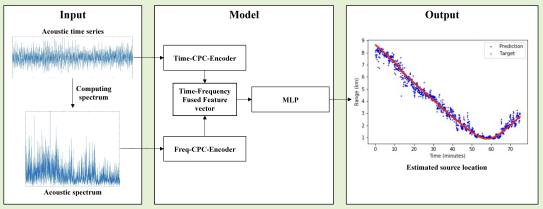
Time-Frequency Fused Underwater Acoustic Source Localization based on Contrastive Predictive Coding

Xiaoyu Zhu, Hefeng Dong, *Member, IEEE*, Pierluigi Salvo Rossi, *Senior Member, IEEE*, and Martin Landrø

***Abstract*— We propose a time-frequency fused underwater acoustic source localization method based on self-supervised learning with contrastive predictive coding. Firstly, two feature extractors are trained to solve the pretext task (predicting the future) based on the unlabeled acoustic signals in the time and frequency domains, respectively. Next, encoders with frozen parameters are taken from the trained feature extractors for extracting the high-level features in the time and frequency domains. During the training stage of the source localizer, features extracted by two encoders are concatenated together as a time-frequency fused feature vector and fed into a 3-layer multi-layer perceptron for solving the downstream task (source localization) based on a tiny labeled dataset. This method is assessed on the SWellEx-96 Experiment and compared with several alternative methods. The performance analysis confirms the promising performance of our proposed method.**



***Index Terms*— Contrastive predictive coding, self-supervised learning, underwater source localization, feature fusion.**

I. Introduction

UNDERWATER acoustic source localization is an active research topic which is gaining relevance due to ocean environment monitoring, navigation and related applications. Recent approaches rely on machine learning, due to the capability to achieve impressive performance with limited prior information (e.g., unknown ocean environment, sound speed profile, and/or seabed parameters) [1]–[7]. A deep neural network trained by supervised learning is an end-to-end model, which can automatically extract useful features and conduct one specific task (e.g, source localization) guided by the labels (source locations). Without enough labels, such a strategy cannot achieve good performance. Unfortunately, the insufficiency of labels is common in underwater acoustics scenarios. Most of the existing works concentrate on utilizing different state-of-the-art architectures of deep neural networks referenced from the field of computer science, such as generalized regression neural network (GRNN) [5] and residual neural network (ResNet) [6], [7], to study their source localization performance. These works used a similar approach to satisfy the required number of labels, which is exploiting acoustic propagation simulation models to create a huge simulation dataset for training the source localizer based on supervised learning. However, this approach has some limitations: (1) creating such a huge simulation dataset is time-consuming and requires large computer storage resources; (2) this simulation dataset is built for one specific ocean area, which means that each time the ocean area is changed, the simulation process needs to be repeated; (3) the numerical acoustic propagation models are built based on simplified theoretical equations which sometimes cannot well describe the real condition; (4) to create the simulation dataset, prior information of the environment is still needed.

These limitations make the aforementioned approach hard to be applied in a real-world ocean monitoring system. For instance, an ocean monitoring system will continuously collect purely acoustic signals over a long period. The collected purely acoustic signals can also provide useful but implicit information about the ocean environment, such as variations in the sound speed profile. This information cannot be sufficiently exploited and absorbed by a purely supervised learning model since the model needs labels. To get rid of the cumbersome simulation process and the difficulties of collecting labels, learning features from unlabeled acoustic signals is an alternative approach potentially capable of mitigating the insufficiency of labeled data. This approach is designed based on self-supervised learning (SSL), which obtains supervisory signals from the unlabeled data and has shown excellent

Part of this paper was presented at the IEEE SENSORS 2021, Virtual Conference, October 2021. The authors would like to acknowledge the Norwegian Research Council and the industry partners of the GAMES consortium at NTNU for financial support (Grant No. 294404). Xiaoyu Zhu would like to acknowledge the China Scholarship Council (CSC) for the fellowship support (No. 201903170205). (Corresponding author: Xiaoyu Zhu.)

Xiaoyu Zhu, Hefeng Dong, Pierluigi Salvo Rossi, and Martin Landrø are with the Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: xiaoyu.zhu@ntnu.no; hefeng.dong@ntnu.no; pierluigi.salvorossi@ntnu.no; martin.landro@ntnu.no).

results in computer vision [8], [9]. The SSL-based acoustic source localization is mostly unexplored with a few exceptions focusing on a reverberant room acoustic environment [10], [11] and on the ocean environment [12]. From the SSL perspective, the existing works used the same method, convolutional autoencoder (CAE) which is a basic method in SSL, to extract useful features from unlabeled data for source localization. Those works confirm the potential of SSL to extract features for source localization especially when the labels are limited. However, with the development of SSL, many methods have been proposed and demonstrated their outperformance over the CAE in the field of computer science, such as contrastive predictive coding (CPC) [13]. These methods have huge potential for underwater acoustic localization. Our motivation is to dig deeper into the application of SSL-based methods for underwater source localization and enhance the related performance. Note that our earlier conference paper has demonstrated the feasibility of applying CPC to extract useful features from acoustic time series for underwater source localization [14].

However, there are some remaining points wroth to be studied:

- The necessity of jointly exploiting the acoustic time series and their corresponding spectra for SSL-based source localization since current SSL-based source localization methods are based on signal processing either in frequency domain [10]–[12] or time domain [14].
- The possibility of exploiting the signals collected among different periods by one specific receiver for enhancing localization performance. This corresponds to a common scenario for a permanent monitoring system based on one specific receiver at a fixed location.
- The possibility of exploiting the signals collected by several receivers at different depths for enhancing localization performance. This corresponds to a common scenario for a monitoring system based on a vertical array.

The *main contributions* of this paper are summarized as follows.

1) We focus on a common type of real-world scenario, that is, a tiny labeled dataset and a large unlabeled dataset (purely acoustic signals) are available. More specifically, the labels provided by the GPS system are paired to the acoustic signals collected by one specific receiver during the same period. Two cases are considered for collecting the unlabeled dataset: (a) one specific receiver and (b) three receivers (containing the specific receiver) at different depths.
2) We propose a CPC-based architecture that exploits joint time-frequency processing for source localization and outperforms the architectures based on single-domain features in terms of localization errors, robustness to receiver depth and generalization capability.
3) A comprehensive performance analysis of the proposed method is presented based on a public database, the SWellEx-96 Experiment [15], and three approaches for exploiting unlabeled data are discussed. Recent alternative methods are used as a benchmark for comparison and assess the value of our proposal.

