can be improved by adjusting the parameters and modifying the network architecture. For example, Zhou [16] proposes the Denoising Representation Recognition (DRR) model that converts the spectrum to a correlation coefficient to generate data for parallel training in the convolutional denoising autoencoder(CDAE) model. Wang [17] proposes a novel stacked convolutional sparse denoising autoencoder (SCSDA) model to complete the blind denoising task of underwater heterogeneous information data. The stacked sparse denoising autoencoder (SSDA) was constructed by three sparse denoising autoencoders (SDA) to extract overcomplete sparse features. The output of the last encoding layer of the SSDA was used as the input of the convolutional neural network (CNN) to extract the features.Othman [18] proposes a residual deep neural network(Resnet), which works onIIR Wiener filter integration to reduce noise. Alberto [19] proposes a novel approach based on a computationally and energy-efficient deep convolutional denoising autoencoder to reduce the noise interference. Qiu [20] proposes a reinforcement learning system that requires sophisticated design and critical parameter choice to meet its oscillatory condition to keep the balance among signals, noise, and the nonlinear system. Li [21] proposes an approach based on relativistic conditional generative adversarial networks (RCGAN) to resolve the conditions of complex marine ambient noise and scarce data. Xing [22] trains and updates the noisy signal via orthogonal matching pursuit (OMP) and method of optimal directions (MOD). The signal reconstruction is completed according to the updated dictionary and sparse coefficients. To summarise, the underwater acoustic signal denoising methods can be broadly classified into three processing modes, time-domain processing, spectrogram mapping, and mask separation, as shown in Figure 1.

The mask separation method is now the dominant approach compared to the first two methods, as shown in Figure 2. Firstly, a set of features (e.g., Mel Frequency Cepstrum Coefficient, MFCC) is learned from the underwater acoustic signal using the auditory properties of the human ear. A filter is used to encode the features into a masking network to estimate the mask parameters for each source. Finally, a decoder is used to recover the masked higher-order features into the underwater acoustic signal. Recurrent neural networks (e.g., Long Short-Term Memory and Gated Recurrent Unit) can be used in many areas of signal processing. Deep learning methods built on recurrent neural networks are essential to modern signal processing. They are often used as a vital module for mask separation in the field of underwater acoustic signal noise reduction [23] [24] [25]. However, the inherent sequential processing sequence mode of the Recurrent Neural Network (RNN) prevents the model from parallelizing the computation during training. After the model feeds the audio file to the encoder via the index log, the RNN-based masking process takes up most of the CUDA memory space. When the natural ambient noise loudness masks the target signal, gradient disappearance and explosion are often encountered during training using RNN models. It is challenging to calculate the multiplicative gradients, which can vary exponentially with the number of layers due to the difficulty in capturing long-term dependencies with RNN.

In order to obtain the denoised acoustic signal, we construct a Dual-Path Transformer Network (DPTN) based on the abovementioned analysis. This network combines the techniques of multi-head attention transformer [26] and Dual-path framework processing. This paper has three contributions:

1. The one-dimensional convolutional neural network is built to construct the encoder module, and a chunking operation is used to convert the feature vector into a three-dimensional tensor. We cascade the encoder and chunk operation as a feature extraction module for underwater acoustic signals.

2. The multi-head attention converter and Dual-path framework construct the separation network to extract the feature mask of the target signal. Further, the denoised signal tensor is restored to the original signal format using overlap-add and transposition network.

3. The Dual-Path transformer network (DPTN) can simultaneously focus on the complete underwater acoustic sequence information and process all time step points in parallel. Relevance between far-off sampling points is achieved by adding a feedforward network and residual connections to the masking network, making it more straightforward for the model to learn the long-term dependencies of the signal.

The specific content is arranged as follows: the Section.2 details the flow of the method, namely Dual-Path Transformer Network (DPTN). The Section.3 uses the ablation experiment and comparison test to analyse the influence of each step of the method in this paper, and the conclusion is in Section.4.

II. MATERIALS AND METHODS

The proposed model is based on the learned-domain masking approach and employs a feature extraction module, masking separation module, and signal reconfiguration module, as shown in Figure 3.

The feature extraction encode module is a one-dimensional convolutional neural network that estimates a learned representation for the input signal. It learns a complete set of state changes and the dynamic parameters of the filter from the

