Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A data-driven method for estimating the target position of low-frequency sound sources in shallow seas*

Xianbin SUN†‡1,2, Xinming JIA†1, Yi ZHENG2, Zhen WANG2

1School of Mechanical and Automobile Engineering, Qingdao University of Technology, Qingdao 266000, China

2Institute of Oceanographic Instrumentation, Shandong Academy of Sciences, Qingdao 266000, China

†E-mail: robin_sun@qut.edu.cn; jiaxinming_123@163.com

Received Apr. 20, 2020; Revision accepted June 4, 2020; Crosschecked Aug. 6, 2020; Published online Jan. 22, 2021

Abstract: Estimating the target position of low-frequency sound sources in a shallow sea environment is difficult due to the high cost of hydrophone placement and the complexity of the propagation model. We propose a compressed recurrent neural network (C-RNN) model that compresses the signal received by a vector hydrophone into a dynamic sound intensity signal and compresses the target position of the sound source into a GeoHash code. Two types of data are used to carry out prior training on the recurrent neural network, and the trained network is subsequently used to estimate the target position of the sound source. Compared with traditional mathematical models, the C-RNN model functions independently under the complex sound field environment and terrain conditions, and allows for real-time positioning of the sound source under low-parameter operating conditions. Experimental results show that the average error of the model is 56 m for estimating the target position of a low-frequency sound source in a shallow sea environment.

Key words: Vector hydrophone; Shallow sea; Low frequency; Location estimation; Recurrent neural network
https://doi.org/10.1631/FITEE.2000181 **CLC number:** TN911.72; P733.23

1 Introduction

Compared with a deep-sea environment, a shallow sea environment has the complex characteristics of spatio-temporal variability, reflection of the signal from the shallow sea bottom, and human offshore activities that cause aliasing of the target signal, thus increasing the difficulty of identification of the sound source signal (Wittekind and Schuster, 2016; Li TC et al., 2017). At the same time, with the development of noise reduction technology and stealth technology, the working frequency band of ship sonar is moving towards lower frequencies. These factors increase the unsuitability of the current, relatively mature position estimation method for a low-frequency acoustic field environment in shallow seas (Zhou et al., 2019). With the increase of the range of human activities and the rapid development of the marine industry, research on estimating the target position of low-frequency sound sources in shallow sea environments, in particular real-time location estimation, has become an urgent priority in most countries.

Vector hydrophones can obtain information of sound pressure and acoustic particle velocity in the ocean sound field at a particular time and point. They have the characteristics of dipole directivity and frequency independence. Consequently, their use has increased in sound field modeling and target azimuth estimation in the ocean environment based on the vector hydrophone's functions, which was realized only when multiple acoustic pressure hydrophone arrays were arranged (D'Spain et al., 2006). Agarwal et al. (2015) used an iterative adaptive approach (IAA)

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 51475249) and the Key Research and Development Program of Shandong Province, China (No. 2018GGX103016)

ORCID: Xianbin SUN, https://orcid.org/0000-0002-2077-5757; Xinming JIA, https://orcid.org/0000-0003-1179-3862

to estimate the direction of arrival (DOA) and power of underwater acoustic emission signals by a vector sensor, and found that the algorithm is robust with partial related or coherent sources in shallow sea conditions. Zhao et al. (2017) proposed an azimuth angle estimation method with an acoustic vector sensor based on an active sonar detection system. However, at present, the application of the vector hydrophone requires the establishment of a complex mathematical model to locate underwater sound sources, and depends largely on background information such as surrounding environmental parameters. Thus, vector hydrophone cannot ascertain the positioning of the sound directly.

As a kind of data-driven model, neural networks have the advantages of self-adaptation, self-organization, and self-learning. At present, they have been used in target recognition, trend prediction, and trajectory tracking (Prieto et al., 2016). Praczyk (2015) used an evolutionary neural network to predict ship behavior. Li GY et al. (2017) predicted the trend of ship movement based on time series using an NARX network. In recent years, neural networks have been increasingly applied to marine ship positioning, but they are used mainly to predict trends after the ship position is known. So far, their use in the field of shallow sea target position estimation has not been reported.

A compressed recurrent neural network (C-RNN) model is proposed for position estimation of low-frequency sound source targets in shallow seas. This model can rapidly output the location of the sound source targets using only the input from the real-time signal received by the vector hydrophone. Because the model discards the step of setting environmental parameters, it is solely data driven. Consequently, it has a wide range of applications for sound source localization in unknown or complex environments.

2 Data fusion and compression

2.1 Characteristics of the signal time domain

As shown in Fig. 1, *O-XYZ* is the absolute coordinate system of the environment, *o-xyz* is the relative coordinate system of the vector hydrophone, and the *XOY* plane represents the sea level. There is likely a certain deviation angle between the *O-XYZ* coordinate system and *o-xyz* relative coordinate system. The signal from point *P* is received by the vector hydrophone at point *Q* after propagation attenuation and ambient noise aliasing. Sound channels of the signal received at point *Q* are: sound pressure *p*, acoustic particle velocity v_x in the *x*-axis direction, acoustic particle velocity v_y in the *y*-axis direction, and acoustic particle velocity v_z in the *z*-axis direction. θ and φ are the pitch and azimuth angles of the sound source relative to the vector hydrophone, respectively, and are both 0° in the *XOY* plane and *X* axis. Since the sound source in this study is a far-field sound source, it is approximated that the sound pressure and the acoustic particle velocities are in the same phase (Dushaw, 2014).

