Recognition of Underwater Acoustic Target Using Sub-pretrained Convolutional Neural Networks

Andi Pan[1], Xi Chen[1], and Wei Li[1,2(✉)]

[1] School of Computer Science, Fudan University, Shanghai 201203, China
weili-fudan@fudan.edu.cn

[2] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 201203, China

Abstract. Underwater acoustic target recognition is the task of classifying targets using ship-radiated noise in the marine environment. It is incredibly hard and complex for the complexity of the marine environment. Before the popularization of deep learning, conventional target recognition methods are mainly based on the audio time-frequency domain analysis. Different targets have obvious variation in some frequency bands, which leads to the inability of traditional methods to make full use of spectral information. In order to extremely extract the information in each frequency bands, this paper proposes a novel Sub-pretrained CNNs. For each frequency band in the spectrogram, a CNN classifier is trained on the training set. Finally, the features extracted by each CNN and the position embedding of the frequency band are concatenated as the input of the global classifier. Compare with state of the art method, the paper achieves better performance. As the experimental results show, the identification performance of UATR can be enhanced by the Sub-pre-trained CNNs method.

Keywords: Convolutional Neural Networks · Audio classification · Underwater acoustic target recognition · Pre-training

1 Introduction

Underwater acoustic target recognition is the task of classifying targets using ship-radiated noise in the marine environment. It is widely used for marine exploration, marine biological surveys, and other research activities. It is incredibly hard and complex for the complexity of the marine environment and the diversity of underwater acoustic targets [1,2].

At present, various UATR methods based on machine learning have been put forward. Commonly, we separate these methods into two kinds: approaches based on artificial feature design and approaches based on automatic feature extraction [1–4]. In general, the most effective method of UATR is based on the characteristics of domain knowledge design, which heavily depends on the

X. Shao et al. (Eds.): CSMT 2020, LNEE 761, pp. 113–123, 2021.
https://doi.org/10.1007/978-981-16-1649-5_10

statistical model [1–3]. MFCC is a widely adopted feature in UATR and speech recognition [3–13]. Nevertheless, the optimal feature of the acoustic target can not be represented by MFCC [8]. To solve the shortcomings of MFCC, other features have been presented. The GFCC was introduced into UATR by Lian [9]. The crux in the process is how to extract the features of underwater acoustic targets.

In recent years, as the solution based on deep learning has made great successes in the field of speech recognition and image classification, people have carried out in-depth research on improving the ability of underwater acoustic target recognition. [14–19] in these studies, the solution based on deep learning shows a strong ability to feature extraction. Compared with the shallow neural network, the deep neural network can extract more abstract and higher-level features from big data [21]. As one of the methods based on deep structure, Deep Boltzmann Machine has better performance in learning and extracting the features of ship radiated noise. Additionally, CNNs [23] are widely used in UATR because of its advantage in processing images [24]. In [25], Yang et al. used ADCNN to simulate the auditory system. Deep learning based methods can extract more information compared with hand-engineering methods.

This paper proposed an Sub-pretrained CNNs based method which combines multi-dimensional feature extracted by CNNs with the position encoding, as the input of the global classifier using fully connected DNN. Firstly, we translate original signals to time-frequency presentations as images. Then, we transform the position of bands in the spectrogram to position encoding. After we concat position encoding and multi-dimensional feature extracted by CNNs, global classifier can recognition underwater targets using the input.

In the second section, the UATR method presented is introduced detailedly. The specific content of the experimental setting is introduced in The third part. The experimental results are addressed in the fourth section. The fifth part summarizes the full paper.

2 Proposed Method

2.1 Framework

For most UATR methods, the process can be divided into feature extraction stage and learning stage. The purpose of CNN is to adopt a deep hidden structure in the perceived signal to produce a great feature presentation. The process of the presented approach for UATR is presented in Fig. 1. As preprocess, we practice STFT to get time-frequency representations of the original signals. Next, we simply utilize each band of time-frequency representation to train each CNN model in the training dataset and train some sub-pre-trained CNNs. The outputs of the last layers of these CNNs can be considered as presentation of the band. Then, we concat vectors as just one vector. Finally, we take the vector as the input of global classifier, which will recognize the target.

