PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Passive Ship Localization in a Shallow Water Using Pre-trained Deep Learning Networks

Dezhi WANG[1*]; Lilun ZHANG[1]; Changchun BAO[1]; Shuqing MA[1]; Yongxian WANG[1];

[1] College of Meteorology and Oceanography,

National University of Defense Technology, China

ABSTRACT

Subject to the lack of detailed environmental information, the classical matched-field processing (MFP) may not be adapted to the accurate localization of underwater acoustic sources. In this paper, a framework that applies deep learning techniques instead of the MFP method is presented for the localization (direction-finding) of ship acoustic sources in a shallow water environment. The original data is recorded from a 128-element vertical array placed in a shallow water. The acquired array data is first processed by the time-domain conventional beamformer (CBF) in order to obtain the beamformed waveform signals corresponding to each direction-of-arrival (DOA) with a resolution of 1 degree. In the meantime, the GPS and recognized DOA information in the diagram of the target ships are employed to generate the labels for these beamformed signals. Base on the labeled data, a framework is proposed to predict the DOA information of target ships in a deep learning (DL) manner using the pre-trained state-of-the-art convolutional neural networks. Driven directly by the array signal data, the proposed method offers a way for the ship localization to overcome the environmental mismatch problem, which is believed to be better than that of conventional MFP method and some other shallow machine learning methods.

Keywords: Ship Localization, Deep Learning, Array Signal Processing

1. INTRODUCTION

Passive sonar acoustic signal processing methods are an established technique to remote detect, localize and monitor marine vessels and their activities. Though a series of related methods have been developed during the past several decades, poor performance on sound-source localization still persists in a complex marine environment, especially when multi-sources exist in the scenario and the signal-to-noise-ratio (SNR) is low. In practical applications, the acoustic characteristics of a shallow water environment are varying in both space and time with high levels of background noise, interferences and multipath effects. The performance of the conventional methods like TDOA estimation (1), cepstral analysis (2) and autocorrelation analysis (3) in these scenarios significantly degrades and cannot effectively detect or localize targets.

Apart from these classical methods, recently data-driven machine learning or deep learning methods become more and more popular for the acoustic source localization, such as range and depth discrimination simulation (4) and seafloor classification (5). The machine learning method is shown to perform significantly better than the conventional matched field processing in ship range estimation in the Santa Barbara Channel Experiment with limited environmental information (6). After achieving state-of-the-art results for image recognition (7), convolutional neural networks (CNNs) become very popular as an automated feature extractor and classifier (8, 9), which combines hierarchical feature extraction and classification at the same time. The CNNs are also applied for the joint detection and ranging of marine vessels (10), which is shown to be able to detect the presence and estimate the range of transiting vessels at greater distances than the conventional method.

In this paper, we aim to develop a scalable system based on the well-developed deep-learning neural networks to predict the DOA information of ships in a shallow water environment using the multi-element hydrophone array data. The main contributions of our work can be summarized as: 1) In order to make full use of raw array signal data, an array data processing approach is proposed to

* wang_dezhi@hotmail.com

generate the labeled dataset, by which the deep learning method can be implemented. 2) It is explored to integrate the well-developed CNN architectures such as ResNext (11) into the framework for more effective feature extraction and prediction. 3) The pre-trained weights of CNN model on a large open-source dataset are employed to enhance the system performance by a transfer learning and fine-tuning strategy. The performance of proposed system is finally validated by using the real data acquired in a shallow water experiment.

The rest of this paper is organized as follows. Section 2 presents details of the proposed system in the study. In Section 3, the dataset and experimental setup are first respectively introduced. Experimental results are then reported and discussed. Finally, Section 4 concludes the paper.

2. METHODOLOGY

2.1 Array Data Processing

The training of deep learning neural networks requires a large number of sample data to reach a convergence. Thus, it is important to first process and convert the raw hydrophone array data into a usable labeled dataset before applying the deep learning techniques. As shown in Figure 1, the processing steps can be listed as follows:

1) Time-domain conventional beamforming. The recorded multi-channel array data is beamformed in the time domain by means of the CBF approach to obtain the waveform signals at each DOA direction with a resolution of 1 degree.
2) Acquisition of GPS and DOA information of target ships. The ship GPS data can be read to provide ground-truth direction-finding information while the DOA information can also be manually recognized in the DOA diagram based on CBF.
3) Audio tagging. On the basis of the GPS and DOA information, the beamformed array signals can be labeled to generate a complete dataset for deep learning applications.