Because the detection target is a point sound source continuously moving and making sound at a certain distance, its position and signal must have

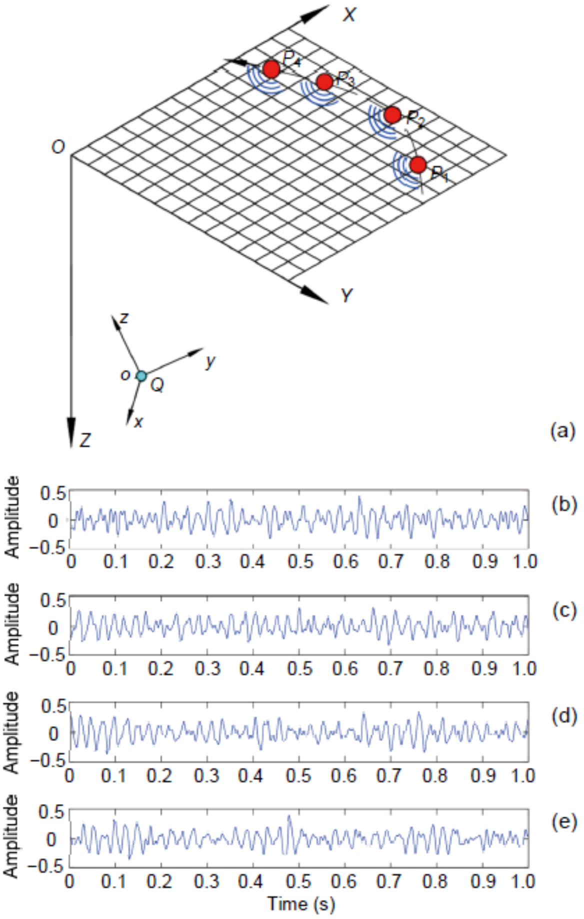

Fig. 1 Sound source *P* at different times along the track (a) and signals received by the vector hydrophone at point *Q* from P_1 (b), P_2 (c), P_3 (d), and P_4 (e)

continuity. Thus, it can be assumed that the signal must have partial similarity to the signal in a certain range in the time domain.

To prove the above hypothesis, a BELLHOP model was used to verify the propagation delay effect of the low-frequency sound source in a shallow sea (Porter, 2011). The sound source is a unit excitation sound source with a frequency of 30 Hz, the position depth is 5 m, the placement depth of the vector hydrophone is 65 m, the horizontal distance between the sound source and vector hydrophone is 4000 m, and the number of transmitted sound rays is 20. The wave height of the sea surface is 2 m, the depth of seawater is 65 m, the density of seawater is 1021 kg/m^3, and the sound velocity in seawater is 1540–1549 m/s. The subsea model is an acoustic elastic half-space model, with a sound velocity of 2000 m/s for sediments and a sediment density of 1810 kg/m^3. The normalized impulse response map and amplitude-delay map of each sound ray are shown in Figs. 2a and 2b, respectively.

In Fig. 2a, there are four echoes, and the last one appears at approximately 0.67 s with a horizontal distance of about 4 km between the sound source and vector hydrophone. Fig. 2b shows that the delay in the signal transmission increases and always stays within a certain scale with the increase in the distance between the sound source and vector hydrophone. This suggests that a signal received by a vector hydrophone has a correlation with the signal in a certain time domain.

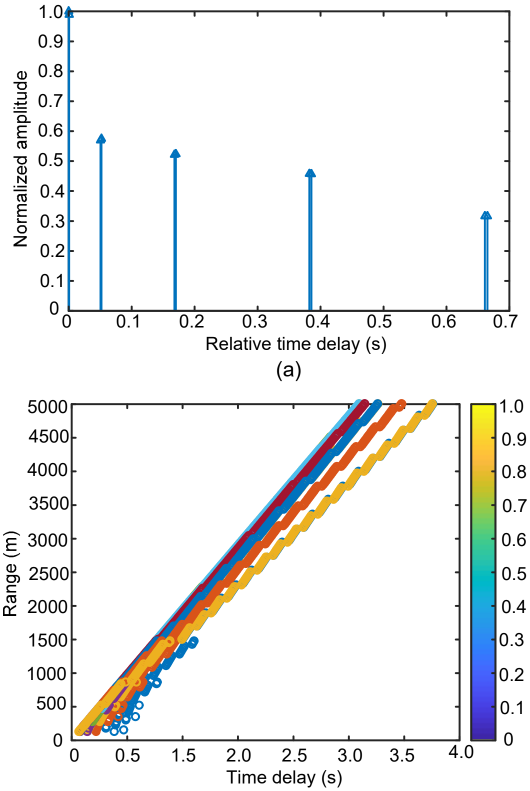

Fig. 2 BELLHOP model results: (a) normalized impulse response; (b) amplitude-delay map of each ray path

To quantify the magnitude of the correlation, a signal with a length of 5 s is randomly cut out from the signal received by the vector hydrophone, and continuously is divided at the length of 1 s to form five signal matrices, i.e., $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{A}_5$. Similarity matrix is used to calculate the correlation coefficient between the signal matrices. Results are shown in Table 1.

Table 1 Correlation coefficient of the adjacent time domain signal matrices

Signal matrices	Correlation coefficient (%)	Signal matrices	Correlation coefficient (%)
$\boldsymbol{A}_1$ & $\boldsymbol{A}_2$	64.4	$\boldsymbol{A}_2$ & $\boldsymbol{A}_5$	58.1
$\boldsymbol{A}_2$ & $\boldsymbol{A}_3$	64.7	$\boldsymbol{A}_3$ & $\boldsymbol{A}_4$	61.2
$\boldsymbol{A}_2$ & $\boldsymbol{A}_4$	63.1	$\boldsymbol{A}_3$ & $\boldsymbol{A}_5$	49.4

It can be seen from Table 1 that the signal matrix correlation coefficient decreases as the time span increases. However, different signal matrices have a high similarity. This suggests that a certain point signal received by the vector hydrophone has a high correlation with a signal in a certain time domain.

