UNDERWATER ACOUSTIC SIGNAL RECOGNITION BASED ON COMBINATION OF MULTI-SCALE CONVOLUTIONAL NEURAL NETWORK AND CONSTANT-Q TRANSFORM

Cong Trang Tran[1], *Van Lam Nguyen*[1], *Xuan Sung Tran*[1], *Thi Huyen Le*[2], *Van Sang Doan*[1,*]

doi:10.56651/lqdtu.jst.v11.n02.537.ict

Abstract

This article proposes a multi-scale deep learning network to classify different underwater acoustic signal sources. The proposed network is cleverly designed with multiple branches, creating a multi-scale block which allows learning various spatial features of Constant-Q Transform spectrograms. The network is trained and tested on the ShipsEAR dataset, which is augmented by the overlap segmentation technique to ensure a balance in ship label classes' data. The experiment results show that our network achieves an average classification accuracy of up to 99.93% and an execution speed of 2.2 ms when configured with two multi-scale blocks and 32 filter channels. In comparison, our network remarkably outperforms other existing networks in terms of accuracy and execution time.

Index terms

Deep neural network, Constant-Q transform, underwater acoustic signal classification, spatial feature.

1. Introduction

Sonar devices play an important role in combat operations in the sea; they are regarded as "ears" to listen and "eyes" to observe the surrounding situation [1], [2]. Certainly, the ability to classify radiated sound of underwater targets is crucial for a sonar operator. It is confident that the operator's knowledge, qualifications, and experience play a key role in making the final prediction of tracked targets based on human auditory perception, as well as spectrogram analysis and other analytical parameters provided by the sonar devices. Therefore, the accuracy of judgment and the efficiency of target classification depend entirely on the operator. Because of the fact that the knowledge and experience of each operator might be different from others, one problem is whether the accuracy of target classification can be maintained if the operator is replaced by another. Moreover,

[1]Naval Academy, Nha Trang City, Khanh Hoa Province, Vietnam
[2]Naval Technical Institute, Hai Phong City, Vietnam
[*]Corresponding author: doansang.g1@gmail.com

the health and psychological status of the operator also has considerable impacts on the classification efficiency.

To assist the operator and facilitate stabilizing the classification efficiency, scientists and researchers have proposed deep learning neural networks in many frameworks. Indeed, neural networks, including a probabilistic neural network (PNN) [3], a multi-layer perceptron (MLP) [4], an adaptive kernel classifier (AKC) [5], and a learning vector quantization (LVQ) [6], have been investigated in [7] to deal with the classification of underwater signals of fishing boats. The work shows that the LVQ approach combined with the Two-Pass Split-Windows (TPSW) algorithm-based preprocessing for extracting tonal features from the average power spectral density (APSD) achieves a better performance than other ones in terms of learning speed and classification accuracy. Despite obtaining a quite high classification accuracy, the dataset used in [7], which contains only 200 patterns of signals, is really small to have an adequate demonstration. In another work [8] in 2012, Feng et al. investigated the characteristics of the simplified fractional Fourier transform (SFRFT) spectrum of chirp periodic signals to reveal the relation between the chirp periodic signal and its chirp harmonics under the conditions of underwater passive detection. Through the simulation and experiment, the signal feature of propeller cavitation noise during the acceleration or deceleration stage has been extracted to passively detect and classify moving vessels and underwater vehicles. However, the recognition approach in [8] is relying on expert feature extraction, which might result in false recognition in the condition of dense underwater sound noises.

As mentioned above, neural networks have been successfully applied to classify underwater acoustic signals radiated from vessel propellers. In this article, we propose to combine constant-Q Transform (CQT) and a multi-scale convolutional neural network (named CQT-CNN) for underwater acoustic signal recognition. While the CQT transform serves as a time-frequency spectral image generator, the multi-scale CNN is responsible for signal classification. The proposed model is trained and tested on a dataset named ShipsEAR [9], and compared with other network models to show off its performance enhancement.

The rest of the article is organized as follows: Section 2 introduces the related works; then, the dataset and CQT processing technique are presented in Section 3. Following this, we describe the proposed multi-scale convolutional neural network for underwater acoustic signal recognition in Section 4. Next, the proposed CQT-CNN model has been experimented on the ShipsEAR dataset. The experimental results are discussed in Section 5. Finally, we conclude the research achievement and define the future research intention related to this work in Section 6.

2. Related works

Underwater acoustic signal recognition is very complex for sonar operators; therefore, expert feature extraction of these signals is always needed, but this process consumes a lot of time and cost. In order to deal with the feature extraction limitation, a convolutional

neural network (CNN) has been proposed in [10] for the recognition and classification of underwater acoustic signals representing the targets of vessels, submarines, and torpedoes. Accordingly, the acoustic signals are transformed into Low-Frequency Array (LOFAR) spectral images for feeding the CNN model. As a result, the model has reached a recognition accuracy of 97.22%. Despite achieving a higher accuracy than other considered networks, the CNN model does not satisfy a generalization for the diversity of underwater acoustic signals because it has been experimented on a relatively small dataset.

In another approach, an anti-noise Power-Normalized Cepstral Coefficients (ia-PNCC) technique based on multi-taper and normalized Gammatone filter banks was proposed in [11] to improve the anti-noise capacity. Consequently, the ia-PNCC features of acoustic signals have mitigated the noise and are well-suited for underwater target recognition using a CNN model. Compared with Mel-scale Frequency Cepstral Coefficients (MFCC) [12] and Linear Prediction Cepstral Coefficients (LPCC) [13] techniques in combining with the CNN model, the ia-PNCC preprocessing technique has offered a higher accuracy for underwater target recognition. However, considering only a three-target-class dataset was insufficient, and the CNN model used for the target recognition was not optimal.