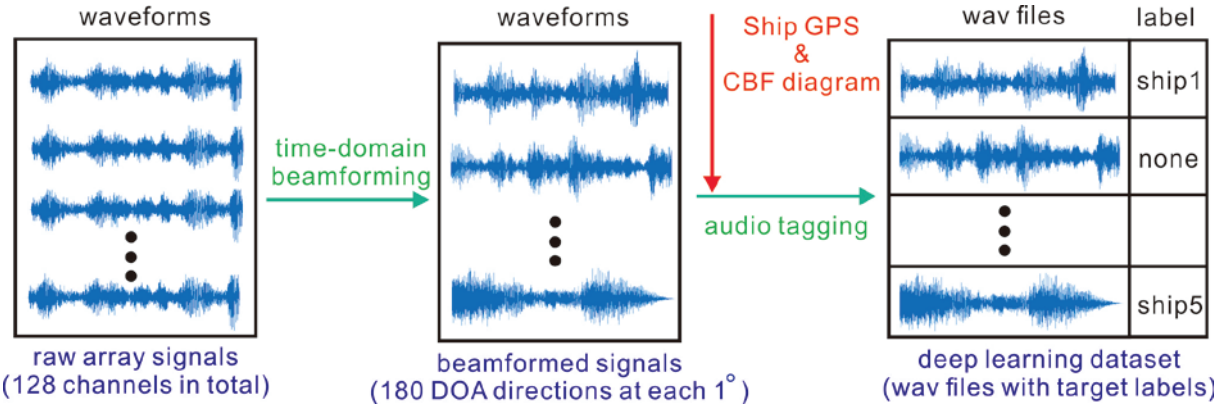

Figure 1 – The diagram of processing array data

2.2 Localization Based on Pre-trained Neural Networks

Based on the generated dataset, a deep neural network architecture is employed to predict the DOAs of target ships. Apart from directly using the raw waveforms as input, the most common choice in audio signal processing is the 2D time-frequency representation, for example, log mel-filter bank features (log-mel), which is achieved based on short-time Fourier transform and usually scaled by a log-like frequency. The log-mel features are considered to be the best time-frequency feature representations for audio signals to be used for deep-learning methods (12, 13), In this study, log-mel features are first obtained from the waveform data and then applied as the input data fed into the neural networks.

As shown in Figure 2, the log-mel features (by 128 filters) and the corresponding delta and delta-delta features are calculated at the same time in order to produce a 3-channel image-type input for deep neural networks. A random selection strategy is also used to stochastically capture a fixed-length segment in time axis from the 3-channel feature maps to generate the equal-length inputs.

The prediction task is implemented by using a state-of-the-art CNN architecture i.e. ResNext101 (14). Due to the fact that low-level semantic features are constant for different tasks like image

recognition and audio tagging, it is more efficient to employ the deep CNN models with pre-trained weights on a large-scale image dataset (e.g. 'ImageNet' (7)) to do the audio tagging in a transfer learning manner (15, 16). Fine-tuning the pre-trained weights of CNN model on the generated array signal dataset will generally allow for a faster training and smaller prediction errors (16).

The ResNext architecture is an extension of the deep residual network where the standard residual block is replaced by a 'split-transform-merge' one (14). As shown in Figure 2, the prediction result is finally achieved after a modified fully-connected layer (FC) and a Softmax layer.