2.2 Signal fusion based on the cross-spectrum method

Despite the neural network's strong self-learning ability and good resolution, the cross-spectrum method is used mainly for information fusion and dimensionality reduction of the multi-dimensional signal while retaining its characteristics. It can also be used to eliminate a large amount of isotropic background noise in the signal (Hildebrand, 2009).

In Fig. 1, the sound pressure p and acoustic particle velocity v received by the vector hydrophone at point Q have the following formulas:

1. Sound pressure formula:

$$p(\boldsymbol{r},t)=A\mathrm{e}^{\mathrm{j}(\omega t-\boldsymbol{kr})}, \tag{1}$$

where $\boldsymbol{r}$ is the vector diameter, ω the acoustic wave frequency, $\boldsymbol{k}$ the wave vector, and A the plane wave sound pressure amplitude.

2. The relationship between sound pressure and acoustic particle velocity:

$$v=-\frac{1}{\rho}\int\nabla p\mathrm{d}t, \tag{2}$$

where ρ is the medium density.

Because the focus of this study is far-field target position estimation, when $\boldsymbol{kr}$>>10, the acoustic impedance ratio Z≈0 and the phase difference is small. It can be approximated that the sound pressure and vibration velocity are in-phase.

Each channel signal received by the vector hydrophone is assumed to be composed of effective signals (including non-isotropic noise) and isotropic noise, which are represented by s and n, respectively.

Taking the x-axis direction as an example, we have

$$\begin{aligned}\overline{I_x}&=\overline{p(t)v_x(t)}=\overline{[p_\mathrm{s}(t)+p_\mathrm{n}(t)][v_{x\mathrm{s}}(t)+v_{x\mathrm{n}}(t)]}\\&=\overline{p_\mathrm{s}(t)v_{x\mathrm{s}}(t)},\end{aligned} \tag{3}$$

where I presents the sound intensity and the superscripted horizontal bar represents the mean value (Felisberto et al., 2010).

To calculate the cross-correlation function of sound pressure p and acoustic particle velocity v in the x direction, we have

$$R_{pv_x}=\int_0^T p(t)p(t-\tau)\cos\theta\cos\varphi\mathrm{d}t, \tag{4}$$

where τ is the translation distance of the signal.

Because the signal collected by the vector hydrophone at point Q is a discrete signal, the cross-correlation function (4) is subject to a fast Fourier transform to obtain a cross-spectrum function:

$$S_{pv_x}=S_p(f)\cos\theta\cos\varphi. \tag{5}$$

Similarly, the cross-spectrum functions of p, v_y, and v_z are expressed as

$$S_{pv_y}=S_p(f)\cos\theta\sin\varphi, \tag{6}$$

$$S_{pv_z}=S_p(f)\sin\theta. \tag{7}$$

Therefore, the target elevation angle θ and azimuth angle φ are expressed as

$$\theta=\arctan\left(\frac{S_{pv_z}(f)}{\sqrt{S_{pv_x}^2(f)+S_{pv_y}^2(f)}}\right), \tag{8}$$

$$\varphi=\arctan\left(\frac{S_{pv_y}(f)}{S_{pv_x}(f)}\right). \tag{9}$$

Combining Eqs. (3)–(9), the sound intensity I can be calculated as

$$\begin{aligned}\overline{I}&=\overline{I_x}\cos\theta\cos\varphi+\overline{I_y}\cos\theta\sin\varphi+\overline{I_z}\sin\theta\\&=\overline{p(t)v_x(t)}\frac{\mathrm{Re}[S_{pv_x}]}{\mathrm{Re}[S_p^2(f)]}+\overline{p(t)v_y(t)}\frac{\mathrm{Re}[S_{pv_y}]}{\mathrm{Re}[S_p^2(f)]}\\&\quad+\overline{p(t)v_z(t)}\frac{\mathrm{Re}[S_{pv_z}]}{\mathrm{Re}[S_p^2(f)]}.\end{aligned} \tag{10}$$

Furthermore, because a large amount of isotropic noise has been eliminated and the subsequent neural network can be used to separate sound rays and weaken non-isotropic noise in the signal, Eq. (10) can fuse and compress multi-channel signals.

2.3 Segmentation of the original signal using a fixed dynamic window based on Shannon entropy

Shannon entropy describes a signal in terms of energy and is a definition of quantification. Shannon entropy is not only a measure of the amount of information required to eliminate uncertainty, but also the amount of information that an unknown event may contain (Chao and Shen, 2003). It is expressed as

$$\mathrm{Shannon}(X)=-\sum_{g=1}^{m}p\log p(x_g), \tag{11}$$

where x_g is the possible value of the random event X and m is the number of pieces of information contained in the signal.

We now introduce the concept of window overlap ratio η, which is a value that indicates the degree of overlap between the next fixed window $W_{\mathrm{f}(i+1)}$ and the previous dynamic window $W_{\mathrm{d}i}$, where i indicates the i^{th} iteration. i=1 represents the initial time and establishes the fixed dynamic window sliding in

Algorithm 1.

Algorithm 1 uses a fixed window to fuse the original signal and initially realizes data size reduction. Then, it uses a dynamic window to find the fastest growing segment of Shannon entropy to achieve data size compression to maximally retain the information contained in the signal. The sliding of the windows allows the reuse of signals, and the design of overlapping and sliding can be used for evidence of fusion in later steps. Fig. 3 is a flowchart of Algorithm 1.