To improve the recognition accuracy, Chen et al. [14] have proposed the use of residual CNN combined with the wavelet transform-based time-frequency images of underwater acoustic signals. The method has achieved an accuracy of over 93%, which is higher than a normal CNN model with five layers. Despite that, the result of the experiment mentioned in [14] was solely based on a two-target dataset, and therefore was not generalized for the recognition problem of multiple underwater acoustic signals. With a larger dataset of ship-radiated sounds, a multi-scale residual unit (MSRU) was proposed to construct a deep convolution stack network, which provided a recognition accuracy of 83.15% [15]. An attention-based neural network (ABNN) was applied for ship recognition in the time-frequency spectrogram with two-source interference [16]. The attention module helps distinguish the closely correlated features in frequency regions to increase the accuracy of the ship's sound recognition.

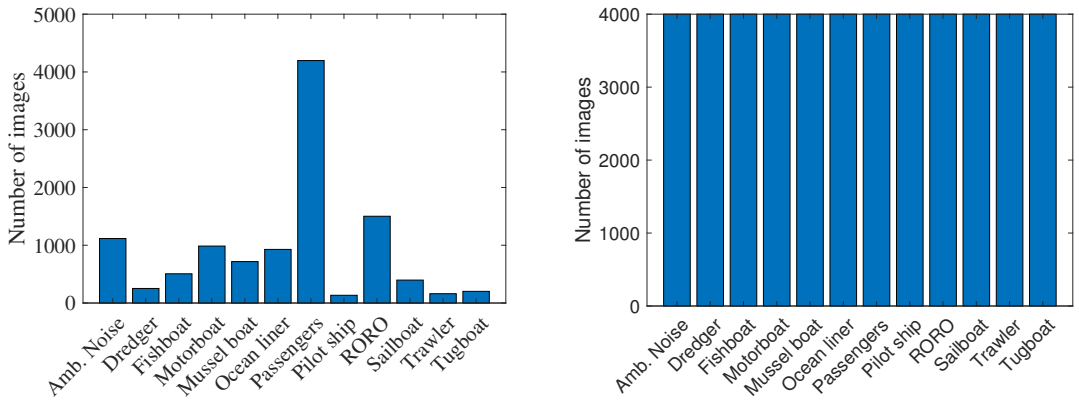
Although the aforementioned research works have succeeded in recognizing the underwater acoustic signals, some limitations still exist in investigating the network structural optimization and computational complexity. To this end, we propose a combination of constant-Q Transform and a multi-scale convolutional neural network for underwater acoustic signal recognition. The CQT transform generates time-frequency spectral images with a high frequency resolution at low frequencies and a high time resolution at high ones, which is well motivated by both musical and perceptual human hearing [17]. The convolutional neural network is designed with multi-scale blocks, which consist of multiple convolutional layers arranged as parallel branches. This design helps the model learn and process feature maps using different filter sizes. As a result, more representative and distinct features will be generated through the multi-scale blocks for improving classification accuracy. The network is trained and tested on a dataset named

ShipsEAR [9]. Subsequently, combined with CQT preprocessing, our network achieves an average classification accuracy of 99.93% and remarkably outperforms several other existing networks in terms of accuracy and execution time.

3. Dataset and Constant-Q Transform based preprocessing technique

3.1. Dataset collection

Dataset plays a crucial role in deep learning-based classification problems; the more diverse the dataset is, the more robust the classifier's performance is. Regarding this research work, datasets of vessel propeller sounds are very scarce due to the expensive acquiring equipment and time cost. Therefore, we collected the datasets from a public source, ShipsEAR, provided by Santos-Domínguez et al. [9] from the University of Vigo, Spain. The dataset was detailed in [18] with the specific water/air environment parameters for sound propagation (such as channel depth, distance from source to sonar, AIS information, locations, surface noise, temperature, etc.). The ShipsEAR dataset consists of 96 audio files of different ships with sufficient information, such as pictures, ship names, locations, and movement situations. Accordingly, by cleaning and analysis, we manually classify and label them into twelve categories, including eleven ship types and one ambient natural noise (Amb. Noise). It is worth noting that there are different movement situations in the same category. Through CQT preprocessing, every three-second frame of sound in the raw waveform is transformed into a CQT spectrogram with a size of 250×500. Consequently, 11,210 spectrograms are generated with the distribution shown in Fig. 1a. It can be observed that the dataset is imbalanced; therefore, we apply an overlap segmentation to augment the dataset, which leads to it being more balanced. This strategy produces a balanced dataset of 48,000 CQT spectrograms (4000 images for each class), whose distribution is in Fig. 1b.



(a) Without augmentation *(b) With time overlapping-based augmentation*

Fig. 1. Distribution of CQT spectrogram images in datasets.

3.2. Constant-Q Transform based preprocessing technique

Constant-Q Transform is a mathematical model used in signal processing, especially musical processing, to transform a signal from the time domain to the frequency domain [17]. If the short-time Fourier transform (STFT) converts a raw waveform signal to a time-frequency spectrogram with linearly spaced bins, the CQT performs this process with equal Q-factor for all frequency bins. It means that CQT provides a high frequency resolution at low frequencies and a high time resolution at high ones, which is well inspired by musical and perceptual human hearing. If $x(n)$ is denoted as a discrete time-domain signal, then its CQT transform $X(k,n)$ is defined as follows:

$$X(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \tag{1}$$

where $k = 1, 2, ..., K$ is the index of frequency bin, $\lfloor \cdot \rfloor$ stands for the rounding operation towards the negative infinity, $a_k^*(n)$ is the complex conjugate of $a_k(n)$ which is so-called time-frequency kernel and defined as follows:

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_k}\right) e^{-j2\pi n \frac{f_k}{f_s}} \tag{2}$$

where f_k is the center frequency of bin k, f_s is the sample rate, $w(\cdot)$ is the window function, and N_k is the window length. The center frequency f_k of bin k is defined as follows:

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{3}$$

where f_1 is the center frequency of the lowest frequency bin, and B is the number of bins per octave, which is the most important parameter of the CQT transform because it defines the trade-off of time and frequency resolution. The Q-factor of bin k is defined by:

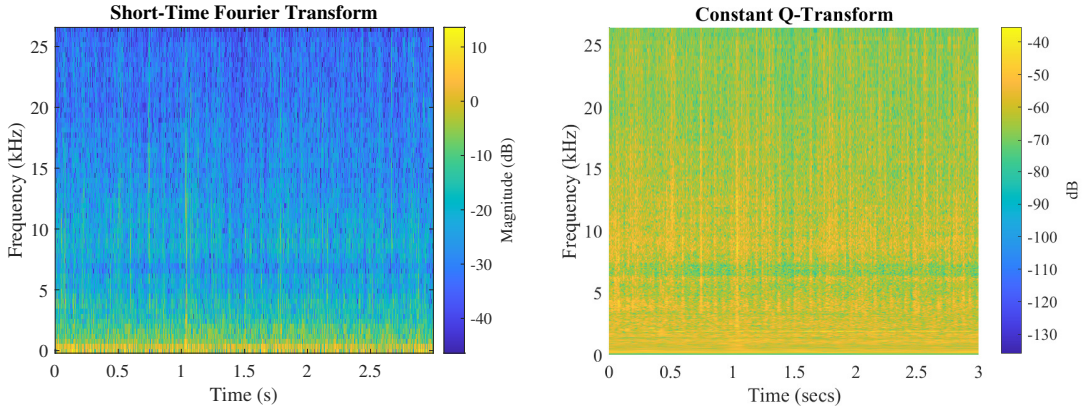
$$Q_k = \frac{f_k}{\Delta f_k} = \frac{N_k f_k}{\Delta \omega f_s} \tag{4}$$

where Δf_k is the bandwidth at -3 dB of the frequency response of the kernel $a_k(n)$, and $\Delta\omega$ is the bandwidth at -3 dB of the main-lobe of the spectrum of the window function $w(t)$. The Q-factors are the same for all frequency bins; therefore, it is recommended to make Q as large as possible:

$$Q = \frac{q}{\Delta\omega(2^{\frac{1}{B}} - 1)} \tag{5}$$

where $q \in (0, 1]$ is the scaling factor. For a computationally efficient implementation, the CQT can be calculated through Fourier transform as follows:

$$X^{CQ}(k, N/2) = \sum_{j=0}^{N} X(j) A_k^*(j) \tag{6}$$



(a) STFT spectrogram *(b) CQT spectrogram*

Fig. 2. Comparing the spectrograms of STFT and CQT.

where $X(j)$ and $A_k(j)$ are the complex-valued Fourier transform of $x(n)$ and $a_k(n)$, respectively. As shown in Fig. 2, we compare spectrograms using traditional STFT (Fig. 2a) and CQT (Fig. 2b), where it can be observed that the CQT spectrogram has a higher resolution at a low frequency than the STFT spectrogram.

4. Multi-scale convolutional neural network for underwater acoustic signal recognition

Inspired by the hearing and learning processes in the human brain, supervised deep neural networks (DNN) were proposed to solve the problem of recognizing underwater acoustic targets based on the sound noise they emit [19]–[21]. The DNN model can automatically identify the targets it observes by being trained with the available dataset. Fundamentally, training is the process of updating the model weights. Thus, the model will perform the tasks rigorously as it was trained. Regarding the underwater acoustic signal classification using DNN, as shown in Fig. 3, supervised learning is conducted based on the dataset labeled with twelve source names, denoted as $\{X|Y\}$, where X is the observed data, and Y is the corresponding labels. In the forward propagation, the input data is processed in the DNN model to provide a predicted class, denoted as $\tilde{Y}$. Then, a specific loss function is applied to calculate the error value between predicted and ground-truth labels. Specifically, the cross-entropy loss function is employed in this work. The cross-entropy loss function is defined as follows:

$$L_{CE} = \frac{1}{C}\sum_{i=1}^{C} y_i \log \tilde{y}_i + (1 - y_i)\log(1 - \tilde{y}_i) \tag{7}$$

where C is the number of output size, y_i and $\tilde{y}_i$ are the score of the i^{th} ground-truth and predicted class, respectively. Afterward, the model weights will be updated according to the error value in the back-propagation. The forward and backward propagation are

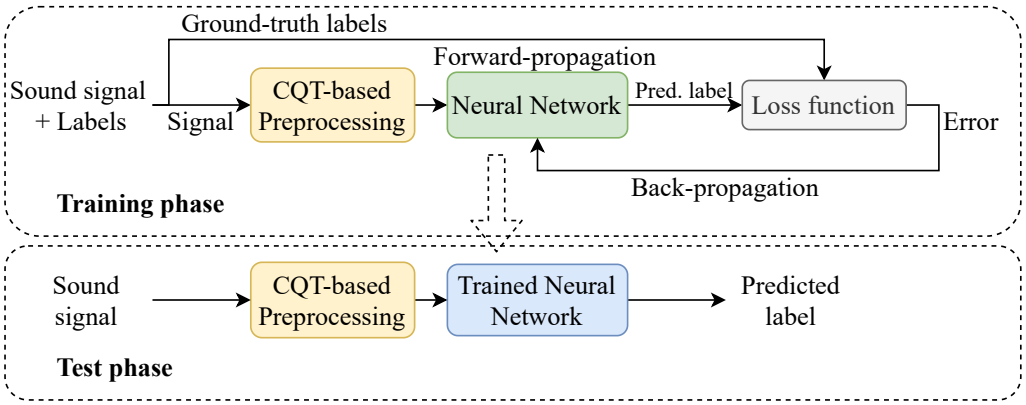


Fig. 3. Training and testing procedure of the proposed neural network.