The rest of the paper is organized as follows: Sec. II states the considered problem; Sec. III describes the theory of CPC, the architectures of the self-supervised feature extractor and time-frequency fused source localizer; Sec. IV presents the SWellEx-96 Experiment, the data preprocessing, and the schemes of building the training and testing datasets; A comprehensive performance analysis on Event S5 is given in Sec. V; In Sec. VI, an analysis of the generalization ability on Event S59 is described; In Sec. VII, the proposed method is compared with a popular joint time-frequency signal processing method, wavelet transform; Finally, the conclusion and future direction are given in Sec. VIII.

Notation: Lowercase bold letters are adopted for vectors, with a_n representing the nth component of $\boldsymbol{a}$. Uppercase bold letters are adopted for matrices. Normal letters express scalars.

II. Problem Formulation

To simplify the problem, we suppose that there is only *one source* present and the source location is the horizontal range between the source and the receiver. The underwater source localization problem is to estimate the corresponding source location y given the collected acoustic signal $\boldsymbol{x}$:

$$y = f_{localizer}(\boldsymbol{x}) \tag{1}$$

where $f_{localizer}(.)$ is the source localizer.

This is a standard regression problem in supervised learning since the source location y is a real number. And the labeled dataset is expressed as a set of data-pairs:

$$[\boldsymbol{X}, \boldsymbol{y}] = [\boldsymbol{x}_i, y_i]_{i=1}^{N} \tag{2}$$

where $\boldsymbol{x}_i$ and y_i are the ith acoustic signal and corresponding source location, respectively. N is the total number of the labeled data.

As an aforementioned scenario, there is only a tiny labeled dataset available which is not enough amount for the purely supervised learning scheme. We apply the SSL-based method to this problem.

SSL strategy consists of a pretext task and a downstream task [8]. The pretext task is a kind of task letting the neural network learn latent features based on unlabeled data, such as reconstructing input [16] and predicting the future [17]. Since the pretext task does not need labels, the unlabeled dataset is expressed as:

$$[\boldsymbol{X}] = [\boldsymbol{x}_j]_{j=1}^{M} \tag{3}$$

where $\boldsymbol{x}_j$ is the jth acoustic signal and M is total number of the available acoustic signals.

Based on the learned features, a downstream task (i.e., underwater source localization as defined in eq. (1)) will be solved by training a model based on a tiny labeled dataset with purely supervised learning scheme. Note that the unlabeled dataset can include the purely acoustic signals that occurred in the labeled dataset since the real objective is source localization which is a supervised regression task and needs labels. Without labels, reusing purely acoustic signals will not influence the fairness of the localization performance analysis.

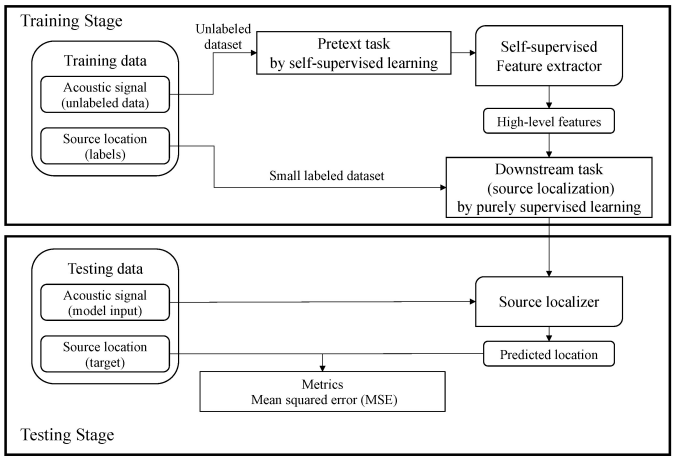


Fig. 1. Workflow of source localization based on SSL.

III. Methodology

A common workflow of source localization based on SSL is shown in Fig. 1, in which the workflow includes two stages, i.e., training and testing stages. The training stage is for constructing and training the source localizer, which consists of two steps. The first step is training the self-supervised feature extractor by solving different pretext tasks based on an unlabeled dataset. The second step is training the source localizer by purely supervised learning based on a tiny labeled dataset. The testing stage is for predicting the source location in real-world scenarios.

A. Self-supervised Feature Extractor based on Contrastive Predictive Coding

In this paper, we use CPC to extract the high-level features from the acoustic signals. The architecture of the self-supervised feature extractor is the same as in [17], [19] and shown in Fig. 2.

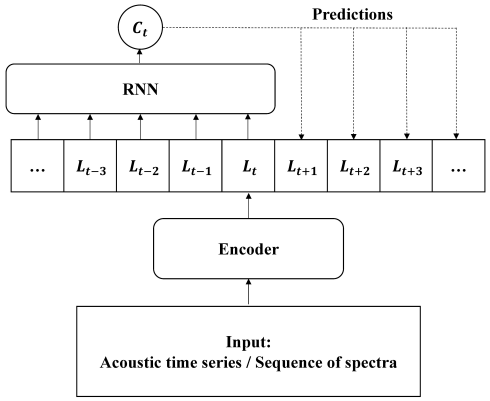


Fig. 2. Architecture of the self-supervised feature extractor.

As shown in Fig. 2, the feature extractor consists of two sub-models. The first sub-model is an non-linear encoder $f_{encoder}$, which compresses the input sequence into a sequence of latent codes $\boldsymbol{L}_t = f_{encoder}(\boldsymbol{x}_t)$. The second sub-model is a recurrent neural network (RNN) acting as an auto-regression model for summarizing all previous latent codes $\boldsymbol{L}_{\leq t}$ into a latent context representation $\boldsymbol{C}_t = f_{RNN}(\boldsymbol{L}_{\leq t})$.

The CPC aims at predicting the future k time steps by modeling a density ratio which preserves the mutual information between $\boldsymbol{x}_{t+k}$ and $\boldsymbol{C}_t$:

$$f_k(\boldsymbol{x}_{t+k}, \boldsymbol{C}_t) = exp(\boldsymbol{L}_{t+k}^T \boldsymbol{W}_k \boldsymbol{C}_t) \tag{4}$$

where a linear transformation $\boldsymbol{W}_k \boldsymbol{C}_t$ is used for the prediction with a different $\boldsymbol{W}_k$ for every timestep k.

During the training stage, the encoder and RNN are trained to jointly optimize a loss function based on Noise-Contrastive Estimation (NCE) [19] for maximizing the mutual information. More details about the theory of CPC can be found in [17], [19].

The feature extractors are named Time-CPC-extractor and Freq-CPC-extractor when the inputs are a time series and a sequence of spectra, respectively.

Based on preliminary configuration experiments, the predicting timestep is chosen as $k = 16$.

