Underwater Acoustic Target Recognition with ResNet18 on ShipsEar Dataset

Conference Paper · May 2021

DOI: 10.1109/ICET51757.2021.9451099

Feng Hong
Institute of Acoustics Chinese Academy of Sciences

Underwater Acoustic Target Recognition with ResNet18 on ShipsEar Dataset

Feng Hong
Shanghai Acoustics Laboratory,
Chinese Academy of Sciences
Shanghai, China
hongfeng@mail.ioa.ac.cn

Chengwei Liu
Shanghai Acoustics Laboratory,
Chinese Academy of Sciences
Shanghai, China
liuchengwei19@mails.ucas.ac.cn

Lijuan Guo
College of Electronic and
Electrical Engineering
Shanghai University of
Engineering Science
Shanghai, China
kerryglj@163.com

Feng Chen
Shanghai Acoustics Laboratory,
Chinese Academy of Sciences
Shanghai, China
chenfeng@mail.ioa.ac.cn

Haihong Feng
Shanghai Acoustics Laboratory,
Chinese Academy of Sciences
Shanghai, China
fhh@mail.ioa.ac.cn

***Abstract*—Underwater Acoustic Target Recognition (UATR) remains one of the most challenging tasks in underwater signal processing due to the lack of labeled data acquisition, the impact of the time-space varying intrinsic characteristics, and the interference from other noise sources. To achieve state-of-the-art accuracy, we propose a novel classification method by using the fusion features and a 16-layer Residual Network (ResNet18). The recognition experiments are conducted on the ship-radiated noise dataset named ShipsEar from a real environment, and the accuracy results of 0.943 show that the proposed method is effective for underwater acoustic recognition problems and outperforms other classification methods.**

Keywords*—ResNet, Underwater acoustics, ShipsEar, Embedding, SpecAugment, UATR, MFCC, combined features

I. Introduction

As a key technology to promote the intelligence of the underwater acoustic equipment system, underwater acoustic target radiated noise recognition is one of the most important complex and changing research directions of underwater acoustic signal processing [1-3]. Developing a robust recognizing system to replace humans' work of identifying ship-radiated noise is of great importance. From a technical perspective, efforts are consistently paid to improve the classification performance in the aspects of feature extraction and classifier training [4-5].

Extracting hand-crafted features from ship-radiated noise and then feeding them into different kinds of classifiers is generally used. On one hand, as for the traditional machine learning feature extraction process, Support Vector Machines (SVM) [6-7]and Principal Component Analysis (PCA) [8] methods are widely used. Features derived from Mel filters of Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel Spectrogram (LM) are two widely used features in Environment Sound Classification (ESC) tasks [9,10] with acceptable performance. Besides, a considerable number of research works indicate that the fusion feature can give a more comprehensive representation of environment sounds [11]. On the other hand, the design of the neural networks plays an important role in achieving a competitive performance together with the optimized feature extraction. For example, a time-delay neural network (TDNN) and convolutional neural network (CNN) are introduced for UATR in [12] respectively. Li et al. [5] introduce a feature optimization approach with Deep Neural Networks (DNN) and optimizing loss function and achieve an accuracy of 84%. Yang et al. [13] propose a so-called competitive Deep Belief Nets (cDBN) for UATR. Luo et al. [14] present a UATR method based on Restricted Boltzmann Machine (RBM), which achieves the accuracy of 93.17% on the dataset of ShipsEar [15]. Ke et al. [4] propose a novel recognition method of four steps including pre-processing, pre-training, fine-tuning, and recognition, which achieves the recognition accuracy of 93.28%.

In this paper, we will introduce the three-dimensional fusion features along with the data augment strategy of SpecAugment and a 18 layer Residual Network (ResNet18) containing the embedding layer to achieve good accuracy.

II. Description of the Classification Method

As depicted in Fig. 1, the classification method mainly contains several steps, i.e., preprocessing, feature extraction, Residual Network training, and embedding layer design. After preprocessing, we perform the feature extraction process to obtain the three-dimensional feature. Afterward, we design a 18 layer Residual Network named ResNet18, which could learn very informative presentations in the training process. To avoid overfitting, we adopt the tactic strategy to train the aggregated features with early stopping and adaptable learning rate. Besides, aiming to increase the distance for inter-class and decrease that for Intra-Class, the embedding layers are well-tuned with the designed center loss function.

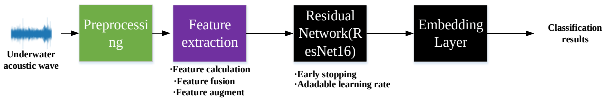


Fig. 1. The depiction of the classification method.