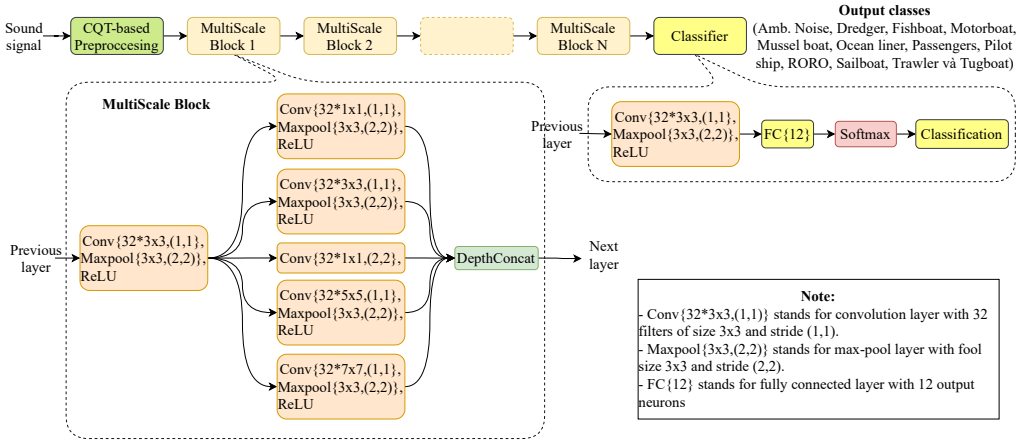


Fig. 4. Structure of CQT-CNN for underwater acoustic signal classification.

repeated in the loop and stopped when the model provides an acceptable minimal loss value or after a defined number of epochs. In the test phase, only unseen data are fed to the trained model in order to predict a class corresponding to the input data.

This study proposes a convolutional neural network (CNN) combined with the CQT preprocessing, so-called CQT-CNN, to classify twelve types of underwater acoustic signals. Eleven sounds are recorded from the surface vessel propellers, and the remainder is the sound recorded from natural water noise. It is worth mentioning that one label can be assigned for multiple vessels of the same type. The structure of CQT-CNN is depicted in Fig. 4, where we can see that CQT-CNN comprises CQT-based Preprocessing, several consecutive connected Multiscale blocks, and a classifier. The detailed parameters of CQT-CNN with two MultiScale blocks are reported in Table 1.

The CQT-based Processing block plays the role of a transformer, which converts the sound signal in the time domain to the CQT spectrogram image as shown Fig. 2b. The detailed process of the CQT-based Processing block has been described in Sub-

Table 1. Detailed parameters of CQT-CNN model with two MultiScale blocks and 32 filter channels

Block	Layer	Output size	Number of weights	Description
	Input	$250 \times 500 \times 1$	0	Size of CQT spectrogram
MultiScale Block 1	Conv-maxpool-ReLU	$125 \times 250 \times 32$	320	Conv{32*(3×3), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	BatchNorm	$125 \times 250 \times 32$	64	Batch normalization
	Branch 1 {Conv-maxpool-ReLU}	$63 \times 125 \times 32$	1056	Conv{32*(1×1), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 2 {Conv-maxpool-ReLU}	$63 \times 125 \times 32$	9248	Conv{32*(3×3), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 3 {Conv}	$63 \times 125 \times 32$	1056	Conv{32*(1×1), stride(2,2)}
	Branch 4 {Conv-maxpool-ReLU}	$63 \times 125 \times 32$	25632	Conv{32*(5×5), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 5 {Conv-maxpool-ReLU}	$63 \times 125 \times 32$	50208	Conv{32*(7×7), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	DepthConcat	$63 \times 125 \times 160$	0	Depth channel concatination
MultiScale Block 2	Conv-maxpool-ReLU	$32 \times 63 \times 32$	46112	Conv{32*(3×3), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	BatchNorm	$32 \times 63 \times 32$	64	
	Branch 1 {Conv-maxpool-ReLU}	$16 \times 32 \times 32$	1056	Conv{32*(1×1), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 2 {Conv-maxpool-ReLU}	$16 \times 32 \times 32$	9248	Conv{32*(3×3), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 3 {Conv}	$16 \times 32 \times 32$	1056	Conv{32*(1×1), stride(2,2)}
	Branch 4 {Conv-maxpool-ReLU}	$16 \times 32 \times 32$	25632	Conv{32*(5×5), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	Branch 5 {Conv-maxpool-ReLU}	$16 \times 32 \times 32$	50208	Conv{32*(7×7), stride(1,1)} Maxpool{3×3, stride(2,2)} ReLU activation function
	DepthConcat	$16 \times 32 \times 160$	0	Depth channel concatination
Classifier	Conv-ReLU	$16 \times 32 \times 32$	46112	Conv{32*(3×3), stride(1,1)} ReLU activation function
	FC-Softmax-Output	$1 \times 1 \times 12$	196620	Number of neurons = number of classes Softmax function
Total			**463692**	

section 2.2. Before being fed to the CQT-CNN network, the CQT image is resized to a specific size, such as 250×500 in this study. Afterward, the resized CQT image is sent to the next block named Multiscale. Each Multiscale block starts with a group of three layers including convolutional (Conv), maxpool, and ReLU activation layers (Conv-Maxpool-ReLU). The Conv layer consists of 32 filter channels. Each channel is assigned with the filter size 3×3 and stride (1,1). Weights in the Conv layer can be

updated in the training process, allowing the network to learn from the provided data. In the forward propagation, the convolution is defined as follows:

$$S_{ij} = (I * K)_{ij} = \sum_{a=\lfloor -\frac{m}{2} \rfloor}^{\lfloor \frac{m}{2} \rfloor} \sum_{b=\lfloor -\frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} I_{i-a,j-b} K_{m/2+a,n/2+b} \tag{8}$$

where I and K stand for two-dimensional (2D) matrices of input data and convolutional kernel, respectively; and the kernel has the size of $m \times n$. The Maxpool layer is employed for the purpose of feature map reduction. Indeed, with the size 3×3 and stride (2,2), the maxpool operation takes the maximum value from each 3×3 sub-matrix and skips two elements to repeat the procedure in turns along row and column. As a result, the output feature map can be halved compared to the input one. Then, the output feature map is activated by the ReLU layer, whose function is defined as follows:

$$y = \max(x, 0) \tag{9}$$

where x and y stand for the input and output of the ReLU function, respectively.

