Attention-based neural network for underwater acoustic target detection and direction-of-arrival estimation

Xu Xiao[1,2,3] Qunyan Ren[1,2,3*] Wenbo Wang[1,2,3] Meng Zhao[1,2,3] Li Ma[1,2,3]

1 Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China; xiaoxu@mail.ioa.ac.cn (X.X.); renqunyan@mail.ioa.ac.cn (Q.R); wangwenbo215@mails.ucas.ac.cn (W.W.); zhaomeng@mail.ioa.ac.cn (M.Z.); mary1968@tom.com (L.M.)
2 Key Laboratory of Underwater Acoustic Environment, Chinese Academy of Sciences, Beijing, 100190, China
3 University of Chinese Academy of Sciences, Beijing, 100049, China
* Correspondence: Qunyan Ren; Institute of Acoustics, Chinese Academy of Sciences; No. 21 North 4th Ring Road, Haidian District, 100190 Beijing, China; Tel.:+86-185-1966-1355; Email Address: renqunyan@mail.ioa.ac.cn.

Abstract: Direction-of-arrival (DOA) estimation for underwater acoustic sources is usually affected by multisource interference and ambient noise. In this study, DOA estimation is achieved by using a conventional beamformer modified by attention mechanism (A-CBF) which explores the spatial spectrum for DOA estimation that can focuses more on the peak of the desired signal while suppressing other peaks caused by interference and noise. The coefficients in A-CBF are learned by a neural network trained by array-received signals. On the basis of the above concept, the neural network determines the presence of the target in the received signals. From data obtained during a 2020 sea trial, the A-CBF model was trained by using a small amount of experiment data. The processing results demonstrate its performance of DOA estimation and target detection through suppressing multisource interference and focusing on the beams of the target ship in the spatial spectrum.

Key words: Underwater acoustic; DOA estimation; attention mechanism; neural network; beamforming.

Attention-based neural network for underwater acoustic target detection and direction-of-arrival estimation

Abstract: Direction-of-arrival (DOA) estimation for underwater acoustic sources is usually affected by multisource interference and ambient noise. In this study, DOA estimation is achieved by using a conventional beamformer modified by attention mechanism (A-CBF) which explores the spatial spectrum for DOA estimation that can focuses more on the peak of the desired signal while suppressing other peaks caused by interference and noise. The coefficients in A-CBF are learned by a neural network trained by array-received signals. On the basis of the above concept, the neural network determines the presence of the target in the received signals. From data obtained during a 2020 sea trial, the A-CBF model was trained by using a small amount of experiment data. The processing results demonstrate its performance of DOA estimation and target detection through suppressing multisource interference and focusing on the beams of the target ship in the spatial spectrum.

Key words: Underwater acoustic; DOA estimation; attention mechanism; neural network; beamforming.

1 Introduction

Underwater acoustic direction-of-arrival (DOA) estimation is a major function of sonar systems. Machine learning approaches have made great progress in DOA estimation of underwater acoustic targets in recent years[1]. The machine learning method usually establishes the mapping relationship between the sample covariance matrix of array signals and the classification labels of arrival directions, thus being trained as a DOA estimator. This approach has been proven to maintain good performance even under low signal-to-noise ratio (SNR) conditions[2].

As a newly machine learning method, deep neural networks (DNNs) have achieved initial success in previous works on underwater acoustic applications, such as source localization[3],[4], DOA estimation[5]-[10], and target recognition[11][12]. However, their high performance is often difficult to explain, because DNN is a “black-box” model whose internal work is not transparent[13]. DNN also lacks physical interpretability, as it works by directly establishing the mapping between the signal feature expression and target attributes.

Attention mechanism[14] is an important technology to inspect the internal work and improve the interpretability of the DNN[15]. It assigns different weights to input features through attention matrixes learned by DNN and displays their contribution to decision-making[16]. Attention-based DNNs are widely used in many fields, such as machine translation[17], image translation[18], speech recognition[19], and image classification[20] In recent years, attention-based DNNs have been proposed in line-spectrum feature extraction[21], target recognition[21], and source localization[22] for underwater acoustics. These models place attention weight on the time-frequency domain features of sonar-received signals to control sensitivities to the desired and interfering targets, thus achieving improved accuracy and interpretability.

In this paper, a conventional beamformer modified by attention mechanism (A-CBF) is proposed. The attention weights are placed on the CBF to ensure that its output is more focused on specific beams and frequencies. This feature results in the spatial spectrum obtained by A-CBF being able to focus more on the peak of the desired signal while suppressing other peaks due to multisource interference and ambient noise.

A convolutional neural network (CNN) is established to learn the attention weight coefficients and is composed of two parts: an attention module and a detection module. The attention module includes the attention weights to be learned, and it computes the A-CBF results and spatial spectrum used for DOA estimation. The detection module analyzes the A-CBF features to determine the presence of the desired target and propagate the detection error back[23] to the attention module to optimize the attention weights. The two modules share weights through CNN to coordinate and promote each other. The model was tested by the South

China Sea experiment conducted in September 2020 to perform DOA estimation and target detection, where multiple interfering sources were found in the experimental area.

The rest of this paper is organized as follows: Section 2 defines the CNN model and the attention mechanism used in this paper; Section 3 describes the details of the at-sea experiment and the experimental data processing; Section 4 analyzes the result of DOA estimation and target detection and compares it with the results of the traditional single-function methods CBF, CNN, and energy detector; and Section 5 concludes the paper.

2 Model Definition

The CNN structure used in this paper is shown in Figure 1. The CNN input is the CBF result, and the CNN output is a binary classification sequence that represents the presence of the target. The process consists of four steps. First, the array signals are processed by CBF. Second, the CBF output is weighted through an attention module based on coefficients learned from training samples to obtain the A-CBF output. Third, the A-CBF output is used to perform spatial spectrum estimation on the one hand and continues to be connected to the detection module on the other hand. Finally, the detection module determines the presence of the desired target and propagates the error back to the attention module to optimize the attention weight of A-CBF.

Several one-snapshot samples are needed to train the model; these samples are labeled as either class 0 (target is not present) or class 1 (target is present). Then, both DOA estimation and target detection are performed when a test is performed on the unlabeled samples, as the following diagram shows:

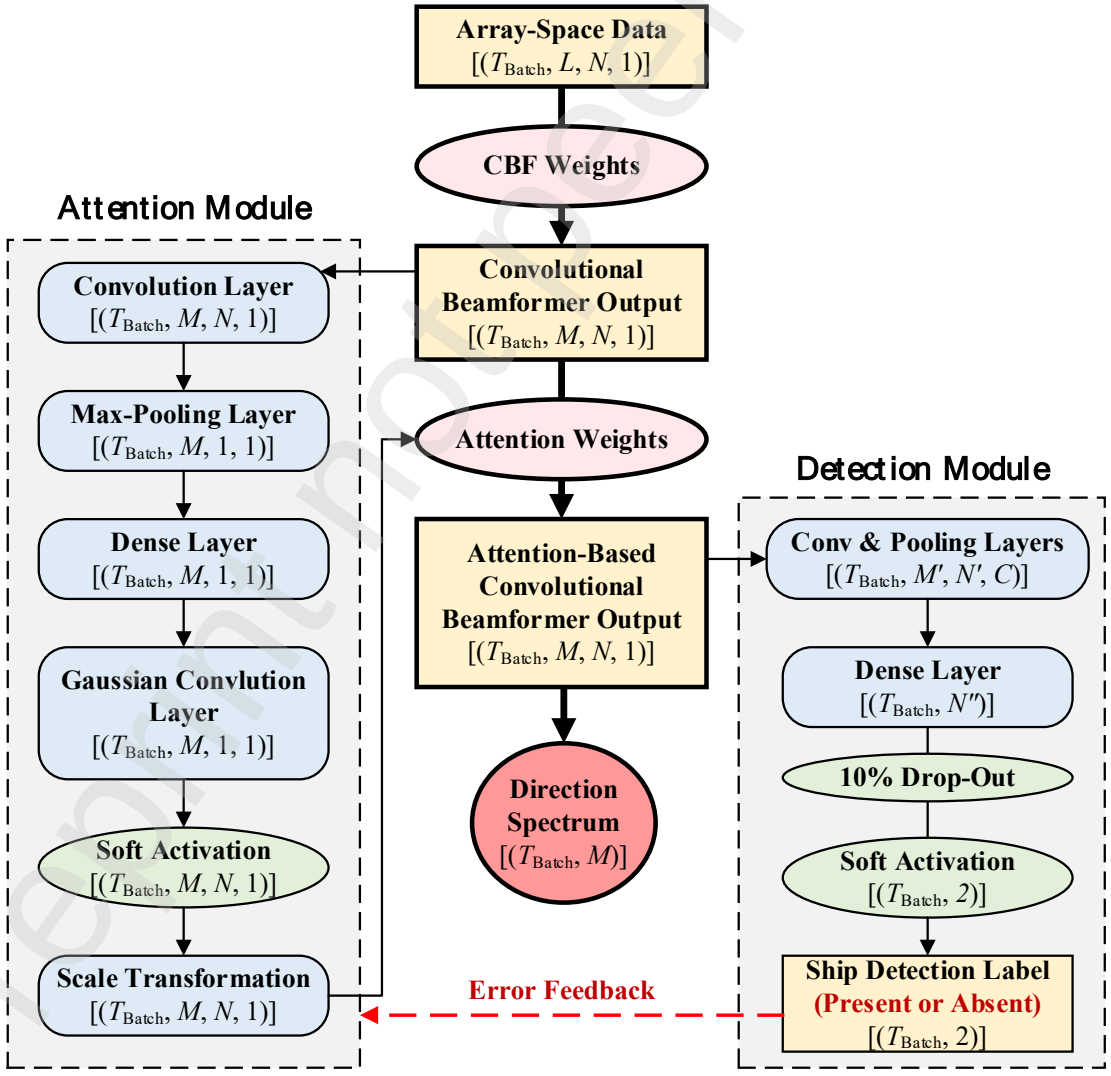


Figure 1. CNN structure for direction spectrum estimation

2.1 A-CBF framework

An L-element horizontal line array (HLA) with uniform receiver separation d is shown in Figure 2. The received array signals are divided into T segments by time. The T segments are the single-snapshot samples that are to be processed for CNN input and then batched with a batch size of T_{Batch}.

The t-segment HLA that received the sound field from a broadband point source is represented by $^{(t)}(\square,l)$ where $l = 1, 2, ..., L$ and $t = 1, 2, ..., T$. The HLA-received sound-field data are preprocessed to eliminate the influence of the source spectrum. To reduce the effect of the source amplitude, $^{(t)}(f,l)$ is normalized by[4]

$$\tilde{\boldsymbol{p}}^{(t)}(f,l)=\frac{\boldsymbol{p}^{(t)}(f,l)}{\sqrt{\sum_{l=1}^{L}\left|\boldsymbol{p}^{(t)}(f,l)\right|^2}} \tag{3}$$

The received HLA sound field is projected into the frequency-beam domain by CBF in the frequency domain. For an HLA with uniform element spacing, the CBF weight is the steering vector

$$\boldsymbol{w}(\theta,f)=\frac{1}{\sqrt{S}}\left[1,e^{j2\pi fd\sin\theta/c},e^{j2\pi f\cdot 2d\sin\theta/c},\cdots,e^{j2\pi f(L-1)d\sin\theta/c}\right]^{\mathrm{T}},\ \ \boldsymbol{w}\in C^{L\times 1} \tag{4}$$

Suppose $\boldsymbol{p}'^{(t)}(f_n)$ is the normalized sound-field vector of size $L \times 1$ composed of the received sound fields of all L elements at frequency f_n, $n = 1, 2, ..., N$, and the L beams point in the directions $\square_m$ where $\square_m = \square_1, \square_2, ..., \square_M$. Thus, the output power of CBF at frequency f_n of segment t is[24]

$$\boldsymbol{B}^{(t)}(\theta_m,f_n)=\left|\boldsymbol{w}(\theta_m,f_n)^{\mathrm{H}}\ \boldsymbol{p}'^{(t)}(f_n)\right|^2,\ \ \boldsymbol{B}\in C^{M\times N},\ \ \boldsymbol{p}'^{(t)}\in C^{L\times 1} \tag{5}$$

where H represents the conjugate transpose. When the beam angle θ_m is in agreement with the target arrival angle, the beam outputs the maximum power. Finally, the broadband CBF accumulates the power of each single-frequency CBF output to obtain the broadband direction spectrum. The DOA of each source can be estimated by finding the peaks on the spatial spectrum.