B. The Time-Frequency Fused Source Localizer

After training the Time-CPC-extractor and Freq-CPC-extractor, only the Time-CPC-Encoder and Freq-CPC-Encoder with frozen parameters are obtained, respectively.

During the training stage of the source localizer, the signals in time and frequency domains are fed into the Time-CPC-Encoder and Freq-CPC-Encoder, respectively. Then the features extracted by the encoders are concatenated together as a time-frequency fused feature vector and fed into a 3-layer multi-layer perceptron (MLP) for the source localization task. Since source localization can be formulated as a regression problem, the mean squared error (MSE) is chosen as the loss function. The architecture of the time-frequency fused source localizer is shown in Fig. 3 where the arrows indicate the direction of the data stream and the input is at the bottom.

IV. Dataset and Preprocessing

A. SWellEx-96 Experiment

SWellEx-96 Experiment was conducted between May 10 and 18, 1996, approximately 12 km from the tip of Point Loma near San Diego, California [15]. The layout of the experiment is shown in Fig. 4. Two acoustic sources were towed by a vessel and simultaneously transmitted various multi-tone signals at frequencies between 50 and 400 Hz. Since the sources were towed simultaneously and no prior information except the source location is known, the vessel and the two sources are treated together as *one source* in this paper.

Events S5 and S59 are source towing events, which were conducted for 75 and 65 minutes on different dates, respectively. There was an extra passing vessel present during the whole period of Event S59. On the contrary, in Event S5, there was no such noise source present. Vertical liner array (VLA) data from Events S5 and S59 are used to evaluate the localization performance. The sampling rate of the acoustic

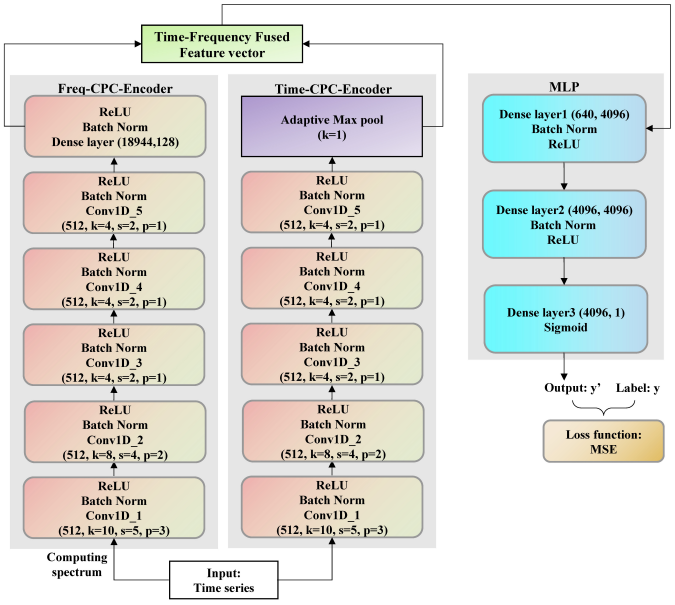


Fig. 3. Architecture of the time-frequency fused source localizer.

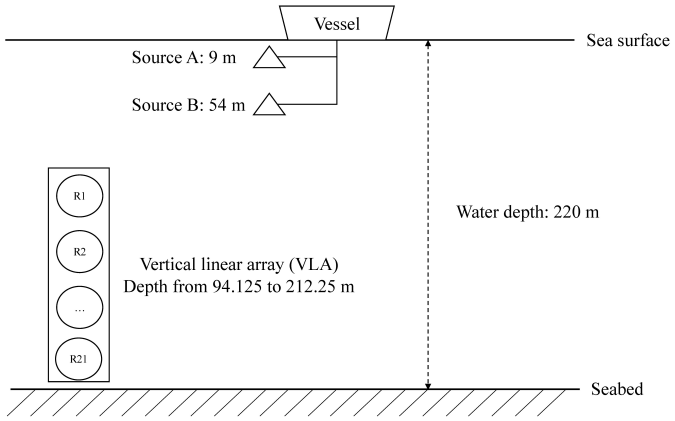


Fig. 4. Layout of the SWellEx-96 Experiment.

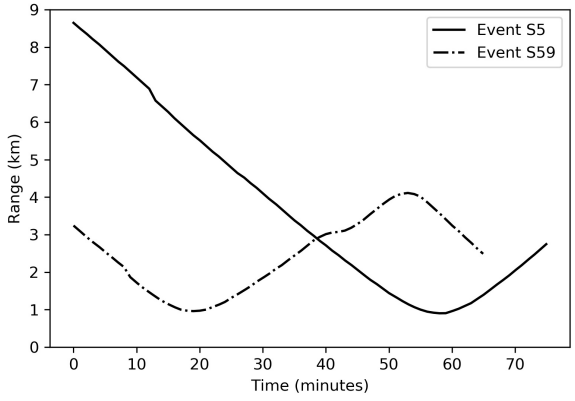


Fig. 5. Horizontal ranges between the source and the VLA of the SWellEx-96 Experiment.

data was 1500 Hz and the VLA contained 21 receivers equally spaced between depth 94.125 m and 212.25 m. Additionally, the labels (source locations) were collected by the GPS system of the vessel and converted to the horizontal ranges between the source and the VLA shown in Fig. 5. The intervals of the range are approximately 1 to 9 km and 1 to 4.5 km for Events S5 and S59, respectively. More detailed information on the SWellEx-96 Experiment can be found in [15].

B. Preprocessing

The acoustic signals collected by a single receiver are cut into slices (four seconds per slice) without overlap and arranged into a time signal matrix $\boldsymbol{X}_{time}$, where each row is related to one slice. Based on $\boldsymbol{X}_{time}$ (without any preprocessing), the spectrum of each row is calculated and arranged into a spectra matrix $\boldsymbol{X}_{spectra}$. In addition, the horizontal ranges between the source and the VLA are known and represented by a label vector $\boldsymbol{y}$.

Several preprocessing methods are applied for the training stability. Based on our previous works [12], [14], min-max scaling (scaling into interval $(0, 1)$) is applied on the spectra matrix $\boldsymbol{X}_{spectra}$ and the label vector $\boldsymbol{y}$. Standardization is applied on the time signal matrix $\boldsymbol{X}_{time}$. Preprocessing and size of the data matrices and the label vector are shown in Table I.

