Deep residual regression network for underwater acoustic source number estimation

Nguyen Ngoc Hoai Phong[1*], Phan Hong Minh[1], Nguyen Manh Cuong[1], Nguyen Van Duc[2], Le Ngoc Hung[3]

[1]Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam;
[2]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, 1 Dai Co Viet, Bach Mai, Hanoi, Vietnam;
[3]Brigade 172, Naval Region 3, Da Nang, Vietnam.
[*]Corresponding author: hphongmta@gmail.com
Received 7 Aug. 2025; Revised 23 Sep. 2025; Accepted 10 Oct. 2025; Published 30 Oct. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.IITE.2025.91-98

ABSTRACT

Source Number Estimation (SNE) is a crucial task in underwater acoustic array signal processing, as it significantly affects the performance of subsequent algorithms. Traditional methods, such as the Akaike Information Criterion (AIC) and Minimum Description Length (MDL), often perform poorly in challenging underwater environments, especially under low Signal-to-Noise Ratio (SNR) conditions, with a limited number of snapshots and complex noise structures. To tackle these issues, this paper presents an Eigenvalue-based Residual Regression Network (EResNet) designed for robust source number estimation. Comprehensive simulations conducted in various complex noise scenarios have shown that EResNet is notably effective. The results reveal that the proposed model achieves higher accuracy and demonstrates greater robustness compared to the AIC and MDL methods and other baseline neural network architectures.

Keywords: Underwater; Source number estimation; Deep learning; Residual network.

1. INTRODUCTION

In array signal processing, particularly for passive sonar in underwater environments, accurately estimating the number of sound sources is crucial for subspace-based direction-of-arrival (DOA) algorithms used in target localization [1-4]. Misestimating the number of sources can lead to unreliable spatial spectrum analysis. Traditional approaches to Source Number Estimation (SNE) mainly rely on Information Theoretic Criteria (ITC) [5, 6], such as the Akaike Information Criterion (AIC) [7] and Minimum Description Length (MDL) [8]. While these methods analyze the eigenvalue distribution of the signal's covariance matrix, they often require a large number of samples for stable estimates, perform poorly at low signal-to-noise ratios (SNR), and usually assume Gaussian white noise - conditions rarely met underwater [9].

The rise of deep learning in this field has led to data-driven methods that can address these limitations [10-14]. Deep learning models can learn complex nonlinear relationships from signal features, and recent studies suggest that using the eigenvalue vector of the covariance matrix as input for neural networks is particularly effective. However, many existing works still focus on simplistic noise models and have not fully utilized more advanced network architectures beyond basic Multi-Layer Perceptrons (MLP).

In this study, we propose EResNet, a Deep Residual Regression Network, to tackle the SNE problem in various scenarios, especially in environments characterized by non-Gaussian underwater noise, low SNR, and a limited number of snapshots. The main contributions of this work are as follows:

1. EResNet Architecture: We develop a deep regression model with residual connections, enabling clearer differentiation between signal and noise eigenvalues, even when closely spaced.

2. Simulation Framework: We create a data generation pipeline to simulate realistic underwater acoustic conditions, including colored and ambient noise based on the Wenz model.

3. Comparative Evaluation: We conduct a performance comparison of EResNet against classical methods (AIC, MDL) and baseline deep learning models (ECNet, ERNet) across various operating conditions, including differing SNRs and snapshot counts.

2. EIGENVALUE-BASED DEEP LEARNING FOR UNDERWATER SNE

2.1. System model

We consider K independent narrowband signals impinging on a Uniform Rectangular Array (URA) with M elements. The distance between elements is denoted by $d_y = d_z = \lambda/2$, and the coordinate origin is placed at the first element of the array (figure 1).