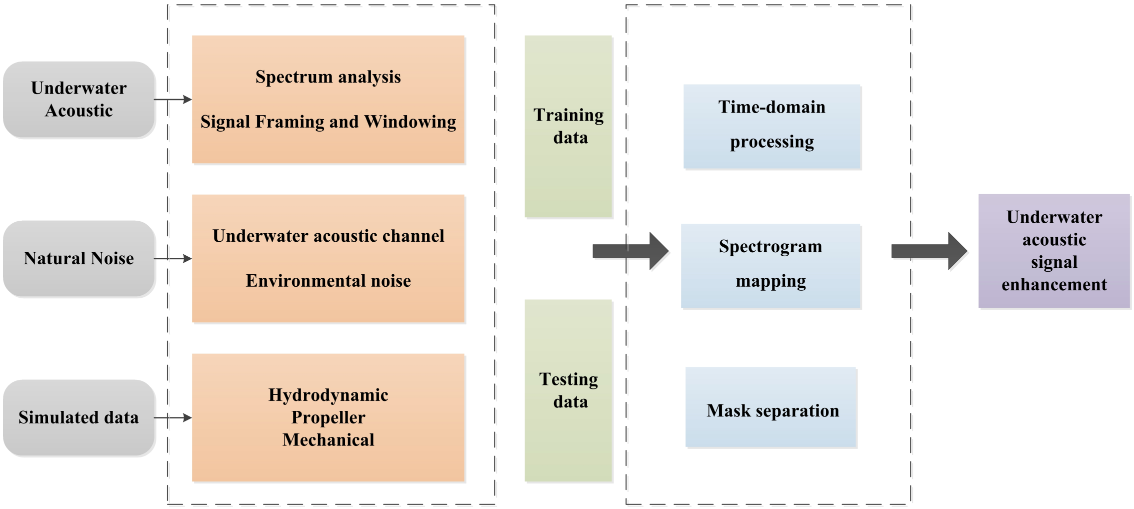

FIGURE 1. A brief description of the acoustic signal denoising methods

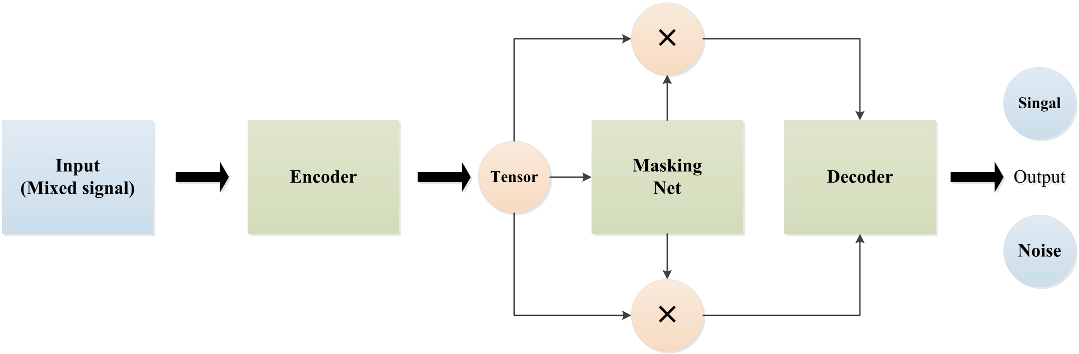

FIGURE 2. A brief description of the mask separation method

existing knowledge. Afterward, the feature vector is sliced into sub-sequences of the same size without changing the dimensionality of the features. Finally, the subsequence is built into a 3D tensor using chunking as the input to the masking network. The masking network employs two transformers embedded inside the Dual-path processing block. The masking network can be divided into two main parts: intra-transformer and inter-transformer, which rely on short-term dependencies and long-term dependencies to learn the target signal and ambient noise, respectively. Moreover, the different characteristics of the signal are learned by short-term and long-term dependencies, respectively. The mask of the target signal is extracted by matrix amplification and compression operations. The mask and the higher-order features of the signal are point multiply calculations by Hadamard functions.Finally, the signal reconfiguration decode module reconstructs the underwater acoustic signals in the time domain.

A. FEATURE EXTRACTION ENCODE MODULE

The encoder takes in time-domain mixture underwater acoustic signal $x \in R^T$ (where T is the time duration of the input signal) as input, which contains target signal and ambient noise signal. It learns an high-dimensional representation $h \in R^{F \times T}$ (where F is the feature dimension of the input signal) using the one-dimensional convolutional neural network (Conv1D) and Relu activation function :

$$h = Relu(Conv1D(x)) \tag{1}$$

B. SIGNAL RECONFIGURATION DECODE MODULE

The decoder uses a transposed convolution network, with the same stride and kernel size of the encoder. The input to the decoder is the element-wise multiplication between the mask m_{target} and the output of the encoder h. Therefore, the transformation of the decoder can be expressed as follows:

$$S_{target} = TransposeConv1D(m_{target} * h) \tag{2}$$

Where S_{target} denotes the clearly target underwater acoustic signal.

C. TRANSFORMER BLOCK

The Dual-path approach is used in the transformer block to model both short-term and long-term dependencies. Figure 4

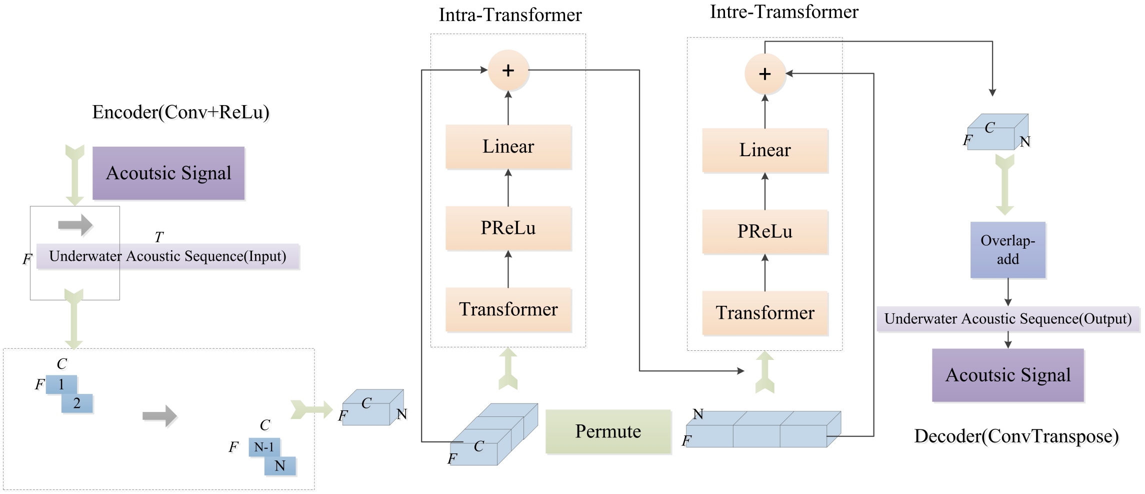

FIGURE 3. A brief description of Dual-Path Transformer Network (DPTN)

illustrates our model's transformer block, intra-Transformer for modeling short-term dependencies and inter-Transformer for modeling longer-term dependencies.

Intra-Transformer processes the second dimension of $h' \in R^{F\times C\times N_c}$, and thus acts on each chunk independently, modeling the short-term dependencies within each chunk. Next, we permute the last two dimensions (which we denote with P) and the inter-Transformer is applied to model the transitions across chunks. This scheme enables effective modelling of long-term dependencies across the chunks. The overall transformation of the transformer is therefore defined as follows:

$$h'' = f_{inter}\left(P\left(f_{intra}\left(h'\right)\right)\right) \tag{3}$$

where f_{inter} is the inter-Transformer operation and f_{intra} is the intra-Transformer operation. Figure 5 shows the architecture of the multi-head attention used to build for both the intra-Transformer and inter-Transformer blocks.

Assume that the feature tensor g is the input to intra-Transformer. First of all, sinusoidal positional encoding (PE) function is added to the input g:

$$g' = g + PE_{pos} \tag{4}$$

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{1000^{(2i)/d_{model}}}\right) \tag{5}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{1000^{(2i+1)/d_{model}}}\right) \tag{6}$$

We must introduce some information about the relative or absolute position of the tokens in the sequence since our model lacks recursive operations and convolution kernels that act on the sequence order. Then, several layers of deformers are applied, and inside each transformer layer $L(.)$, we first apply layer normalization, followed by the multi-head self-attention (MSA):

$$g'' = MSA\left(LayerNormalazation\left(g'\right)\right) \tag{7}$$

The attention allows the model to jointly attend to information from different representation subspaces at different positions. Each attention head computes the scaled dot-product attention between all the sequence elements. The input consists of queries (Q), keys (K), and values (V) of the dimension. We compute the dot products of the query with all keys, divide each by and apply a softmax function to obtain weights on the values. We simultaneously compute the attention function on a set of queries packed into a matrix Q. The keys and values are also packed into matrices K and V. We compute the parameters matrix of outputs as:

$$Attention\left(Q,K,V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{8}$$

The multi-head self-attention (MSA) (9) allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this, which enhances the expressiveness of each attention layer without changing the number of parameters.

