Self-supervised Underwater Source Localization based on Contrastive Predictive Coding

Xiaoyu Zhu*, Hefeng Dong*, Pierluigi Salvo Rossi*, and Martin Landrø*
*Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway
Email: xiaoyu.zhu@ntnu.no; hefeng.dong@ntnu.no; pierluigi.salvorossi@ntnu.no; martin.landro@ntnu.no

2021 IEEE Sensors | 978-1-7281-9501-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/SENSORS47087.2021.9639566

***Abstract*—This work introduces a two-step self-supervised learning scheme, namely contrastive predictive coding (CPC), for underwater source localization. In the first step, a CPC-based self-supervised feature extractor is trained with the acoustic signals. In the second step, the encoder with frozen parameters is taken from the trained feature extractor and connected with a multi-layer perceptron (MLP) trained for source localization on a small labeled dataset. This approach is evaluated on a public dataset, SWellEx-96 Event S5, against an autoencoder (AE) scheme and a purely supervised scheme. The results indicate that the CPC scheme has the best performance and can extract the slow-changing features related to the source.**

***Index Terms*—Underwater source localization, contrastive predictive coding, self-supervised learning**

I. Introduction

Underwater source localization is an important task in underwater acoustics and many related fields. Due to the limitations (i.e. the requirement of high accuracy ocean environmental information and seabed parameters) of the conventional methods, machine learning (ML)-based methods have been applied recently since they do not require significant prior information [1]. However, most ML-based methods for underwater source localization are based on supervised learning scheme [2]–[4], which needs large amount of labeled data. Unfortunately, in real scenarios, the amount of labeled data is extremely limited due to the high cost and difficulty for data collection and labeling. This limitation has inspired methods based on self-supervised learning (SSL) [5]–[7]. SSL defines pretext tasks that are formulated using only unlabeled data to learn high-level semantic features [8]. Based on the learned features, a downstream task, such as underwater source localization in this work, can be solved by training a model based on a small labeled dataset.

We propose an SSL underwater source localization approach based on contrastive predictive coding (CPC) [9]. CPC can learn the high-level features from unlabeled data for the downstream tasks by predicting the future in latent space with a probabilistic contrastive loss. The performance of CPC-based methods has been demonstrated in the fields of image and speech processing [10], [11].

More specifically, a CPC-based feature extractor is trained on the acoustic signals collected as time series by a single hydrophone to learn high-level features. Then, the encoder with frozen parameters is taken from the trained feature extractor and connected with a 3-layer multi-layer perceptron (MLP) for the underwater source localization (together namely, Encoder-MLP). The Encoder-MLP is trained on a small labeled dataset with a purely supervised learning scheme. The performance of our approach is assessed based on a public dataset against two Encoder-MLPs with the same architecture trained by an autoencoder (AE) scheme (pixel-by-pixel reconstruction of the input) and a purely supervised scheme, respectively.

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).

II. Methodology

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

We use CPC [9] scheme of self-supervised learning to extract the high-level features from the acoustic time series. The architecture of the self-supervised feature extractor is the same as in [9], [12] and shown in Fig. 1.