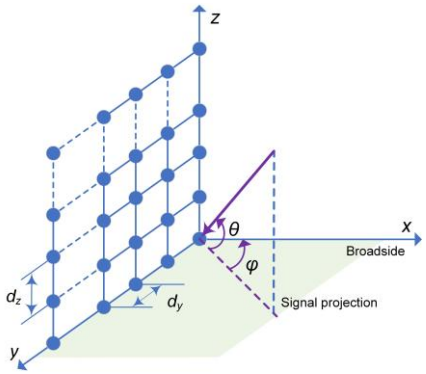


Figure 1. *Structure of a URA.*

The observed signal matrix at the array for N snapshots is modeled as follows.

$$\mathbf{X} = \mathbf{AS} + \mathbf{W} \tag{1}$$

Where, $\mathbf{X} \in \mathbb{C}^{M \times N}$ is the observed signal matrix at the array, $\mathbf{A} = \left[\mathbf{a}(\theta_1, \varphi_1), \ldots, \mathbf{a}(\theta_K, \varphi_K)\right]$ is the steering matrix, where the steering vector of the k-th source $\mathbf{a}(\theta_k, \varphi_k) \in \mathbb{C}^{M \times 1}$ (elevation θ_k and azimuth φ_k), $\mathbf{S} \in \mathbb{C}^{K \times N}$ is the signal matrix of K sources, $\mathbf{W} \in \mathbb{C}^{M \times N}$ is the noise matrix (white, colored, or underwater ambient). The steering vector $\boldsymbol{\alpha}(\theta_k, \varphi_k)$ is calculated by

$$\mathbf{a}(\theta_k, \varphi_k) = e^{-j\mathbf{r}_k^t \mathbf{C}} \tag{2}$$

Where, $\mathbf{C} \in \mathbb{R}^{M \times 3}$ is the matrix of the geometric coordinates of the URA elements, and $\mathbf{r}_k$ is the wave number vector of the k-th source. We consider a signal wavelength $\lambda = f/1500$ where f is the operating frequency of the antenna. The wave number vector $\mathbf{r}_k$ is then calculated by

$$\mathbf{r}_k = \frac{2\pi}{\lambda} \begin{bmatrix} \sin(\theta_k)\cos(\varphi_k) \\ \sin(\theta_k)\sin(\varphi_k) \\ \cos(\theta_k) \end{bmatrix} \tag{3}$$

The Sample Covariance Matrix (SCM) with N snapshots is calculated by

$$\hat{\mathbf{R}} = \frac{1}{N}\mathbf{X}\mathbf{X}^H = \frac{1}{N}\sum_{t=1}^{N} \mathbf{x}(t)\mathbf{x}^H(t) \tag{4}$$

Where, $\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{a}(\theta_k, \varphi_k)\, \mathbf{s}_k(t) + \mathbf{w}(t)$, with $\mathbf{s}_k(t)$ and $\mathbf{w}(t)$ are the signal of the k-th source and the noise, respectively.

The eigenvalues of the SCM are calculated by

$$\hat{\mathbf{R}} = \mathbf{XLX}^H = \sum_{m=1}^{M} l_m \mathbf{x}_m \mathbf{x}_m^H \tag{5}$$

Where, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]$ is the matrix containing the eigenvectors $\mathbf{x}_m$, and $\mathbf{L} = \text{diag}(l_1, l_2, ...l_M)$ is the diagonal matrix containing the eigenvalues l_m. After decomposition, the eigenvalues are sorted in descending order as follows.

$$l_1 \geq l_2 \geq \ldots l_K \geq l_{K+1} \geq \ldots l_M \tag{6}$$

Under ideal conditions, the eigenvalues l_m will separate into two distinct subspaces. The signal subspace contains the *K* largest eigenvalues $(l_1, ..., l_K)$, while the noise subspace consists of the *M* - *K* smallest eigenvalues $(l_{K+1}, ..., l_M)$.

To create a diverse dataset for training deep learning models, we construct the noise matrix **W**, as described in equation (1), using three noise models: Gaussian white noise, time-correlated colored noise, and underwater ambient noise.

Gaussian white noise, characterized by its noise power of σ_n^2, is represented as

$$S_{wh}(f) = 2\sigma_n^2 \tag{7}$$

Colored noise, with a noise sample $w(n)$ at time *n*, is calculated according to the following autoregressive equation.

$$w(n) = \alpha w(n-1) + v(n) \tag{8}$$

Where, $v(n) \sim \mathcal{N}(0, \sigma_v^2)$ is Gaussian white noise, α is the autoregressive coefficient, with $|\alpha| < 1$ to ensure stability.