$$MSA(Q,K,V) = Concat(head_1, ..., head_i)W^o \tag{9}$$

$$head_i = attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{10}$$

Where the projections are parameter matrices $W_i^Q \subset R^{d_{model}\times d_k}$, $W_i^K \subset R^{d_{model}\times d_k}$, and $W_i^V \subset R^{d_{model}\times d_k}$. To enhance gradient backpropagation, we add residual connections between transformer layers and across the transformer architecture. Last but not least, the transformer uses a feed-forward network (FFW), which is applied to each position separately:

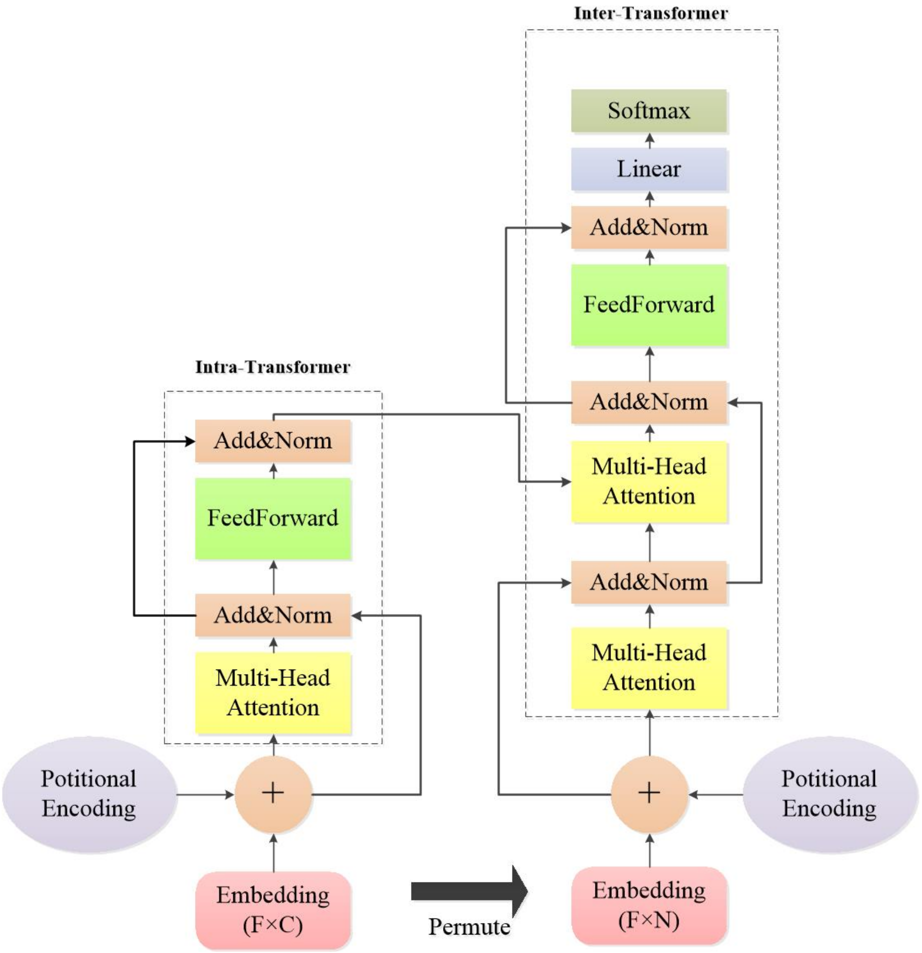

FIGURE 4. **Transformer block**

$$g''' = FeedForward\left(Norm(g'' + g')\right) + g'' + g\prime \quad (11)$$

A fully connected feed-forward network is present in each layer of the encoder and decoder, and it is applied to each transformer network layer identically and separately. This is composed of two linear transformations with an activation function called ReLU in the middle.

$$FeedForward(\alpha) = max(0, W_1x + b_1)W_2 + b_2 \quad (12)$$

The network structure of the linear transformation is the same on different modules when faced with different underwater ambient noise, but the model can be designed with different parameters at different layers.

III. EVALUATION

The experimental results will be presented and discussed in this section on the ShipsEar dataset [27] and the Deepship dataset [28]. Some ablation experiments and comparison tests show some of our explorations for underwater acoustic noise reduction.

A. DATASET

The ShipsEar dataset was collected with recordings made by hydrophones deployed from docks to capture different vessel speed noises and cavitation noises corresponding to docking or undocking maneuvers. The recordings are of actual vessel sounds captured in a natural environment. Therefore, anthropogenic and natural background noise and vocalization are present in marine mammals. The ShipsEar comprises 90 recordings in wav format with five significant classes. Where each primary class contains one or more subclasses (e.g., Class A is composed of fishing, trawlers, mussel, tugboats, and dredgers), the duration of each audio segment varies from 15 seconds to 10 minutes. Each class is divided, as shown in Table 1.

The Deepship dataset consists of 47 h and 4 min of real-world underwater recordings of 265 different ships belonging to four classes, as shown in Table 2. The dataset includes recordings from different sea states and yearly noise levels. The dataset is beneficial for evaluating the performance of existing algorithms and researching deep learning methods.

To better verify the performance of the model. All signals were segmented according to a fixed time of 5 seconds. We

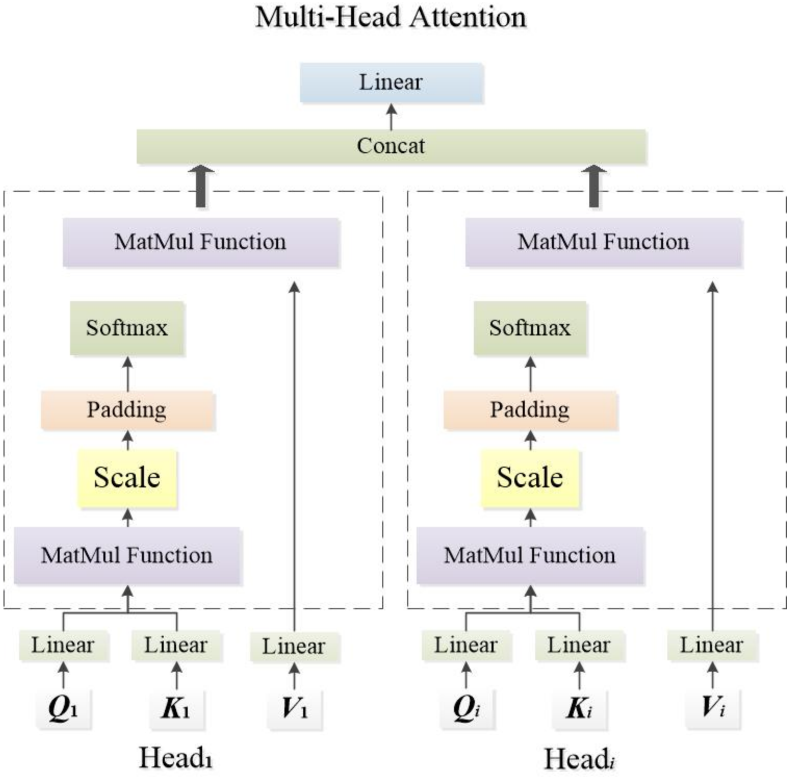

FIGURE 5. The architecture of the multi-head attention

TABLE 1. ShipsEar Dataset

Ship Type	Targets(Seconds)
Class A	Dredger(262s), Fishboat(514s), Musselboat(730s), Trawler(163s), Tugboat(206s)
Class B	Motorboat(1014s), Pilotship(138s), Sailboat(408s)
Class C	Passengers(4270s)
Class D	Oceanliner(942s), Roro(1483)
Class E	Natural Noise(1160s)