A. *Preprocessing*

Preprocessing is a prerequisite before feeding the signals of different lengths to the designed networks. One feasible way is to split the underwater acoustic signal into several frames of fixed length using a sliding designed window of appropriate width. Considering that an underwater acoustic signal can be seemed as stable in a very short time, the framing of the recorded signal could be performed. This naturally increases the number of samples as some parts of the underwater acoustic signal are reused and that can be viewed as some sort of data augmentation [20]. For underwater acoustic, a sampling rate of 20,480 Hz may be considered a good trade-off between the quality of the input sample and the computational cost of the model.

B. *Feature extraction process*

The procedure for extracting the final feature contains several steps, i.e., feature calculation, feature fusion, and feature augment, as illustrated in Fig. 2. Considering that different features can capture different characteristics of the underwater acoustic signal, therefore, multiple features can be combined to exploit the complementary information. At the stage of feature calculation and feature fusion, we extract the LM as the first channel, the MFCC as the second channel, and the composition of Chroma, Contrast, Tonnetz, and Zero-cross ratio called CCTZ as the third channel. The first two channels both have the shape of (60,41), where 60 denotes the number of the bands and 41 denotes the number of frames, respectively. As for the third channel, the shape of CCTZ are (24,41), (6,41), (6,41), and (1,41) with the remaining lines paddled by zero. Afterward, at the stage of feature augment, we apply the feature augment method of SpecAugment [19] on the LM feature five times, leaving the other channels unchanged. The resulting three-dimensional feature matrix is composed of the three channels as mentioned.

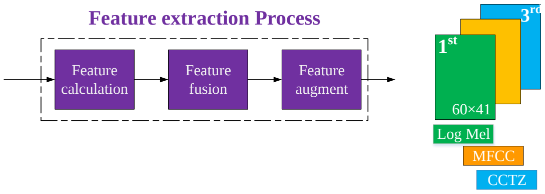


Fig. 2. The feature extraction process.

C. *Structure of the ResNet18*

The number of convolutional layers plays a key role in detecting high-level concepts [20]. Residual learning framework could decrease the difficulty of the training of deep neural networks with many layers [18]. Here, ResNet18 as a modified version of the general ResNet50 is shown in Fig. 3. The details are listed as follows:

- ***Stage 0:*** The input layer is zero-paddled with a shape of 3 × 3. The shape changes from (60×41×3) to (66×47×3).
- ***Stage 1:*** The first stage consists of a convolutional layer of 64 filters of the rectangular shape of 3 × 3 and stride of 2 × 2. The batch normalization layer is applied, followed by a ReLU activation function and max-pooling with the shape of 3 × 3 and stride of 2 × 2.
- ***Stage 2:*** The second stage consists of a convolutional block named block-2a, an identity block named block-2b, and an identity block named block-2c. The structure of the convolutional block is shown in Fig. 3. Compared with that, the identity block lacks the convolution layer marked with the dashed red box, respectively.
- ***Stage 3:*** The third stage consists of a convolutional block named block-3a and an identity block named block-3b.
- ***Average pooling and flatten layer:*** The average pooling layer is applied with the shape of 2 × 2. The shape of the output of the flatten layer changes from (15×11×512) to (7×5×512). The flatten layer

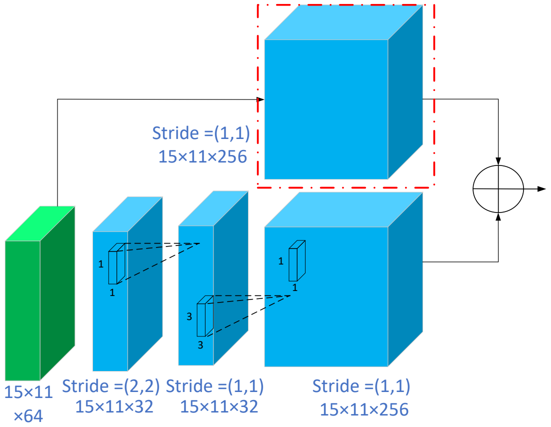


Fig. 3. The convolutional block.

As for the identity block, non-linear activation helps the designed networks learn more complex features. We take the PreLU as the activation function which contains a learnable coefficient controlling the slope of the negative part.

III. Experiments and Analysis

A. *Dataset description and preparation*

Before The detailed description of the dataset of the ship-radiated noise called ShipsEar (available at http://atlanttic.uvigo.es/underwaternoise/), which contains a total of 91 records of 11 vessel types and one background noise class, is presented in [15]. To keep consistent with other classification methods, the 11 vessel types are merged into four experiment classes and the background noise type is as one class, as shown in Tab. 1.