TABLE I
OVERVIEW OF THE DATABASE AND PREPROCESSING

	Event S5	Event S59	Preprocessing
$\boldsymbol{X}_{time}$	1125×6000	975×6000	Standardization
$\boldsymbol{X}_{spectra}$	1125×3000	975×3000	Min-max scaling
$\boldsymbol{y}$	1125×1	975×1	Min-max scaling

C. Schemes of Building the Training and Testing Datasets

For training the feature extractors, all of the unlabeled acoustic signals are used to build the training dataset. During each round of training, the input for Time-CPC-extractor is each row in $\boldsymbol{X}_{time}$. Since the spectrum cannot provide the time information which CPC needs, the input for Freq-CPC-extractor is a sequence of rows from $\boldsymbol{X}_{spectra}$ with continuous indexes. We used one implementation trick to make inputs the same time-span (i.e., four seconds per input) for Time-CPC-Encoder and Freq-CPC-Encoder to solve the downstream task in Fig. 3.

In the case study, three scenarios about the collected purely acoustic signals are defined:

- Case 1: Purely acoustic signals collected by a single receiver during Event S5 are available.
- Case 2: Purely acoustic signals collected by a single receiver among periods of Events S5 and S59 are available. The dataset is constructed by concatenating the data matrix from each event along the time axis.
- Case 3: Purely acoustic signals collected by multi-receivers (assuming three receivers at different depths in this paper) during Event S5 are available. The dataset is constructed by concatenating the data matrix collected by each receiver along the time axis.

The details of the training datasets for the feature extractors are shown in Table II, where the numbers of training samples are 1125, 2100, and 3375 in Cases 1, 2, and 3, respectively.

TABLE II
TRAINING DATASETS FOR THE FEATURE EXTRACTORS IN CASE STUDY

	Case 1	Case 2	Case 3
Time-CPC-extractor	1125×6000	2100×6000	3375×6000
Freq-CPC-extractor	1125×3000	2100×3000	3375×3000

For training the source localizer, only 141 labeled data of Event S5 are used to build the training dataset for mimicking the scenario of extremely limited labels. During each round of training, the inputs for the trained Time-CPC-Encoder and Freq-CPC-Encoder are one row in $\boldsymbol{X}_{time}$ and the corresponding row in $\boldsymbol{X}_{spectra}$, respectively. Since source localization can be formulated as a regression problem, labels in the training dataset should cover the whole interval of range. The samples in Event S5 are expressed by the index $i \in (1, 1125)$.

The scheme of building the training dataset for the source localizer is:

$$(\boldsymbol{x}_i, y_i) \quad \forall i : mod(i, 8) = 1 \tag{5}$$

where $\boldsymbol{x}_i$ is the ith row in data matrices $\boldsymbol{X}_{time}$ or $\boldsymbol{X}_{spectra}$. y_i is the ith element in the label vector $\boldsymbol{y}$.

The scheme of building the testing dataset is:

$$\begin{aligned} \text{Receiver for training} &: \forall i : mod(i, 8) \neq 1 \\ \text{Other receivers} &: \forall i \in (1, 1125) \end{aligned} \tag{6}$$

To show the influence of different receiver-depth, receivers no. 1 (top), no. 10 (middle), and no. 21 (bottom) are chosen to build the training datasets (for both feature extractors and the source localizer), respectively.

V. EX 1: PERFORMANCE ANALYSIS ON EVENT S5

Three cases are designed for evaluating the proposed method. For a comprehensive analysis, all candidate localizers are tested on the data collected by all receivers. The performance metric is the averaged MSE tested on all receivers.

A. Case 1: Solving Pretext Task based on a Single Receiver in Event S5

Case 1 is the basic performance analysis. In this case, the proposed method is compared with several alternative methods which have been evaluated using the datasets of the SWellEx-96 Experiment [5], [12], [14]. In addition, the pretext task is solved based on the unlabeled data collected by a single receiver in Event S5.

The candidate methods for Case 1 are:

- A GRNN-based localizer based on $\boldsymbol{X}_{spectra}$ (namely, GRNN) [5].
- A CAE-based localizer based on $\boldsymbol{X}_{spectra}$ (namely, CAE) [12].
- A Freq-CPC-based localizer based on $\boldsymbol{X}_{spectra}$ (namely, Freq-CPC).
- A Time-CPC-based localizer based on $\boldsymbol{X}_{time}$ (namely, Time-CPC) [14].
- The proposed localizer based on both $\boldsymbol{X}_{spectra}$ and $\boldsymbol{X}_{time}$ (namely, Time-Freq-CPC).

The performance of source localization for Case 1 is shown in Fig. 6, where the legend shows the relationship between colors and candidate localizers. In the abscissa, R1, R10, and R21 are related to the top, middle, and bottom receivers, respectively. The ordinate expresses the performance metric, i.e., MSE.

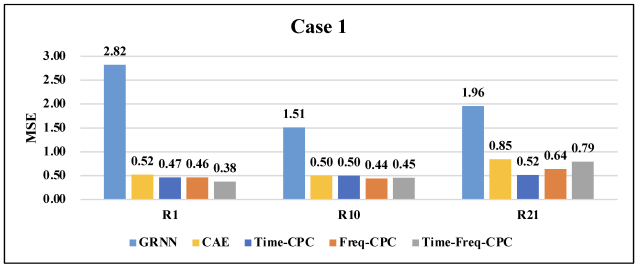


Fig. 6. Performance of the candidate localizers for Case 1.

From Fig. 6, interesting phenomena can be found:

- The GRNN-based localizer has the worst performance since it is a purely supervised model. As mentioned in Sec. I, a vast labeled dataset is always required for sufficiently training a purely supervised model. This phenomenon reveals the drawback of the purely supervised learning scheme when only an extremely limited labeled dataset is available.
- All CPC-based localizers show better performance than CAE, which illustrates that the pretext task of CPC is better than that of CAE. A similar phenomenon has already been mentioned in our previous work [14] and the works in the field of computer vision [13].
- Trained on R1 and R10, Freq-CPC is better than Time-CPC. On the contrary, trained on R21, Freq-CPC is worse than Time-CPC. This phenomenon inspired an intuition: **A combination of features extracted from data both in the time and frequency domains may improve the localization performance**.
- According to our intuition, Time-Freq-CPC is also tested. Trained on R1, Time-Freq-CPC provides a significant improvement in performance, which is consistent with our intuition. However, trained on R10 or R21, Time-Freq-CPC doesn't show overwhelmed performance than others. A possible reason for this irregular phenomenon may be that Time-Freq-CPC needs more unlabeled data to learn better features. Hence more cases are studied as follows.

In short, Case 1 provides a basic comparison of the methods proposed in recent years. The results show that the CPC-based localizers can provide better localization performance due to a better design of the pretext task. However, the performance of Time-Freq-CPC shows an irregular pattern, which needs more cases to discuss.

Since the superior performance of the CPC-based localizers is demonstrated in Case 1, the further performance analysis will only focus on the CPC-based localizers.