TABLE 2. DeepShip Dataset

Ship Type	Targets(No.of Ships)	Total Time)	Total Recordings
Cargo Ship	69	10 h 40 min	110
Tug	17	11 $h17$ min	70
Passenger ship	46	12 $h22$ min	193
Tanker	133	12 $h45$ min	240

set up three tasks to verify the denoising effect of the model:

Task 1: Samples without noise classes are randomly selected from the ShipsEar, and the noise classes in ShipsEar are superimposed with these extracted samples. The signal-to-noise ratio of the mixed signals is adjusted by the signal fusion method to 0 dB. The dataset is then divided into a test set, validation set, and training set according to the ratio of 6:2:2.

Task 2: The first 20% of each audio file in ShipsEar and DeepShip is extracted and used as the training set, and the remaining 80% of the segments of the audio files are divided into two parts: the validation set and the test set. In Task 2, noise is added the same way as in Task 1.

Task 3: Debug the denoising model individually by training all the data in the ShipsEar. The dataset in DeepShip is used as a test set and input to the trained denoising model according to the predefined classes to achieve signal noise reduction. Use the evaluation metrics to compare the change in the signal-to-noise ratio of the signal after noise reduction. In Task 3, noise is added the same way as in Task 1.

The difficulty from Task 1 to Task 3 gradually increases and at the same time it also becomes more suitable for practical applications.

B. OPTIMISATION

Table 3 shows the Dual-Path Transformer Network (DPTN) model used in this article compiled by CUDA11.3 [29]. The encoder is based on 256 convolutional filters with a kernel size of 4 and a stride factor of 2. The decoder uses the same number of permutation convolution filters. The encoder was chosen to have the same kernel size and move a step to maintain consistency before and after the audio duration. In ablation experiments, where the input and output channels of the network are kept the exact size before and after the feature tensor transfer, we test the number of repetitions necessary for the transformer process in the masking network to fit the denoising network better. We applied 8 parallel attention heads and 1024-dimensional positional feed-forward networks within each transformer layer. The learning rate is initialized to 1e-5 and decays for every 10 epochs by 0.98 Early stop training is introduced when there has been no improvement for 5 epochs. The weight decay has defaulted to an $L2$ penalty. Adam [30] is used as the optimizer, and the dynamic mixing (DM) [31] is introduced as the data augmentation. The gradual increase in the difficulty of the two tasks exercises the generalization ability and robustness while making the model conditions more realistic. The masking network processes chunks of size C = 250 with a 50% overlap. We employ $N1$ and $N2$ layers of transformers in both intra-Transformer and inter-Transformer.

C. TRAINING OBJECTIVE

The objective of training the end-to-end learning framework is to maximize the signal-to-noise ratio (SNR) [32]and the scale-invariant signal-to-noise ratio (SI-SNR) [33]. They are commonly used as the evaluation metric for signal noise reduction. SNR requires both the target signal and the enhanced signal to know. It is an energy ratio expressed in dB between the energy of the target signal contained in the enhanced signal and the energy of the error. SI-SNR uses a single coefficient to account for scaling discrepancies compared to SNR. The scale invariance is ensured by normalizing the signal to zero-mean before the calculation. So the higher it is, the better. SNR and SI-SNR are defined as:

$$SNR = 10log_{10}\frac{\left\|S_{target}\right\|^2}{\left\|\widehat{S}_{target} - S_{target}\right\|^2} \tag{13}$$

$$SI - SNR = 10log_{10}\frac{\left\|\theta s_{target}\right\|^2}{\left\|\widehat{s}_{target} - \theta s_{target}\right\|^2} \tag{14}$$

$$\theta = \frac{\langle \widehat{s}_{target}, s_{target}\rangle}{\left\|s_{target}\right\|^2} \tag{15}$$

Where $\widehat{s}_{target} \in R^{1\times T}$ and $s_{target} \in R^{1\times T}$ are the estimated and original clearly sources and $\left\|S\right\|^2 = \langle S, S\rangle$ denotes the signal power.

D. RESULTS

1) Waveform and spectrum display before and after noise reduction in **Task 1**

The waveform and spectrum display before and after noise reduction in Task 1 are shown in Figures 6 through 10. The original signals of various classes(Class A to D) are allowed to become distorted by ambient noise(Class E) using the Luo method [34]. It is clear that when the various underwater acoustic signals are interfered with by ambient noise, the characteristics of the underwater acoustic signals themselves are significantly altered. After using the proposed DPTN algorithm, most of the noise is eliminated, and the underwater acoustic signal retains the detailed information of the original underwater acoustic signal.

2) Comparison of ablation experiments in **Task 1**

In Task 1, we investigate the effects of various hyperparameters and network structures on the Dual-Path Transformer Network (DPTN) performance. Table 4 provides a summary of the findings, and the experiment is used for the test set noise reduction.

We note that the number of intra- and inter-transformers significantly impacts the performance. When both transformers are iterated 8 times, the best result is obtained. Instead, we find a slight reduction in the model's effectiveness after 16 iterations on each transformer. When we examine the cause, we can see that when the model design is too complex, the Dual-Path Transformer Network (DPTN) may overfit. After training, the model loses some generalization ability but retains a robust fitting ability. In addition, we let the inter-transformer use a single-time transformer, which has a performance of 15.59dB. We can observe a significant weakening of the model effect when using a single-layer transformer for the intra-transformer, indicating that local processing, or the intra-transformer, has a more significant impact on the denoising performance. The intra-Transformer is the initial transformer of the masking network. It has the ability to modify the model's hyperparameters to alter the performance of the subsequent network structure in addition to learning the mask. Therefore, the effectiveness of the DPTN depends on selecting the proper network structure for each layer. Finally, it is clear that this paper's positional coding was beneficial. We notice a slight performance difference between 8 and 16 heads when considering the number of heads. In order to make the model more lightweight, the proposed DPTN selects an attention technique with 8 heads.