Underwater ambient noise, according to the Wenz model [15], includes turbulence noise $N_t(f)$, shipping noise $N_s(f)$, wind-driven noise $N_w(f)$, and thermal noise $N_{th}(f)$. The power spectral density of each type of noise is frequency-dependent (measured in kHz), and is expressed in decibels referenced to one micropascal (dB re μPa).

$$\begin{gathered} N_t(f) = 27 - 30\log f \\ N_s(f) = 40 + 20(s - 0.5) + 26\log f - 60\log(f + 0.03) \\ N_w(f) = 50 + 7.5\sqrt{w} + 20\log f - 40\log(f + 0.4) \\ N_{th}(f) = -15 + 20\log f \end{gathered} \tag{9}$$

Where, $s = 0 \div 1$ is the shipping activity factor, and *w* is the wind speed. The composite noise power spectral density is given by

$$N(f) = N_t(f) + N_s(f) + N_w(f) + N_{th}(f) \tag{10}$$

To achieve a stable and effective training process for neural networks, the raw eigenvalue vectors obtained from the sample covariance matrix are subjected to a two-step preprocessing procedure: logarithmic transformation followed by data standardization.

Logarithmic transformation compresses the dynamic range of eigenvalues, bringing values closer together on a new scale. Similar transformations have been applied in other studies to enhance model performance [13, 16].

The raw eigenvalues $l = [l_1, l_2, ...l_M]^T$ are transformed into new eigenvalues as follows.

$$l_m^{'} = \log(l_m + \varepsilon) \tag{11}$$

Where, l_m is the m-th eigenvalue, and $\varepsilon = 10^{-10}$ is added to ensure numerical stability by preventing the logarithm of zero from being calculated.

Data Standardization not only accelerates the convergence rate but also helps avoid bias during the training process. In this study, we utilize the StandardScaler method to transform the elements $l_m^{'}$ into new values according to a specific formula.

$$l_m^{''} = \frac{l_m^{'} - \mu_m}{\sigma_m} \tag{12}$$

Where, μ_m and σ_m represent the mean and standard deviation of the j-th feature, respectively, and are calculated solely from the training dataset.

2.2. Architecture of EResNet

The EResNet network proposed in this study is a deep neural network that is based on an MLP structure and incorporates residual connections. This design enables the construction of deeper networks while maintaining an effective learning process, which helps address the vanishing gradient problem often faced in deep neural networks.

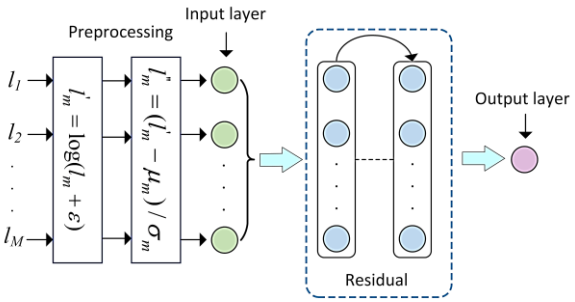


Figure 2. *Architecture of EresNet.*

The network architecture is depicted in Figure 2, starting with an input layer that receives the preprocessed eigenvalue vector $\mathbf{x} = [l_1^{''}, ..., l_M^{''}]^T$. This feature vector $\mathbf{x}$ is then passed through a series of residual blocks. Each residual block allows information from the previous layer to bypass the main path through a shortcut connection, which is added directly to the output of the block. The output of the i-th residual block, denoted as $\mathbf{x}_{i+1}$, is defined as

$$\mathbf{x}_{i+1} = \mathrm{ReLU}\left(F\left(\mathbf{x}_i, \Theta_i\right) + shortcut\left(\mathbf{x}_i\right)\right) \tag{13}$$

Where, $\mathbf{x}_i$ represents the input to the i-th layer, and Θ_i denotes the parameters of that layer. The function $F\left(\mathbf{x}_i, \Theta_i\right)$ is a nonlinear mapping that consists of the following layers *Dense* → *BatchNormalization* → *ReLU* → *Dropout*. The term $shortcut\left(\mathbf{x}_i\right)$ refers to the shortcut connection.