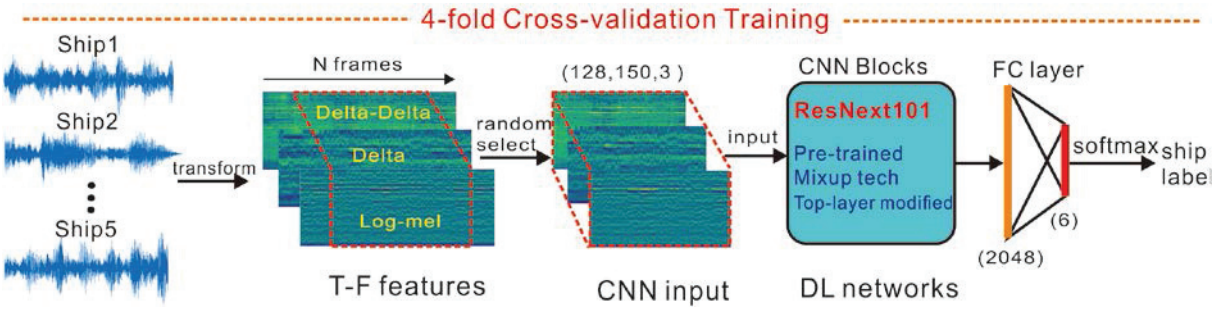

Figure 2 – Illustration of the proposed architecture

3. EXPERIMENT

3.1 Data of the Shallow Water Experiment

The experimental data used in this study comes from a shore hydrophone array test in a shallow sea area. The array is a line fiber-optic hydrophone array having a total of 128 array elements with an array element spacing of 6.25 m and a sampling frequency of 20 kHz. The experiment depth is about 95 m. There are several ships transiting in the area and the GPS information of some ships are also recorded. A subset of the array recording data is selected to be used in this study.

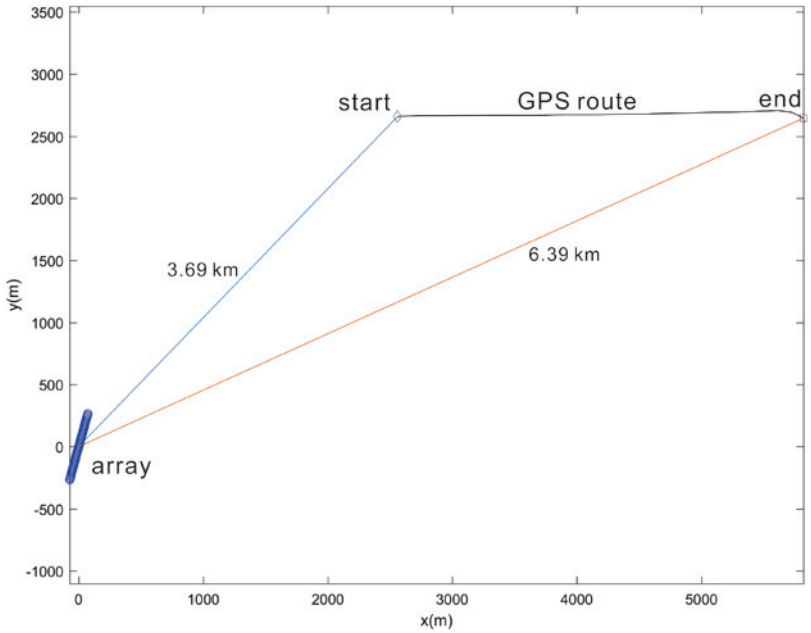

Figure 3 – The GPS route of Ship1

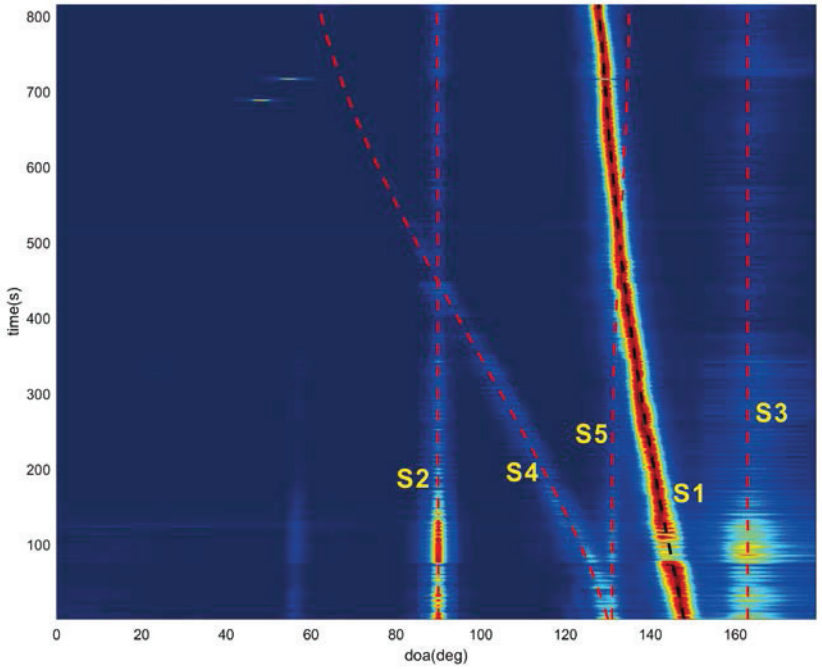

Figure 4 – Evolution of DOA based on the CBF

As shown in Figure 3, a ship target travels from a 3.69 km distance to a 6.39 km distance from the array according to the GPS data during this period. Base on the DOA diagram shown in Figure 4, the routes of 5 ship targets are recognized manually, which provide the ground-truth label information for the beamformed data. According to the aforementioned data processing steps, the raw array data are transformed into a labeled dataset. As shown in Table 1, the distribution of the beamformed array data with different labels is illustrated where each wav file has a same time length of about 1.63 s.