We test the Dual-Path Transformer Network's (DPTN) signal enhancement speed with various network structures during training. The training curve for the model in Task 1 is depicted in Figures 11 and 12. We plot the performance versus time for the first 150 training iterations on the validation set. We utilized the same computer and GPU for every model to ensure an accurate comparison. Additionally, a batch size of 1 was used for training all systems. When the hyperparameters are selected to make the model more lightweight, the model converges faster during training. However, the model

TABLE 3. Dual-Path Transformer Network Structure

Intra-Transformer	**Inter-Transformer**	**Masking network**	**DPTN**
Positional Encoding	Positional Encoding	Intra-Transformer	Encoder
Normalization	Normalization	Inter-Transformer	Chunking
Multi-head attention	Multi-head attention		Masking network
Normalization	Normalization		Overlap-add
Feedforward	Feedforward		Decoder
PReLu	PReLu		
Linear	Linear		

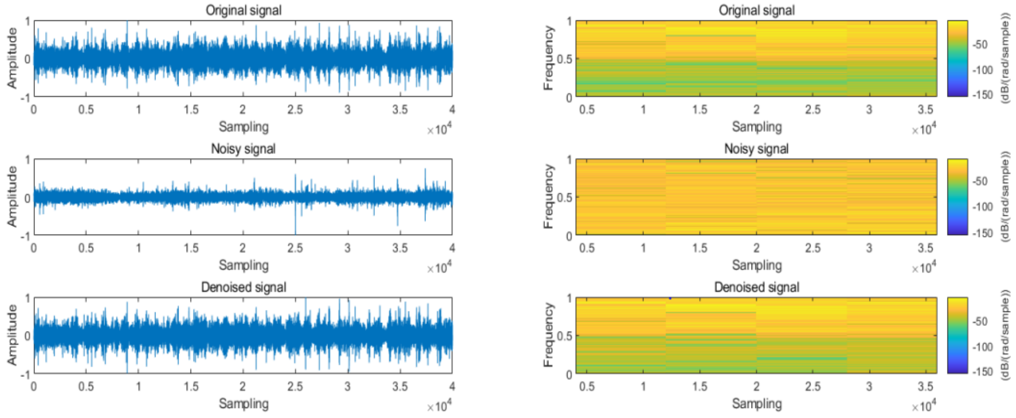

FIGURE 6. The waveform and spectrum of cargoship

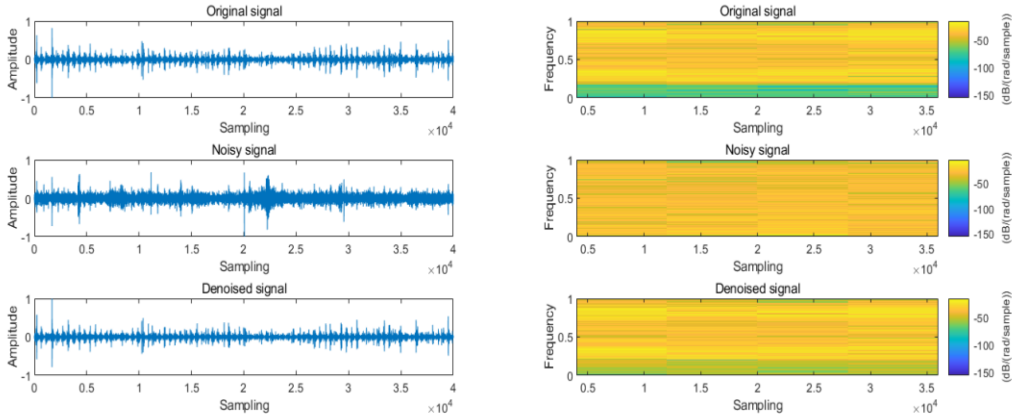

FIGURE 7. The waveform and spectrum of fishship

TABLE 4. Ablation experiment of the DPTN(N_{intra} is the number of the intra-transformer, N_{inter} is the number of the inter-transformer, Heads is the number of the multi-head, D is the dimension of feedforward, PE is the positional encoding.)

Model	SNR/SI-SNR	N_{intra}	N_{inter}	Heads	D	PE
DPTN	**16.68/15.82**	**8**	**8**	**8**	**1024**	**Yes**
DPTN(No PE)	14.61/13.28	8	8	8	1024	No
DPTN(D=2048)	14.69/13.25	8	8	8	2048	Yes
DPTN(Heads=4)	14.37/13.05	8	8	4	1024	Yes
DPTN(Heads=16)	15.61/13.09	8	16	4	1024	Yes
DPTN(N_{inter}=1)	15.59/13.89	8	1	8	1024	Yes
DPTN(N_{intra}=1)	13.73/10.2	1	8	8	1024	Yes
DPTN(N=4)	15.07/12.26	4	4	8	1024	Yes
DPTN(N=16)	14.81/12.89	16	16	8	1024	Yes