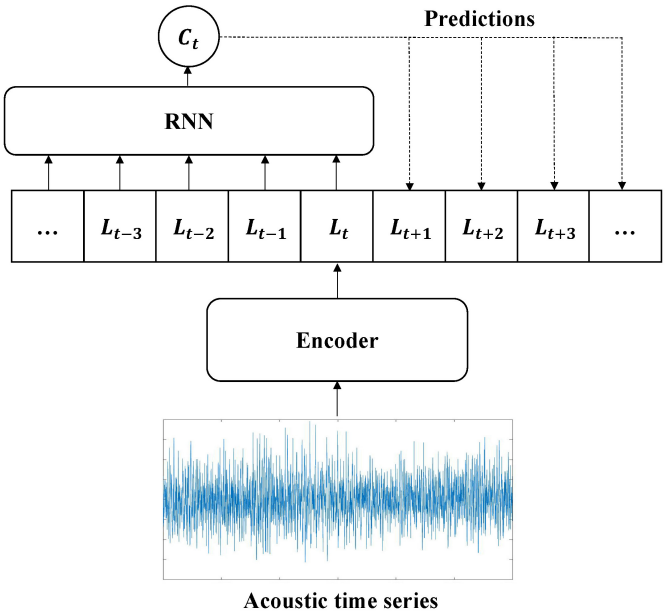

Fig. 1. Architecture of the self-supervised feature extractor

First, a non-linear encoder $f_{encoder}$ maps a raw acoustic time sequence into a sequence of latent features $L_t = f_{encoder}(x_t)$ with a lower temporal resolution. Next, a recurrent neural network (RNN) is chosen as an autoregression

model g_{ar} which summarizes all $L_{\leq t}$ in the latent space and produces a context latent representation $C_t = g_{ar}(L_{\leq t})$.

The goal of CPC is to predict the future k time steps by modeling a density ratio which preserves the mutual information between x_{t+k} and C_t:

$$f_k(x_{t+k}, C_t) = exp(L_{t+k}^T W_k C_t) \quad (1)$$

where a linear transformation $W_k C_t$ is used for the prediction with a different W_k for every timestep k.

The encoder and RNN are trained jointly to optimize a loss function based on Noise-Contrastive Estimation (NCE) [12] for maximizing the mutual information.

Note that, according to some configuration experiments, the predicting timestep is chosen as $k = 16$ in this work.

B. The Encoder-MLP for Underwater Source Localization

After training the self-supervised feature extractor, the parameters of the encoder are frozen. The encoder is connected with a 3-layer MLP for the underwater source localization. Since the source localization task is a regression task, the mean squared error (MSE) is chosen as the loss function. The architecture of the Encoder-MLP is shown in Fig.2 where the arrows indicate the direction of the data stream.

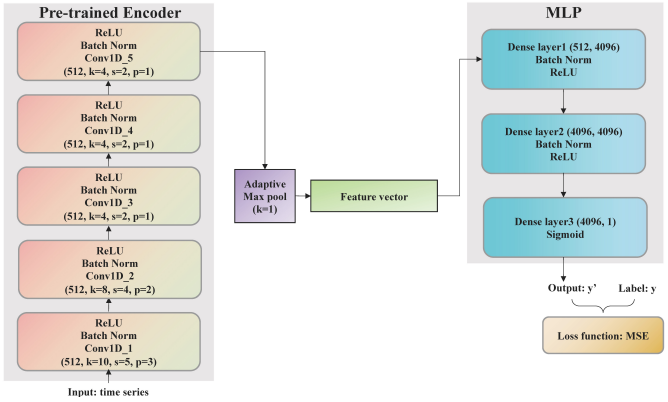

Fig. 2. Architecture of the Encoder-MLP

III. Experiments

We assess the source localization performance of 3 Encoder-MLPs with the same architecture trained by CPC, AE, and purely supervised schemes.

A. Dataset and Preprocessing

Vertical linear array (VLA) data from SWellEx-96 Event S5 are used to asses the localization performance. The sampling rate of the acoustic data was 1500 Hz and the recording time of the data was 75 minutes. The VLA contained 21 receivers equally spaced between depth 94.125 m and 212.25 m. Furthermore, the horizontal range between the source and the VLA was also provided in the dataset. More detailed information of this dataset can be found in [7].

In this paper, the acoustic signals (time series) collected by a single receiver are directly used to train the models. The acoustic signals are cut into slices (4 seconds per slice) and arrange into a signal matrix $\boldsymbol{X}$ format with the shape of 1125×6000, where each row is related to once slice. More specifically, 1125 is the total number of time steps and 6000 is the length of each slice. The horizontal range between the source and the VLA can be represent by a label vector $\boldsymbol{y}$ with the shape of 1125×1.