After processing by the Conv-Maxpool-ReLU layer, the data is delivered to five branches: four Conv-Maxpool-ReLU branches and one Conv branch. The four Conv-Maxpool-ReLU branches are assigned with the same number of filters but have different filter sizes, including 1×1, 3×3, 5×5, and 7×7. This design helps to induce diverse filtering for extracting more robust and representative information from various sound characteristics. The Conv branch plays a skip-connection role as in the residual module, where the former features from the previous layer are reused for the subsequent blocks. Finally, outputs of all branches are depth-wise concatenated by using the DepthConcat layer. The stride parameter of the Conv layer in the skip connection must be set to (2,2) to ensure the application of deep concatenation.

After processing through several Multiscale blocks, the data goes to the Classifier block, where a Conv-Maxpool-ReLU group, fully connected layer (FC), Softmax, and Classification layers are designed. The FC layer has twelve output neurons corresponding to the number of label classes in the dataset. The Softmax layer uses the softmax function to provide scores or probabilities of each class, helping to judge the output label. The softmax function is described as follows:

$$\rho_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{10}$$

where z is the output features of the FC layer. Finally, the CQT-CNN model predicts the sound signal regarding the class having the highest probability, which means:

$$Source_{predicted} = \arg\max\{\rho(z)\} \tag{11}$$

5. Experimental results and discussion

This section reports experimental results evaluated on the twelve-label ShipsEAR dataset to demonstrate the efficiency of the CQT-CNN model. The training parameters are as follows: employment of the stochastic gradient descent with momentum (SGDM), the mini-batch size of 16, the maximum epoch of 20, the initial learning rate of 0.001, and reduced 10 times for every ten epochs. The training and testing processes and other measurements are conducted on a laptop with a configuration of CPU Corei5 9300H, RAM 16GB, and GPU NVIDIA GeForce GTX 1660ti. The dataset is randomly split into 2 separated portions, including 80% portion for training and 20% one for validation and testing. Fig. 5 shows the convergence process of training and validation losses of the CQT-CNN model.

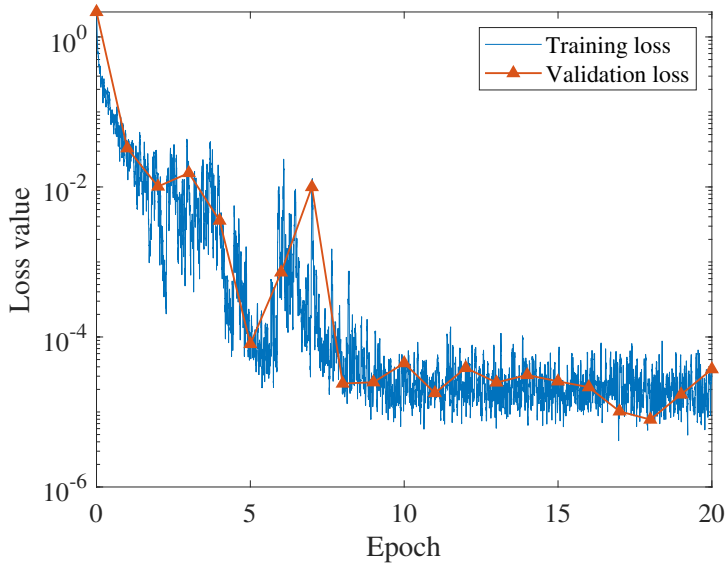
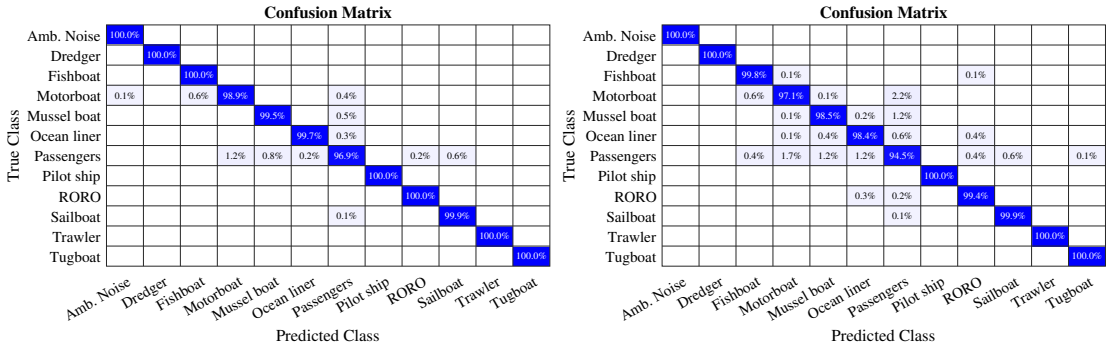


Fig. 5. Training process of the CQT-CNN model.

First and foremost, we evaluate the multiscale CQT-CNN compared to a non-multiscale model in terms of performance accuracy. Accordingly, all Conv layers of the non-multiscale model are assigned a filter size of 3×3. In this experiment, every Conv layer of both models is set with four filter channels. The confusion matrices taken into account for comparison are shown in Fig. 6, where we can observe that the proposed multi-scale design provides higher classification accuracy than the non-multiscale one thanks to the various filters in the Multi-scale blocks, which help to learn more robust and representative features from underwater acoustic sounds. Specifically, the multiscale CQT-CNN model achieves the lowest accuracy of 96.9% for Passenger labels, which is about 2.6% higher than the non-multiscale model. The highest misclassification occurs between Passengers and Motorboat.

In the second experiment, the classification accuracy and inference time of multiscale



(a) Multiscale model *(b) Non-multiscale model*

Fig. 6. Performance comparison between the multiscale model and the non-multiscale one.