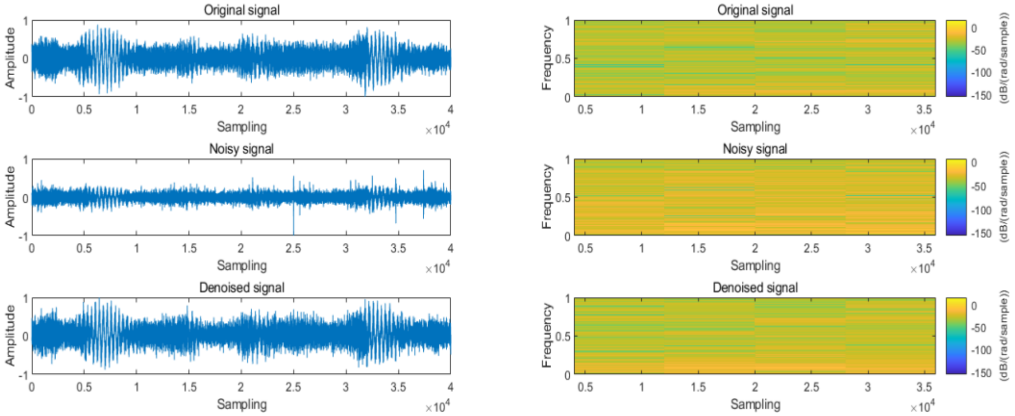

FIGURE 8. The Waveform and spectrum of motorboat

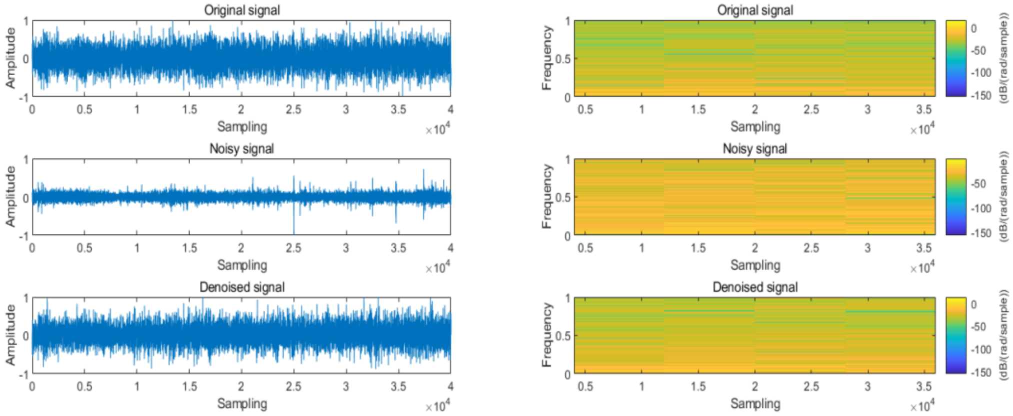

FIGURE 9. The Waveform and spectrum of oceanliner

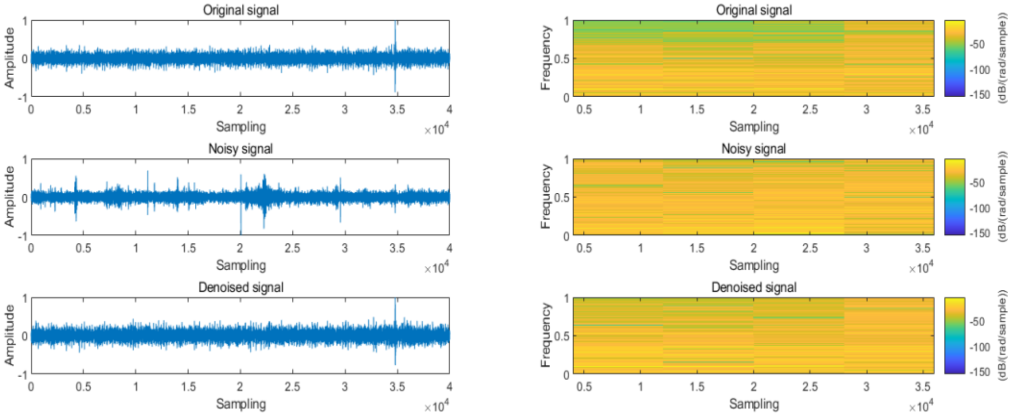

FIGURE 10. The Waveform and spectrum of passenger

TABLE 5. Results of different noise reduction methods.

Method	Task 1(SNR/SI-SNR)	Task 2	Task 3
RNN	10.21/9.9	8.37/8.6	7.89/6.67
FCN	9.68/10.93	8.25/8.8	7.30/7.62
DPTN	**16.68/15.82**	**13.71/12.53**	**10.37/11.69**
Wavelet denoising	9.64/8.98		
Interval-dependent denoising	8.67/7.33		

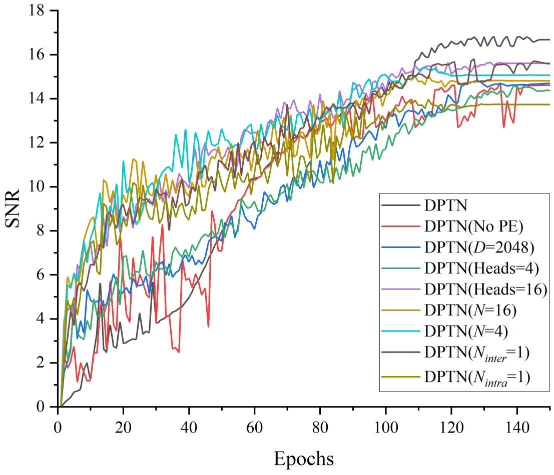

FIGURE 11. The SNR variation cure of ablation experiment

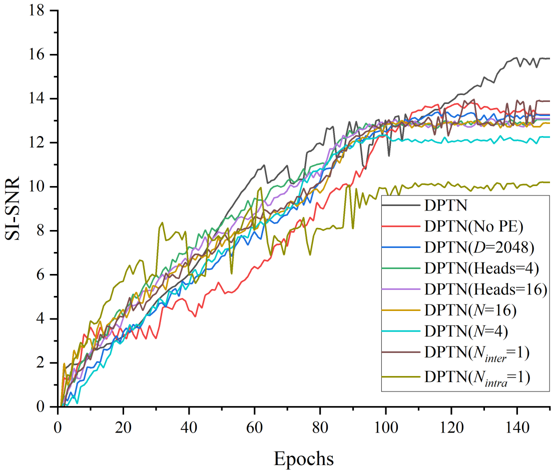

FIGURE 12. The SI-SNR variation cure of ablation experiment

testing results are not as good as DPTN. For example, the denoising effect of DPTN is improved by 1.09 over DPTN (N_{inter}=1), but the model convergence time is prolonged by about 3 hours. Therefore, different collocation methods can be invoked according to different experimental needs.

3) Comparison of the denoising performance of different models in **Task 1, 2 and 3**

Table 5 compares the performance of the Dual-Path Transformer Network (DPTN) with other deep learning-based approaches to three tasks in noise reduction. The denoising methods include Recurrent Neural Network (RNN) [25], Fully Connected Neural Network-based model (FCN) [16], wavelet denoising method [12] and interval correlation denoising method [13]. Since the wavelet denoising and interval denoising methods feed the noisy signal directly into the model and reduce the noise by transform decomposition methods, so the whole process can be implemented without training the model. By evaluating three tasks on the test dataset, the signal-to-noise ratio improvement of our proposed DPTN model is 16.68db, 13.71db, and 10.37db in task 1, task 2, and task 3, respectively.

4) RMSE, Spectral entropy, and Phase diagram display before and after noise reduction by DPTN

To more thoroughly assess the results of the denoising underwater acoustic signal using DPTN and the change in features before and after denoising. We calculated the change in the short root mean square error (RMSE) before and after signal denoising, which is a way of responding to the change in signal energy. By comparing the changes of RMSE before and after noise reduction, we know that the clear underwater acoustic signal will not change drastically in a short period. However, when ambient noise is added, it will cause the RMSE of the signal to change. As shown in Figure 13 and Figure 14, we can find that the value of RMSE of the noisy signal is around 0.1 in the first 2 seconds. The value of RMSE can increase to about 0.15 when the change in RMSE from 4S to 5S is significant. On the other hand, the RMSE of the denoised signal is approximately 0.075 in the first two seconds and can continue to be approximately 0.075 when the change in RMSE is minimal in the period between 4 and 5 seconds. The results show that the RMSE of the target signal can be maintained in a relatively stable state after using the DPTN method for noise reduction.