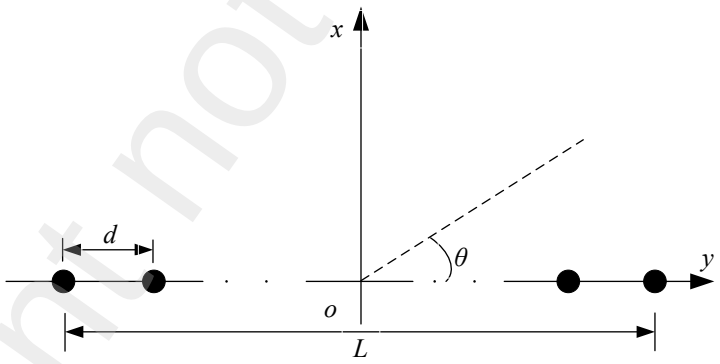


Figure 2. Sketch map of an L-element HLA

In introducing the attention mechanism, the CBF output $\boldsymbol{B}$ should be weighted with an attention weight matrix to control sensitivity to different frequency and spatial components. Suppose that the attention weight matrix is $\boldsymbol{A}$. Thus, the A-CBF output is the Hadamar product of weight matrix $\boldsymbol{A}$ and CBF output $\boldsymbol{B}$ (represented as $\odot$), and the A-CBF output is defined as

$$\tilde{\boldsymbol{B}}^{(t)}(\theta_m,f_n)=\boldsymbol{A}^{*}\odot\boldsymbol{B}^{(t)}(\theta_m,f_n). \tag{6}$$

where $\boldsymbol{A}^*$ represents the optimal value of $\boldsymbol{A}$, which is to be searched in a CNN through the chain rule of backpropagation (BP)[23]. The optimization problem can be described as

$$\boldsymbol{y}=\mathrm{CNN}\left(\boldsymbol{A},\boldsymbol{w}_{\mathrm{DNN}},\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)},\cdots,\boldsymbol{B}^{(T)}\right) \tag{7}$$

$$\boldsymbol{A}^{*}=\underset{\boldsymbol{A},\boldsymbol{w}\in\mathbb{R}^{M\times N}}{\arg\min}\,J\left(\boldsymbol{A},\boldsymbol{w}_{\mathrm{DNN}},\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)},\cdots,\boldsymbol{B}^{(T)},\boldsymbol{S}^{(1)},\boldsymbol{S}^{(2)},\cdots,\boldsymbol{S}^{(T)}\right) \tag{8}$$

where $\boldsymbol{w}_{\mathrm{DNN}}$ is the CNN weights, and $\boldsymbol{S}^{(t)}=[s_0{}^{(t)}, s_1{}^{(t)}]$ is the binary classification label of each data segment, whose value is [1, 0], which represents the presence of the target signal, or [0, 1], which represents the absence of the target signal in a data segment; $\boldsymbol{y}^{(t)}=[y_0{}^{(t)}, y_1{}^{(t)}]$ is the output sequence of the CNN activated by the softmax function[25], where the two entries represent the present and absent probability of the target, respectively, and their value range is between [0, 1]; J is the cost function, which is defined by binary cross entropy

$$J=-\frac{1}{T}\sum_{t=1}^{T}\left(s_0^{(t)}\log y_0^{(t)}+s_1^{(t)}\log y_1^{(t)}\right) \tag{9}$$

In CNN training, the BP algorithm is used to find the gradients for CNN weights and neurons with respect to the cost function, which needs an optimizer to update the weights iteratively by using those gradients. The traditional gradient descent method often has difficulty converging due to the complex underwater environment, as the cost function is usually nonconvex and easily falls into the local optimum or tends to overlearn with a fixed learning rate. Thus, an adaptive optimization method called adaptive moment estimation algorithm (Adam)[26] is used dynamically adjust the learning rate for CNN weights, which uses first and second moment estimation of the gradient to keep the learning rate within a certain range by bias correction iteratively, thus obtaining a stable parameter update.

Suppose t is the number of iterations, and u_t is any of the element of $\boldsymbol{w}_{\mathrm{DNN}}$ to be estimated in DNN in the t^{th} interation. First, the exponential moving averages of the gradient m_t, whose initial value m_0 is 0, are calculated. Considering the gradient momentum of the previous iteration, suppose the hyperparameter β_1 is the exponential decay rates

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)\nabla_u J(u_{t-1}) \tag{10}$$

The exponential moving average of the squared gradient v_t is then calculated, whose initial value v_0 is 0, and suppose the hyperparameter β_1 is the exponential decay rates

$$v_t=\beta_{t-1}+(1-\beta_2)\nabla_u J(u_{t-1})^2 \tag{11}$$

First, the bias is corrected, and then the parameters are updated

$$\begin{cases}\hat{m}_t=m_t/(1-\beta_1^t)\\ \hat{v}_t=v_t/(1-\beta_2^t)\\ u_t=u_{t-1}-\eta\cdot\hat{m}_t/\sqrt{\hat{v}_t}\end{cases} \tag{12}$$

In Equation (10), η is the initial learning rate. The Adam algorithm adaptively adjusts the update value from both the mean and uncentered variance of the gradient, thus improving the convergence efficiency.

2.2 Soft attention mechanism

The soft attention mechanism[20] is applied to define the attention weight matrix $\boldsymbol{A}$. A soft attention mask should consist of at least one convolutional pooling layer activated by the softmax function[20]. The CBF output $\boldsymbol{B}(\theta_m f_n)$ should be convolved with a convolution kernel $\boldsymbol{R}$ of size $d \times d$ whose entries are represented by R_{jk}, where $j, k = 1, 2..., K$, and the output offset is b. The dimension of the matrix resulting from a convolution operation is $(M - K + 1) \times (N - K + 1)$. Thus, the border elements are lost from the output every time a convolution operation is performed, precluding the building of deeper networks. Thus, $\mu = (K - 1)/2$ numbers of layers of zeros are padded to the border of the matrix, expanding the dimension to $(M + K - 1) \times (N + K - 1)$. The expanded matrix is denoted as $\boldsymbol{B'}$. Therefore, the output of the convolutional layer becomes a matrix $\boldsymbol{Z}$ with the dimension of $M \times N$. In forward propagation, the values of each entry z_{uv} are

$$\begin{cases}z_{uv}=\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\boldsymbol{B'}\cdot r_{jk}\cdot\xi(j,k)+b,\\ \xi(j,k)=\begin{cases}1, & 0\le j,k\le K\\ 0, & \text{others}\end{cases}.\end{cases} \tag{13}$$

where $u = 1, 2,..., M$, $v = 1, 2, ..., N$. Then, an average pooling layer with a pooling core with a size

of 1 × N and a step size of 1 × 1 is defined. The matrix is pooled into a vector $\boldsymbol{\psi}$ of length M that corresponds to M beam angles, which is expressed as

$$\boldsymbol{\psi} = \left[\sum_{v=1}^{N} \frac{z_{1v}}{N} \quad \sum_{v=1}^{N} \frac{z_{2v}}{N} \quad \cdots \quad \sum_{v=1}^{N} \frac{z_{uv}}{N} \quad \cdots \quad \sum_{v=1}^{N} \frac{z_{Mv}}{N} \right]^{T}. \tag{14}$$

Then, $\boldsymbol{\psi}$ is activated by the softmax function[25] to obtain the original attention weight vector $\boldsymbol{\alpha}$ of length S, which is expressed as

$$\boldsymbol{\alpha} = \mathrm{Softmax}(\boldsymbol{W}\boldsymbol{\psi}), \tag{15}$$

where $\boldsymbol{W}$ is an attention score matrix with a size of $M \times M$, in which the values of each element are obtained through CNN learning. Thus, the attention effect on the beam domain is applied.

Considering the correlation of the sound field between adjacent beams, the attention on beam domain should maintain a beamwidth. Thus, a Gaussian layer[21] is defined. $\boldsymbol{\alpha}$ is multiplied with the Gaussian kernel function, as represented by the following matrix operations:

$$\boldsymbol{H} = \left(-\frac{1}{2\sigma^2} \begin{bmatrix} h_{11}^2 & h_{12}^2 & \cdots & h_{1M}^2 \\ h_{21}^2 & h_{22}^2 & & \\ \vdots & & \ddots & \\ h_{M1}^2 & & & h_{MM}^2 \end{bmatrix} \right), \quad h_{jk} = k - j, \quad j,k \in [1, M] \tag{16}$$

$$\boldsymbol{g} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{H} \right) \boldsymbol{\alpha}, \tag{17}$$

where $\boldsymbol{g}$ is the output vector of size M of the Gaussian layer, which is the attention weight layer and σ is the standard deviation of the Gaussian function. Larger σ gives more distracted attention[21], which ensures that the attention is not overly focused on a single beam. Finally, suppose $\boldsymbol{\varepsilon}$ is a row coefficient vector of length N. The output matrix $\boldsymbol{A}$ is expressed as

$$\boldsymbol{A} = \boldsymbol{g} \times \boldsymbol{\varepsilon}, \tag{18}$$

As a variable that could affect attention in the frequency domain, vector ε could be set as a trainable or untrainable parameter in training configuration. In this work, vector ε is simply set as an untrainable layer of an array of ones. Finally, the A-CBF outputs are given by Equation (16) and are accumulated according to frequency to obtain the broadband result for each single-snapshot sample, which is the A-CBF spatial spectrum. The DOA estimation of the source can be obtained by finding the peak.

3 Experiment

In September 2020, the ship-radiated noise measurement experiment was conducted in the northern South China Sea. The topography map of the experimental area is shown in Figure 3(a). This experiment was undertaken on the same voyage as in literature[21], while the data were recorded at a different time. As the target ship, an experiment ship sails along the track at a speed of 6 knots or is moored at stations A and B, and the tow ship with a towed HLA sails along the track C–D at a speed of 6 knots. The number of HLA elements is 179, and the element spacing is from 0.2 to 4 m. The sea floor is relatively flat, and the water depth is 125–140 m. The length of the HLA recorded data is 1200 seconds, and the sampling rate is 8 kHz. Figure 3(b) shows the time-frequency spectrograms of the signal obtained by a short-time Fourier transform of the whole data at the 90th element (intermediate element). In the first 585 seconds, the target ship is moored at the station with its main and auxiliary engines shutting down. At about 300 seconds, the tow ship begins to accelerate, as some of the line spectrum in the spectrum changes, and the ship's noise level increases. At 586 seconds, the target ship begins to sail along the track as the low-frequency energy increased.

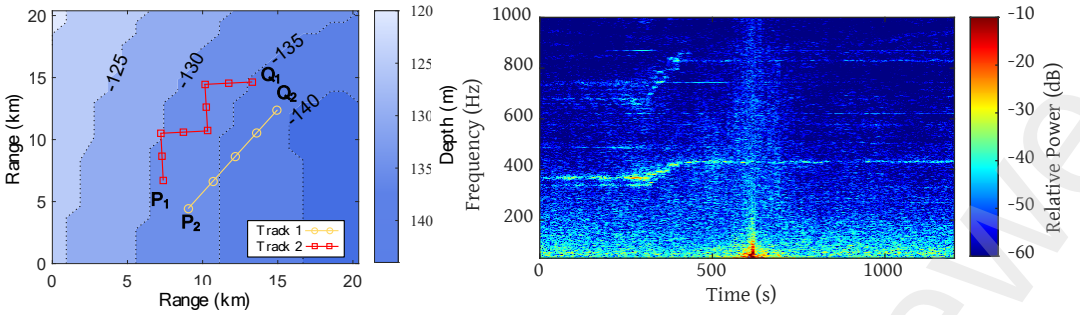
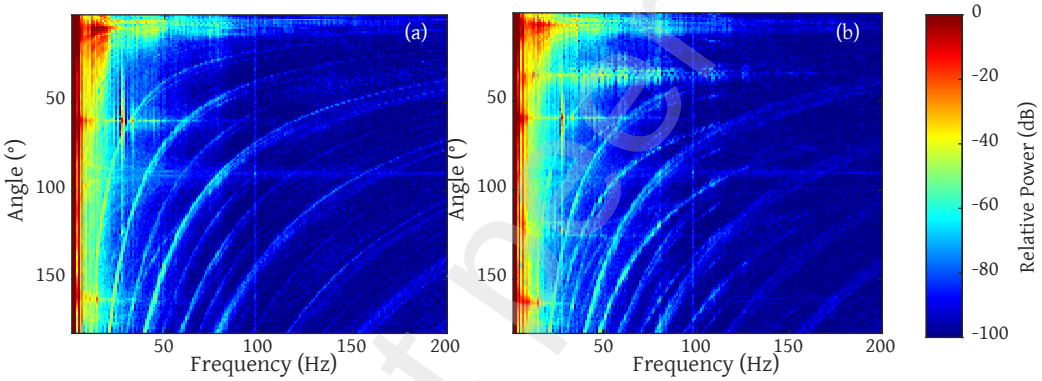


Figure 3. (a) Seafloor topography with track 1 (P1–Q1) of the target ship and track 2 (P2–Q2) of the tow ship as also presented in [21] (b) Spectrograms of signals of the 90th element of the HLA in the whole 1200 seconds.

The data were divided into 1200 segments, with a time length of each segment of 1 second, and all segments were continuous without overlap. The CBF output of each segment of the signal is calculated. The lower and upper limit frequencies of broadband CBF are 1 and 1000 Hz, respectively, with an interval of 1 Hz, and the range of the beam angle is 0°–180°. Figure 4 presents the samples of the CBF output of the 200th and 800th data segments at each frequency point in the low-frequency band (1–200 Hz). The middle-high frequency (200–1000 Hz) energy is relatively weaker but is still used as CNN input to mine potential features.