2.4 GeoHash encoding for location label compression

GeoHash is a positioning method that converts two-dimensional (2D) labels into one-dimensional (1D) strings. The longer the string length, the higher the accuracy of the area represented (Fox et al., 2013). Because this geocoding system uses a hierarchical data structure and provides attributes of arbitrary precision, it is widely used in the fields of transportation and logistics (Moussalli et al., 2015).

For simplification, we convert the 2D labels into decimal labels through GeoHash encoding. Fig. 4 is the diagram of GeoHash encoding.

As shown in Fig. 4, L_i and B_i (i=1, 2, 3) are the longitudes and latitudes of the i^{th} iteration, respectively, i is the number of iterations in a rectangular area composed of a latitude interval [Lat′, Lat] and a longitude interval [Lng′, Lng], and P is the target position. Algorithm 2 is the encoding process of GeoHash.

In Algorithm 2, the position obtained in each iteration will increasingly approach the point P. The positioning accuracy can be controlled by controlling the number of iterations.

We use GeoHash coding to compress 2D label data into 1D data, avoiding the complexity of subsequent steps needed to index two labels at the same time, thus improving the calculation speed. This makes GeoHash suitable for real-time monitoring of target orientation.

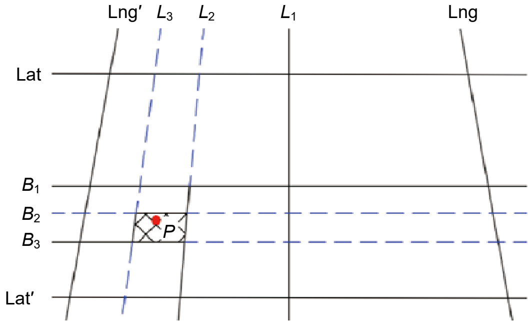

Fig. 4 GeoHash coding

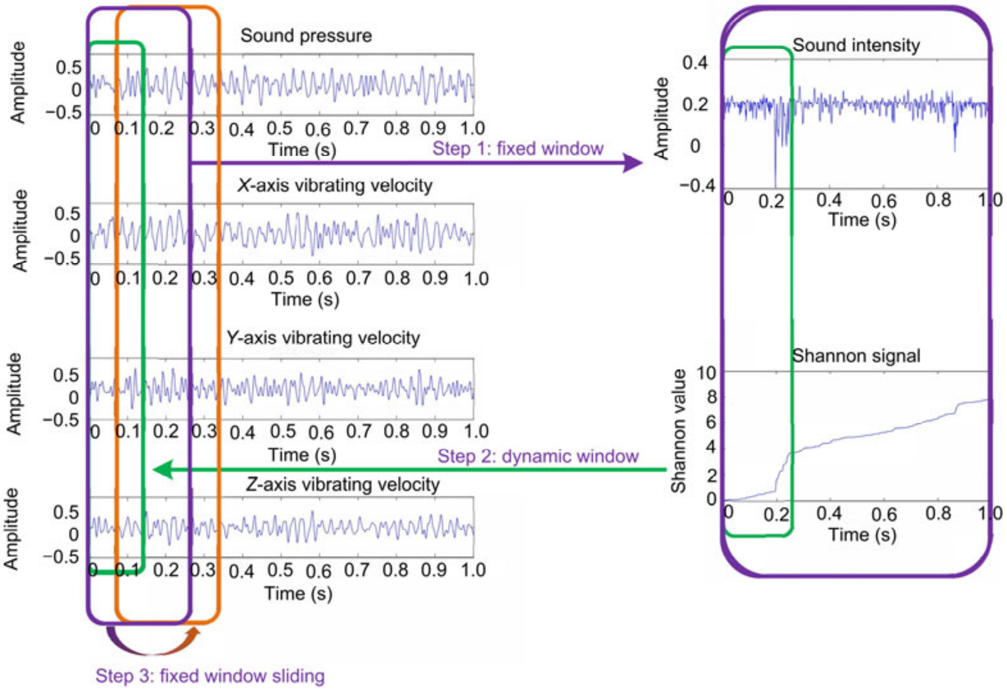

Fig. 3 Flowchart of the fixed dynamic window sliding process

Algorithm 1 Fixed dynamic window sliding

```
Input: Four-channel signals received by vector hydrophone: $p$, $v_x$, $v_y$, and $v_z$; number of signal interceptions num; length of the fixed window $l_f$; initial starting point of the fixed window $t_{f1}$; threshold $\gamma$ of the fastest increase in Shannon entropy; useful Shannon entropy threshold interval [$\zeta$, $\xi$]; window overlap ratio $\eta$; minimum length $l_0$ and maximum length $l_1$ of Shannon entropy.
Output: Signal in the dynamic window $W_{di}$.
while $i$<num do
  With the four-channel original signals received by the vector hydrophone, a fixed window $W_{fi}$ of starting point $t_{fi}$ and length $l_f$ is generated. Four-channel signals in the window are fused into a monophonic sound signal $I_{fi}$ by Eq. (10)
  while $j$<length($I_{fi}$) do
    In the fused sound intensity signal segment $I_{fi}$, a dynamic window $W_{d(i,j)}$ of length $l_{d(i,j)}$ is intercepted from the signal starting point, and its Shannon entropy $S_{d(i,j)}$ is calculated using Eq. (11)
    if $S'_{d(i,j)}>\gamma$ then
      It can be considered that the fastest growing segment of Shannon entropy is found and its length is labeled as $l_{di}$
    else if max($S_{d(i,j)}$)<$\zeta$ && min($S_{d(i,j)}$)>$\xi$ then
      Signal segment $I_{fi}$ is considered as an invalid signal segment, and $l_{di}=l_0$
    else $l_{di}=l_1$
    end if
  end while
  $t_{f(i+1)}=t_{fi}+(1-\eta)l_{di}$
  $i=i+1$
end while
```