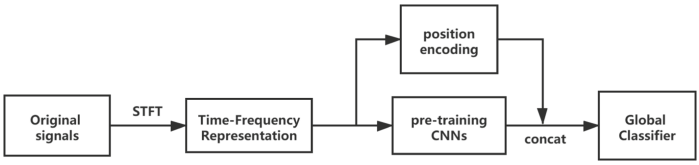

Fig. 1. The process of the presented UATR

2.2 Sub-pretrained CNNs

Spectrograms are 2D representations like an image comprising time and frequency dimensions, although very distinct from the original images. There exists an obvious diversification during the frequency dimension. As shown in Fig. 2, in the spectrograms obtained, we observed a clear variation of the magnitude of different frequency bands, particularly specific to every kind of target. For instance, the "B" class owns more extra power in higher frequency bins; the "C" class has more energy in mid-frequency bins and less energy in higher frequency bins; for "E" class Background noise recordings, energy is well-distributed in frequency bands. We utilize these observations to put forward Sub-pre-trained CNNs, which is talked about in the accompanying.

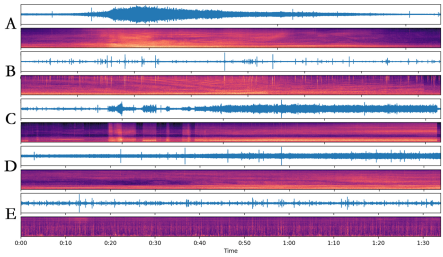

Fig. 2. Time-frequency presentation

To extremely extract the information in each frequency band and fully take advantage of variation of the magnitude of different frequency bands, we propose the Sub-pretrained CNNs method. The process of this method can be illustrated in Fig. 3. Firstly, we extract the spectrogram for the N samples and perform normalization. Then we split the spectrogram into several bands. It takes spectrogram to F*T dimension, bands size is the number of bands. These bands are independently inputted into 2 conv-layers. Kernel-size is set (5, 5), which has large receptive field. After conv-layer, sigmoid activation and max-pooling follow. Then, we flatten the output of CNNs and concat these vectors as just one vector. Finally, to capture the global relations between frequency bands, we use MLP as classifier to classify the input using diversified information.

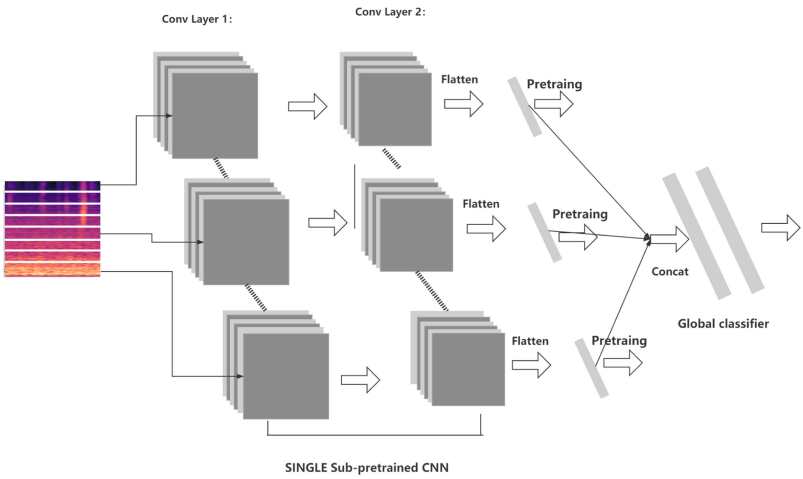

Fig. 3. Sub-pretrained CNNs

2.3 Position Encoding

Position and order of bands are the essential parts of any spectrogram. They define the high and low frequency and thus the actual characteristics of an acoustic target. Convolutional Neural Networks (CNNs) rarely take the order of bands into account. They parse a spectrogram band by band in a sequential manner. This will integrate the bands' order.