B. Case 2: Solving Pretext Task based on a Single Receiver in both Events S5 and S59

Case 2 investigates the practicability of using data collected among different periods by a single receiver which is a common real-world scenario for a permanent ocean monitoring system. In this case, the pretext task is solved based on the unlabeled data collected by a single receiver among periods of Events S5 and S59. The candidate localizers are Time-CPC, Freq-CPC, and Time-Freq-CPC.

The source localization performance for Case 2 is shown in Fig. 7 with the same abscissa and ordinate as in Fig. 6. The blue, orange, and gray bars are related to the Time-CPC, Freq-CPC, and Time-Freq-CPC, respectively.

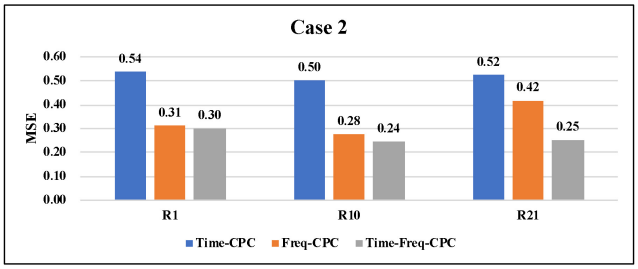


Fig. 7. Performance of candidate localizers for Case 2.

The following phenomena can be observed from Fig. 7:

- Compared to Case 1, the performance of Time-CPC keeps almost the same. However, the performance of Freq-CPC shows a significant improvement. It illustrates that Freq-CPC can get more benefits than Time-CPC from the data collected by a specific receiver among different periods.
- Time-Freq-CPC shows the best performance among all candidate localizers. It is consistent with our intuition that using the fused features from both time and frequency domains can significantly improve the performance, especially when localizers are trained on R10 and R21.

Case 2 provides a practical approach of sufficiently using the unlabeled data collected by a single receiver among different periods (Events S5 and S59) and demonstrates the benefit of using the time-frequency fused features when the number of available unlabeled data increases.

C. Case 3: Solving Pretext Task based on Multi-receivers (R1, R10, and R21) in Event S5

Case 3 analyzes the practicability of using data collected by multi-receivers. In this case, the pretext task is solved based on the unlabeled data collected by multi-receivers (R1, R10, and R21) in Event S5. The candidate localizers are the same as in Case 2.

The performance of source localization for Case 3 is shown in Fig. 8, where the legend, abscissa, and ordinate are the same as in Fig. 7.

From Fig. 8, we can find:

- Compared to Cases 1 and 2, the performances of Time-CPC and Freq-CPC have been improved, which illustrates that both Time-CPC and Freq-CPC can benefit from the data collected by multi-receivers. This is an intuitive phenomenon since the selected three receivers were located at different depths providing depth-related information in the water column. However, in Cases 1 and 2, the pretext task is solved based on only the single receiver's data, which cannot provide depth-related information.
- Trained on R1 and R10, Time-Freq-CPC shows similar performance as Freq-CPC. However, trained on R21, Time-Freq-CPC is the best among all candidate localizers. This phenomenon illustrates that Time-Freq-CPC can also increase the robustness of the receiver-depth selection since most localizers perform the worst when they are trained based on R21.

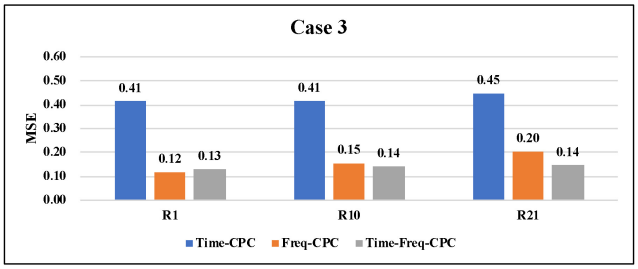


Fig. 8. Performance of candidate localizers for Case 3.

In short, Case 3 provides a practical approach of sufficiently using unlabeled data collected by multi-receivers among different receiver-depths and demonstrates again the benefits of using the time-frequency fused features.

It should be noted that different configurations of selecting three receivers have been tested, whose detailed discussion is not shown in this paper since our proposed localizer has already been demonstrated to have a superior performance by the aforementioned case study. The results of the configuration tests show that the performance becomes better when the distance between the receivers is larger (i.e., R1, R10, and R21) than that when the distance between the receivers is smaller (for instance, R1, R2, and R3). It is because the larger distance between receivers, the more information about the whole water column will be obtained. However, as for the generalization ability on Event S59 to be discussed in Sec. VI, these configurations show a minor difference.

To have an intuitive view of localization performance, Fig. 9 shows the comparison of localization results among three CPC-based localizers in Case 3. To solve the downstream task, localizers are trained on R1 and tested on R2.

D. Discussion for EX 1

Time-CPC only gets benefits from the unlabeled data containing more space variations (collected by multi-receivers among different receiver-depths) as shown in Case 3. Freq-CPC can also get benefits from the unlabeled data containing more time variations (collected among different periods) as shown in Case 2. As shown in all cases, Time-Freq-CPC can further improve the performance by fusing the features from both time and frequency domains.

To sum up, EX 1 provides a comprehensive analysis for different real-world scenarios and demonstrates the superior