Algorithm 2 GeoHash encoding

```
Input: Initial latitude interval [Lat′, Lat] and longitude interval [Lng′, Lng]; position of target $P$; maximum number of iterations num; number of iterations $i$; binary sequence $x$, where the subscript of $x$ indicates the position of the characters in the sequence.
Output: GeoHash encoded decimal position label $y$.
while $i$<num do
  $B_i = (\mathrm{Lat}'_i + \mathrm{Lat}'_{i+1})/2,\ L_i = (\mathrm{Lng}'_i + \mathrm{Lng}'_{i+1})/2$
  if $P \in [\mathrm{Lat}'_i, B_i)$ then
    $x_{2(\mathrm{num}-i)-1}=0$
  else
    $x_{2(\mathrm{num}-i)-1}=1$
  end if
  if $P \in [\mathrm{Lng}'_i, L_i)$ then
    $x_{2(\mathrm{num}-i)}=0$
  else
    $x_{2(\mathrm{num}-i)}=1$
  end if
  Convert binary sequence $x$ to decimal label $y$
end while
```

3 Establishment of the compressed recurrent neural network model

3.1 Recurrent neural networks

The premise of traditional neural networks such as the artificial neural network (ANN) and convolutional neural network (CNN) is that each element in the network, including the input and output of the network, is independent of each other. However, in practical applications, because of many interconnected factors, traditional neural networks are less able to handle time-series problems (van Gerven and Bohte, 2017). In this study, we use a recurrent neural network (RNN) based on time series. This network addresses the shortcomings of other networks, where the information flow can achieve only one-way propagation by adding a step of back propagation (Gu et al., 2018). This propagation method improves the sensitivity of RNN to time; that is, every decision made will be affected by the residual information from the previous moment.

Fig. 5 shows the RNN calculation process, which includes the input layer $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \ldots\}$, hidden layer $\{\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, \boldsymbol{s}_{t+1}, \ldots\}$, and output layer $\{\boldsymbol{o}_0, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, \boldsymbol{o}_{t+1}, \ldots\}$. The final output $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \ldots\}$ of RNN is obtained by the activation of function $g(\boldsymbol{o}_t)$ of the output layer.

In Fig. 5, $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{V}$ are the weights of RNN, which represent the connection of the input neuron $\boldsymbol{x}$ to the hidden neuron $\boldsymbol{s}$, the self-loop connection of the hidden neuron $\boldsymbol{s}$, and the connection of hidden neuron $\boldsymbol{s}$ to output neuron $\boldsymbol{o}$, respectively. The right part of Fig. 5 is the calculation process of RNN (left part) after expansion.

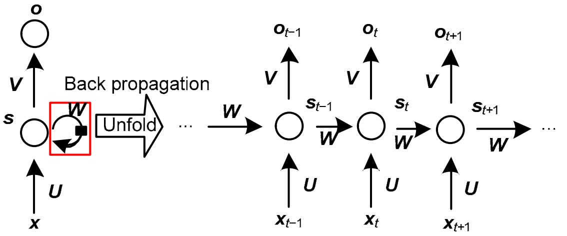

Fig. 5 Recurrent neural network calculation process

As time t changes, the forward propagation equations are expressed as

$$\boldsymbol{s}_t = f(\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{s}_{t-1} + \boldsymbol{b}), \tag{12}$$

$$\boldsymbol{o}_t = \boldsymbol{V}\boldsymbol{s}_t + \boldsymbol{c}, \tag{13}$$

$$\boldsymbol{y}_t = g(\boldsymbol{o}_t) = g(\boldsymbol{V}\boldsymbol{s}_t + \boldsymbol{c}), \tag{14}$$

where $\boldsymbol{s}_t$ is the memory at time t and is related to the memory of $\boldsymbol{s}_{t-1}$ at the previous time, f is the activation function of hidden neurons, g is the activation function of output, and $\boldsymbol{b}$ and $\boldsymbol{c}$ are biases. The deviation of $\boldsymbol{y}_t$ from the real result is the cost function $L(t)$.

During the back propagation process in Fig. 5, the network has the following gradient update formulas (Ian et al., 2016):

$$\nabla_{\boldsymbol{c}} L = \sum_t \left(\frac{\partial \boldsymbol{o}_t}{\partial \boldsymbol{c}}\right)^{\mathrm{T}} \nabla_{\boldsymbol{o}_t} L, \tag{15}$$

$$\nabla_{\boldsymbol{b}} L = \sum_t \mathrm{diag}(1 - \boldsymbol{s}_t^2) \nabla_{\boldsymbol{s}_t} L, \tag{16}$$

$$\nabla_{\boldsymbol{V}} L = \sum_t (\nabla_{\boldsymbol{o}_t} L) \boldsymbol{s}_t^{\mathrm{T}}, \tag{17}$$

$$\nabla_{\boldsymbol{W}} L = \sum_t \mathrm{diag}(1 - \boldsymbol{s}_t^2)(\nabla_{\boldsymbol{s}_t} L) \boldsymbol{s}_{t-1}^{\mathrm{T}}, \tag{18}$$

$$\nabla_{\boldsymbol{U}} L = \sum_t \mathrm{diag}(1 - \boldsymbol{s}_t^2)(\nabla_{\boldsymbol{s}_t} L) \boldsymbol{x}_t^{\mathrm{T}}. \tag{19}$$

It can be seen from Eqs. (12)–(19) that the output value y of RNN is related to the previous input values $\boldsymbol{x}_t$, $\boldsymbol{x}_{t-1}$, …, so that the network can realize the correlation between the output and input elements based on time series through forward propagation and back propagation. Because the signals from the sound source and received by the vector hydrophone are continuous, the use of RNN may show better performance.