This paper use the position encoding method proposed in Transfomer [26], which is a simple yet efficient tool. Firstly, it is not just a number. Instead, it's a d-dimensional vector that incorporates information about a specific position in a spectrogram. Secondly, this vector is not integrated into the classifier itself. Instead, this vector is used to equip each word with information about its position in a spectrogram. Basically, we enhance the classifier's input to inject the order of bands.

$$PE(pos) = sin(\frac{pos}{length}) \tag{1}$$

2.4 Classifier

We test SVM, Decision Tree and MLP as classifier. The performance of classifiers are shown in Sect. 4. The methods are implemented with scikit tools. The principles of these algorithms are introduced as follows.

Support Vector Machine. SVM [27] is a very classical and commonly used model. Because it has very good classification ability and strong interpretability, it has a good effect on small samples. For linearly separable data, linear support vector machine strives to find a segmentation line to maximize the distance

between positive and negative samples. When the data is approximately separable but not completely separable and not completely separable, there are a small number of abnormal samples. Using soft margin maximization, we can fit a classifier that basically separates the samples but can not completely separate them. When the data set can not be divided by the interval represented by the linear function, someone put forward the kernel function to convert the original data space where the training set samples exist toward a higher dimensional feature space, formerly the data set converts separable. In order to train a nonlinear classifier, the principle is shown in the figure. The common kernel functions are Gaussian kernel and so on.

Decision Tree. Decision Tree [28] is a model that accords with human judgment intuition and has strong explanation. After abstraction, the decision tree model is generally more like a tree, so it is named decision tree. As shown in Fig. 4, the segmentation part of the branches in this structure is to select a feature in the sample features to segment the data set. The decision book belongs to supervised learning.

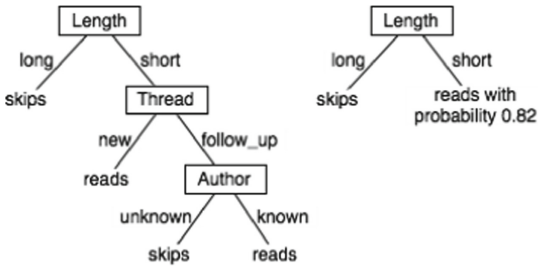

Fig. 4. Two decision trees.

Multi-layer Perceptron. Perceptron (Perceptron) is the origin of deep learning. Through the weight w and the offset term b, it can map a multi-dimensional input X to a binary value, through which a simple binary classification can be achieved. Multilayer perceptrons are in the form of multiple functions. As shown in Fig. 5, the multilayer perceptron is the superimposed multiple function of the function represented by the perceptron, which is divided into input, output, concealment and multiple perceptrons according to function and position. At the same time, if each unit of the multilayer perceptron is linear, then any multilayer perceptron can be equivalent to a single layer perceptron. Therefore, the multilayer perceptron is essentially the superposition of multiple nonlinear functions. Finally, the model is used to measure the fitting degree of the training set, and the variables in the model are taken as the loss function of the parameters. Through the back propagation algorithm, a multi-layer perceptron can be fitted on the training set.

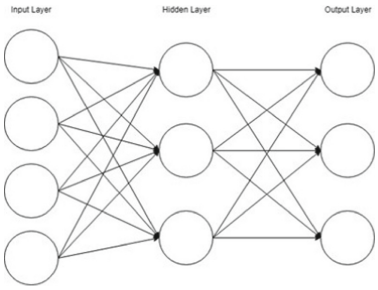

Fig. 5. Structure of Multi-layer Perceptron.

3 Experiments Setup

3.1 Experimental Datasets

The ship target dataset used in this paper is the ShipsEar [30] dataset recorded in different regions of the Spanish coast from 2012 to 2013. The dataset has a total of 90 records of 11 ship types within 15 s to 10 min. According to the original labels of the dataset, they can be merged into 4 large groups in accordance with the type of ship. Class and E class: background noise recordings, The detailed division is shown in Table 1 below:

Table 1. ShipEar dataset details.