(a) 200th segment (b) 800th segment

Figure 4. CBF output of the 200th and 800th data segments

Figure 5(a) shows the spatial spectrum obtained by the broadband CBF for the whole data, in which five sources with strong energy and some other sources with weak energy can be observed. Among them, the bright line within beam angle of 10° is the self-radiated noise of the tow ship. In the 1st to 584th seconds, the target ship is moored at the station with the main and auxiliary engine shutting down. At about 585 seconds, the target ship starts to sail, corresponding to a beam angle of about 30°–50°. In addition, 3 bright lines with strong energy can be observed at the beam angle of about 50°–180°. DOA estimation of each sources was obtained by finding the peaks of these angular sectors on the CBF spatial spectrum. Figure 5(b) shows the DOA of each ship estimated by CBF. The estimated directions of the target ship are verified to be in agreement with its GPS directions.

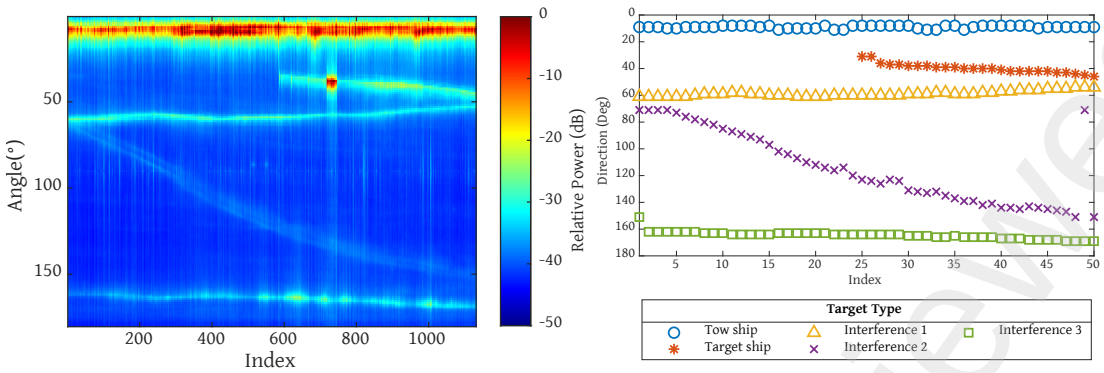


Figure 5. (a) Spatial spectrum via the index of the data segments estimated by CBF; (b) DOA estimation of the tow ship (blue circle), target ship (orange asterisk), and the interference vessel 1 (yellow triangle), 2 (purple cross), and 3 (green square)

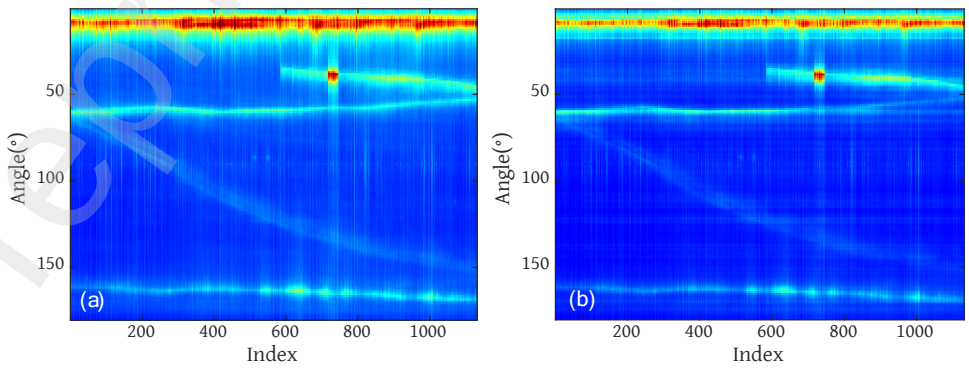
4 Result Analysis

BP and Adam algorithm are applied in the neural network to train the attention parameters. The initial learning rate is 0.001, and other parameters follow the default values as provided in the paper[26]. The loss function is the cross-entropy function. Dropout regularization[27] with a 90% probability of neuron activation is used in each iteration. The initial weight of the neural network is generated by the truncated Gaussian distribution model with a standard deviation of 0.1. The fully connected layers in detection module are activated by the ReLU function, and the output layer contains two softmax nodes, which represents the present and absent probability of the target, respectively, whose value is between [0, 1]. The training batch size is set to 128 as decided by the GPU computing performance, and the number of iterations of model training is 30.

In Section 4.1, the HLA data segments are randomly separated into two parts, where 80% of samples are selected for training the attention module of A-CBF model, and the spatial spectrum of whole data was obtained at each training iteration. In Section 4.2, the other 20% samples are used to test the detection module of the A-CBF model and for comparison with conventional energy detector[28].

4.1 DOA estimation of attention module

After 30 iterations of network training, the direction spectrum of all samples output by CBF and A-CBF with 5^{th}, 10^{th}, and 30^{th} iterations is shown according to time in Figure 6. The figure shows that the A-CBF output energy is mainly distributed on the beams where the target ship is located. In addition, A-CBF shows an obvious suppression effect on the noise of other interference vessels, with the strong noise of the tow ship in particular being significantly suppressed in the spatial spectrum. The DOA estimation of the target ship can be directly obtained by searching for the peak in Figure 6(d). The result is verified to be in agreement with the GPS records.



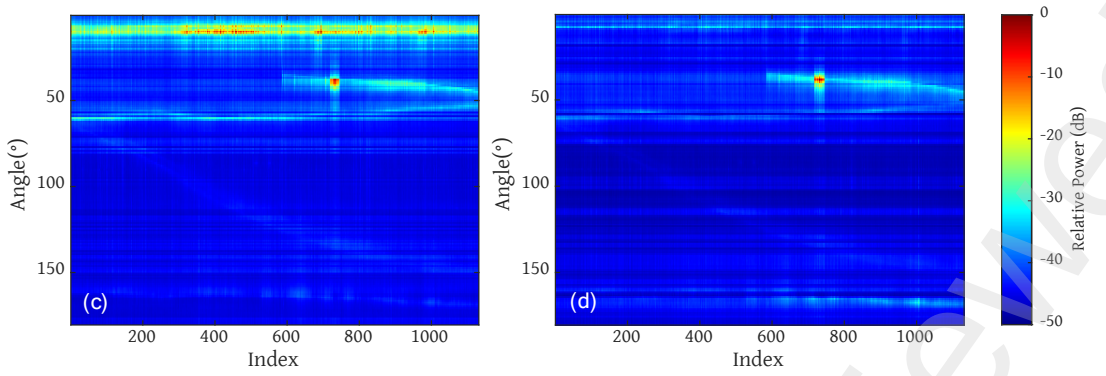
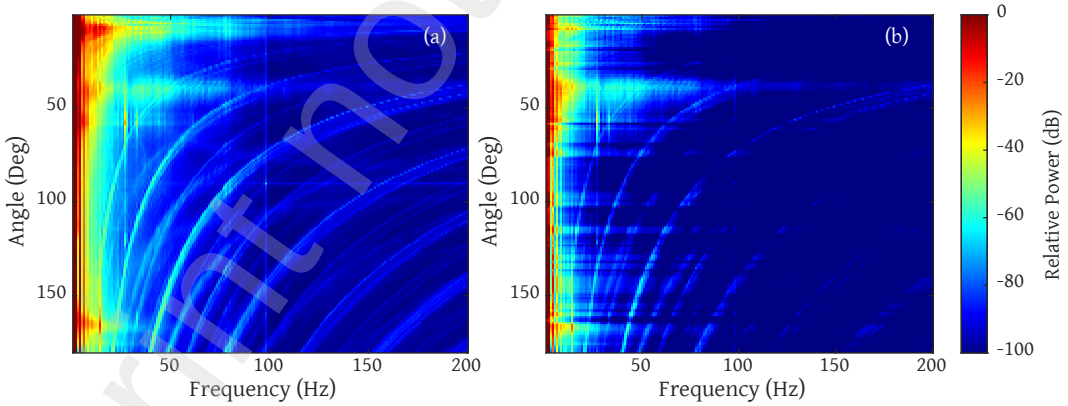


Figure 6. Spatial spectrum via the index of the data segments obtained by (a) CBF and (b-d) A-CBF at (b) 5th iteration, (c) 10th iteration, and (d) 30th iteration

The average CBF and A-CBF output among the 585th–1200th segments in the frequency-beam domain are shown in Figures 7(a) and 7(b), respectively. As can be seen from the figure, the A-CBF serves as a frequency filter, suppressing the frequency component of the interference. For CBF, the tow ship noise (~10°) has a very strong low-frequency radiated noise at 0–50 Hz, which is much stronger than that of the target ship (~40°). However, for A-CBF, the tow ship noise above 10 Hz is obviously suppressed, which means that it becomes weaker than that of the target ship. The frequency of high energy with some beam angles above 50° is also suppressed.

The average result of 585th–1200th segments for the CBF and A-CBF direction spectrum is shown in Figure 7(c). This result indicates that the A-CBF output power in the target direction is about 10 dB higher than that of traditional CBF. In addition, the tow ship's noise energy is reduced by about 8 dB.

However, A-CBF has a larger energy gain in the directions above 160° probably because of overfitting, as the features of these beams benefit the loss reduction on our limited dataset. Moreover, A-CBF excessively suppressed some beam angles that did not contribute to loss reduction, probably because the CNN was overlearned. To avoid these problems, a larger training set is needed to improve the generalization ability of the CNN model, and better end conditions should be used.



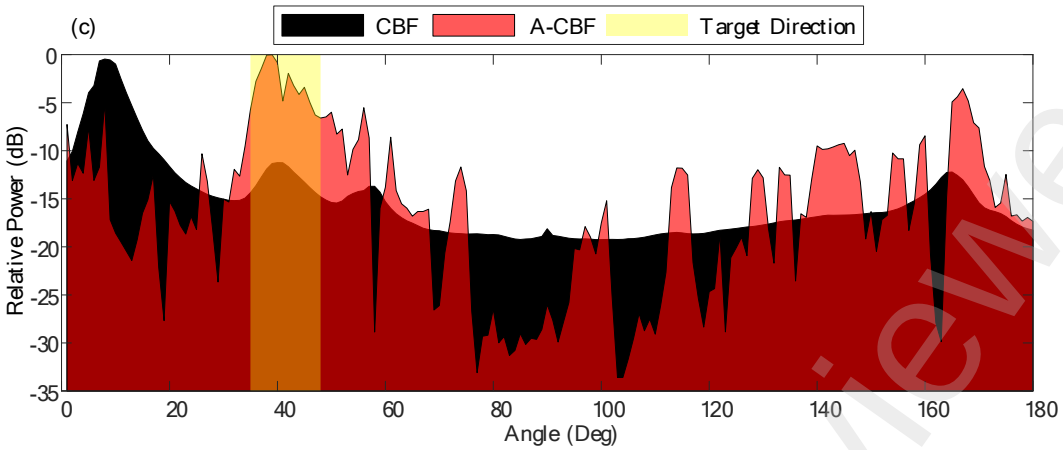


Figure 7. (a) Average CBF output of the 585–1200th segments (b) normalized average A-CBF output at 585–1200 seconds (c) comparison of spatial spectra between the CBF and A-CBF

4.2 Target detection of detection module

In the attention-based CNN, the detection module determines the presence of the desired target and propagates the error back to the attention module to optimize the attention weight of A-CBF. Therefore, the accuracy of target detection can directly affect the performance of A-CBF.

The loss curve and accuracy of target detection via training iteration for the attention-based CNN are shown in Figure 8. With the increase in the iterations, the neural network gradually focuses on the target ship, and the model sensitivity to the interference target and background noise decreases, thus achieving high accuracy and low cross-entropy error on the training set (98.62%, J = 0.014) and test set (97.92%, J = 0.016).

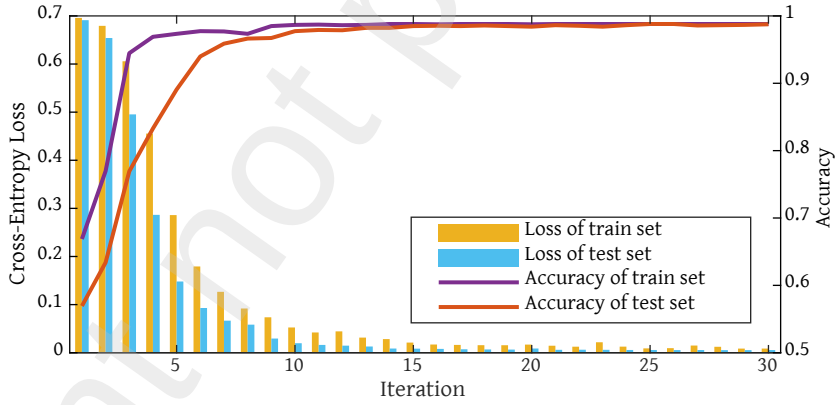


Figure 8. Loss curve and target detection accuracy of the A-CBF CNN

The detection performance of A-CBF is compared with that of a traditional energy detector[28] and a traditional CNN. First, for a comparison with the energy detector, broadband CBF was performed on the 240 test samples in the angular sector and frequency band of the target signal (30°–50° and 0–1000 Hz). The signal is in-phase stacked in the estimated DOA to increase the SNR, and then the energy is compared with a detection threshold (DT) required by the sonar system. Theoretically, the DT is selected to maximize the probability of detection for a given probability of false alarm according to the Neyman–Pearson criterion[29]. Here a DT is simply adopted with the highest accuracy (97.50%) resulting from this test dataset to compare with the CNN results. Then, to compare with a traditional CNN without the attention mechanism, the attention module in Figure 1 is eliminated, and the CBF output is directly fed to the detection module for target determination. The number of layers and neurons and the activation function are the same as those in a detection module. Moreover, the training strategies, including loss function, optimizer, regularization, and iteration number, are all the

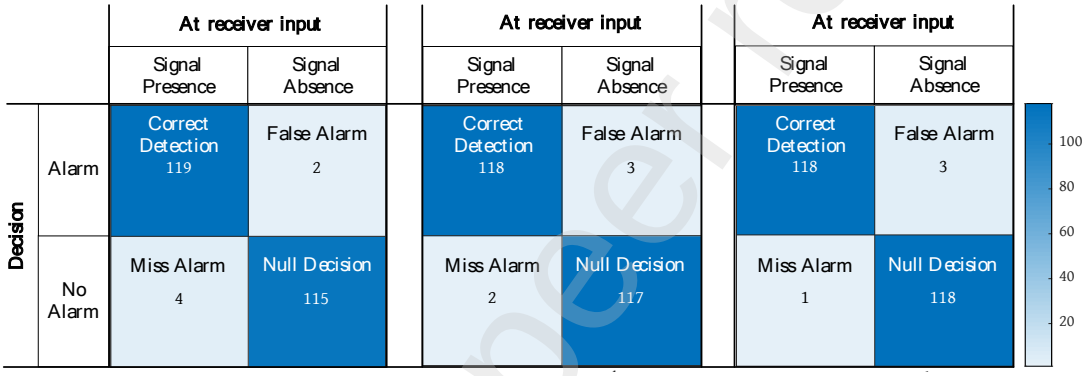
same.