3.2 Compressed recurrent neural network model

RNNs have better time-series analysis capabilities. However, they increase the interconnection of neurons in the same layer and back propagation of neurons, which increases calculations in the RNNs compared with other neural networks. Therefore, it is necessary to compress the original data to improve the calculation speed of the model and achieve the goal of real-time monitoring of the sound source position.

The cross-spectrum method used in Section 2 can compress the four-channel original signals into a mono sound intensity signal. The GeoHash encoding compresses the 2D sound source position labels into a 1D decimal value. The fixed dynamic window design based on Shannon entropy can reduce the length of the signal segment on the premise that the window contains sufficient information. The window overlap design is introduced to ensure the use of redundant information, and the results of the segment signal position estimation are used for cross validation to ensure the accuracy of the model. All the above-listed design elements are aimed at achieving the compression of data with the goal of reducing information loss, thus facilitating the use of RNN.

This specific method is designed to achieve data size compression of the original signal and the sound source location through fixed dynamic window sliding and GeoHash encoding, respectively, and then to use these two types of data to carry out initial training for RNN. Subsequently, the trained RNN is used to carry out regression analysis of the unknown signal. Finally, the regression results are decoded and fused to obtain the location of the target sound source. Fig. 6 is the flowchart of the C-RNN model.

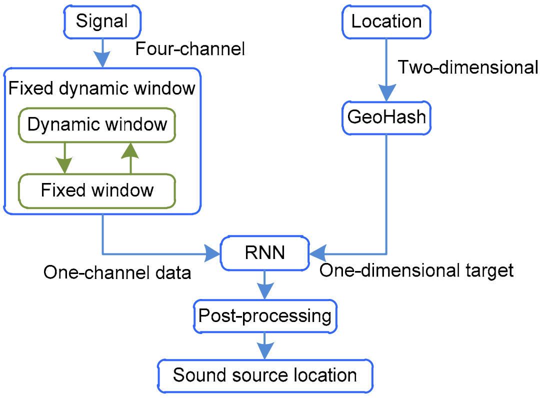

Fig. 6 Compressed recurrent neural network model

4 Experiments and analysis

4.1 Source of experimental data

The data used comes from a seesaw experiment in a shallow sea area of the South China Sea on August 2, 2018. The seawater environment is an isothermal layer environment, and the temperature varies with the water depth in the range of 28–4 °C. The sound source signal is a mixture of low-frequency signals at 23, 27, 33, 39, and 42 Hz. The sound source target travels at a speed of 4 km/h, and simulates the signals generated by low-frequency ships in shallow waters. The real-time position and track are given by the Global Positioning System (GPS). The vector hydrophone is placed at 65-m depth, and the sampling frequency is 1024 Hz.

Considering the effects of computer performance, data size, and iteration speed, a 10-min continuous signal is selected as the experimental signal from the 1-h data collected in the seesaw experiment. Fifty segments with a length of 1 s are randomly cut out from the original signal set, and are randomly divided into a training set, a validation set, and a test set at a ratio of 7:2:1.

4.2 Setting of model parameters

Reasonable setting of model parameters has direct impact on experimental results. To improve the accuracy and real-time performance of the estimation, a fixed parameter method was used in the parameter setting process. Contrast experiments were performed on parameters such as the fixed window size, window overlap ratio, network type, and network structure, to test the training results under short-term iterations.

The EarlyStopping mechanism (Demuth et al., 2014) was used to monitor the training results, and the regression evaluation index root-mean-square error (RMSE) (Chai and Draxler, 2014) was used to evaluate the training results. RMSE is expressed as

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{h=1}^{m}\left(O_h-\hat{O}_h\right)^2}, \tag{20}$$

where O is the model regression result, $\hat{O}$ the real result, m the sample length of $\hat{O}$, and h the corresponding sample point.

After testing the C-RNN model with different parameters, we obtained the optimal values of some parameters of the model (Table 2). The model's optimizer is Adam.

4.3 Comparative model

To compare the impact of each step and the influence of different structures on the experimental results of the dataset OS, a hierarchical structure (Fig. 7) was established based on the C-RNN model. Each layer of the structure can be replaced by other structures, and a total of 14 comparative models were established. Table 3 is a correspondence table of each structure in Fig. 7.

To study the effect of different step combinations on experimental results, five models were established: an SPVV+LSTM model using the original signal and LSTM; an SPVV+RNN model using the original signal and RNN; an SPVV+O+RNN model intercepting the original signal using overlapping fixed dynamic window and RNN; an SI+O+RNN model

Table 2 Optimal values of some parameters of the C-RNN model

Parameter	Value
Fixed window size (s)	0.1
Window overlap ratio	0.4
Number of digits after GeoHash encoding	6
Number of neurons	150
Epoch size	100
Value of EarlyStopping	6

Table 3 Correspondence table of each structure in Fig. 7

Step	Structure	Description
Interception method	NW	No window
	FW	Fixed window
	O-FW	Overlapping fixed window
	DW	Dynamic window
	FDW	Fixed dynamic window
	O-FDW	Overlapping fixed dynamic window
Used signals	SP	Sound pressure signal
	VV	Vector velocity signal
	SPVV	Combination of SP and VV signals
	SI	Sound intensity signal after fusion
Network structure	RNN	Recurrent neural network
	LSTM	Long short-term memory
	One-BiLSTM	One hidden layer of bidirectional LSTM
	Two-BiLSTM	Two hidden layers of bidirectional LSTM
Average results	Average	Average results of the same signal segments

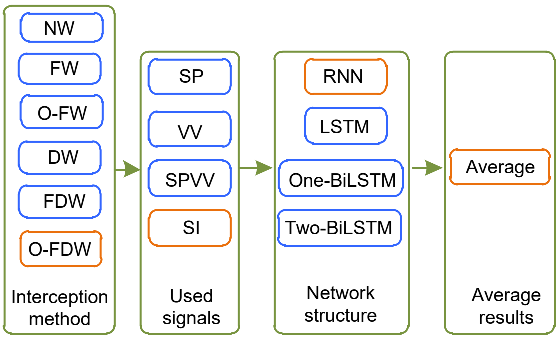

Fig. 7 Comparative experimental model

using the sound intensity signal of an overlapping fixed dynamic window and RNN; a GeoHash+SI+O+RNN model using the fused sound intensity signal and the GeoHash encoding target position of RNN.