A	Fishing boats.Trawlers.Mussel boats.Tugboats.Drafgers
B	Motorboats.Pilot boats.Sailboats
C	Passenger ferries
D	Ocean liner.Ro-Ro vessels
E	Background noise recordings

3.2 Training Setup

We choose 52,734 Hz as the target audio signal sampling rate, and a 90 ms Hamming window as windowing function with a 50% overlap is used. The output Mel spectrum is stored in a $3 \times 224 \times 224$ image format for subsequent operations. In addition, we downsampled the experimental audio data. The window length is 25 ms, the overlap length (Hop size) is 10 ms, the output spectrum is 96×64, and the embedding code size is 128.

We implement Sub-pretrained CNNs in Pytorch. Most experiments have been carried out with sklearn [31].

3.3 Evaluation Indexes

We compare the predicted results of the model with the labels to obtain the number of TP, FP, TN, and FN in the evaluation. And for each experimental result, the accuracy rate, recall rate, and F1 function are calculated separately to measure the experimental results comprehensively and accurately. These indicators can be expressed by the following formula:

$$Accuracy = \frac{TP}{TP + FP + TN + FN} \tag{2}$$

4 Experiments Results

4.1 Bands Size Setting

To find optimal Bands size, the experiment was designed. We set the optional values of band size to 10, 20, 30, 40. In the contrast experiment, the classification accuracy reaches the highest when band size equal to 20. Therefor, we set the band size to 20 in the following experiments.

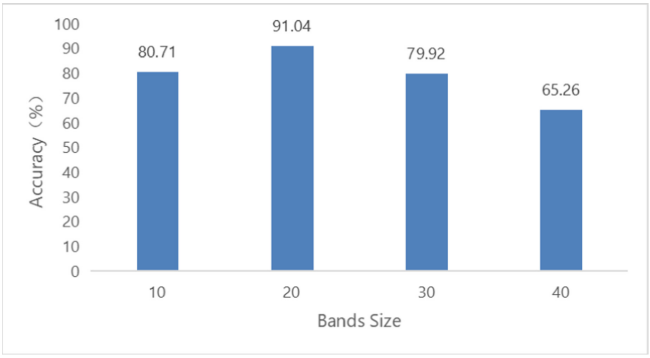

Fig. 6. The recognition accuracy with different bands size

4.2 Evaluation of Position Encoding

To illustrate the importance of position encoding, classification performance of MLP with encoding and without are measured using the classification accuracy. The comparison between MLP with position encoding and MLP without position is shown in Fig. 7. It is clear that position encoding can introduce more structure information in spectrogram, which contributes to improving the performance.

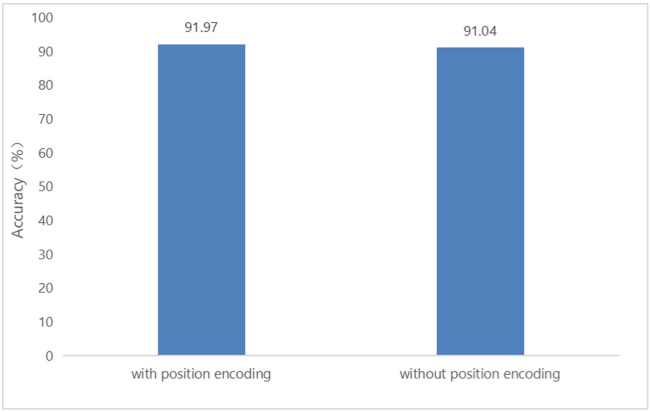

Fig. 7. The comparison between MLP with position encoding and MLP without position encoding

4.3 A Comparison of Three Kind of Classifiers

To find the optimal classifier, we compare three kinds of classifiers. As illustrated in Fig. 8, MLP classifier has the highest accuracy over Decision Tree and SVM. In contrast recognition, it is clear that the MLP classifier is more suitable for underwater target recognition. We speculate that this might be due to the advantage of MLP in classification.