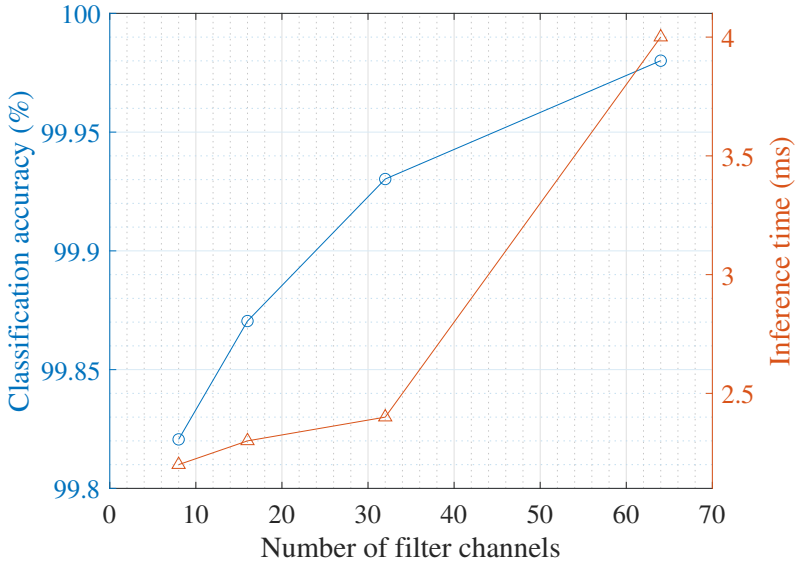
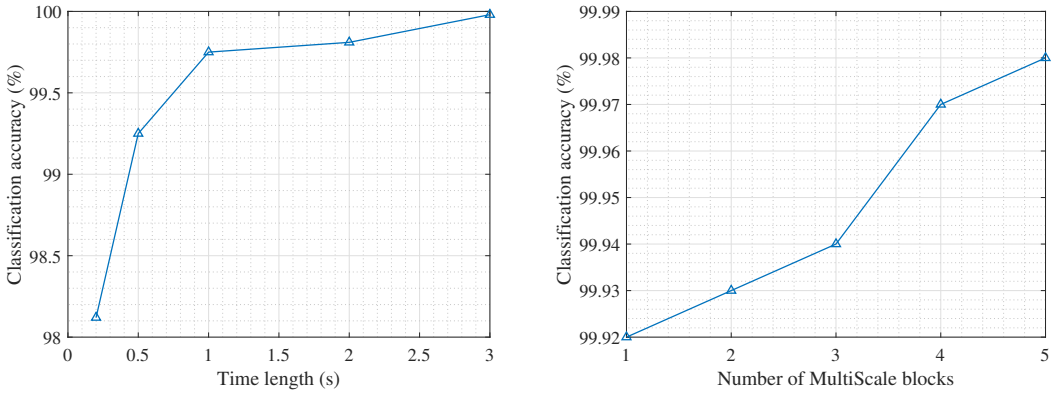


Fig. 7. Performance of CQT-CNN with different numbers of filter channels.

CQT-CNN are investigated when changing the number of filter channels in the Conv layers. Accordingly, the number of filter channels is designated with several values of {8, 16, 32, 64} to observe the accuracy and execution time trend. It is worth noting that this experiment selects two Multiscale blocks for CQT-CNN. Results in Fig. 7 show that the classification accuracy of CQT-CNN gradually increases with the number of filter channels. However, the increment of filter channels results in a slower processing speed of the model. Specifically, CQT-CNN of 64 filters takes the inference time of 4 ms, which is much slower than the model of 32 filters (only 0.24 ms). Meanwhile, the difference in execution time for 8, 16, and 32 filters are insignificant. This result is due to a significant change in the number of learnable parameters of CQT-CNN when adjusting the number of filters from 32 to 64.



(a) Different time lengths *(b) Different numbers of Multiscale blocks*

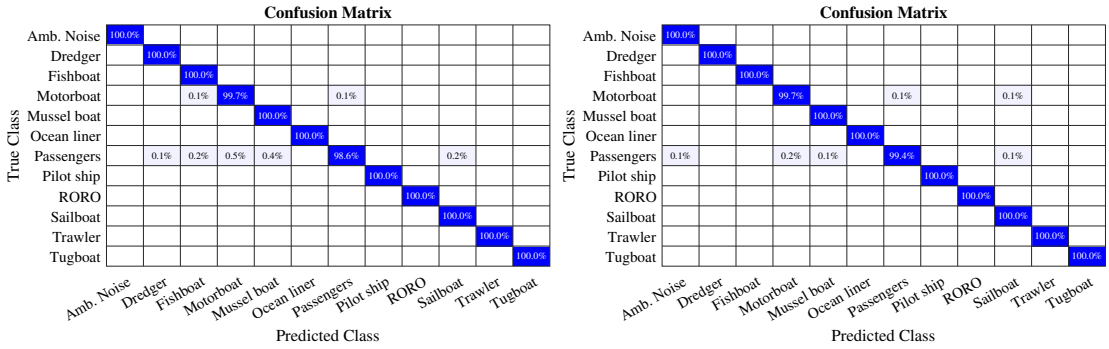
Fig. 8. Performance of CQT-CNN with different time lengths and different numbers of Multiscale blocks.

In the next experiment, we measure the dependence of classification accuracy on the time length of input sound data. Accordingly, not only do we segment the audio signals by 3 s but also into smaller segments, such as 0.2, 0.5, 1.0, and 2.0 s, then preprocess using CQT transform before feeding them to the proposed network. In this experiment, the network is designed with two Multiscale blocks and 32 filters. The experimental results in Fig. 8a indicate that the longer the signal fed to the network is, the higher the accuracy of the network is. Evidently, the vessel propellers radiate the sound signals, which have energy concentrated at lower frequencies. Therefore, more representative information will be processed in the network when a longer signal is captured.