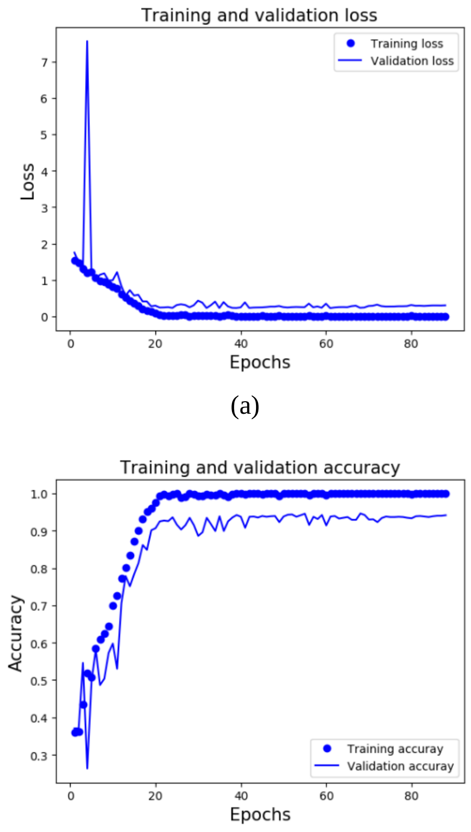
TABLE 1. A DETAILED DESCRIPTION OF THE FIVE CLASSES.

Class A	Class B	Class C	Class D	Class E
background noise	fishing boats, trawlers, mussel boats, tugboats, and the dredger	motorboats, pilot boats, and sailboats	passenger ferries	ocean liners and ro-ro vessels

Identify applicable funding agency here. If none, delete this text box.

B. *Experimental Result*

The proposed method is verified by a computer with four GPUs of Nvidia GeForce RTX 2080Ti and Core i7-6900K CPU. The batch size and the maximum number of epochs are set to be 128 and 200, respectively. To accelerate the training process, the early stopping strategy is that the training will be stopped if the validation loss is reduced by larger than 0.00005 in 20 successive epochs. Besides, the adaptable strategy of learning rate is adopted, where the initial value is 0.001 and the value will be 60% of the former value every 20 epochs. By using such a strategy, the practical number of epochs used is 88. As shown in Fig. 4(a), the training loss and validation loss are rapidly reduced within about 20 epochs and they are gradually decreased without overfitting due to the fact of the design of the adaptable learning rate. Meanwhile, Fig. 4(b) also shows that the accuracy is improved at a varied speed. Only small improvements are obtained after the turn and the best accuracy of the validation data is 0.946. We can observe that Figs. 4(a) and 4(b) describe the same process.



(a)

(b)

Fig. 4. The training and validation accuracy and loss when training the model: (a) The loss; (b) The accuracy.

Here, we evaluate our model of ResNet18 with the three-dimensional feature. The split proportional of the train set, validation set, and the test set is 0.7, 0.2, and 0.1, respectively. Specifically, each network is trained with 0.7 of the dataset with the augmentation. Tab. 2 depicts the detailed results of the classification system at the aspect of precision, recall, and f1-score. It is clear that all the accuracy of the recognition of each class is higher than 0.900 and the average precision or recall or f1-score is 0.941, where the support denotes the number of each test class. For the convenience of comparison, classifier performance is measured using the classification accuracy, defined as the average precision. The ability of the described classifier to identify different vessels is indicated by the fact that there is no confusion between background noise class E and four vessel classes A–D [15]. The vessel classes with the best results are A (background noise) and B (fishing boats, trawlers, mussel boats, tugboats, and the dredger), with classification rates of 0.970 and 0.958, respectively. The poorest results are obtained for C (motorboats, pilot boats, and sailboats). Although the acoustic dataset contains high background noise in shallow-water, the overall performance is still satisfactory.

TABLE 2. THE RESULTS OF RESNET18 WITH THE THREE-DIMENSIONAL FEATURE.

Class	Precision	Recall	F1-score	Support
class A	0.970	0.958	0.964	166
class B	0.958	0.936	0.947	220
class C	0.917	0.914	0.915	243
class D	0.945	0.956	0.950	501
class E	0.935	0.941	0.938	337
average	0.943	0.943	0.943	1467

The baseline design [15] shows that using the basic machine learning method the accuracy is 0.754. Besides, the accuracy achieved by ResNet18 model as well as that achieved by other state-of-the-art approaches of RBM + BP [14] and RSSD+ [4] described in the literature are presented in Tab. 3. Our method achieves an accuracy of 0.943. To our knowledge, The best results are obtained by the proposed method for ShipsEar.