The EResNet network addresses the SNE problem as a regression task. Consequently, the output layer of the network consists of a single neuron with a linear activation function. The network produces a scalar output, denoted as $\hat{K}$, which directly predicts the number of sources. To train the network, we minimize the Mean Squared Error (MSE) between the predicted values $\hat{K}$ and the actual number of sources $\hat{K}$ over a mini-batch of B samples as

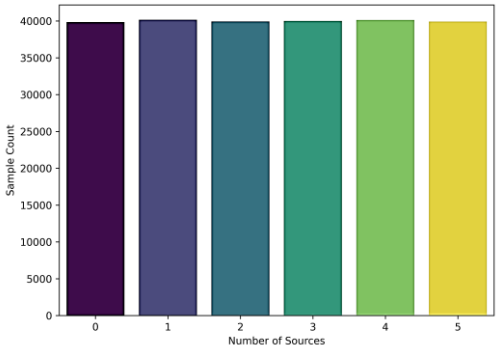
$$LMSE(\Theta) = \frac{1}{B}\sum_{i=1}^{B}(K_i - \hat{K}_i)^2 \tag{14}$$

The EResNet model is trained offline using synthetic data, and its parameters are optimized with the Adam (Adaptive Moment Estimation) algorithm [17].

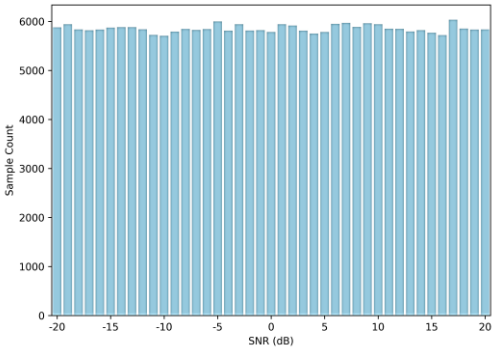
3. SIMULATION RESULTS AND DISCUSSION

3.1. Simulation preparation

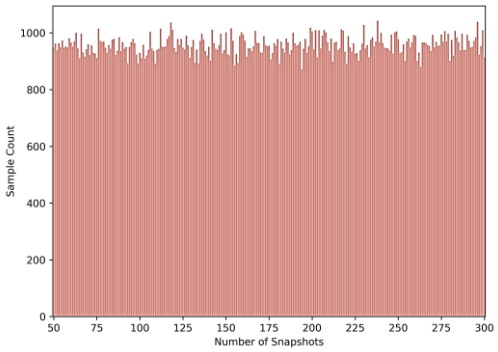
In this section, we utilize an 8x8 URA with $M = 64$ to generate our simulation dataset. The parameters for the simulation are set as follows: the number of sources is randomly selected from 0 to 5; the SNR ranges from 0 to 20 dB; the number of snapshots varies between 50 and 300; and the noise is generated from three types: white noise, colored noise ($\alpha = 0.85$), and underwater ambient noise ($s = 0.5$, $w = 10$ m/s). For each sample, parameters are randomly chosen, resulting in a total of 240,000 samples. The synthesized dataset is then analyzed to assess its diversity and balance. The evaluation results, presented in figure 3, demonstrate that the data samples are evenly distributed in terms of the number of sources, SNR, number of snapshots, and noise characteristics.



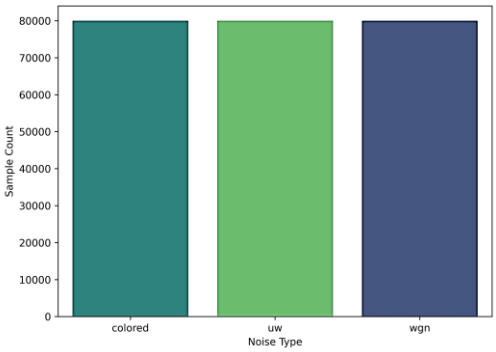
(a) Number of sources versus sample index.



(b) SNR versus sample index.



(c) Number of snapshots versus sample index.



(d) Noise type versus sample index.

Figure 3. *Evaluation of the dataset.*

The simulation dataset was divided into three subsets: 70% for the training set, 15% for the validation set, and 15% for the test set. After conducting experiments and making adjustments, we selected a model architecture that includes 2 input and output layers (with sizes 64 and 1, respectively) and 6 hidden layers (with sizes 128, 128, 64, 64, 32, and 32). The model was built and trained using the Keras library version 3.12.0, with an initial learning rate of 0.0005, a batch size of 512, and a total of 150 epochs. We employed the ADAM optimization algorithm, along with two callback mechanisms: EarlyStopping and ReduceLROnPlateau.