For the training stability, standardization and min-max scaling (scaling into interval $(0, 1)$) are applied on the signal matrix $\boldsymbol{X}$ and the label vector $\boldsymbol{y}$, respectively.

In real scenarios, the number of labeled data could be extremely limited. To mimic this situation, only 12.5% labeled data are used to build the training dataset for the Encoder-MLP.

To show the influence of different receiver-depths, receivers no. 1 (top), no. 10 (middle), and no. 21 (bottom) are chosen to build the dataset, respectively.

B. Training Strategy and Hyperparameters

The strategy of training the Encoder-MLP for source localization consists of two steps. In the first step, the self-supervised feature extractor is trained based on the purely acoustic signals collected during 75 minutes without any labels for extracting the high-level features. In the second step, the parameters of the trained encoder are frozen and the Encoder-MLP is trained based on the small labeled dataset for source localization.

The learning rates for first and second steps are 1×10^{-4} and 5×10^{-5}, respectively. The optimization scheme is Adam [13]. The epoch and the batch-size are 100 and 50 for each step, respectively.

C. Performance Analysis

To make a comprehensive comparison, 3 Encoder-MLPs are tested on the data collected by all receivers and trained separately based on the data collected by receivers no. 1, no. 10, and no. 21.

The performance of source localization is shown in Fig. 3 where the metric is MSE. In the figure, the blue, orange, and gray bars are related to the Encoder-MLPs trained based on CPC, AE, and purely supervised schemes, respectively. 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.

From Fig. 3, interesting phenomena can be found:

- The CPC based model shows the best performance, followed by the purely supervised learning and AE based models. This illustrates the advantage of the CPC scheme, which can learn better features than AE. This is because there is no guarantee that the pixel-by-pixel reconstruction of the input is a good metric for learning generalized features [14].
- There is a slight difference in performance among different receiver-depths. Trained on the data from R21, the performance is the worst. The similar phenomenon was found in our previous work [7].

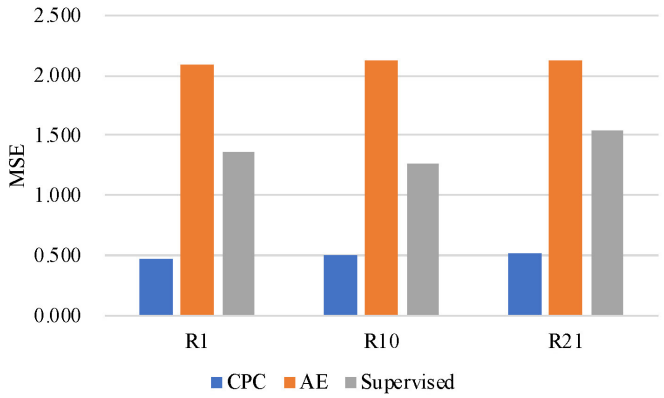

Fig. 3. Performance of Encoder-MLPs based on CPC, AE, and purely supervised schemes.

D. Importance of Slow-changing Features

To have an intuitive understanding of the advantages of CPC compared to AE, Fig. 4 shows the comparison of localization results between Encoder-MLPs based on CPC and AE. The models are trained on receiver no. 1 and tested on receiver no. 2.

In the figure, the abscissa indicates the time during the whole event, and the ordinate indicates the horizontal range between the source and the VLA. In the legend, the blue and red dots express the prediction and target of the range, respectively.

From Fig. 4, phenomena can be found:

- In the plot of AE, there is a wide ambiguity interval of prediction from 3 to 7 km during the first 40 minutes. This time period relates to the far-field condition, which means that the high-frequency components of ship noise cannot provide enough contribution for the signal and the source-related contribution will dominate the signal. More details of the spectrum analysis can be found in [15].
- From the acoustic perspective, the model based on AE cannot extract the features related to the source directly from the acoustic signal in the time domain. However, in the plot of CPC, there is no such ambiguity interval. It means that the model based on CPC can extract the features related to the source. Based on the spectrum analysis, the frequency of the source is lower than that of ship noise. The lower frequency component corresponds to the slower changing feature in the time domain. This shows some consistency with the characteristic of CPC which is aiming to extract the slow-changing features [9].