Further, we use spectral entropy to investigate the changes in underwater acoustic signals before and after denoising. Spectral entropy reflects the relationship between the power

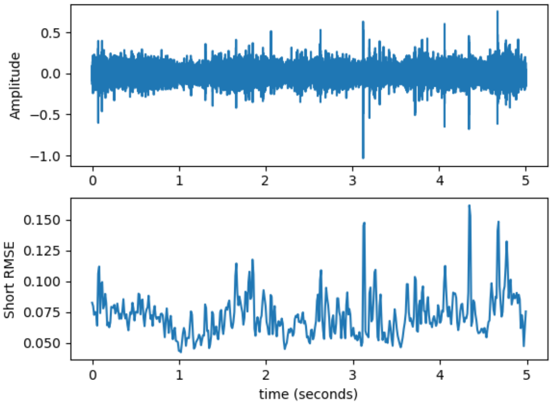

FIGURE 13. The Noisy signal Short RMSE

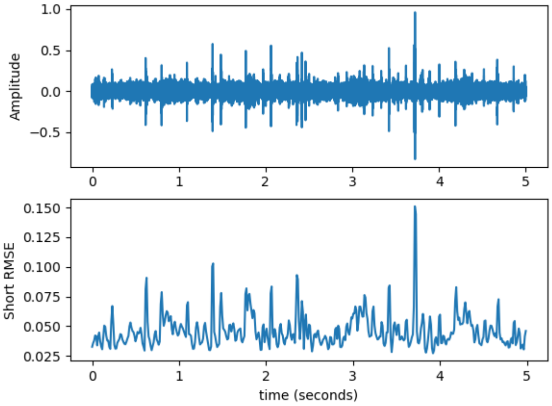

FIGURE 14. The Denoised Signal Short RMSE

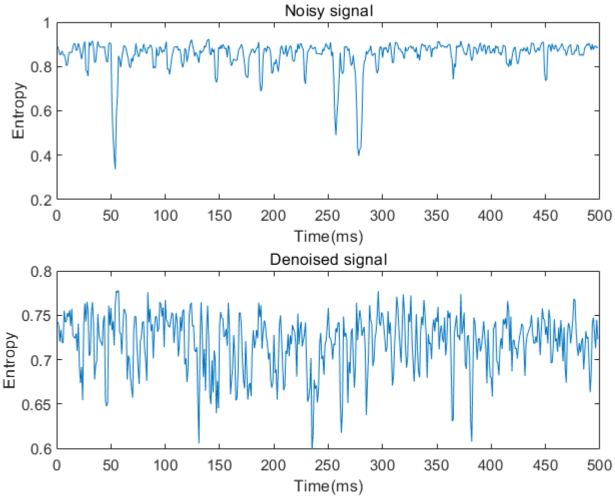

FIGURE 15. Spectral Entropy

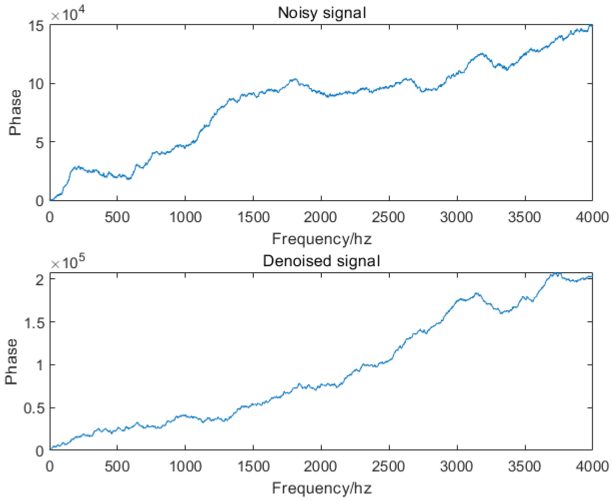

FIGURE 16. Phase diagram

spectrum and entropy rate. Entropy is a measure of the degree of uncertainty of various random tests. The natural ambient noise in the underwater acoustic signal experiments has a high degree of randomness and confusion, which causes the spectral entropy to fluctuate widely. As shown in Figure 15, the noisy signal's spectral entropy fluctuates between 0.3 and 0.9, and when the DPTN denoising method is applied, it fluctuates between 0.6 and 0.8.

Finally, we compared the phase diagram's transformation before and after denoising. The phase diagram is an essential feature in identifying the signal of the underwater acoustic signal. In the identification task, the signal's phase change can be used to determine the type of underwater acoustic signal. As shown in Figure 16 and Figure 17, the two original signals fall under the same class (Class A) in the ShipsEar data. However, different phase change is produced when natural ambient noise is added. Phase change makes it difficult for the signal classification system to differentiate between the classes of two signals. The phase change of the two denoised signals is similar after using the DPTN model to eliminate background noise, allowing us to distinguish between the two denoised signals that belonged to the same class.

IV. CONCLUSIONS

This paper proposes an end-to-end method of masking pattern denoising. First, the method reconstructs the noisy signal by coding and learns the feature mapping by deep learning. Then, in the high-dimensional feature space of deep neural networks, the transformer between features allows the model to acquire more knowledge. Finally, the denoised signal is recovered by decoding. To validate the effectiveness of our method, we verify the model's performance on the ShipsEar and DeepShip datasets. The experimental results show that our proposed model is more competitive than other deep learning techniques. The signal-to-noise ratio improvement

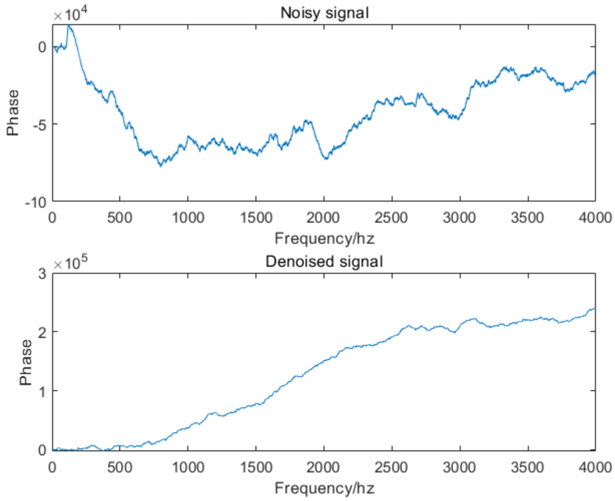

FIGURE 17. Phase diagram

of our suggested DPTN model is 16.68 dB, 13.71 dB, and 10.37 dB in Task 1, Task 2, and Task 3, respectively. In future work, we apply migration learning methods and GAN to underwater acoustic signal denoising.

REFERENCES

[1] Railey K , Dibiaso D , Schmidt H . An acoustic remote sensing method for high-precision propeller rotation and speed estimation of unmanned underwater vehicles[J]. The Journal of the Acoustical Society of America, 2020, 148(6):3942-3950.

[2] Jiang F, Zhang Z. An improved underwater TDOA/AOA joint localisation algorithm[J]. IET Communications, 2021, 15(6): 802-814.

[3] Vieira V , Coelho R , FMD Assis. Hilbert–Huang–Hurst-based non-linear acoustic feature vector for emotion classification with stochastic models and learning systems[J]. IET Signal Processing, 2020, 14(8):522-532.

[4] Li G, Yang Z, Yang H. Noise reduction method of underwater acoustic signals based on uniform phase empirical mode decomposition, amplitude-aware permutation entropy, and Pearson correlation coefficient[J]. Entropy, 2018, 20(12): 918.