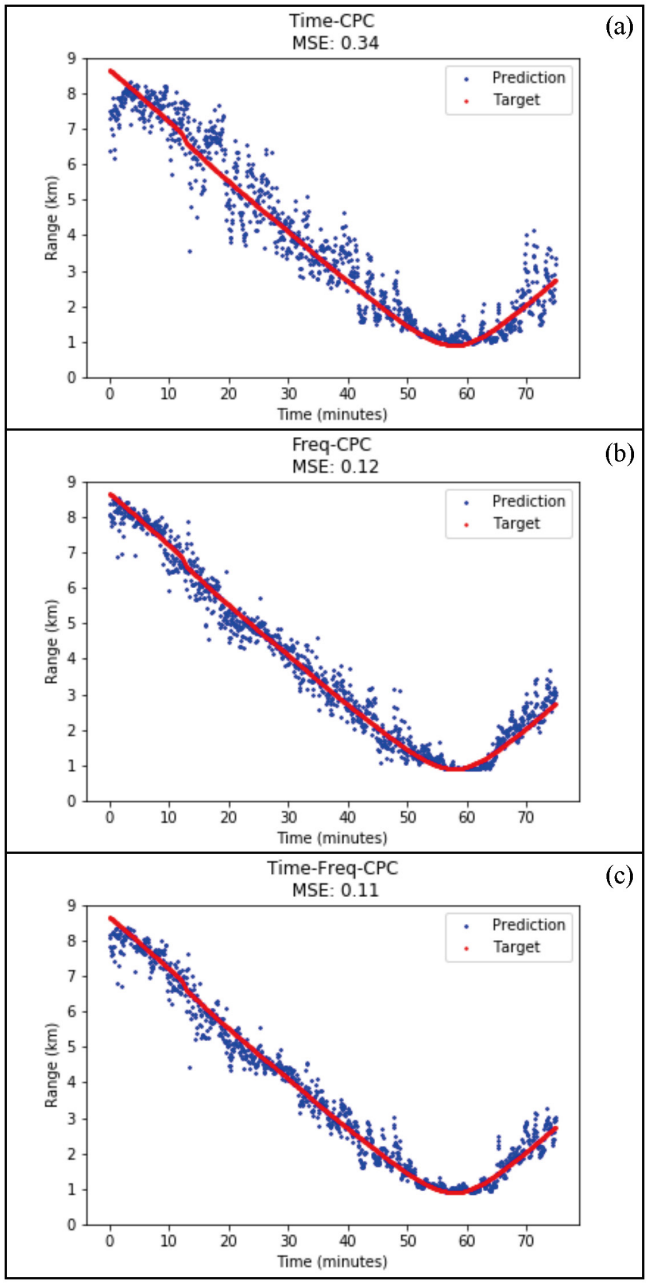


Fig. 9. Comparison of the source localization results among three CPC-based localizers on Event S5.

performance of the proposed method.

VI. EX 2: Generalization Ability Analysis on Event S59

The downstream task of all localizers in EX 1 is solved based on the labeled dataset in Event S5, which means that no localizer can access any labels in Event S59. In EX 2, three CPC-based localizers (Time-CPC, Freq-CPC, and Time-Freq-CPC) are chosen for investigating their generalization ability on Event S59.

The performance of source localization for EX 2 is shown in Fig. 10, in which the legend, abscissa, and ordinate are the same as in Fig. 7.

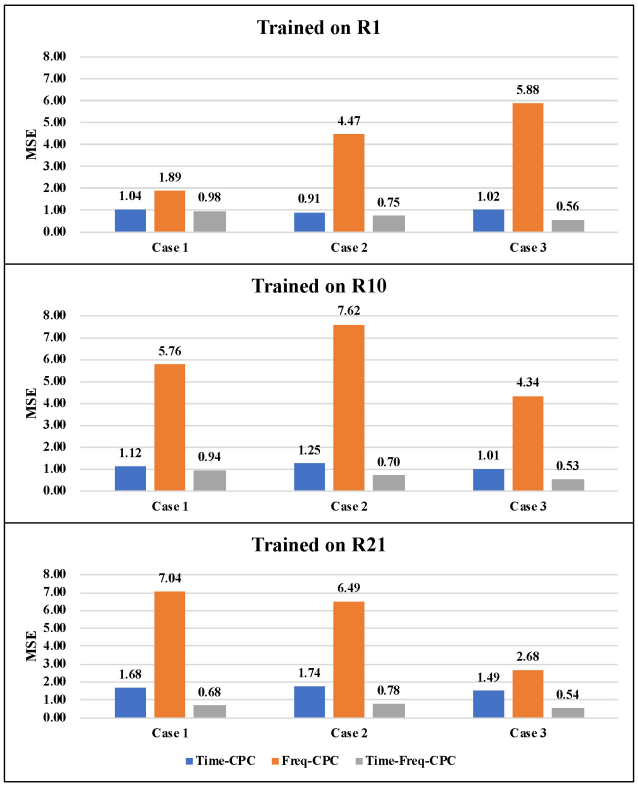


Fig. 10. Performance of candidate localizers for EX 2.

Fig. 10 reveals the following phenomena:

- Time-Freq-CPC shows an overwhelmed performance followed by Time-CPC, and Freq-CPC keeps the worst performance.
- Comparing the three sub-figures, Time-CPC and Freq-CPC show a biased performance when they are trained on the labeled dataset collected at different receiver-depth. However, compared to others, the performance of Time-Freq-CPC shows less bias among different receiver-depth.
- In each sub-figure, the performances of Time-CPC and Time-Freq-CPC show a similar ranking pattern as in EX 1, i.e., Case 3 is the best, followed by Case 2 and Case 1. This is because the number of unlabeled data for Case 3 is the greatest, followed by that for Case 2 and Case 1.

In short, EX 2 illustrates the generalization performance of the three candidates localizers and demonstrates that the proposed method (Time-Freq-CPC) shows the best ability of generalization on the unseen dataset.

To have an intuitive view of the generalization performance, Fig. 11 shows the comparison of localization results among the three CPC-based localizers tested on R1 in Case 3.

VII. Comparison between the Proposed Time-Frequency Fusion Method and the Wavelet Transform

The localization performance between the proposed method and the wavelet transform for joint time-frequency signal processing are compared in this section. Continuous wavelet transform (CWT) based on the Morlet wavelet with 512 scale