The labeled dataset is divided into the development and evaluation datasets where the development and evaluation datasets respectively consist of 50% (the first half and the other half) of the total number of samples. A 4-fold cross-validation setup is used on the development dataset. The averaged prediction accuracy across the four folds on the evaluation dataset is reported for performance analysis. The accuracy metric used is defined as the number of correctly predicted audio recordings divided by the total number of recordings.

Table 1 – List of recordings of beamformed array signals

Label	Number of development samples (5-fold cross validation)	Number of evaluation samples	Total number of samples	Sampling rate
Ship1	1123	1123	2246	20k Hz
Ship2	1198	1197	2395	20k Hz
Ship3	2805	2805	5610	20k Hz
Ship4	1173	1172	2345	20k Hz
Ship5	588	588	1176	20k Hz
Ship4 & Ship5	25	25	50	20k Hz
Ship1 & Ship5	152	152	304	20k Hz
Ship2 & Ship4	78	77	155	20k Hz
None	38505	38505	77010	20k Hz

3.2 Experimental Setup

The production of the log-mel, delta and delta-delta feature maps are obtained by Librosa toolbox in Python and stored in the hard disk in forms of pickle files. As shown in Figure 5, an example of the 3-channel input feature maps for ResNext model is presented. The training process is carried out based on a 4-fold cross-validation setup on the development dataset and the performance is then tested on the evaluation dataset. Since the class imbalance problem is really significant for the dataset, a simple data balancing technique is developed just in terms of ensuring at least one sample to be selected for each class in a training batch.

Different configurations of the proposed architecture and related parameters are tested in this work according to the heuristic experience. The adaptive moment estimation (ADAM) algorithm is employed to optimize the cross-entropy loss objective function. The learning rate is set as multi-steps, where 0.001 is set for the first 50 epochs and multiplied by 0.1 for every next 20 epochs. A batch size of 32 is applied based on the hardware capacity. The models are implemented by PyTorch using GPU acceleration on a hardware resource consisting of Xeon E5 2683V3 CPU and 2 GTX 1080Ti GPU cards driven by CUDA with cuDNN (17).

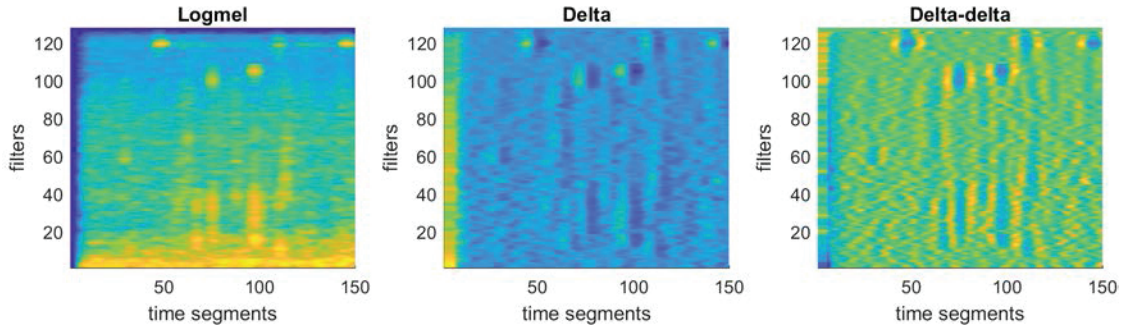

Figure 5 – An example of the 3-channel input (logmel) for DL model, the 3 channels are respectively logmel, delta feature of logmel and delta-delta feature of logmel

3.3 Results and Discussion

As shown in Table 2, after completing the model training process, class-wise validation accuracies are averaged over all the four cross-validation folds. The class-wise evaluation accuracies are also obtained on the evaluation dataset. It is shown that the evaluation accuracies are a little lower than the counterpart validation accuracies, which indicates the generalization capability of the proposed method is relatively good. It is worth noting that the training dataset has a significant class imbalance problem, where the number of samples in the largest class is about 100 times of the number of the smallest class. Since a class balancing technique has been utilized in the training, the model produces relatively small difference on the accuracy for different classes.