To evaluate the performance of the proposed network architecture, we established various evaluation scenarios and compared the accuracy of the EResNet network with previously studied models, specifically ECNet, ERNet [11], and traditional source estimation algorithms MDL and AIC [6].

Due to the impact of colored and underwater ambient noise, the eigenvalues associated with the noise space diverge, which can lead to the failure of the MDL and AIC methods. To address

this issue during the evaluation process, we combined the Diagonal Loading (DL) technique with the MDL and AIC methods. This modification adjusts the eigenvalues to reduce the spread within the noise space [18]. Specifically, the modified eigenvalues are represented by

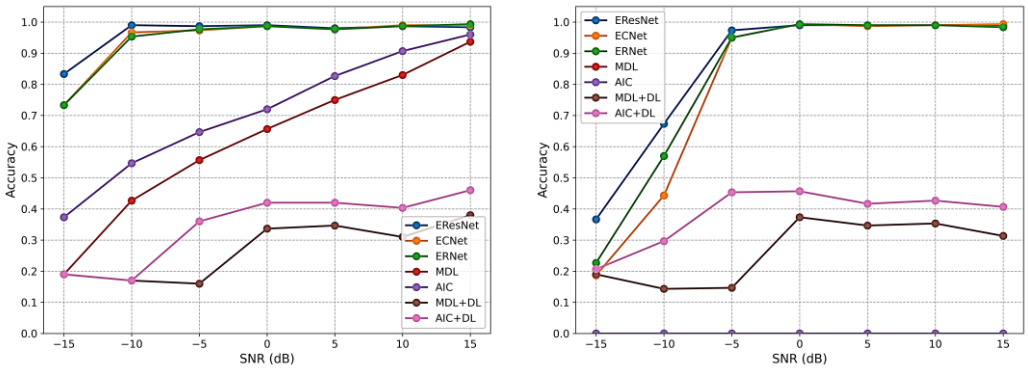
$$\hat{l}_m = l_m + \sqrt{\sum_{m=1}^{M} l_m} \tag{15}$$

In this study, we use the accuracy value, denoted as *Acc*, which represents the ratio of correctly predicted samples to the total number of test samples. The accuracy is calculated by

$$Acc = \frac{N_c}{N_t} \tag{16}$$

where N_c is the total number of test samples for which the predicted source number matches the actual source number and N_t is the total number of samples in the test set. For this study, the test set comprised 300 samples for each specific case.

3.2. Performance evaluation versus SNR



(a) White noise. *(b) Underwater ambient noise.*

Figure 4. *The accuracy performance versus SNR, where N = 100.*

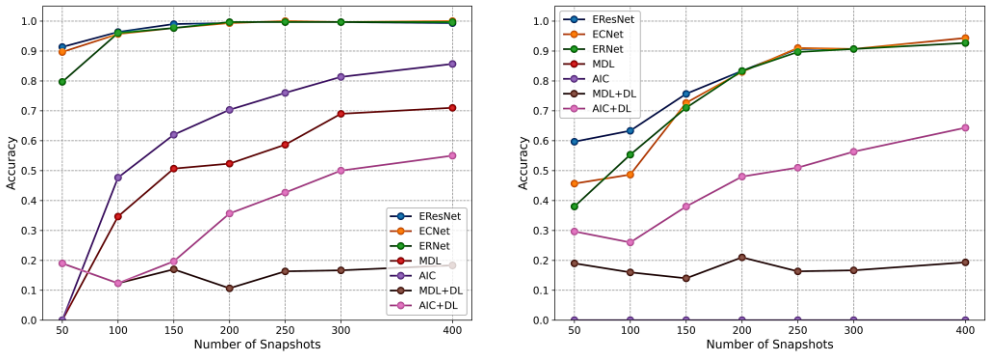
The performance evaluation of the algorithms as a function of SNR is illustrated in figure 4. Overall, the deep learning models demonstrate highly effective estimation performance, achieving high accuracy quickly as the SNR increases. EresNet consistently outperforms the others in the region where SNR is below -5 dB, particularly in the presence of underwater ambient noise.