The comparison models above all use the OS dataset consisting of 50 segments of 1-s signals randomly cut from the original 10-min signal. To verify the performance of C-RNN, the following datasets were established: a 30-min+50-sample dataset that intercepts 50 samples from a 30-min signal; a 30-min+200-sample dataset that intercepts 200 samples from a 30-min signal; a 60-min+500-sample dataset that intercepts 500 samples from a 60-min signal; a 60-min+1000-sample dataset that intercepts 1000 samples from a 60-min signal. The C-RNN model with the parameters in Table 2 was used to perform experiments on these sample sets.

4.4 Experimental results and analysis

Table 4 shows the experimental results of a total of 24 experiments. Because the EarlyStopping mechanism was used to monitor the training results, the time in Table 4 is the average time of each iteration, the data in the remaining columns is the deviation from the actual position after each experimental result is converted to a position label, and the unit is meter.

Table 4 shows that the C-RNN model has better performance in terms of the positioning accuracy and learning speed compared with other models. It ultimately achieved a positioning accuracy of 56 m through mutual verification of the position estimation results. After expanding the data size of the sample set, the performance can be stabilized within a certain degree of confidence. This shows that the model is

Table 4 Experimental results of different sound source localization models

Experiment		Lat (m)				Lng (m)				Iteration time (s)
		Before		After		Before		After		
Type	Structure	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	
Interception method	NW	116.0	135.7	116.0	135.7	487.4	584.9	487.4	584.9	**0.04**
	FW	36.3	47.2	19.7	23.8	250.7	330.3	209.1	228.8	0.53
	O-FW	25.9	35.6	14.7	18.4	154.2	240.2	124.8	146.8	0.26
	DW	130.6	155.2	130.6	155.2	304.2	365.7	304.2	365.7	0.17
	FDW	18.4	**21.7**	18.4	21.7	151.3	180.8	151.3	180.8	0.34
	O-FDW	**17.1**	23.3	**8.9**	**10.7**	**104.8**	**177.5**	**57.7**	**70.8**	0.15
Information fusion	SP	19.7	28.0	10.4	15.0	124.7	208.3	92.7	112.5	0.16
	VV	25.5	36.1	14.5	17.5	140.4	221.6	103.8	116.1	**0.13**
	SPVV	26.8	37.6	15.8	19.7	175.6	270.6	148.0	175.9	0.15
	SI	**8.9**	**10.7**	**8.9**	**10.7**	**57.7**	**70.8**	**57.7**	**70.8**	0.16
Network	RNN	**20.4**	**26.6**	**10.7**	**13.0**	**111.3**	**185.7**	**69.0**	**83.3**	**0.24**
	LSTM	60.1	65.5	59.3	64.5	244.1	276.8	239.3	271.0	0.28
	One-BiLSTM	59.7	66.2	59.2	65.0	242.2	273.2	238.7	266.6	0.50
	Two-BiLSTM	57.9	63.3	57.9	63.3	226.4	260.7	226.4	260.7	0.73
Structure	SPVV+LSTM	**10.3**	**12.0**	10.3	12.0	**89.8**	**99.4**	89.8	99.4	**0.08**
	SPVV+RNN	24.1	29.0	24.1	29.0	133.7	171.4	133.7	171.4	0.21
	SPVV+O+RNN	15.5	19.2	15.2	18.6	115.4	135.8	111.5	129.0	0.60
	SI+O+RNN	14.2	16.7	14.2	16.7	106.5	123.5	106.5	123.5	0.63
	GEO+SI+O+RNN	17.1	23.3	**8.9**	**10.7**	104.8	177.5	**57.7**	**70.8**	0.16
Sample	10-min+50-sample	**14.6**	**19.5**	**8.7**	**9.9**	**76.1**	**143.5**	**56.4**	**68.1**	0.20
	30-min+50-sample	63.2	90.4	48.2	57.8	223.1	330.4	145.3	162.9	**0.15**
	30-min+200-sample	53.6	79.4	37.5	48.8	193.5	302.8	131.5	184.6	0.18
	60-min+500-sample	35.9	44.5	35.9	44.5	188.4	300.0	188.4	300.0	**0.15**
	60-min+1000-sample	46.6	66.0	29.1	38.6	229.2	365.0	167.6	231.5	0.17

Before: before the averaging; After: after the averaging. Best results are in bold effective in estimating the location of low-frequency sound sources in shallow waters.

To further demonstrate the advantages of the data-driven model, we added Bharathi and Mohanty (2018)'s proposition for underwater sound source localization by the EMD-based maximum likelihood method (EMD ML TDE). The signal set used in the model was the test set of the C-RNN model, and the positioning result of the model was converted equally into the C-RNN model result. It was concluded that the average positioning accuracy of EMD ML TDE at latitude was 584 m, and the average positioning accuracy at longitude was 239 m. Compared with the C-RNN model, a higher positioning error was obtained. We also found that the model did not have good resolution for the movement of the sound source in a short time.