To measure the detection performance, the definition of evaluation indexes of accuracy, precision, recall, false-alarm and missing-alarm rate, and the confusion matrix are given[30]. Suppose N_{TP} represents the number of correctly determined positive samples, N_{TN} represents the number of correctly determined negative samples, N_{FP} represents the number of false-alarm negative samples, and N_{FN} represents the number of missing-alarm positive samples. These four parameters constitute the 2 × 2 confusion matrix of target detection. Figure 9 shows the confusion matrix of the energy detector, traditional CNN, and attention-based CNN, for a total of 240 test samples.

Then, the accuracy rate $R_{ACC,}$ precision rate R_{PCS}, recall rate R_{TP}, false-alarm rate R_{FP}, and missing-alarm rate R_{FN} is calculated by using the confusion matrix and are respectively defined as

$$R_{ACC} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \quad R_{PCS} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad R_{TP} = \frac{N_{TP}}{N_{TP} + N_{FN}},$$

$$R_{FP} = \frac{N_{FP}}{N_{TN} + N_{FP}}, \quad R_{FN} = \frac{N_{FN}}{N_{TP} + N_{FN}}$$



(a) Energy detector (b) Traditional CNN (c) Attention-based CNN

Figure 9. Comparison of confusion matrix

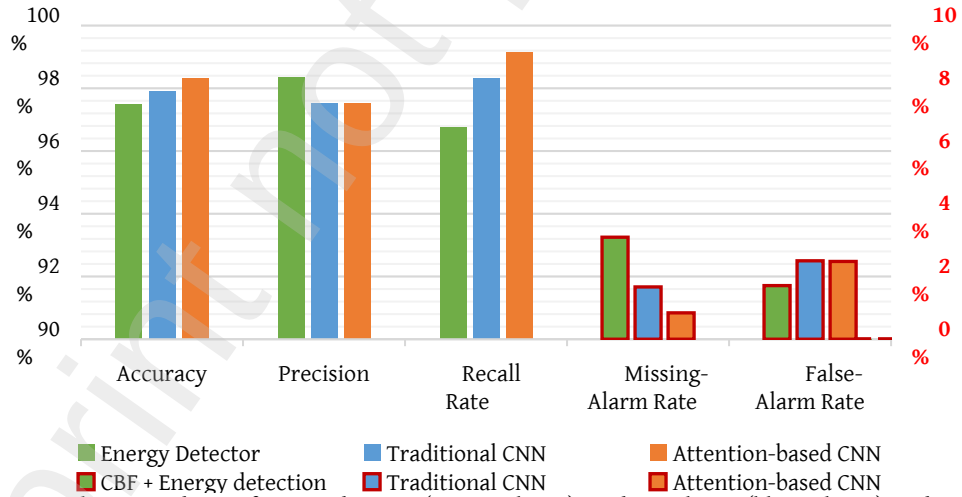


Figure 10. Evaluation indexes of energy detector (green column), traditional CNN (blue column), and attention-based CNN (orange column), and the red-framed columns correspond to the auxiliary coordinate axes on the right

Table 1. Detection accuracy

Method	Accuracy	Precision	Recall Rate	Missing-Alarm Rate	False-Alarm Rate
Energy Detector	97.50%	**98.35%**	96.75%	3.25%	**1.71%**
Traditional CNN	97.92%	97.52%	98.33%	1.67%	2.50%
Attention-based CNN	**98.33%**	97.52%	**99.16%**	**0.84%**	2.48%

Figure 10 and Table 1 show the comparison of evaluation indexes in the test data among the energy detector, traditional CNN, and the attention-based CNN.

When compared with those of the energy detector, the accuracy rate and recall rate of attention-based CNN (98.33% and 99.16%) were 0.83% and 2.41% higher than those of the energy detector (97.50% and 96.75%), respectively. This result occurred probably because the training data used for attention-based CNN have a similar feature distribution to the test data. Thus, the test error can decrease and converge to a low value with each iteration, as shown in Figure 8. However, the precision of attention-based CNN (98.35%) was 0.83% lower than that of the energy detector (97.52%) because, without any cofrequency and codirection interference, the energy detector can precisely determine if the target is present through threshold comparison, while the attention-based CNN may be overfitted due to the diversity of data and the complexity of the model. In addition, A-CBF method has a lower missing-alarm rate (2.41% lower) but a higher false-alarm rate (0.77% higher), probably because the attention-based CNN uses data mining for deep features rather than a single energy value, thus achieving a lower rate of missed detection. As for the energy detector, the simple threshold detection method has a higher confidence level and fewer false alarms in the case of no cofrequency or codirection interference. Moreover, for attention-based CNN, all training samples contribute equally to the cost function in CNN training. Thus, the proportion of positive and negative samples determines the possibility of false alarm and missing alarm. However, for the energy detector, the preset value of DT affects the correlation between false alarm and missing alarm.

When compared with traditional CNN, the attention-based CNN has a similar performance, as indicated by the less than 1% difference in each index (Figure 10). However, unlike traditional CNN, attention-based CNN is able to visualize the spatial orientation and frequency domain it focuses on, as shown in Figures 6 and 7, thus explaining its principle of determining the presence of targets. The attention mechanism intuitively shows the target features to which the CNN pays attention, thus making the decision more credible. More importantly, the overfitting of CNN can be directly reflected in the distribution of attention, as mentioned above. Therefore, the attention mechanism can be used to evaluate the reliability of models and decisions.

5 Conclusion

In this paper, attention mechanism is used for DOA estimation based on CBF. The spatial spectrum obtained by attention-based CBF model is more focused on the energy peak of the desired target, and multisource interference is suppressed, thus achieving a higher target detection rate simultaneously. The model is characterized by the following features:

- It focuses on the energy peak of the desired target in the spatial spectrum and suppresses those of multisource interference.
- It conducts target detection in the meantime.
- It uses detection errors to optimize the A-CBF coefficients adaptively.
- It has a more interpretable CNN structure.

The model was tested during a South China Sea experiment conducted in September 2020. The results show that the spatial spectrum obtained by A-CBF suppresses the noise of interfering ships and presents a clear energy peak of the target ship for better DOA estimation. The CNN also performs target detection on the test data. Results show that the proposed model has higher detection accuracy and recall rate and fewer false alarms than the traditional energy

detector and CNN. However, this model has a slightly higher missing-alarm rate than the traditional energy detector does, thereby indicating overfitting caused by data shortage, although the proposed CNN structure has reduced many unnecessary parameters. To further reduce the dependence on the amount of data, data enhancement techniques such as simulation expansion could be used to enrich the dataset in future work.

References

[1] Michael J. Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A. Roch, Sharon G annot, and Charles-Alban Deledalle , "Machine learning in acoustics: Theory and applica tions", The Journal of the Acoustical Society of America 146, 3590-3628 (2019) https://doi.org/10.1121/1.5133944

[2] Emma Ozanich, Peter Gerstoft, and Haiqiang Niu , "A feedforward neural network for di rection-of-arrival estimation", The Journal of the Acoustical Society of America 147, 203 5-2048 (2020) https://doi.org/10.1121/10.0000944

[3] Haiqiang Niu, Zaixiao Gong, Emma Ozanich, Peter Gerstoft, Haibin Wang, and Zhenglin Li , "Deep-learning source localization using multi-frequency magnitude-only data", The Journal of the Acoustical Society of America 146, 211-222 (2019) https://doi.org/10.1121/1.5116016

[4] Wenbo Wang, Haiyan Ni, Lin Su, Tao Hu, Qunyan Ren, Peter Gerstoft, and Li Ma , "Dee p transfer learning for source ranging: Deep-sea experiment results", The Journal of the Acoustical Society of America 146, EL317-EL322 (2019) https://doi.org/10.1121/1.5126923

[5] E. Ozanich, P. Gerstoft and H. Niu, "A Deep Network for Single-Snapshot Direction of A rrival Estimation," 2019 IEEE 29th International Workshop on Machine Learning for Sign al Processing (MLSP), 2019, pp. 1-6, doi: 10.1109/MLSP.2019.8918746.

[6] Huaigang Cao, Wenbo Wang, Lin Su, Haiyan Ni, Peter Gerstoft, Qunyan Ren, and Li Ma , "Deep transfer learning for underwater direction of arrival using one vector sensor", The Journal of the Acoustical Society of America 149, 1699-1711 (2021) https://doi.org/10.1121/10.0003645

[7] Liu, Yuji, Huixiu Chen, and Biao Wang. "DOA estimation based on CNN for underwater acoustic array." Applied Acoustics 172 (2021): 107594.

[8] Junjun Jiang, ZhenningWu, Min Huang, and Zhongzhe Xiao. "Detection of underwater ac oustic target using beamforming and neural network in shallow water." Applied Acousti cs 189 (2022): 108626.

[9] H. Cao, W. Wang, H. Ni, Q. Ren and L. Ma, "Deep Learning for DOA Estimation Using a Vector Hydrophone," OCEANS 2019 MTS/IEEE SEATTLE, 2019, pp. 1-4, doi: 10.23919/OCEANS40490.2019.8962679.

[10] M. Wajid, B. Kumar, A. Goel, A. Kumar and R. Bahl, "Direction of Arrival Estimation wi th Uniform Linear Array based on Recurrent Neural Network," 2019 5th International C onference on Signal Processing, Computing and Control (ISPCC), 2019, pp. 361-365, doi: 10.1109/ISPCC48220.2019.8988399.

[11] Shen S, Yang H, Yao X, Li J, Xu G, Sheng M. Ship Type Classification by Convolutional Neural Networks with Auditory-Like Mechanisms. Sensors. 2020; 20(1):253. https://doi.org/10.3390/s20010253

[12] C. Li, Z. Huang, J. Xu and Y. Yan, "Underwater target classification using deep learning, " OCEANS 2018 MTS/IEEE Charleston, 2018, pp. 1-5, doi: 10.1109/OCEANS.2018.8604906.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A surve y of methods for explaining black box mod-els,” ACM Comput. Surv. 51, 1–42 (2019).

[14] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv:1409.0473 (2014).

[15] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, “An attentive survey of attentio n models,” arXiv:1904.02874 (2019).
[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “ Show, attend and tell: Neural image captiongeneration with visual attention,” in Procee dings of the International Conference on Machine Learning, Lille, France (July 6–11, 201 5).
[17] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to a ttention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
[18] Poplin, R., Varadarajan, A.V., Blumer, K. et al. Prediction of cardiovascular risk factors f rom retinal fundus photographs via deep learning. Nat Biomed Eng 2, 158–164 (2018). h ttps://doi.org/10.1038/s41551-018-0195-0
[19] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network fo r large vocabulary conversational speech recog-nition,” in Proceedings of the 2016 ICAS SP, Shanghai, China (March 20–25, 2016), pp. 4960–4964.
[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, and X. Tang, “Residual attention n etwork for image classification,” inProceedings of the IEEE Conference on Computer Visi on and Pattern Recognition, Honolulu, HI (July 21–26, 2017), pp. 3156–3164.
[21] Xu Xiao, Wenbo Wang, Qunyan Ren, Peter Gerstoft, and Li Ma , "Underwater acoustic t arget recognition using attention-based deep neural network", JASA Express Letters 1, 1 06001 (2021) https://doi.org/10.1121/10.0006299
[22] X. Xiao, W. Wang, Q. Ren, M. Zhao and L. Ma, "Source Ranging Using Attention-Based Convolutional Neural Network," 2021 OES China Ocean Acoustics (COA), 2021, pp. 1038-1 042, doi: 10.1109/COA50123.2021.9519915.
[23] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representatio ns by back-propagating errors." nature 323.6088 (1986): 533-536.
[24] Wenbo Wang, Zhen Wang, Lin Su, Tao Hu, Qunyan Ren, Peter Gerstoft, and Li Ma , "S ource depth estimation using spectral transformations and convolutional neural network in a deep-sea environment", The Journal of the Acoustical Society of America 148, 3633 -3644 (2020) https://doi.org/10.1121/10.0002911
[25] T. Hastie, R. Tibshriani, and J. Friedman, The Elements of Statistical Learning, Springer Series in Statis-tics Springer New York Inc., New York, 2ndedition, 2009, Chap. 11, p. 3 93.
[26] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXi v preprint arXiv:1412.6980 (2014).
[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networksfrom overfitting,” J. Mach. Learn. Res. 15, 1929 –1958 (2014).
[28] Ziomek L J. Fundamentals of acoustic field theory and space-time signal processing[M]. CRC press, 2020.
[29] Neyman, Jerzy, and Egon Sharpe Pearson. "IX. On the problem of the most efficient tes ts of statistical hypotheses." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231.694-706 (1933): 289-337.
[30] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

Attention-based neural network for underwater acoustic target detection and direction-of-arrival estimation

Abstract: Direction-of-arrival (DOA) estimation for underwater acoustic sources is usually affected by multisource interference and ambient noise. In this study, DOA estimation is achieved by using a conventional beamformer modified by attention mechanism (A-CBF) which explores the spatial spectrum for DOA estimation that can focuses more on the peak of the desired signal while suppressing other peaks caused by interference and noise. The coefficients in A-CBF are learned by a neural network trained by array-received signals. On the basis of the above concept, the neural network determines the presence of the target in the received signals. From data obtained during a 2020 sea trial, the A-CBF model was trained by using a small amount of experiment data. The processing results demonstrate its performance of DOA estimation and target detection through suppressing multisource interference and focusing on the beams of the target ship in the spatial spectrum.