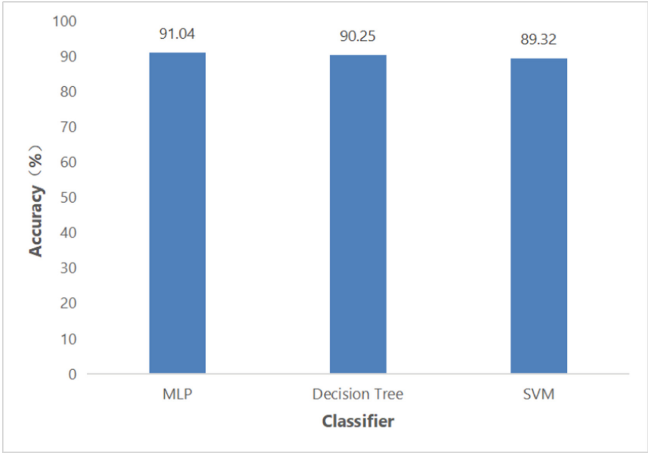

Fig. 8. The recognition accuracy with different bands size

4.4 Evaluation of Sub-pretrained CNNs with Position Encoding

Considering that methods with position encoding can achieve better performance than without, we train a Sub-pretrained CNNs with position encoding and using MLP as the classifier. As a result, accuracy is 91.97%. This best result shows in the confusion matrix. Table 3 shows the confusion matrix of the proposed UATR methods obtained from testing data. Compare with state of the art method, the paper achieves better performance (Table 2).

Table 2. Comparison of performance between Pretrained CNNs and DBM based.

Method	Accuracy
Sub-pretrained CNNs with position encoding	91.97%
DBM [22]	90.70%
VGGISH [32]	89.22%

Table 3. Confusion matrix of the proposed model.

True predicted	A	B	C	D	E
A	0.92	0.01	0.02	0.01	0.00
B	0.01	0.85	0.03	0.02	0.00
C	0.02	0.00	0.93	0.01	0.00
D	0.01	0.03	0.00	0.92	0.01
E	0.00	0.01	0.00	0.00	0.98

5 Conclusions

In the work, a new UATR algorithm based on regional pre-training convolution neural network is introduced, in order to fully extract the information contained in different frequency bands in the spectrum. The output of the last hidden layer of each sub-network is spliced and connected with the position vector as the input of the total classifier, and then the general classifier is trained. Compare with state of the normal training convolution neural network model, the proposed UATR algorithm achieves better performance, the sub-pre-trained CNN is introduced to learn more information, and the classification accuracy is 91.97%. This method proposes an innovative model training method, which can be effectively applied to UATR tasks, also give inspiration to other similar tasks.

Acknowledgement. This work was supported by National Key R&D Program of China (2019YFC1711800) and NSFC (61671156).