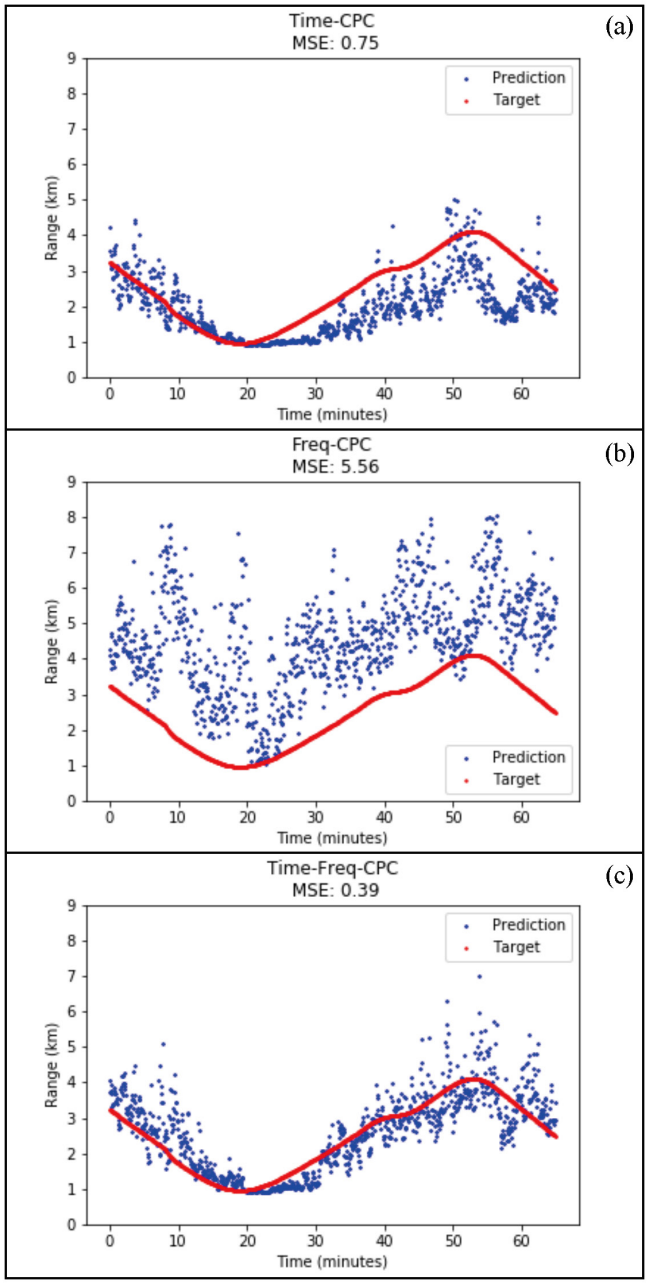


Fig. 11. Comparison of the generalization performance among three CPC-based localizers on Event S59.

factors is used to obtain the time-frequency representation from the collected acoustic signal.

The signal processing based on wavelet is expressed as follows:

1) CWT: The acoustic signals collected by a single receiver are processed by CWT based on the Morlet wavelet with 512 scale factors. The obtained time-frequency representation is expressed by a matrix $\boldsymbol{X}_{wavelet}$ of size $512 \times N$, where 512 refers to the total number of scale factors related to the frequency and N is the total length of the collected acoustic signals ($N = 6750000$ and 5850000 for Event S5 and Event S59, respectively).
2) Preprocessing: Both min-max scaling and standardization have been tested. Standardization is chosen for the final comparison due to its better localization performance. For a simpler description, we reuse $\boldsymbol{X}_{wavelet}$ to express the preprocessed matrix.

The two-step workflow for training the wavelet-based localizer is shown in Fig. 12 and expressed as follows.

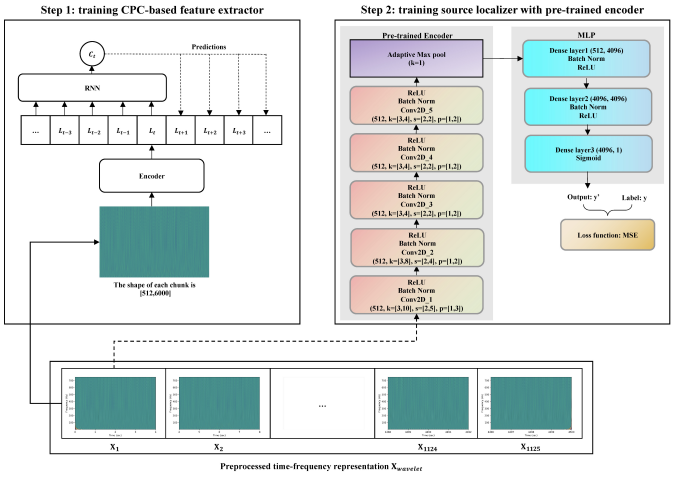


Fig. 12. Workflow for wavelet-based localizer.

After preprocessing, the acoustic signals collected by a single receiver are transformed into time-frequency representations which are shown at the bottom of Fig. 12. Here we use Event S5 as an example, but the processing workflow for data in Event S59 is the same. To keep the consistency of the input time-span (i.e., four seconds per input data) as in the paper, $\boldsymbol{X}_{wavelet}$ is cut into 1125 chunks $[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{1125}]$ (four seconds per chunk). The strategy of training the wavelet-based localizer is totally the same as in the paper. Note that the architecture of the encoder is slightly different from that in the paper since the input data has two dimensions.

The performance analysis is shown in Fig. 13 and expressed as follows. The purple and gray bars are related to the wavelet-based localizer and Time-Freq-CPC, respectively.

Fig. 13. Performance comparison between proposed method and wavelet-based localizer.

- The performance on Event S5:
  - In Case 1, wavelet-based localizer performs significantly better when trained on R21 and almost the same when trained on R1 and R10.
  - In Cases 2 and 3, wavelet-based localizer performs significantly worse than the proposed method.

- The generalized performance on Event S59:
  - The wavelet-based localizer has a worse overall generalized performance, especially in Case 3. With a few exceptions: trained on R10 in Case 1, trained on R21 in Case 2.

Our metrics for evaluating the performance consist of the performance on the basic testing dataset that has the same origin as the training dataset (i.e., the performance on Event S5) and the generalized performance on an unseen dataset (i.e., the performance on Event S59). Intuitively, before discussing the generalized performance, the model should at least perform well enough on the basic testing dataset. Based on this logic, the proposed method has a better comprehensive performance than the wavelet-based localizer especially when the number of unlabeled data increases (i.e., in Case 2 and Case 3).

In addition, there are some points worth to be emphasized from the implementation perspective:

- 512 scale factors correspond to the maximum limit of our computing resources for preprocessing and training neural networks.
- Compared to the proposed method, the wavelet-based localizer needs much more computing resources.

VIII. Conclusion and Future Directions

In this paper, we demonstrate the necessity of joint time-frequency signal processing for underwater source localization and propose a self-supervised source localizer whose performance is evaluated on the SWellEx-96 Experiment. The comprehensive analysis demonstrates the proposed method is superior to alternative methods in terms of localization performance, the robustness of receiver-depth selection, and generalization capabilities on an unseen dataset. Two common scenarios (Case 2 and Case 3) for a permanent ocean monitoring system are discussed and the possibility of enhancing localization performance when more unlabeled data are available is demonstrated.

Future directions of research are:

- The possibility of combing the dataset collected at different ocean areas for effective underwater source localization in a wider ocean area based on the proposed method.
- The interpretation of the extracted features since the meaning of these features may provide benefits for the theoretical acoustic scientist to design a better simulation model.

Implementation code availability:

The whole implementation code will be available on GitHub (https://github.com/Xiaoyu-freshman) after the manuscript is published.

References

[1] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," The Journal of the Acoustical Society of America, vol. 146, pp. 3590–3628, October 2019.

[2] X. Cao, R. Togneri, X. Zhang, and Y. Yu, "Convolutional Neural Network With Second-Order Pooling for Underwater Target Classification," IEEE Sensors Journal, vol. 19, pp. 3058–3066, December 2018.