The MDL and AIC algorithms perform reasonably well, aligning with theoretical expectations. In environments with white noise, both MDL and AIC show a gradual and stable increase in accuracy as the SNR improves. However, when confronted with underwater ambient noise, these methods struggle to maintain their estimation capability, even at higher SNR levels. This limitation arises because the MDL/AIC algorithms assume that the noise eigenvalues are equal, which is only valid for white noise with a flat power spectrum. In contrast, underwater ambient noise, as described by the Wenz model, has a non-flat power spectrum concentrated in specific frequency bands, particularly low frequencies due to shipping activities.

When the MDL/AIC algorithms are combined with the DL technique, they manage to maintain some estimation capability in the underwater ambient noise scenario, despite their poor performance in the white noise case. The DL technique alters the spectral structure of the eigenvalues, making it more challenging for MDL/AIC to establish a clear discrimination threshold between the signal and noise subspaces. For underwater ambient noise, adding a large constant to all eigenvalues stabilizes the distorted spectral structure. While this adjustment does not restore the ideal structure, it creates a new one that, although not perfectly accurate, is more stable. This stability allows the MDL/AIC algorithms to identify a threshold for making predictions.

3.3. Performance evaluation versus number of snapshots

Figure 5 illustrates the accuracy of various algorithms as the number of data samples increases from 50 to 400. In the presence of white noise, as shown in figure 5(a), deep learning algorithms are the most effective, particularly with a limited number of snapshots. The performance of these deep learning models increases rapidly, approaching a nearly perfect level (accuracy ≈ 1.0) once the number of snapshots reaches 150 - 200 or more. This indicates their strong ability to learn effectively from a relatively small dataset. In contrast, the MDL/AIC group of algorithms is highly reliant on the number of snapshots. The MDL algorithm starts with very low accuracy but steadily improves, reaching over 70% accuracy at 400 snapshots. AIC shows a similar trend, increasing from about 20% to over 85% as the number of snapshots grows from 50 to 400.



(a) White noise. *(b) Underwater ambient noise.*

Figure 5. *The accuracy performance versus SNR, where SNR = -10 dB.*

Under underwater ambient noise conditions, as illustrated in figure 5(b), the performance of deep learning algorithms does decline compared to the white noise scenario at low snapshot counts; however, this group still exhibits impressive robustness. EResNet remains the leading model, beginning with an accuracy of around 60% at 50 snapshots and rising to nearly 95% at 400 snapshots.

Similar to the analysis in section 3.1, the MDL/AIC algorithms struggle to maintain their estimation capabilities in the presence of underwater ambient noise. When combined with deep learning techniques, they still retain some estimation capability. However, the accuracy for both methods remains relatively low, around 30 - 40%.

4. CONCLUSIONS

In this study, we introduced the EResNet architecture to address the problem of estimating the number of underwater sources. Our experimental results demonstrated that EResNet significantly outperforms traditional methods such as AIC and MDL, especially in challenging conditions like low SNR, limited snapshots, and complex noise structures. The incorporation of residual connections in EResNet has allowed it to learn features more effectively compared to previously studied network architectures, resulting in consistent performance improvements.

However, the proposed model does have some inherent limitations. Since EResNet was trained entirely on simulated data, its performance on experimental data, which tends to exhibit more complex noise and signal propagation characteristics, has yet to be validated. Additionally, the computational complexity of the deep neural network poses a challenge for applications that require real-time processing.

Future development will aim to bridge the gap between simulation and real-world applications. One promising approach is to integrate advanced signal denoising filters before eigenvalue decomposition to improve the quality of the input features. Moreover, it is essential to collect and fine-tune the model using experimental datasets to validate its performance and ensure that EResNet can be effectively deployed in practical settings.

REFERENCES

[1]. J. Barabell, *"Improving the resolution performance of eigenstructure-based direction-finding algorithms,"* Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pp. 336–339, (1983).

[2]. F. Gao and A. B. Gershman, *"A generalized ESPRIT approach to direction-of-arrival estimation",* IEEE Signal Process. Lett., Vol. 12, No. 3, pp. 254–257, (2005).

[3]. Qian, L. Huang, N. Sidiropoulos, and H. So, *"Enhanced PUMA for direction-of-arrival estimation and its performance analysis,"* IEEE Trans. Signal Process., Vol. 64, No. 16, pp. 4127–4137, (2016).

[4]. P. Stoica and A. Nehorai, *"Performance study of conditional and unconditional direction-of-arrival estimation,"* IEEE Trans. Acoust., Speech Signal Process., Vol. 38, No. 10, pp. 1783–1795, (1990).