Key words: Underwater acoustic; DOA estimation; attention mechanism; neural network; beamforming.

1 Introduction

Underwater acoustic direction-of-arrival (DOA) estimation is a major function of sonar systems. Machine learning approaches have made great progress in DOA estimation of underwater acoustic targets in recent years[1]. The machine learning method usually establishes the mapping relationship between the sample covariance matrix of array signals and the classification labels of arrival directions, thus being trained as a DOA estimator. This approach has been proven to maintain good performance even under low signal-to-noise ratio (SNR) conditions[2].

As a newly machine learning method, deep neural networks (DNNs) have achieved initial success in previous works on underwater acoustic applications, such as source localization[3],[4], DOA estimation[5]-[10], and target recognition[11][12]. However, their high performance is often difficult to explain, because DNN is a “black-box” model whose internal work is not transparent[13]. DNN also lacks physical interpretability, as it works by directly establishing the mapping between the signal feature expression and target attributes.

Attention mechanism[14] is an important technology to inspect the internal work and improve the interpretability of the DNN[15]. It assigns different weights to input features through attention matrixes learned by DNN and displays their contribution to decision-making[16]. Attention-based DNNs are widely used in many fields, such as machine translation[17], image translation[18], speech recognition[19], and image classification[20] In recent years, attention-based DNNs have been proposed in line-spectrum feature extraction[21], target recognition[21], and source localization[22] for underwater acoustics. These models place attention weight on the time-frequency domain features of sonar-received signals to control sensitivities to the desired and interfering targets, thus achieving improved accuracy and interpretability.

In this paper, a conventional beamformer modified by attention mechanism (A-CBF) is proposed. The attention weights are placed on the CBF to ensure that its output is more focused on specific beams and frequencies. This feature results in the spatial spectrum obtained by A-CBF being able to focus more on the peak of the desired signal while suppressing other peaks due to multisource interference and ambient noise.

A convolutional neural network (CNN) is established to learn the attention weight coefficients and is composed of two parts: an attention module and a detection module. The attention module includes the attention weights to be learned, and it computes the A-CBF results and spatial spectrum used for DOA estimation. The detection module analyzes the A-CBF features to determine the presence of the desired target and propagate the detection error back[23] to the attention module to optimize the attention weights. The two modules share

weights through CNN to coordinate and promote each other. The model was tested by the South China Sea experiment conducted in September 2020 to perform DOA estimation and target detection, where multiple interfering sources were found in the experimental area.

The rest of this paper is organized as follows: Section 2 defines the CNN model and the attention mechanism used in this paper; Section 3 describes the details of the at-sea experiment and the experimental data processing; Section 4 analyzes the result of DOA estimation and target detection and compares it with the results of the traditional single-function methods CBF, CNN, and energy detector; and Section 5 concludes the paper.

2 Model Definition

The CNN structure used in this paper is shown in Figure 1. The CNN input is the CBF result, and the CNN output is a binary classification sequence that represents the presence of the target. The process consists of four steps. First, the array signals are processed by CBF. Second, the CBF output is weighted through an attention module based on coefficients learned from training samples to obtain the A-CBF output. Third, the A-CBF output is used to perform spatial spectrum estimation on the one hand and continues to be connected to the detection module on the other hand. Finally, the detection module determines the presence of the desired target and propagates the error back to the attention module to optimize the attention weight of A-CBF.

Several one-snapshot samples are needed to train the model; these samples are labeled as either class 0 (target is not present) or class 1 (target is present). Then, both DOA estimation and target detection are performed when a test is performed on the unlabeled samples, as the following diagram shows:

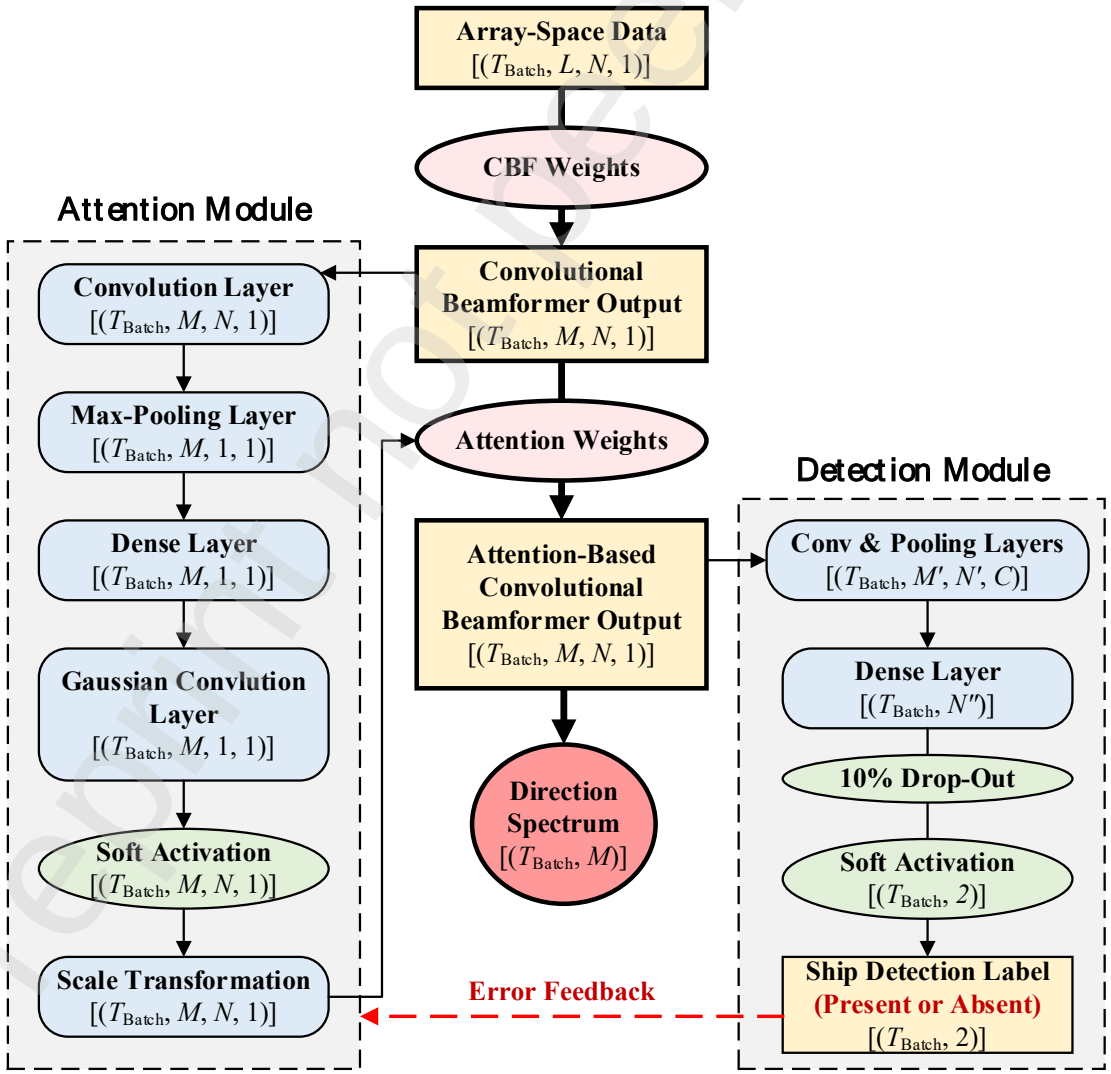


Figure 1. CNN structure for direction spectrum estimation

2.1 A-CBF framework

An L-element horizontal line array (HLA) with uniform receiver separation d is shown in Figure 2. The received array signals are divided into T segments by time. The T segments are the single-snapshot samples that are to be processed for CNN input and then batched with a batch size of T_{Batch}.

The t-segment HLA that received the sound field from a broadband point source is represented by $^{(t)}(\square,l)$ where $l = 1, 2, ..., L$ and $t = 1, 2, ..., T$. The HLA-received sound-field data are preprocessed to eliminate the influence of the source spectrum. To reduce the effect of the source amplitude, $^{(t)}(f,l)$ is normalized by[4]

$$\tilde{\boldsymbol{p}}^{(t)}(f,l) = \frac{\boldsymbol{p}^{(t)}(f,l)}{\sqrt{\sum_{l=1}^{L}\left|\boldsymbol{p}^{(t)}(f,l)\right|^2}} \tag{2}$$

The received HLA sound field is projected into the frequency-beam domain by CBF in the frequency domain. For an HLA with uniform element spacing, the CBF weight is the steering vector

$$\boldsymbol{w}(\theta,f) = \frac{1}{\sqrt{S}}\left[1, e^{j2\pi fd\sin\theta/c}, e^{j2\pi f\cdot 2d\sin\theta/c}, \cdots, e^{j2\pi f(L-1)d\sin\theta/c}\right]^{\mathrm{T}}, \quad \boldsymbol{w}\in C^{L\times 1} \tag{3}$$

Suppose $\boldsymbol{p}'^{(t)}(f_n)$ is the normalized sound-field vector of size $L \times 1$ composed of the received sound fields of all L elements at frequency f_n, $n = 1, 2, ..., N$, and the L beams point in the directions $\square_m$ where $\square_m = \square_1, \square_2, ..., \square_M$. Thus, the output power of CBF at frequency f_n of segment t is[24]

$$\boldsymbol{B}^{(t)}(\theta_m,f_n) = \left|\boldsymbol{w}(\theta_m,f_n)^{\mathrm{H}}\,\boldsymbol{p}'^{(t)}(f_n)\right|^2, \quad \boldsymbol{B}\in C^{M\times N}, \quad \boldsymbol{p}'^{(t)}\in C^{L\times 1} \tag{4}$$

where H represents the conjugate transpose. When the beam angle θ_m is in agreement with the target arrival angle, the beam outputs the maximum power. Finally, the broadband CBF accumulates the power of each single-frequency CBF output to obtain the broadband direction spectrum. The DOA of each source can be estimated by finding the peaks on the spatial spectrum.