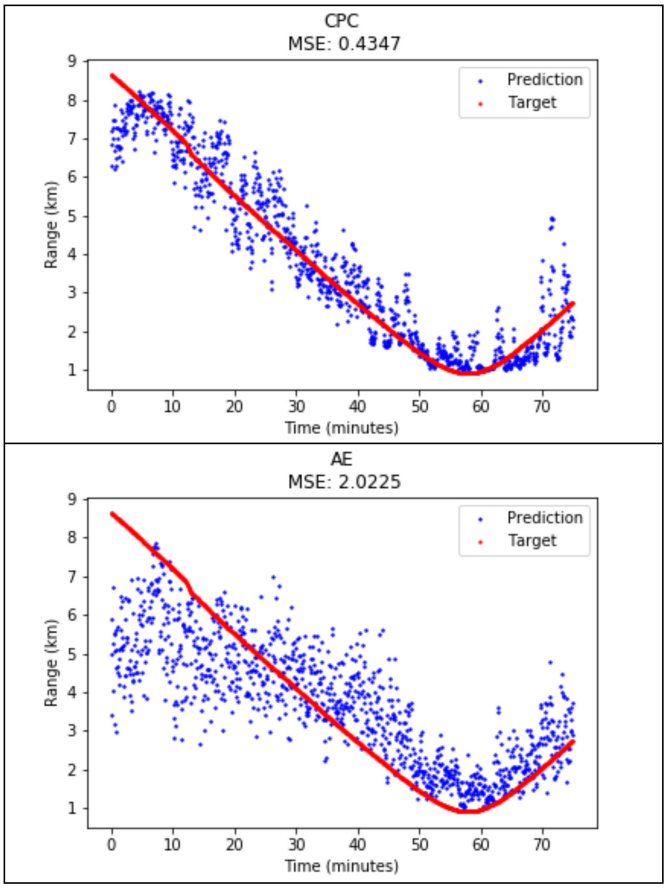

Fig. 4. Comparison of the source localization results between CPC (upper) and AE (bottom).

IV. Conclusions

This work investigates the application of contrastive predictive coding (CPC) for underwater source localization based on self-supervised learning. The CPC scheme is assessed on a public dataset, SWellEx-96 Event S5, against an autoencoder (AE) and purely supervised schemes. To mimic real scenarios, only 12.5% labeled data are used to build the training dataset for the Encoder-MLP. According to the performance analysis, the CPC-based Encoder-MLP shows the best performance among different receiver-depths. This can be explained from the acoustic perspective that the CPC can extract the slow-changing features corresponding to the contribution of the source from the acoustic signals in the time domain.

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] R. Lefort, G. Real, and A. Drémeau, "Direct regressions for underwater acoustic source localization in fluctuating oceans," Applied Acoustics, vol. 116, pp. 303–310, January 2017.

[3] 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.

[4] Y. Liu, H. Niu, and Z. Li, "Source ranging using ensemble convolutional networks in the direct zone of deep water," Chinese Physics Letters, vol. 36, pp. 044302, January 2019.

[5] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 1393-1407, August 2016.

[6] 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.

[7] 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.

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

[10] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," Knowledge-Based Systems, vol. 224, pp. 107090, April 2021.

[11] 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.

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

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[14] 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.

[15] Y. J. Du, Z. W. Liu, and L. G. Lü, "Range Localization of a Moving Source Based on Synthetic Aperture Beamforming Using a Single Hydrophone in Shallow Water," Applied Sciences, vol. 10, pp. 1005, February 2020.