TABLE 3. THE COMPARISON OF RECOGNITION ACCURACY WITH OTHER MODELS ON SHIPSEAR

Model	Accuracy
Baseline [15]	0.754
RBM + BP [14]	0.932
RSSD [4]	0.933
ResNet18+3D	0.943

IV. CONCLUSIONS

We designed a Residual Network called ResNet18 and the optimizing feature extraction method. The results of 0.943 on the ShipsEar dataset indicated that the proposed method achieved state-of-the-art accuracy. The proposed method provided good technical support for the target classification and recognition function of the Sonar system.

ACKNOWLEDGMENT

This work reported herein was funded jointly by the National Natural Science Foundation of China for Young Scholar (Grant No. 61801471), National Key R&D Program of China (Grant No. 2016YFC0302000), Youth Innovation Promotion Association CAS, the development fund for Shanghai talents, and Jiading Youth Talents Program.

REFERENCES

[1] C. Chen and J. Fu, "Recognition method of underwater target radiated noise based on convolutional neural network", Acoustic technology, vol. 38, no. 2, pp. 424, 2019.

[2] A. Pezeshki, M. R. Azimi-Sadjadi, L. L. Scharf and M. Robinson, "Underwater target classification using canonical correlations", Proc. Oceans Celebrating Past Teaming Toward Future, pp. 1906-1911, Sep. 2003.

[3] A. Pezeshki, M. R. Azimi-Sadjadi and L. L. Scharf, "Undersea target classification using canonical correlation analysis", IEEE J. Ocean. Eng., vol. 32, pp. 948-955, 2007.

[4] X. Ke, F. Yuan, and E. Cheng. "Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm", Sensors 2018, 12: 4318.

[5] C. Li, Z. Liu, J. Ren. "A Feature Optimization Approach Based on Inter-Class and Intra-Class Distance for Ship Type Classification", Sensors 2020, 20: 5429.

[6] Q. Meng, S. Yang, S. Piao. "The classification of underwater acoustic target signals based on wave structure and support vector machine", J. Acoust. Soc. Am. 2014, 136, 2265.

[7] L. Jian, H.Yang, L. Zhong. “Underwater target recognition based on line spectrum and support vector machine”. In Proceedings of the International Conference on Mechatronics, Control and Electronic Engineering (MCE2014), Shenyang, China, 29–31 August 2014; Atlantis Press: Paris, France, 2014; pp. 79–84.

[8] L. Zhang, D. Wu, X. Han, Z. Zhu. “Feature extraction of underwater target signal using Mel frequency cepstrum coefficients based on acoustic vector”. J. Sens. 2016, 7864213.

[9] Z. Zhang, S. Xu, S. Cao, S. Zhang. “Deep Convolutional Neural Network with Mixup for Environmental Sound Classification”. arXiv 2018, arXiv:1808.08405.

[10] J. Li, W. Dai, F. Metze. “A comparison of deep learning methods for environmental sound detection”. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 126–130.

[11] Y. Su,; K. Zhang, J. Wang. “ Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion”. Sensors 2019, 19: 1733.

[12] G. Hu, K. Wang, Y. Peng. “Deep learning methods for underwater target feature extraction and recognition”. Comput. Intell. Neurosci. 2018, 2018, 1214301.

[13] H. Yang, S. Shen, X. Yao. “Competitive Deep-Belief Networks for Underwater Acoustic Target Recognition”. Sensors 2018, 18, 952.

[14] X. Luo, Y. Feng. “An Underwater Acoustic Target Recognition Method Based on Restricted Boltzmann Machine”. Sensors 2020, 20: 5399.

[15] D. Santos-Domínguez, et al. "ShipsEar: An underwater vessel noise database." Applied Acoustics 113(2016):64-69.

[16] Y. Wen, K. Zhang, Z. Li. “A discriminative feature learning approach for deep face recognition‟. In ECCV; Springer: Cham, Switzerland, 2016; pp. 499–515.

[17] H. Yang, S. Shen, X. Yao, M. Sheng, C. Wang, “Competitive deep-belief networks for underwater acoustictarget recognition”. Sensors 2018, 18: 952.

[18] K. He, X. Zhang, S. Renl. “Deep Residual Learning for Image Recognition”. IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.

[19] D. S. Park , W. Chan, Y. Zhang. “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”. Interspeech 2019.

[20] S. Abdoli, P. Cardinal, A. L. Koerich. “End-to-end environmental sound classification using a 1D convolutional neural network”. Expert Systems with Applications, 136, 252-263.

AUTHORS’ BACKGROUND

1. This form helps us to understand your paper better, the form itself will not be published.

2. Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor

Your Name	Position	Research Field	Personal Webpage
Feng Hong	associate professor	underwater acoustic target recognition	none