[5]. M. Wax and T. Kailath, *"Detection of signals by information theoretic criteria,"* IEEE Trans. Acoust., Speech Signal Process., Vol. ASSP-33, No. 2, pp. 387–392, (1985).

[6]. Nadler, *"Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator,"* IEEE Trans. Signal Process., Vol. 58, No. 5, pp. 2746–2756, (2010).

[7]. H. Akaike, *"A new look at the statistical model identification,"* IEEE Trans. Autom. Control, Vol. AC-19, No. 6, pp. 716–723, (1974).

[8]. J. Rissanen, *"Modeling by shortest data description,"* Automatica, Vol. 14, No. 5, pp. 465–471, (1978).

[9]. R. Coates, *"Underwater Acoustic Systems,"* New York: Wiley, (1989).

[10]. W. Hu, R. Liu, X. Lin, Y. Li, X. Zhou, and X. He, *"A deep learning method to estimate independent source number,"* Proc. 4th Int. Conf. Syst. Informat. (ICSAI), pp. 1055–1059, (2017).

[11]. Y. Yang, F. Gao, C. Qian, and G. Liao, *"Model-aided deep neural network for source number detection,"* IEEE Signal Process. Lett., Vol. 27, pp. 91–95, (2020).

[12]. S. Zhou, T. Li, Y. Li, R. Zhang, and Y. Ruan, *"Source number estimation via machine learning based on eigenvalue preprocessing",* IEEE Commun. Lett., Vol. 26, No. 10, pp. 2360–2364, (2022).

[13]. T. Hoang and K. Lee, *"Coherent signal enumeration based on deep learning and the FTMR algorithm,"* Proc. IEEE Int. Conf. Commun., Seoul, (2022).

[14]. A. Barthelme, R. Wiesmayr, and W. Utschick, *"Model order selection in DoA scenarios via cross-entropy based machine learning techniques,"* Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Barcelona, Spain, (2020).

[15]. G. M. Wenz, *"Acoustic ambient noise in the ocean: Spectra and sources,"* J. Acoust. Soc. Am., Vol. 34, No. 12, pp. 1936–1956, (1962).

[16]. S. Zhou, T. Li, Y. Li, R. Zhang, and Y. Ruan, *"Machine-learning-based source number estimation under unknown spatially correlated noise,"* IEEE Sensors J., Vol. 24, No. 9, pp. 14800–14811, (2024).

[17]. Kingma and J. Ba, *"Adam: A method for stochastic optimization,"* arXiv:1412.6980, (2014).

[18]. J. Xie and X. S., *"Determining the number of sources based on diagonal loading to the covariance matrix,"* Syst. Eng. Electron., Vol. 29, No. 5, pp. 596–600, (2008).

TÓM TẮT

Mạng hồi quy dư sâu để ước tính số lượng nguồn tín hiệu thủy âm

Ước tính số lượng nguồn (SNE) là một bài toán nền tảng trong xử lý tín hiệu mảng thủy âm, có vai trò quyết định đến hiệu suất của các thuật toán xử lý tiếp theo. Các phương pháp kinh điển như Tiêu chí Thông tin Akaike (AIC) và Độ dài Mô tả Tối thiểu (MDL), thường suy giảm hiệu suất nghiêm trọng trong các điều kiện bất lợi của môi trường dưới nước, đặc biệt là khi tỷ số tín hiệu trên nhiễu (SNR) thấp, số lượng mẫu (snapshots) hạn chế, và nhiễu có cấu trúc phức tạp. Để giải quyết những thách thức này, bài báo này giới thiệu một kiến trúc mạng hồi quy kết nối dư dựa trên giá trị riêng (EResNet) để ước lượng số nguồn một cách bền vững. Thông qua các mô phỏng toàn diện với nhiều kịch bản nhiễu phức tạp EResNet đã chứng tỏ hiệu quả vượt trội. Kết quả cho thấy mô hình đề xuất đạt độ chính xác cao hơn và thể hiện tính bền vững so với các thuật toán AIC, MDL cũng như các kiến trúc mạng nơ-ron cơ sở khác.

Từ khoá: Thủy âm; Ước lượng số nguồn; Học sâu; Mạng kết nối dư.