Next, the CQT-CNN model is analyzed with different numbers of Multiscale blocks, including 1, 2, 3, 4, and 5. All Conv layers are set in this experiment with 32 filters of size 3×3. Numerical results in Fig. 8b reveal that despite improving classification accuracy when increasing the number of Multiscale blocks, these differences in accuracy are tiny because the CQT-CNN model of one Multiscale block seems to be satisfied with the ShipsEAR dataset. Specifically, the accuracy enhances only 0.06% for changing from 1 to 5 Multiscale blocks.

The preprocessing technique also affects the classification performance; therefore, the experiment to compare the classification accuracy of the proposed network with STFT and CQT transforms is carried out. As a result, due to higher resolution at low frequencies, the network with CQT transform yields a higher classification accuracy than that with STFT one (99.93% average accuracy for CQT compared to 99.86% for STFT). The confusion matrices of two networks are shown in Fig. 9, where the left one is of STFT transform, and the right one is of CQT transform.

Finally, the CQT-CNN network of two Multiscale blocks, 32 filter channels of size 3×3 is taken into account for competing with other existing state-of-the-art models, including EfficientNet [22], SqueezeNet [23], GoogLeNet [24], MobileNet-V3 [25], and ShufleNet [26] in the same task of underwater acoustic signal classification. The



(a) STFT transform *(b) CQT transform*

Fig. 9. Confusion matrix comparison of the proposed network when using STFT and CQT transform.

Table 2. The comparison between CQT-CNN and other state-of-the-art models

Model	Accuracy	Execution time
EfficientNet	98.84 %	13.8 ms
SqueezeNet	86.84 %	**2.1 ms**
GoogLeNet	98.26 %	3.5 ms
MobileNet-V3	99.06 %	4.8 ms
ShufleNet	96.78 %	10.4 ms
CQT-CNN	**99.93 %**	**2.2 ms**

performance comparison in terms of classification accuracy and inference time is reported in Table 2, where we can observe that the CQT-CNN model obtains the highest accuracy of 99.93% and the second fastest execution time of 2.2 ms, which is slower than SqueezeNet by about 0.1 ms. However, SqueezeNet provides the lowest classification accuracy of 86.84%. Overall, the CQT-CNN model remarkably outperforms other considered networks.

6. Conclusion

Our study has demonstrated the superiority of multi-scale CQT-CNN by evaluating different hyper-parameter configurations for underwater acoustic signal classification and comparing CQT-CNN with other existing models. With the clever design of multiple branches to create the multi-scale block, the CQT-CNN has been capable of learning various spatial features of Constant-Q Transform spectrograms. Through fine-tuning some parameters of the network, we have found a trade-off for the classification accuracy and execution time of CQT-CNN. Specifically, the network of two multiscale blocks and 32 filter channels in Conv layers has yielded an accuracy of 99.93% and a running time of 2.2 ms. This performance has remarkably outperformed other considered networks. In the future, we will focus on the dataset collected from real-world environments using our own designed devices. Next, we continue developing the network to implement the underwater acoustic signal classification task in real-time.

References

[1] J. A. Fornshell and A. Tesei, "The development of SONAR as a tool in marine biological research in the twentieth century," *International Journal of Oceanography*, vol. 2013, pp. 1–9, Nov. 2013. doi: 10.1155/2013/678621.

[2] H. Peyvandi, M. Farrokhrooz, H. Roufarshbaf, and S.-J. Park, "SONAR systems and underwater signal processing: Classic and modern approaches," in *Sonar Systems*. InTech, Sep. 2011. doi: 10.5772/17505.

[3] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, Jan. 1990. doi: 10.1016/0893-6080(90)90049-q.

[4] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Trans. Cir. and Sys.*, vol. 8, no. 7, p. 579–588, Jul. 2009. doi: 10.5555/1639537.1639542.

[5] F. B. Tek, İ. Çam, and D. Karlı, "Adaptive convolution kernel for artificial neural networks," *Journal of Visual Communication and Image Representation*, vol. 75, p. 103015, Feb. 2021. doi: 10.1016/j.jvcir.2020.103015.

[6] A. Sato and K. Yamada, "Generalized learning vector quantization," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[7] J.-D. L. C.-H. Chen and M.-C. Lin, "Classification of underwater signals using neural networks," *Tamkang Journal of Science and Engineering*, vol. 3, no. 1, pp. 31–48, Jun. 2000. doi: 10.6180/jase.2000.3.1.04.

[8] Y. Feng, R. Tao, and Y. Wang, "Modeling and characteristic analysis of underwater acoustic signal of the accelerating propeller," *Science China Information Sciences*, vol. 55, no. 2, pp. 270–280, Jun. 2011. doi: 10.1007/s11432-011-4285-9.

[9] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," *Applied Acoustics*, vol. 113, pp. 64–69, Dec. 2016. doi: 10.1016/j.apacoust.2016.06.008.

[10] H. Wu, Q. Song, and G. Jin, "Deep learning based framework for underwater acoustic signal recognition and classification," in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence - CSAI '18*. ACM Press, 2018. doi: 10.1145/3297156.3297180.

[11] N. Wang, M. He, J. Sun, H. Wang, L. Zhou, C. Chu, and L. Chen, "ia-PNCC: Noise processing method for underwater target recognition convolutional neural network," *Computers, Materials &amp Continua*, vol. 58, no. 1, pp. 169–181, 2019. doi: 10.32604/cmc.2019.03709.

[12] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4th International Conference on Signal Processing and Communication Systems*. IEEE, Dec. 2010. doi: 10.1109/icspcs.2010.5709752.

[13] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*. IEEE, Jan. 2016. doi: 10.1109/confluence.2016.7508171.

[14] Y. Chen, S. Du, H. Quan, and B. Zhou, "Underwater target recognition method based on convolution residual network," *MATEC Web of Conferences*, vol. 283, p. 04011, 2019. doi: 10.1051/matecconf/201928304011.