References

1. Yang, H., Shen, S., Yao, X., Sheng, M., Wang, C.: Competitive deep-belief networks for underwater acoustic target recognition. Sensors **18**, 952 (2018)
2. Wang, X., Jiao, J., Yin, J., Zhao, W., Han, X., Sun, B.: Underwater sonar image classification using adaptive weights convolutional neural network. Appl. Acoust. **146**, 145–154 (2018)
3. Wang, W., Li, S., Yang, J., Liu, Z., Zhou, W.: Feature extraction of underwater target in auditory sensation area based on MFCC. In: 2016 IEEE/OES China Ocean Acoustics (COA), pp. 1–6. IEEE (2016)
4. Yue, H., Zhang, L., Wang, D., Wang, Y., Lu, Z.: The classification of underwater acoustic targets based on deep learning methods. In: 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017), pp. 526–529. Atlantis Press (2017)
5. Lu, Z., Zhang, X., Zhu, J.: Feature extraction of ship-radiated noise based on mel frequency cepstrum coefficients. Ship Sci. Technol. **26**(2), 51–54 (2004)
6. Ke, X., Yuan, F., Cheng, E.: Underwater acoustic target recognition based on supervised feature-separation algorithm. Sensors **18**(12), 4318 (2018)
7. Zhang, L., Wu, D., Han, X., Zhu, Z.: Feature extraction of underwater target signal using mel frequency cepstrum coefficients based on acoustic vector sensor. J. Sens. **2016**, 1–11 (2016)
8. Sharma, R., Vignolo, L., Schlotthauer, G., Colominas, M., Rufiffiner, H.L., Prasanna, S.: Empirical mode decomposition for adaptive AM-FM analysis of speech: a review. Speech Commun. **88**, 39–64 (2017)
9. Lian, Z., Xu, K., Wan, J., Li, G.: Underwater acoustic target classification based on modified GFCC features. In: Proceedings of the IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 25–26 March 2017, pp. 258–262 (2017)
10. Lim, T., Bae, K., Hwang, C., Lee, H.: Underwater transient signal classification using binary pattern image of MFCC and neural network. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E91A**, 772–774 (2008)
11. Jankowski Jr., C., Quatieri, T., Reynolds, D.: Measuring fine structure in speech: Application to speaker identification. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995, pp. 325–328. IEEE, Piscataway (1995)
12. Guo, Y., Gas, B.: Underwater transient and non transient signals classification using predictive neural networks. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009, pp. 2283–2288 (2009)
13. Hu, G., Wang, K., Peng, Y., Qiu, M., Shi, J., Liu, L.: Deep learning methods for underwater target feature extraction and recognition. Comput. Intell. Neurosci. **2018**, 1214301 (2018)
14. Jiang, Y., Wang, D.L., Liu, R.S., Feng, Z.M.: Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2112–2121 (2014)
15. Lee, H., Yan, L., Pham, P., Ng, A.Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), vol. 9, pp. 1096–1104, December 2009

16. Jaitly, N., Hinton, G.: Learning a better representation of speech soundwaves using restricted Boltzmann machines. In: Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2011), pp. 5884–5887, May 2011
17. Palaz, D., Collobert, R., Magimai-Doss, M.: Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH 2013, pp. 1766–1770, August 2013
18. Huang, G., Huang, G.-B., Song, S., You, K.: Trends in extreme learning machines: a review. Neural Netw. **61**, 32–48 (2015)
19. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014)
20. Bisot, V., Serizel, R., Essid, S., et al.: Acoustic scene classification with matrix factorization for unsupervised feature learning. In: IEEE International Conference on Acoustics. IEEE (2016)
21. Kamal, S., Mohammed, S.K., Pillai, P.R.S., Supriya, M.H.: Deep learning architectures for underwater target recognition. In: Proceedings of Ocean Electronics (SYMPOL), October 2013, pp. 48–54 (2013)
22. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
23. Deng, L., Abdel-Hamid, O., Yu, D.: A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013, pp. 6669–6673 (2013)
24. Swietojanski, P., Ghoshal, A., Renals, S.: Convolutional neural networks for distant speech recognition. IEEE Signal Process. Lett. **21**(9), 1120–1124 (2014)
25. Yang, H., Li, J., Shen, S., Xu, G.: A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition. Sensors **19**, 1104 (2019)
26. Ott, M., Edunov, S., Baevski, A., et al.: FAIRSEQ: a Fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North (2019)
27. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
28. Zhou, B., Cao, C., Li, C., et al.: Hybrid islanding detection method based on decision tree and positive feedback for distributed generations. IET Gen. Transm. Distrib. **9**, 1819–1825 (2015)
29. Yue, H., Zhang, L., Wang, D., Wang, Y., Lu, Z.: The classification of underwater acoustic targets based on deep learning methods. Adv. Intell. Syst. Res. **134**, 526–529 (2017)
30. Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., et al.: Shipsear: an underwater vessel noise database. Appl. Acoust. **113**, 64–69 (2016)
31. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
32. Deng, J., Pan, A., Xiao, C., Chen, S.: Transfer learning for acoustic target recognition. Comput. Syst. Appl. **29**(10), 255–261 (2020)