[3] W. Kong, J. Hong, M. Jia, J. Yao, W. Cong, H. Hu, and H. Zhang, "YOLOv3-DPFIN: A Dual-Path Feature Fusion Neural Network for Robust Real-Time Sonar Target Detection," IEEE Sensors Journal, vol. 20, pp. 3745–3756, December 2019.

[4] X. Cao, R. Togneri, X. Zhang and Y. Yu, "Convolutional neural network with second-order pooling for underwater target classification", IEEE Sensors Journal, vol. 19, pp. 3058-3066, April 2019.

[5] Y. Wang and H. Peng, "Underwater acoustic source localization using generalized regression neural network," The Journal of the Acoustical Society of America, vol. 143, pp. 2321–2331, April 2018.

[6] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," The Journal of the Acoustical Society of America, vol. 142, pp. 1176-1188, September 2017.

[7] Y. Lin, M. Zhu, Y. Wu, and W. Zhang, " Passive Source Ranging Using Residual Neural Network With One Hydrophone in Shallow Water," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), 2020, pp. 122-125.

[8] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1476-1485, 2019.

[9] I. Misra and L. Maaten, "Self-supervised learning of pretext-invariant representations," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707-6717, 2020.

[10] M. J. Bianco, S. Gannot, and P. Gerstoft, "Semi-supervised source localization with deep generative modeling," 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6, September 2020.

[11] M. J. Bianco, S. Gannot, E. Fernandez-Grande and P. Gerstoft, "Semi-Supervised Source Localization in Reverberant Environments With Deep Generative Modeling," IEEE Access, vol. 9, pp. 84956-84970, 2021.

[12] X. Zhu, H. Dong, P. Salvo Rossi, and M. Landrø, "Feature Selection Based on Principal Component Regression for Underwater Source Localization by Deep Learning," Remote Sensing, vol. 13, pp. 1486, April 2021.

[13] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, pp. 2, December 2020.

[14] X. Zhu, H. Dong, P. S. Rossi, and M. Landrø, "Self-supervised Underwater Source Localization based on Contrastive Predictive Coding," 2021 IEEE Sensors, 2021, pp. 1-4.

[15] J. Murray and D. Ensberg. The SWellEx-96 Experiment. http://swellex96.ucsd.edu/index.htm (accessed March, 2021).

[16] M. Chen, X. Shi, Y. Zhang, D. Wu and M. Guizani, "Deep Feature Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network," in IEEE Transactions on Big Data, vol. 7, pp. 750-758, September 2021.

[17] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, arXiv:1807.03748.

[18] M. A. Ranzato, Y. L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," Advances in neural information processing systems, vol. 20, pp. 1185-1192, 2007.

[19] C. I. Lai, "Contrastive predictive coding based feature for automatic speaker verification," 2019, arXiv:1904.01575.

Xiaoyu Zhu received the B.Sc. degree in atmospheric science from the Lanzhou University, China in 2016 and the M.Sc. degree in computer science and technology from the National University of Defense Technology, China in 2018, respectively. He is currently pursuing the Ph.D. degree in electronics and telecommunications with the Department of Electronic Systems, Norwegian University of Science and Technology (NTNU), Norway, within the Acoustics Research Group.

His research interests include underwater source localization and geoacoustic inversion based on machine learning.

Hefeng Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in physics and theoretical physics from the Northeast Normal University, China in 1983 and 1986, respectively, and the Ph. D. degree in geoacoustics from the Jilin University, China, in 1994.

From 1986 to 1994, she was a lecture of physics with the Northeast Normal University, China where she was associate professor from 1995 to 2000. She was visiting scholar and post doctorial fellow at the Norwegian University of Science and Technology, Norway between 1999 and 2001. From 2001 to 2002 she worked as a research scientist at the SINTEF Petroleum Research, Norway. Since 2002 she has been Professor in Acoustic Remote Sensing with the Norwegian University of Science and Technology, Norway. She was on sabbatical with the Underwater Acoustics Laboratory, University of Victoria, Canada, the College of Earth, Ocean and Environment, University of Delaware, USA, and the Laboratory of Mechanics and Acoustics, France in the periods of 2008-2009, 2014-2015, and 2019-2020, respectively. Her research interests include wave propagation, passive acoustics, geoacoustic modelling and inversion, and signal processing in ocean acoustics and seismic. Dr. Dong is a member of the Acoustical Society of America and IEEE.

Pierluigi Salvo Rossi (SM'11) was born in Naples, Italy, in 1977. He received the Dr.Eng. degree (*summa cum laude*) in telecommunications engineering and the Ph.D. degree in computer engineering from the University of Naples "Federico II", Italy, in 2002 and 2005, respectively. He is currently a Full Professor and the Deputy Head with the Department Electronic Systems, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He is also a part-time Research Scientist with the Department Gas Technology, SINTEF Energy Research, Norway.

Previously, he worked with the University of Naples "Federico II", Italy, with the Second University of Naples, Italy, with NTNU, Norway, and with Kongsberg Digital AS, Norway. He held visiting appointments with Drexel University, USA, Lund University, Sweden, NTNU, Norway, and Uppsala University, Sweden.

His research interests fall within the areas of communication theory, data fusion, machine learning, and signal processing. Prof. Salvo Rossi was awarded as an Exemplary Senior Editor of the IEEE COMMUNICATIONS LETTERS in 2018. He is (or has been) in the Editorial Board of the IEEE COMMUNICATIONS LETTERS, the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, the IEEE SENSORS JOURNAL, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Martin Landrø received an M.S. (1983) and Ph.D. (1986) in physics from the Norwegian University of Science and Technology.

From 1986 to 1989, he worked at SERES A/S. From 1989 to 1996, he was employed at IKU Petroleum Research as a research geophysicist and manager. From 1996 to 1998, he worked as a specialist at Statoil's research center in Trondheim. Since 1998, Landrø has been a professor at the Norwegian University of Science and Technology, Department of Petroleum Engineering and Applied Geophysics.

He received the Norman Falcon award from EAGE in 2000 and the award for best paper in GEOPHYSICS in 2001. In 2004 he received the Norwegian Geophysical award, and in 2007 Statoil's researcher prize. He received the SINTEF award for outstanding pedagogical activity in 2009. In 2010 he received the Louis Cagniard award from EAGE and in 2011 the Eni award (New Frontiers in Hydrocarbons). In 2012 he received the Conrad Schlumberger award from EAGE. Landrø's research interests include seismic inversion, marine seismic acquisition, and 4D and 4C seismic. This includes geophysical monitoring of CO2 storage. In 2014 he received the IOR award from the Norwegian Petroleum Directorate. He is a member of EAGE, SEG, The Norwegian Academy of Technological Sciences and The Royal Norwegian Society of Sciences and Letters.