[15] S. Tian, D. Chen, H. Wang, and J. Liu, "Deep convolution stack for waveform in underwater acoustic target recognition," *Scientific Reports*, vol. 11, no. 1, May 2021. doi: 10.1038/s41598-021-88799-z.

[16] X. Xiao, W. Wang, Q. Ren, P. Gerstoft, and L. Ma, "Underwater acoustic target recognition using attention-based deep neural network," *JASA Express Letters*, vol. 1, no. 10, p. 106001, Oct. 2021. doi: 10.1121/10.0006299.

[17] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," 2010. doi: 10.5281/ZENODO.849740.

[18] *ShipsEAR: an underwater vessel noise database*. [Online]. Available: http://atlanttic.uvigo.es/underwaternoise/.

[19] X. Cheng and H. Zhang, "Underwater target signal classification using the hybrid routing neural network," *Sensors*, vol. 21, no. 23, p. 7799, Nov. 2021. doi: 10.3390/s21237799.

[20] V.-S. Doan, T. Huynh-The, and D.-S. Kim, "Underwater acoustic target classification based on dense convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, Oct. 2022. doi: 10.1109/lgrs.2020.3029584.

[21] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, and K. Sammut, "A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance," *Sensors*, vol. 22, no. 6, p. 2181, Mar. 2022. doi: 10.3390/s22062181.

[22] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019. doi: 10.48550/ARXIV.1905.11946.

[23] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," 2016. doi: 10.48550/ARXIV.1602.07360.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014. doi: 10.48550/ARXIV.1409.4842.

[25] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. doi: 10.48550/ARXIV.1905.02244.
[26] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017. doi: 10.48550/ARXIV.1707.01083.

Manuscript received 31-08-2022; Accepted 11-11-2022.

Cong Trang Tran was born in 1972 in Ninh Binh Province, Vietnam. He received Bachelor's degree in Electronic Engineering from Naval Academy, Nha Trang City, Vietnam; Master of Science degree from Da Nang University in 2009; and Ph.D. degree in Radar and Navigation Engineering at the Institute of Military Science and Technology, Hanoi, Vietnam. He is currently the Dean of the Faculty of Communication and Radar, Naval Academy. His research field includes underwater acoustic and sonar, radar and navigation, and electronic warfare.
Email: trancongtrang@gmail.com.

Van Lam Nguyen was born in 1965 in Nam Dinh, Vietnam. He received Engineer's degree at Caspian Higher Naval Red Banner School named after Sergei Kirov, Soviet Union (now Azerbaijan Higher Naval Academy) in 1988; Ph.D. degree at Baltic State Technical University, Russia in 2005; and became an associate professor in 2018. He is now the Rector of Naval Academy, Nha Trang City, Khanh Hoa Province, Vietnam. His research field includes system analysis and integration, information processing, and control of engineering systems.
Email: lamhvhq1965@gmail.com.

Xuan Sung Tran received his degree of engineer in technical telecommunication from Naval Academy, Nha Trang City, Khanh Hoa province, Vietnam in 2020. He is currently a lecturer working in Faculty of Communication and Radar, Naval Academy. His current research interests include radar and sonar systems as well as signal processing.
Email: sungtranna@gmail.com.

Thi Huyen Le received her engineer's degree in Electrical and Electronic Engineering from Le Quy Don Technical University in Hanoi, Vietnam in 2014. She is currently a researcher at Naval Technical Institute. Her current research interests include communication and sonar systems, and signal processing.
Email: lehuyen.mta.204@gmail.com.

Van Sang Doan received his M.Sc. and Ph.D. degrees in electronic systems and devices from Faculty of Military Technology, University of Defence, Brno, Czech Republic, in 2013 and 2016, respectively. He was awarded three Honor medals by the Faculty of Military Technology of the University of Defence in 2011, 2013, and 2016, respectively. From 2019 to 2020, he was a postdoctoral research fellow at ICT Convergence Research Center, Kumoh National Institute of Technology, South Korea. He is currently a lecturer at Faculty of Communication and Radar, Naval Academy, Nha Trang City, Vietnam. His current research interests include communication, radar and sonar systems, signal processing, and deep learning.
Email: doansang.g1@gmail.com.

NHẬN DẠNG TÍN HIỆU THỦY ÂM DỰA TRÊN SỰ KẾT HỢP CỦA MẠNG NƠ-RON TÍCH CHẬP ĐA KÍCH THƯỚC VÀ PHÉP BIẾN ĐỔI HẰNG SỐ Q

Trần Công Tráng, Nguyễn Văn Lâm, Trần Xuân Sùng, Lê Thị Huyền, Đoàn Văn Sáng

Tóm tắt

Bài báo đề xuất một mạng nơ-ron học sâu đa kích thước để phân loại các nguồn tín hiệu thủy âm khác nhau. Mạng đề xuất được thiết kế khéo léo với nhiều nhánh kết nối tạo ra một khối tích chập đa kích thước, cho phép mô hình học được các đặc trưng không gian khác nhau của biểu đồ phổ CQT (Constant-Q Transform). Mạng được huấn luyện và thử nghiệm trên tập dữ liệu ShipsEAR. Để đảm bảo tính cân bằng giữa dữ liệu của các lớp nhãn mục tiêu, tệp này được củng cố bằng kỹ thuật phân đoạn chồng lấn. Kết quả thử nghiệm cho thấy mạng đề xuất đạt được độ chính xác phân loại trung bình lên tới 99,93% và tốc độ thực thi là 2,2 ms khi được chỉ định cấu hình với hai khối tích chập đa kích thước và 32 kênh lọc. Khi được so sánh, mạng đề xuất cho thấy hiệu năng vượt trội hơn về độ chính xác và thời gian thực thi so với một số mạng hiện có khác.

Từ khóa

Mạng nơ-ron sâu, biến đổi hằng số Q, phân loại tín hiệu thuỷ âm, đặc tính không gian.