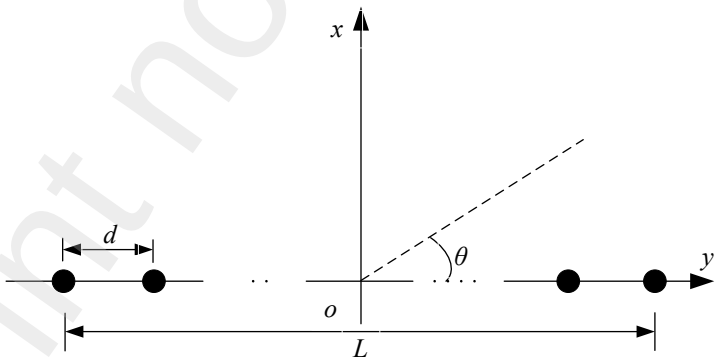


Figure 2. Sketch map of an L-element HLA

In introducing the attention mechanism, the CBF output $\boldsymbol{B}$ should be weighted with an attention weight matrix to control sensitivity to different frequency and spatial components. Suppose that the attention weight matrix is $\boldsymbol{A}$. Thus, the A-CBF output is the Hadamar product of weight matrix $\boldsymbol{A}$ and CBF output $\boldsymbol{B}$ (represented as $\circledcirc$), and the A-CBF output is defined as

$$\tilde{\boldsymbol{B}}^{(t)}(\theta_m,f_n) = \boldsymbol{A}^{*}\odot\boldsymbol{B}^{(t)}(\theta_m,f_n). \tag{5}$$

where $\boldsymbol{A}^*$ represents the optimal value of $\boldsymbol{A}$, which is to be searched in a CNN through the chain rule of backpropagation (BP)[23]. The optimization problem can be described as

$$\boldsymbol{y} = \mathrm{CNN}\left(\boldsymbol{A}, \boldsymbol{w}_{\mathrm{DNN}}, \boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}, \cdots, \boldsymbol{B}^{(T)}\right) \tag{6}$$

$$\boldsymbol{A}^{*}=\arg\min_{\boldsymbol{A},\boldsymbol{w}\in\mathbb{R}^{M\times N}} J\left(\boldsymbol{A},\boldsymbol{w}_{\mathrm{DNN}},\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)},\cdots,\boldsymbol{B}^{(T)},\boldsymbol{S}^{(1)},\boldsymbol{S}^{(2)},\cdots,\boldsymbol{S}^{(T)}\right) \tag{7}$$

where $\boldsymbol{w}_{\mathrm{DNN}}$ is the CNN weights, and $\boldsymbol{S}^{(t)}=[s_0^{(t)}, s_1^{(t)}]$ is the binary classification label of each data segment, whose value is [1, 0], which represents the presence of the target signal, or [0, 1], which represents the absence of the target signal in a data segment; $\boldsymbol{y}^{(t)}=[y_0^{(t)}, y_1^{(t)}]$ is the output sequence of the CNN activated by the softmax function[25], where the two entries represent the present and absent probability of the target, respectively, and their value range is between [0, 1]; J is the cost function, which is defined by binary cross entropy

$$J=-\frac{1}{T}\sum_{t=1}^{T}\left(s_0^{(t)}\log y_0^{(t)}+s_1^{(t)}\log y_1^{(t)}\right) \tag{8}$$

In CNN training, the BP algorithm is used to find the gradients for CNN weights and neurons with respect to the cost function, which needs an optimizer to update the weights iteratively by using those gradients. The traditional gradient descent method often has difficulty converging due to the complex underwater environment, as the cost function is usually nonconvex and easily falls into the local optimum or tends to overlearn with a fixed learning rate. Thus, an adaptive optimization method called adaptive moment estimation algorithm (Adam)[26] is used dynamically adjust the learning rate for CNN weights, which uses first and second moment estimation of the gradient to keep the learning rate within a certain range by bias correction iteratively, thus obtaining a stable parameter update.

Suppose t is the number of iterations, and u_t is any of the element of $\boldsymbol{w}_{\mathrm{DNN}}$ to be estimated in DNN in the t^{th} interation. First, the exponential moving averages of the gradient m_t, whose initial value m_0 is 0, are calculated. Considering the gradient momentum of the previous iteration, suppose the hyperparameter β_1 is the exponential decay rates

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)\nabla_u J(u_{t-1}) \tag{9}$$

The exponential moving average of the squared gradient v_t is then calculated, whose initial value v_0 is 0, and suppose the hyperparameter β_1 is the exponential decay rates

$$v_t=\beta_{t-1}+(1-\beta_2)\nabla_u J(u_{t-1})^2 \tag{10}$$

First, the bias is corrected, and then the parameters are updated

$$\begin{cases}\hat{m}_t=m_t/(1-\beta_1^t)\\ \hat{v}_t=v_t/(1-\beta_2^t)\\ u_t=u_{t-1}-\eta\cdot\hat{m}_t/\sqrt{\hat{v}_t}\end{cases} \tag{11}$$

In Equation (10), η is the initial learning rate. The Adam algorithm adaptively adjusts the update value from both the mean and uncentered variance of the gradient, thus improving the convergence efficiency.

2.2 Soft attention mechanism

The soft attention mechanism[20] is applied to define the attention weight matrix $\boldsymbol{A}$. A soft attention mask should consist of at least one convolutional pooling layer activated by the softmax function[20]. The CBF output $\boldsymbol{B}(\theta_m f_n)$ should be convolved with a convolution kernel $\boldsymbol{R}$ of size $d \times d$ whose entries are represented by R_{jk}, where $j, k = 1, 2..., K$, and the output offset is b. The dimension of the matrix resulting from a convolution operation is $(M - K + 1) \times (N - K + 1)$. Thus, the border elements are lost from the output every time a convolution operation is performed, precluding the building of deeper networks. Thus, $\mu = (K - 1)/2$ numbers of layers of zeros are padded to the border of the matrix, expanding the dimension to $(M + K - 1) \times (N + K - 1)$. The expanded matrix is denoted as $\boldsymbol{B'}$. Therefore, the output of the convolutional layer becomes a matrix $\boldsymbol{Z}$ with the dimension of $M \times N$. In forward propagation, the values of each entry z_{uv} are

$$\begin{cases}z_{uv}=\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\boldsymbol{B'}\cdot r_{jk}\cdot\xi(j,k)+b,\\ \xi(j,k)=\begin{cases}1, & 0\le j,k\le K\\ 0, & \text{others}\end{cases}.\end{cases} \tag{12}$$

where $u = 1, 2, \ldots, M$, $v = 1, 2, \ldots, N$. Then, an average pooling layer with a pooling core with a size of $1 \times N$ and a step size of 1×1 is defined. The matrix is pooled into a vector $\boldsymbol{\psi}$ of length M that corresponds to M beam angles, which is expressed as

$$\boldsymbol{\psi} = \left[\sum_{v=1}^{N} \frac{z_{1v}}{N} \quad \sum_{v=1}^{N} \frac{z_{2v}}{N} \quad \cdots \quad \sum_{v=1}^{N} \frac{z_{uv}}{N} \quad \cdots \quad \sum_{v=1}^{N} \frac{z_{Mv}}{N} \right]^T. \tag{13}$$

Then, $\boldsymbol{\psi}$ is activated by the softmax function[25] to obtain the original attention weight vector $\boldsymbol{\alpha}$ of length S, which is expressed as

$$\boldsymbol{\alpha} = \mathrm{Softmax}(\boldsymbol{W}\boldsymbol{\psi}), \tag{14}$$

where $\boldsymbol{W}$ is an attention score matrix with a size of $M \times M$, in which the values of each element are obtained through CNN learning. Thus, the attention effect on the beam domain is applied.

Considering the correlation of the sound field between adjacent beams, the attention on beam domain should maintain a beamwidth. Thus, a Gaussian layer[21] is defined. $\boldsymbol{\alpha}$ is multiplied with the Gaussian kernel function, as represented by the following matrix operations:

$$\boldsymbol{H} = \left(-\frac{1}{2\sigma^2} \begin{bmatrix} h_{11}^2 & h_{12}^2 & \cdots & h_{1M}^2 \\ h_{21}^2 & h_{22}^2 & & \\ \vdots & & \ddots & \\ h_{M1}^2 & & & h_{MM}^2 \end{bmatrix} \right), \quad h_{jk} = k - j, \quad j,k \in [1, M] \tag{15}$$

$$\boldsymbol{g} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{H} \right) \boldsymbol{\alpha}, \tag{16}$$

where $\boldsymbol{g}$ is the output vector of size M of the Gaussian layer, which is the attention weight layer and σ is the standard deviation of the Gaussian function. Larger σ gives more distracted attention[21], which ensures that the attention is not overly focused on a single beam. Finally, suppose $\boldsymbol{\varepsilon}$ is a row coefficient vector of length N. The output matrix $\boldsymbol{A}$ is expressed as

$$\boldsymbol{A} = \boldsymbol{g} \times \boldsymbol{\varepsilon}, \tag{17}$$

As a variable that could affect attention in the frequency domain, vector $\boldsymbol{\varepsilon}$ could be set as a trainable or untrainable parameter in training configuration. In this work, vector $\boldsymbol{\varepsilon}$ is simply set as an untrainable layer of an array of ones. Finally, the A-CBF outputs are given by Equation (16) and are accumulated according to frequency to obtain the broadband result for each single-snapshot sample, which is the A-CBF spatial spectrum. The DOA estimation of the source can be obtained by finding the peak.

3 Experiment

In September 2020, the ship-radiated noise measurement experiment was conducted in the northern South China Sea. The topography map of the experimental area is shown in Figure 3(a). This experiment was undertaken on the same voyage as in literature[21], while the data were recorded at a different time. As the target ship, an experiment ship sails along the track at a speed of 6 knots or is moored at stations A and B, and the tow ship with a towed HLA sails along the track C–D at a speed of 6 knots. The number of HLA elements is 179, and the element spacing is from 0.2 to 4 m. The sea floor is relatively flat, and the water depth is 125–140 m. The length of the HLA recorded data is 1200 seconds, and the sampling rate is 8 kHz. Figure 3(b) shows the time-frequency spectrograms of the signal obtained by a short-time Fourier transform of the whole data at the 90th element (intermediate element). In the first 585 seconds, the target ship is moored at the station with its main and auxiliary engines shutting down. At about 300 seconds, the tow ship begins to accelerate, as some of the line spectrum in the spectrum changes, and the ship's noise level increases. At 586 seconds, the target ship begins to sail along the track as the low-frequency energy increased.

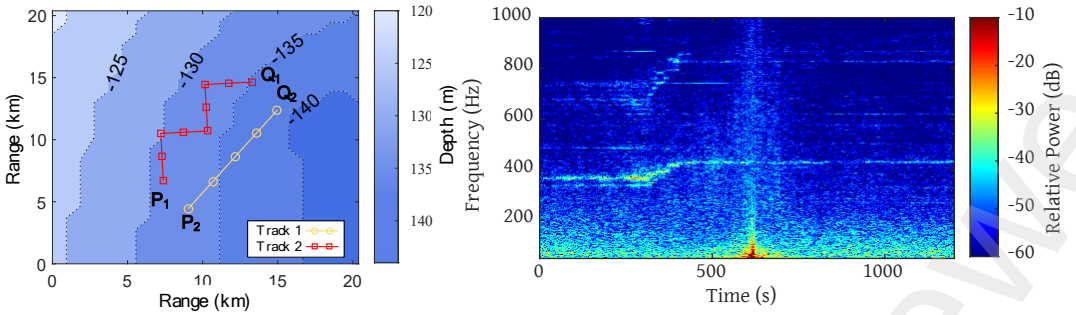
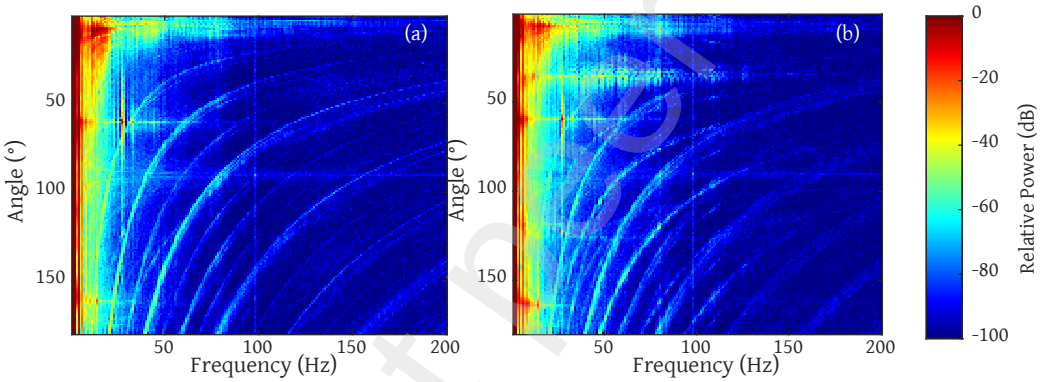


Figure 3. (a) Seafloor topography with track 1 (P1–Q1) of the target ship and track 2 (P2–Q2) of the tow ship as also presented in [21] (b) Spectrograms of signals of the 90th element of the HLA in the whole 1200 seconds.

The data were divided into 1200 segments, with a time length of each segment of 1 second, and all segments were continuous without overlap. The CBF output of each segment of the signal is calculated. The lower and upper limit frequencies of broadband CBF are 1 and 1000 Hz, respectively, with an interval of 1 Hz, and the range of the beam angle is 0°–180°. Figure 4 presents the samples of the CBF output of the 200th and 800th data segments at each frequency point in the low-frequency band (1–200 Hz). The middle-high frequency (200–1000 Hz) energy is relatively weaker but is still used as CNN input to mine potential features.



(a) 200th segment (b) 800th segment

Figure 4. CBF output of the 200th and 800th data segments

Figure 5(a) shows the spatial spectrum obtained by the broadband CBF for the whole data, in which five sources with strong energy and some other sources with weak energy can be observed. Among them, the bright line within beam angle of 10° is the self-radiated noise of the tow ship. In the 1st to 584th seconds, the target ship is moored at the station with the main and auxiliary engine shutting down. At about 585 seconds, the target ship starts to sail, corresponding to a beam angle of about 30°–50°. In addition, 3 bright lines with strong energy can be observed at the beam angle of about 50°–180°. DOA estimation of each sources was obtained by finding the peaks of these angular sectors on the CBF spatial spectrum. Figure 5(b) shows the DOA of each ship estimated by CBF. The estimated directions of the target ship are verified to be in agreement with its GPS directions.