[5] Moreaud U, Courmontagne P, Chaillan F, et al. Performance assessment of noise reduction methods applied to underwater acoustic signals[C]//OCEANS 2016 MTS/IEEE Monterey. IEEE, 2016: 1-15.

[6] Ou H, Allen J S, Syrmos V L. Frame-based time-scale filters for underwater acoustic noise reduction[J]. IEEE Journal of oceanic engineering, 2011, 36(2): 285-297.

[7] Moreaud U, Courmontagne P, Chaillan F, et al. A new way for underwater acoustic signal analysis: The morphological filtering[C]//OCEANS 2015-Genova. IEEE, 2015: 1-9.

[8] Tzeng Y , Too G . On the alternative objective functions for minimum variance distortionless response technique in order to restore the original source[J]. Journal of the Acoustical Society of America, 2016, 139(4):2084-2084.

[9] Chen D, Chu X, Ma F, et al. A variational approach for adaptive underwater sonar image denoising[C]//2017 4th International Conference on Transportation Information and Safety (ICTIS). IEEE, 2017: 1177-1181.

[10] Taroudakis M, Smaragdakis C, Ross Chapman N. Denoising underwater acoustic signals for applications in acoustical oceanography[J]. Journal of Computational Acoustics, 2017, 25(02): 1750015.

[11] Li Y X . A novel noise reduction technique for underwater acoustic signals based on complete ensemble empirical mode decomposition with adaptive noise, minimum mean square variance criterion and least mean square adaptive filter[J]. Defence Technology, 2020, 16(3):12.

[12] Li Y , Li Y , Chen X , et al. A New Underwater Acoustic Signal Denoising Technique Based on CEEMDAN, Mutual Information, Permutation Entropy, and Wavelet Threshold Denoising[J]. Entropy, 2018, 20(8):563

[13] Yan H, Xu T, Wang P, et al. MEMS hydrophone signal denoising and baseline drift removal algorithm based on parameter-optimized variational mode decomposition and correlation coefficient[J]. Sensors, 2019, 19(21): 4622.

[14] Qi Q, Chen H, Yan Z, et al. Holographic reconstruction research on the radiated acoustic field of the underwater vehicle[C]//OCEANS 2019-Marseille. IEEE, 2019: 1-5.

[15] Lee K C. Underwater acoustic localisation by GMM fingerprinting with noise reduction[J]. International Journal of Sensor Networks, 2019, 31(1): 1-9.

[16] Zhou X , Yang K . A denoising representation framework for underwater acoustic signal recognition[J]. The Journal of the Acoustical Society of America, 2020, 147(4):EL377-EL383.

[17] Wang X, Zhao Y, Teng X, et al. A stacked convolutional sparse denoising autoencoder model for underwater heterogeneous information data[J]. Applied Acoustics, 2020, 167: 107391.

[18] Othman A , Iqbal N , Hanafy S M , et al. Automated Event Detection and Denoising Method for Passive Seismic Data Using Residual Deep Convolutional Neural Networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, PP(99):1-11.

[19] Testolin A, Diamant R. Underwater acoustic detection and localization with a convolutional denoising autoencoder[C]//2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). IEEE, 2019: 281-285.

[20] Qiu Y, Yuan F, Ji S, et al. Stochastic resonance with reinforcement learning for underwater acoustic communication signal[J]. Applied Acoustics, 2021, 173: 107688.173:107688.

[21] LI Y, WANG B, SHAO G, et al. A Method of Noise Reduction for Underwater Acoustic Communication Signal Based on RCGAN[J]. ACTA ELECTONICA SINICA, 2022, 50(1): 54.

[22] Xing C, Wu Y, Xie L, et al. A sparse dictionary learning-based denoising method for underwater acoustic sensors[J]. Applied Acoustics, 2021, 180: 108140.

[23] Zhang W, Li X, Zhou A, et al. Underwater acoustic source separation with deep Bi-LSTM networks[C]//2021 4th International Conference on Information Communication and Signal Processing (ICICSP). IEEE, 2021: 254-258.

[24] Liu L, Cai L, Ma L, et al. Channel state information prediction for adaptive underwater acoustic downlink OFDMA system: deep neural networks based approach[J]. IEEE Transactions on Vehicular Technology, 2021, 70(9): 9063-9076.

[25] Zhang,Fumin and Zhang,Ziqiao and Tong,Feng et al. Modeling and learning underwater acoustic channel parameters through deep recursive neural networks[J]. The Journal of the Acoustical Society of America, 2022, 151(4):A233-A234.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[27] D Santos-Domínguez, Torres-Guijarro S , A Cardenal-López, et al. ShipsEar: An underwater vessel noise database[J]. Applied Acoustics, 2016, 113:64-69.

[28] Irfan,Muhammad and Jiangbin,Zheng and Ali,Shahid and Iqbal,Muhammad and Masood,Zafar and Hamid,Umar, et al. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification[J]. Expert systems with applications, 2021, 183:115270.

[29] Qi J , Li K C , Jiang H , et al. GPU-accelerated DEM implementation with CUDA[J]. International Journal of Computational Science and Engineering, 2015, 11(3):330-337.

[30] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[31] Zeghidour N, Grangier D. Wavesplit: End-to-end speech separation by speaker clustering[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2840-2849.

[32] Bogale T E , Vandendorpe L . Max-Min SNR Signal Energy Based Spectrum Sensing Algorithms for Cognitive Radio Networks with Noise Variance Uncertainty[J]. IEEE Transactions on Wireless Communications, 2014, 13(1):280-290.

[33] Le Roux J, Wisdom S, Erdogan H, et al. SDR–half-baked or well done?[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 626-630.

[34] Hershey J R, Chen Z, Le Roux J, et al. Deep clustering: Discriminative embeddings for segmentation and separation[C]//2016 IEEE interna-

tional conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 31-35.

YONG QIANG SONG

He received the bachelor's and master's degrees in Shandong Normal University,JINAN, China, in 2017 and 2020, respectively, and he is currently pursuing the doctor's degree in National Innovation Institute of Defense Technology, Beijing, China. He research interests include underwater acoustic signal noise reduction and chaotic signal processing.

PLACE
PHOTO
HERE

FENG LIU Assistant Researcher

National Innovation Institute of Defense Technology, Chinese Academy of Military Science, China. He received the bachelor's degrees in Huazhong University of Science and Technology He received the master's degrees in Naval Aviation Engineering University. He received the doctor's degrees in NavalAviation Engineering University. He research interests include underwater acoustic signal noise reduction and chaotic signal processing.

PLACE
PHOTO
HERE

TONGSHENG SHEN Researcher

National Innovation Institute of Defense Technology, Chinese Academy of Military Science, China He research interests include underwater acoustic signal noise reduction and chaotic signal processing.

• • •