Table 2 – Results of the class-wise accuracies

Class	Validation	Evaluation
Ship1	95.3%	87.2%
Ship2	96.7%	87.7%
Ship3	97.5%	92.0%
Ship4	86.1%	72.5%
Ship5	79.0%	70.0%
Ship4 & Ship5	68.0%	56.0%
Ship1 & Ship5	78.9%	75.6%
Ship2 & Ship4	70.5%	64.9%
None	92.9%	91.8%
Averaged	85.0%	77.5%

Due to the fact that the target ships actually travel far away from the array in the test, the SNR of the samples in the evaluation dataset are generally lower than that in the training dataset. Thus the averaged accuracies of the proposed system can still be considered as good. In the future study, the improvements can be carried out to more efficiently solve the class imbalance problem and further increase the model generalization capability. Moreover, the achievement of ground-truth DOA labels can be refined to eliminate the possible errors coming from array configuration and other factors.

4. CONCLUSION

Despite deep learning techniques have shown to be very effective than other methods in a variety of research fields, they still have not been widely explored in the domain of underwater passive ship localization. In this work, we have investigated the use of a scalable deep learning system integrated with the well-developed CNN architectures to predict the DOA information of target ships. A list of the array data processing steps is proposed to make the full use of the large-scale array recordings for training deep neural networks. The developed system shows a great potential on the problem solution of the ship acoustic localization in a complex marine environment. In future work, the proposed system will be carefully improved and widely tested based on different underwater acoustic scenarios. In addition, some efforts will be made to give an analysis of the physical meanings of what have been learned and extracted by the deep learning neural networks in order to obtain an appropriate interpretation.

ACKNOWLEDGEMENTS

This study was funded by the National Natural Science Foundation of China (No. 61806214), the Science and Technology Foundation of State Key Laboratory of Sonar Technology (No. 6142109180204), the Scientific Research Project of NUDT (No.ZK17-03-31) and the Double Top Construction Foundation of NUDT (No.qnrc02).

REFERENCES

1. Hamilton M, Schultheiss PM. Passive ranging in multipath dominant environments. I. Known multipath parameters. IEEE Transactions on Signal Processing. 1992;40(1):1-12.
2. Gao Y, Clark M, Cooper P, editors. Time delay estimate using cepstrum analysis in a shallow littoral environment. Conf Undersea Defence Technology; 2008.
3. Benesty J, Chen J, Huang Y. Time-delay estimation via linear interpolation and cross correlation. IEEE Transactions on speech and audio processing. 2004;12(5):509-19.
4. Ozard JM, Zakarauskas P, Ko P. An artificial neural network for range and depth discrimination in matched field processing. The Journal of the Acoustical Society of America. 1991;90(5):2658-63.
5. Michalopoulou Z-H, Alexandrou D, De Moustier C. Application of neural and statistical classifiers to the problem of seafloor characterization. IEEE Journal of Oceanic Engineering. 1995;20(3):190-7.
6. Niu H, Ozanich E, Gerstoft P. Ship localization in Santa Barbara Channel using machine learning classifiers. The Journal of the Acoustical Society of America. 2017;142(5):EL455-EL60.
7. Krizhevsky A, Sutskever I, Hinton GE, editors. ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems; 2012.
8. Piczak KJ. Environmental sound classification with convolutional neural networks. 2015:1-6.
9. Salamon J, Bello J. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters. 2016;PP(99):1-.
10. Ferguson EL, Ramakrishnan R, Williams SB, Jin CT, editors. Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017: IEEE.
11. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-8.
12. Li J, Wei D, Metze F, Qu S, Das S, editors. A Comparison of deep learning methods for environmental sound. IEEE International Conference on Acoustics; 2017.
13. Huzaifah M. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. arXivorg. 2016.
14. Hitawala S. Evaluating ResNeXt Model Architecture for Image Classification. arXivorg. 2018.
15. Dan CC, Meier U, Schmidhuber J. Transfer learning for Latin and Chinese characters with Deep Neural Networks. 2012;20:1-6.

16. Amaral T, Kandaswamy C, Silva LM, Alexandre LA, Sá JMD, Santos JM, editors. Improving Performance on Problems with Few Labelled Data by Reusing Stacked Auto-Encoders. International Conference on Machine Learning and Applications; 2014.
17. Serrano J. Nvidia Introduces cuDNN, a CUDA-based library for Deep Neural Networks. http://infoqcom. 2015.