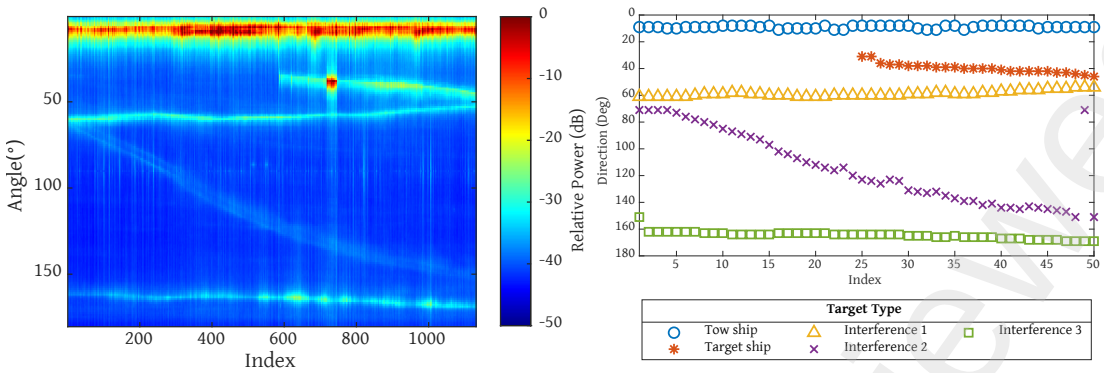


Figure 5. (a) Spatial spectrum via the index of the data segments estimated by CBF; (b) DOA estimation of the tow ship (blue circle), target ship (orange asterisk), and the interference vessel 1 (yellow triangle), 2 (purple cross), and 3 (green square)

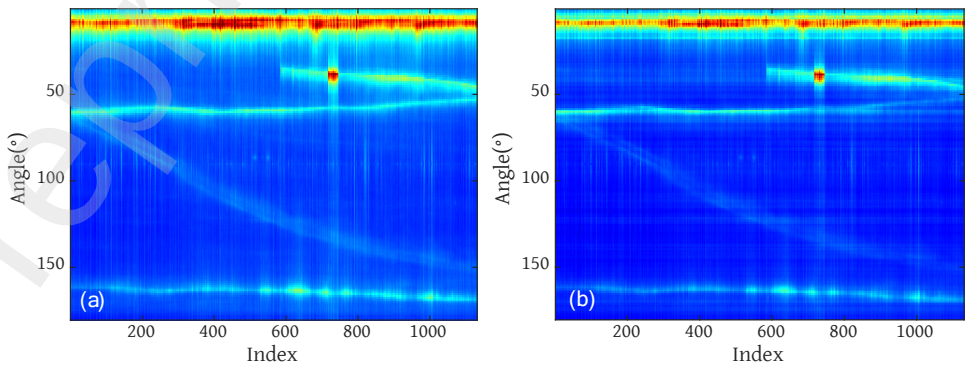
4 Result Analysis

BP and Adam algorithm are applied in the neural network to train the attention parameters. The initial learning rate is 0.001, and other parameters follow the default values as provided in the paper[26]. The loss function is the cross-entropy function. Dropout regularization[27] with a 90% probability of neuron activation is used in each iteration. The initial weight of the neural network is generated by the truncated Gaussian distribution model with a standard deviation of 0.1. The fully connected layers in detection module are activated by the ReLU function, and the output layer contains two softmax nodes, which represents the present and absent probability of the target, respectively, whose value is between [0, 1]. The training batch size is set to 128 as decided by the GPU computing performance, and the number of iterations of model training is 30.

In Section 4.1, the HLA data segments are randomly separated into two parts, where 80% of samples are selected for training the attention module of A-CBF model, and the spatial spectrum of whole data was obtained at each training iteration. In Section 4.2, the other 20% samples are used to test the detection module of the A-CBF model and for comparison with conventional energy detector[28].

4.1 DOA estimation of attention module

After 30 iterations of network training, the direction spectrum of all samples output by CBF and A-CBF with 5^{th}, 10^{th}, and 30^{th} iterations is shown according to time in Figure 6. The figure shows that the A-CBF output energy is mainly distributed on the beams where the target ship is located. In addition, A-CBF shows an obvious suppression effect on the noise of other interference vessels, with the strong noise of the tow ship in particular being significantly suppressed in the spatial spectrum. The DOA estimation of the target ship can be directly obtained by searching for the peak in Figure 6(d). The result is verified to be in agreement with the GPS records.



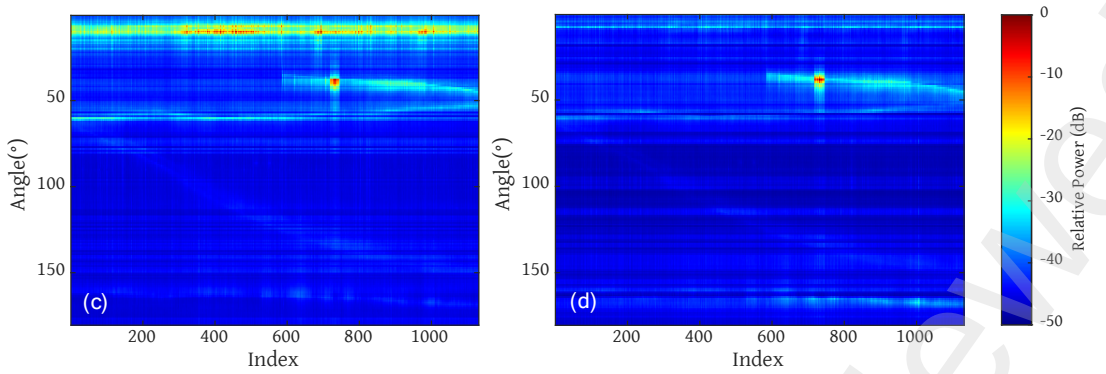
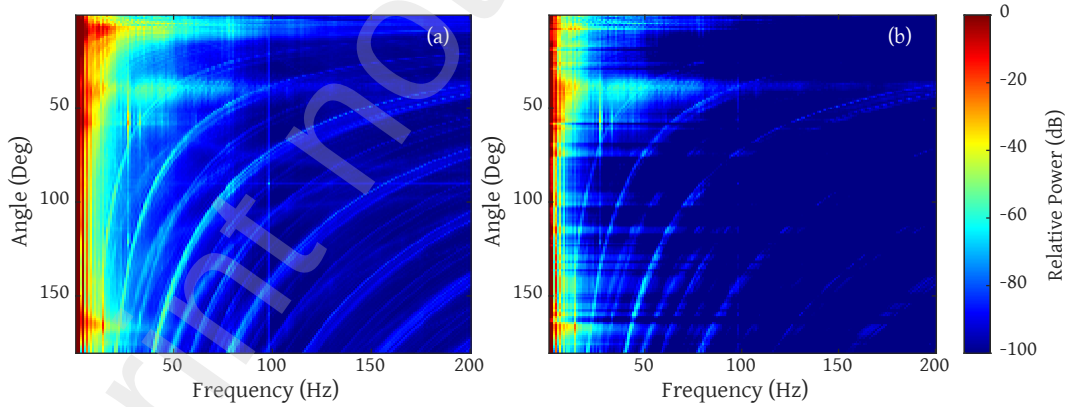


Figure 6. Spatial spectrum via the index of the data segments obtained by (a) CBF and (b-d) A-CBF at (b) 5th iteration, (c) 10th iteration, and (d) 30th iteration

The average CBF and A-CBF output among the 585th–1200th segments in the frequency-beam domain are shown in Figures 7(a) and 7(b), respectively. As can be seen from the figure, the A-CBF serves as a frequency filter, suppressing the frequency component of the interference. For CBF, the tow ship noise (~10°) has a very strong low-frequency radiated noise at 0–50 Hz, which is much stronger than that of the target ship (~40°). However, for A-CBF, the tow ship noise above 10 Hz is obviously suppressed, which means that it becomes weaker than that of the target ship. The frequency of high energy with some beam angles above 50° is also suppressed.

The average result of 585th–1200th segments for the CBF and A-CBF direction spectrum is shown in Figure 7(c). This result indicates that the A-CBF output power in the target direction is about 10 dB higher than that of traditional CBF. In addition, the tow ship's noise energy is reduced by about 8 dB.

However, A-CBF has a larger energy gain in the directions above 160° probably because of overfitting, as the features of these beams benefit the loss reduction on our limited dataset. Moreover, A-CBF excessively suppressed some beam angles that did not contribute to loss reduction, probably because the CNN was overlearned. To avoid these problems, a larger training set is needed to improve the generalization ability of the CNN model, and better end conditions should be used.



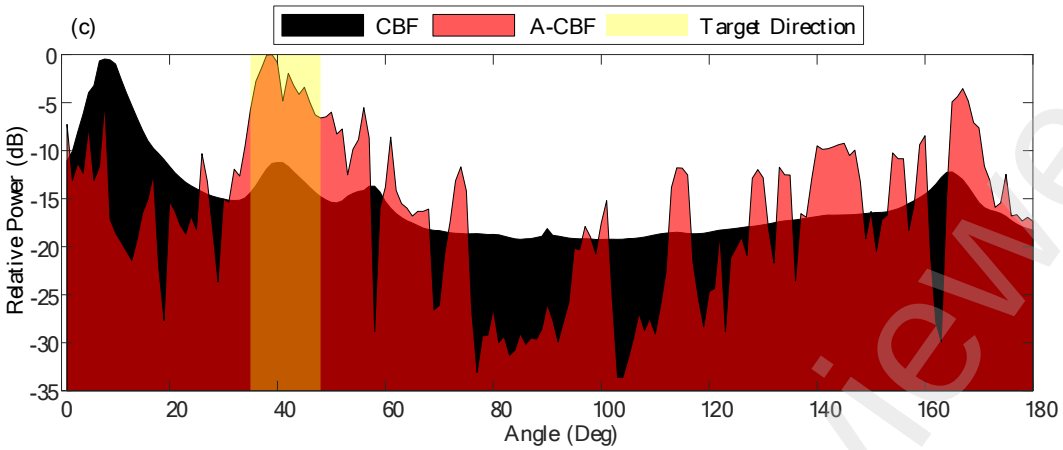


Figure 7. (a) Average CBF output of the 585–1200th segments (b) normalized average A-CBF output at 585–1200 seconds (c) comparison of spatial spectra between the CBF and A-CBF

4.2 Target detection of detection module

In the attention-based CNN, the detection module determines the presence of the desired target and propagates the error back to the attention module to optimize the attention weight of A-CBF. Therefore, the accuracy of target detection can directly affect the performance of A-CBF.

The loss curve and accuracy of target detection via training iteration for the attention-based CNN are shown in Figure 8. With the increase in the iterations, the neural network gradually focuses on the target ship, and the model sensitivity to the interference target and background noise decreases, thus achieving high accuracy and low cross-entropy error on the training set (98.62%, J = 0.014) and test set (97.92%, J = 0.016).

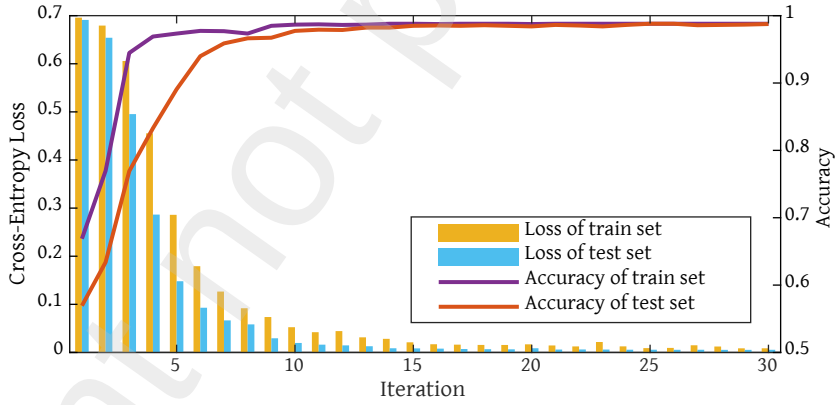


Figure 8. Loss curve and target detection accuracy of the A-CBF CNN

The detection performance of A-CBF is compared with that of a traditional energy detector[28] and a traditional CNN. First, for a comparison with the energy detector, broadband CBF was performed on the 240 test samples in the angular sector and frequency band of the target signal (30°–50° and 0–1000 Hz). The signal is in-phase stacked in the estimated DOA to increase the SNR, and then the energy is compared with a detection threshold (DT) required by the sonar system. Theoretically, the DT is selected to maximize the probability of detection for a given probability of false alarm according to the Neyman–Pearson criterion[29]. Here a DT is simply adopted with the highest accuracy (97.50%) resulting from this test dataset to compare with the CNN results. Then, to compare with a traditional CNN without the attention mechanism, the attention module in Figure 1 is eliminated, and the CBF output is directly fed to the detection module for target determination. The number of layers and neurons and the activation function are the same as those in a detection module. Moreover, the training strategies, including loss function, optimizer, regularization, and iteration number, are all the

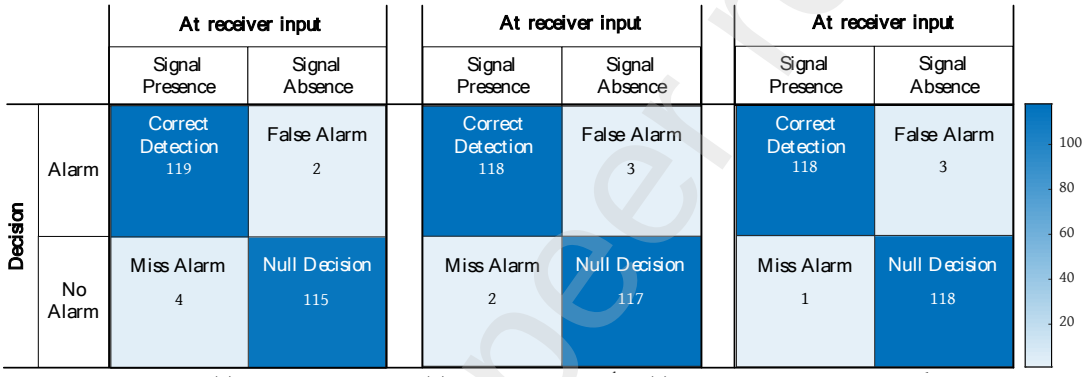
same.