5 Conclusions

In this study, we have proposed a compressed recurrent neural network (C-RNN) model for real-time position estimation of sound sources in shallow waters. The model considers that the chirp signal received by the vector hydrophone must be related to the signal in a certain time domain. The four-channel signal is compressed into a single-channel dynamic sound intensity signal through a dynamic fixed window. The two-dimensional position information is compressed into a decimal code, and then the recurrent neural network is used to perform regression analysis of the relationship between the compressed sound intensity signal and the sound source position. Finally, the trained C-RNN model can be used to achieve localization of unknown signals.

Through experimental comparison, it was found that the C-RNN model improves the operating speed for low-frequency sound source signals, and can provide accurate positioning. The average error radius was approximately 56 m.

Contributors

Xianbin SUN designed the research. Xinming JIA and Zhen WANG processed the data. Xinming JIA drafted the manuscript. Xianbin SUN and Yi ZHENG helped organize the manuscript. Xianbin SUN and Xinming JIA revised and finalized the paper.

Compliance with ethics guidelines

Xianbin Sun, Xinming JIA, Yi ZHENG, and Zhen WANG declare that they have no conflict of interest.

References

Agarwal A, Kumar A, Agrawal M, 2015. Iterative adaptive approach to DOA estimation with acoustic vector sensors. OCEANS, p.1-8. https://doi.org/10.1109/OCEANS-Genova.2015.7271605

Bharathi BMR, Mohanty AR, 2018. Underwater sound source localization by EMD-based maximum likelihood method. *Acoust Aust*, 46(2):193-203. https://doi.org/10.1007/s40857-018-0129-8

Chai T, Draxler RR, 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci Model Dev*, 7(3): 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chao A, Shen TJ, 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ Ecol Stat*, 10(4):429-443. https://doi.org/10.1023/A:1026096204727

D'Spain GL, Luby JC, Wilson GR, et al., 2006. Vector sensors and vector sensor line arrays: comments on optimal array gain and detection. *J Acoust Soc Am*, 120(1):171-185. https://doi.org/10.1121/1.2207573

Demuth HB, Beale MH, Jess OD, et al., 2014. Neural Network Design. Martin Hagan.

Dushaw B, 2014. Acoustic Tomography, Ocean. In: Njoku EG (Ed.), Encyclopedia of Remote Sensing. Springer, New York, USA, p.4-11. https://doi.org/10.1007/978-0-387-36699-9_211

Felisberto P, Santos P, Jesus SM, 2010. Tracking source azimuth using a single vector sensor. Int Conf on Sensor Technologies and Applications, p.416-421. https://doi.org/10.1109/SENSORCOMM.2010.66

Fox A, Eichelberger C, Hughes J, et al., 2013. Spatio-temporal indexing in non-relational distributed databases. Int Conf on Big Data, p.291-299. https://doi.org/10.1109/BigData.2013.6691586

Gu JX, Wang ZH, Kuen J, et al., 2018. Recent advances in convolutional neural networks. *Patt Recogn*, 77:354-377. https://doi.org/10.1016/j.patcog.2017.10.013

Hildebrand JA, 2009. Anthropogenic and natural sources of ambient noise in the ocean. *Mar Ecol Progr Ser*, 395:5-20. https://doi.org/10.3354/meps08353

Ian G, Yoshua B, Aaron C, 2016. Deep Learning. The MIT Press, USA.

Li GY, Kawan B, Wang H, et al., 2017. Neural-network-based modelling and analysis for time series prediction of ship motion. *Ship Technol Res*, 64(1):30-39. https://doi.org/10.1080/09377255.2017.1309786

Li TC, Su JY, Liu W, et al., 2017. Approximate Gaussian conjugacy: parametric recursive filtering under nonlinearity, multimodality, uncertainty, and constraint, and beyond. *Front Inform Technol Electron Eng*, 18(12):1913-1939. https://doi.org/10.1631/FITEE.1700379

Moussalli R, Srivatsa M, Asaad S, 2015. Fast and flexible conversion of Geohash codes to and from latitude/longitude coordinates. Proc IEEE 23rd Annual Int Symp on Field-Programmable Custom Computing Machines, p.179-186. https://doi.org/10.1109/FCCM.2015.18

Porter MB, 2011. The BELLHOP Manual and User's Guide: Preliminary Draft. Technology Report, USA.

Praczyk T, 2015. Using evolutionary neural networks to predict spatial orientation of a ship. *Neurocomputing*, 166:229-243. https://doi.org/10.1016/j.neucom.2015.03.075

Prieto A, Prieto B, Ortigosa EM, et al., 2016. Neural networks: an overview of early research, current frameworks and new challenges. *Neurocomputing*, 214:242-268. https://doi.org/10.1016/j.neucom.2016.06.014

van Gerven MAJ, Bohte SM, 2017. Editorial: artificial neural networks as models of neural information processing. *Front Comput Neurosci*, 11:114. https://doi.org/10.3389/fncom.2017.00114

Wittekind D, Schuster M, 2016. Propeller cavitation noise and background noise in the sea. *Ocean Eng*, 120:116-121. https://doi.org/10.1016/j.oceaneng.2015.12.060

Zhao AB, Ma L, Ma XF, et al., 2017. An improved azimuth angle estimation method with a single acoustic vector sensor based on an active sonar detection system. *Sensors*, 17(2):412. https://doi.org/10.3390/s17020412

Zhou JB, Zhang MH, Piao SC, et al., 2019. Low frequency ambient noise modeling and comparison with field measurements in the South China Sea. *Appl Acoust*, 148:34-39. https://doi.org/10.1016/j.apacoust.2018.11.013