To measure the detection performance, the definition of evaluation indexes of accuracy, precision, recall, false-alarm and missing-alarm rate, and the confusion matrix are given[30]. Suppose N_{TP} represents the number of correctly determined positive samples, N_{TN} represents the number of correctly determined negative samples, N_{FP} represents the number of false-alarm negative samples, and N_{FN} represents the number of missing-alarm positive samples. These four parameters constitute the 2 × 2 confusion matrix of target detection. Figure 9 shows the confusion matrix of the energy detector, traditional CNN, and attention-based CNN, for a total of 240 test samples.

Then, the accuracy rate $R_{ACC,}$ precision rate R_{PCS}, recall rate R_{TP}, false-alarm rate R_{FP}, and missing-alarm rate R_{FN} is calculated by using the confusion matrix and are respectively defined as

$$R_{ACC} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \quad R_{PCS} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad R_{TP} = \frac{N_{TP}}{N_{TP} + N_{FN}},$$

$$R_{FP} = \frac{N_{FP}}{N_{TN} + N_{FP}}, \quad R_{FN} = \frac{N_{FN}}{N_{TP} + N_{FN}}$$



(a) Energy detector (b) Traditional CNN (c) Attention-based CNN

Figure 9. Comparison of confusion matrix

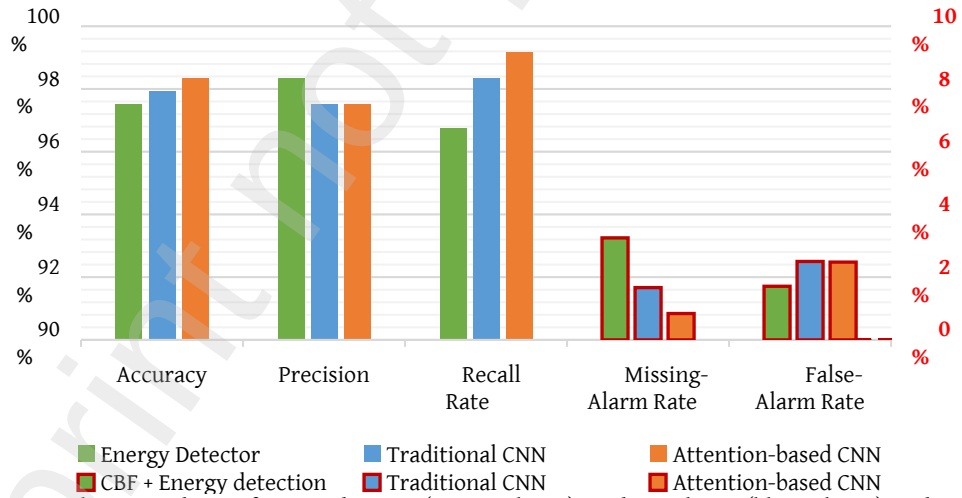


Figure 10. Evaluation indexes of energy detector (green column), traditional CNN (blue column), and attention-based CNN (orange column), and the red-framed columns correspond to the auxiliary coordinate axes on the right

Table 1. Detection accuracy

Method	Accuracy	Precision	Recall Rate	Missing-Alarm Rate	False-Alarm Rate
Energy Detector	97.50%	**98.35%**	96.75%	3.25%	**1.71%**
Traditional CNN	97.92%	97.52%	98.33%	1.67%	2.50%
Attention-based CNN	**98.33%**	97.52%	**99.16%**	**0.84%**	2.48%

Figure 10 and Table 1 show the comparison of evaluation indexes in the test data among the energy detector, traditional CNN, and the attention-based CNN.

When compared with those of the energy detector, the accuracy rate and recall rate of attention-based CNN (98.33% and 99.16%) were 0.83% and 2.41% higher than those of the energy detector (97.50% and 96.75%), respectively. This result occurred probably because the training data used for attention-based CNN have a similar feature distribution to the test data. Thus, the test error can decrease and converge to a low value with each iteration, as shown in Figure 8. However, the precision of attention-based CNN (98.35%) was 0.83% lower than that of the energy detector (97.52%) because, without any cofrequency and codirection interference, the energy detector can precisely determine if the target is present through threshold comparison, while the attention-based CNN may be overfitted due to the diversity of data and the complexity of the model. In addition, A-CBF method has a lower missing-alarm rate (2.41% lower) but a higher false-alarm rate (0.77% higher), probably because the attention-based CNN uses data mining for deep features rather than a single energy value, thus achieving a lower rate of missed detection. As for the energy detector, the simple threshold detection method has a higher confidence level and fewer false alarms in the case of no cofrequency or codirection interference. Moreover, for attention-based CNN, all training samples contribute equally to the cost function in CNN training. Thus, the proportion of positive and negative samples determines the possibility of false alarm and missing alarm. However, for the energy detector, the preset value of DT affects the correlation between false alarm and missing alarm.

When compared with traditional CNN, the attention-based CNN has a similar performance, as indicated by the less than 1% difference in each index (Figure 10). However, unlike traditional CNN, attention-based CNN is able to visualize the spatial orientation and frequency domain it focuses on, as shown in Figures 6 and 7, thus explaining its principle of determining the presence of targets. The attention mechanism intuitively shows the target features to which the CNN pays attention, thus making the decision more credible. More importantly, the overfitting of CNN can be directly reflected in the distribution of attention, as mentioned above. Therefore, the attention mechanism can be used to evaluate the reliability of models and decisions.

5 Conclusion

In this paper, attention mechanism is used for DOA estimation based on CBF. The spatial spectrum obtained by attention-based CBF model is more focused on the energy peak of the desired target, and multisource interference is suppressed, thus achieving a higher target detection rate simultaneously. The model is characterized by the following features:

- It focuses on the energy peak of the desired target in the spatial spectrum and suppresses those of multisource interference.
- It conducts target detection in the meantime.
- It uses detection errors to optimize the A-CBF coefficients adaptively.
- It has a more interpretable CNN structure.

The model was tested during a South China Sea experiment conducted in September 2020. The results show that the spatial spectrum obtained by A-CBF suppresses the noise of interfering ships and presents a clear energy peak of the target ship for better DOA estimation. The CNN also performs target detection on the test data. Results show that the proposed model has higher detection accuracy and recall rate and fewer false alarms than the traditional energy

detector and CNN. However, this model has a slightly higher missing-alarm rate than the traditional energy detector does, thereby indicating overfitting caused by data shortage, although the proposed CNN structure has reduced many unnecessary parameters. To further reduce the dependence on the amount of data, data enhancement techniques such as simulation expansion could be used to enrich the dataset in future work.

References

[1] Michael J. Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A. Roch, Sharon G annot, and Charles-Alban Deledalle , "Machine learning in acoustics: Theory and applica tions", The Journal of the Acoustical Society of America 146, 3590-3628 (2019) https://doi.org/10.1121/1.5133944

[2] Emma Ozanich, Peter Gerstoft, and Haiqiang Niu , "A feedforward neural network for di rection-of-arrival estimation", The Journal of the Acoustical Society of America 147, 203 5-2048 (2020) https://doi.org/10.1121/10.0000944

[3] Haiqiang Niu, Zaixiao Gong, Emma Ozanich, Peter Gerstoft, Haibin Wang, and Zhenglin Li , "Deep-learning source localization using multi-frequency magnitude-only data", The Journal of the Acoustical Society of America 146, 211-222 (2019) https://doi.org/10.1121/1.5116016

[4] Wenbo Wang, Haiyan Ni, Lin Su, Tao Hu, Qunyan Ren, Peter Gerstoft, and Li Ma , "Dee p transfer learning for source ranging: Deep-sea experiment results", The Journal of the Acoustical Society of America 146, EL317-EL322 (2019) https://doi.org/10.1121/1.5126923

[5] E. Ozanich, P. Gerstoft and H. Niu, "A Deep Network for Single-Snapshot Direction of A rrival Estimation," 2019 IEEE 29th International Workshop on Machine Learning for Sign al Processing (MLSP), 2019, pp. 1-6, doi: 10.1109/MLSP.2019.8918746.

[6] Huaigang Cao, Wenbo Wang, Lin Su, Haiyan Ni, Peter Gerstoft, Qunyan Ren, and Li Ma , "Deep transfer learning for underwater direction of arrival using one vector sensor", The Journal of the Acoustical Society of America 149, 1699-1711 (2021) https://doi.org/10.1121/10.0003645

[7] Liu, Yuji, Huixiu Chen, and Biao Wang. "DOA estimation based on CNN for underwater acoustic array." Applied Acoustics 172 (2021): 107594.

[8] Junjun Jiang, ZhenningWu, Min Huang, and Zhongzhe Xiao. "Detection of underwater ac oustic target using beamforming and neural network in shallow water." Applied Acousti cs 189 (2022): 108626.

[9] H. Cao, W. Wang, H. Ni, Q. Ren and L. Ma, "Deep Learning for DOA Estimation Using a Vector Hydrophone," OCEANS 2019 MTS/IEEE SEATTLE, 2019, pp. 1-4, doi: 10.23919/OCEANS40490.2019.8962679.

[10] M. Wajid, B. Kumar, A. Goel, A. Kumar and R. Bahl, "Direction of Arrival Estimation wi th Uniform Linear Array based on Recurrent Neural Network," 2019 5th International C onference on Signal Processing, Computing and Control (ISPCC), 2019, pp. 361-365, doi: 10.1109/ISPCC48220.2019.8988399.

[11] Shen S, Yang H, Yao X, Li J, Xu G, Sheng M. Ship Type Classification by Convolutional Neural Networks with Auditory-Like Mechanisms. Sensors. 2020; 20(1):253. https://doi.org/10.3390/s20010253

[12] C. Li, Z. Huang, J. Xu and Y. Yan, "Underwater target classification using deep learning, " OCEANS 2018 MTS/IEEE Charleston, 2018, pp. 1-5, doi: 10.1109/OCEANS.2018.8604906.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A surve y of methods for explaining black box mod-els,” ACM Comput. Surv. 51, 1–42 (2019).

[14] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv:1409.0473 (2014).

[15] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, “An attentive survey of attentio n models,” arXiv:1904.02874 (2019).
[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “ Show, attend and tell: Neural image captiongeneration with visual attention,” in Procee dings of the International Conference on Machine Learning, Lille, France (July 6–11, 201 5).
[17] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to a ttention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
[18] Poplin, R., Varadarajan, A.V., Blumer, K. et al. Prediction of cardiovascular risk factors f rom retinal fundus photographs via deep learning. Nat Biomed Eng 2, 158–164 (2018). h ttps://doi.org/10.1038/s41551-018-0195-0
[19] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network fo r large vocabulary conversational speech recog-nition,” in Proceedings of the 2016 ICAS SP, Shanghai, China (March 20–25, 2016), pp. 4960–4964.
[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, and X. Tang, “Residual attention n etwork for image classification,” inProceedings of the IEEE Conference on Computer Visi on and Pattern Recognition, Honolulu, HI (July 21–26, 2017), pp. 3156–3164.
[21] Xu Xiao, Wenbo Wang, Qunyan Ren, Peter Gerstoft, and Li Ma , "Underwater acoustic t arget recognition using attention-based deep neural network", JASA Express Letters 1, 1 06001 (2021) https://doi.org/10.1121/10.0006299
[22] X. Xiao, W. Wang, Q. Ren, M. Zhao and L. Ma, "Source Ranging Using Attention-Based Convolutional Neural Network," 2021 OES China Ocean Acoustics (COA), 2021, pp. 1038-1 042, doi: 10.1109/COA50123.2021.9519915.
[23] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representatio ns by back-propagating errors." nature 323.6088 (1986): 533-536.
[24] Wenbo Wang, Zhen Wang, Lin Su, Tao Hu, Qunyan Ren, Peter Gerstoft, and Li Ma , "S ource depth estimation using spectral transformations and convolutional neural network in a deep-sea environment", The Journal of the Acoustical Society of America 148, 3633 -3644 (2020) https://doi.org/10.1121/10.0002911
[25] T. Hastie, R. Tibshriani, and J. Friedman, The Elements of Statistical Learning, Springer Series in Statis-tics Springer New York Inc., New York, 2ndedition, 2009, Chap. 11, p. 3 93.
[26] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXi v preprint arXiv:1412.6980 (2014).
[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networksfrom overfitting,” J. Mach. Learn. Res. 15, 1929 –1958 (2014).
[28] Ziomek L J. Fundamentals of acoustic field theory and space-time signal processing[M]. CRC press, 2020.
[29] Neyman, Jerzy, and Egon Sharpe Pearson. "IX. On the problem of the most efficient tes ts of statistical hypotheses." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231.694-706 (1933): 289-337.
[30] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.