A DEEP NEURAL NETWORK BASED METHOD OF SOURCE LOCALIZATION IN A SHALLOW WATER ENVIRONMENT

Zhaoqiong Huang[1,2], *Ji Xu*[1,2], *Zaixiao Gong*[2,3], *Haibin Wang*[2,3], *Yonghong Yan*[1,2]

[1] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences

ABSTRACT

This paper applies deep neural network (DNN) to source localization in a shallow water environment because of its powerful modeling capability and the little dependence on the prior knowledge of environmental parameters. The classical two-stage scheme is adopted, in which feature extraction and DNN analysis are independent steps. It firstly extracts the input feature from the observed signal received by underwater hydrophones. The eigenvectors associated with the modal signal space are decomposed from the covariance matrices of the data field at different frequencies, which are used as the input feature of DNN. The time delay neural network (TDNN) is exploited to model the long term feature representation and construct the regression model. The output is the source range-depth estimate. Several experiments using simulation and experimental data are conducted to evaluate the performance of the proposed method. The results demonstrate the effectiveness and potential of DNN for source localization. Particularly, experiments show that simulation data can be merged to train a general model for experimental data when lacking of sufficient training data in real-world environment.

Index Terms— Source localization, shallow water environment, modal signal space, time delay neural network.

1. INTRODUCTION

Source localization in a shallow water environment has attracted a lot of attention of many scholars in the past several decades. Matched field processing (MFP) is a famous model based method for underwater source detection and localization [1]–[9]. MFP calculates the replicas using the propagation model, then the location, where the modeled field best matches with the experimental field, is taken as the source location estimation. MFP usually requires the accurate information of environmental parameters to calculate the modeled field. However, the environmental parameters are usually variant and imprecise, which may lead to incorrect or inaccurate localization results.

To reduce the dependence on the environmental information, many data-based localization methods are presented. The array/waveguide invariant is proposed for robust source-range estimation [10]–[12]. Machine learning is a famous data-driven technique, which has been introduced to source localization [13]–[17]. These methods can localize source successfully by taking the source localization as a classification or regression task, however, they are commonly based on conventional classifiers or shallow feed-forward neural networks (FNNs) [14]–[18]. Although the convolutional neural network (CNN) is taken for passive acoustic ranging [19], it is designed for the near-field scenario using a single sensor, so that the time delay between the direct and indirect sound propagation path can be measured. To our best knowledge, few methods are reported for source localization based on deep neural network (DNN) using underwater multi-sensor arrays for far-field scenario. Compared to shallow FNNs, DNN [20, 21] is advantageous to represent the complex nonlinear relationship. In this paper, we take advantage of DNN to estimate the wide-band source location using a vertical linear array (VLA) in a shallow water environment.

The proposed method adopts the two-stage scheme that incorporates feature extraction and DNN analysis. The eigenvectors associated with the modal signal space decomposed from the covariance matrices of the data field at different frequencies are taken as the input feature of DNN. Then, the estimates of range and depth are given by regression network. Since the time delay neural network (TDNN) [22] ic capable of modeling the temporal dynamics in sound signal, it is used as the basic network architecture.

In contrast to previous methods, there are two major characteristics for the proposed method. First, DNN directly learns the mapping relationship from the original data, rather than construct the acoustic model in advance. Second, the simulation data is proven to be a feasible alternative for our method in real application, where training data collection is very costly. It enables sufficient training data to guarantee the performance of the trained model. Our experiments have shown the model trained by simulation data can also achieve a good performance on real experimental data.

2. SIGNAL MODEL

Let's consider a single wide-band sound source impinges on a VLA of K sensors in a far-field scenario. The source location is denoted as a two-dimensional vector, (r_s, z_s). Using the matrix notation, the pressure field received by the sensors is described as [9]

$$\boldsymbol{P} = \boldsymbol{H}\boldsymbol{S} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{S} \in \mathbb{C}^M$ with $S_m(r_s, z_s) = a(2\pi/k_m r_s)^{1/2}\Psi_m(z_s)e^{jk_m r_s}$, a denotes the complex Gaussian random amplitude of the source, M ($M < K$) denotes the mode number in the water column (higher

This work is partially supported by the National Natural Science Foundation of China (Nos. 11590770-4) and the Innovation Foundation of Chinese Academy of Sciences (No. CXQZ201701).

modes are treated as noise), k_m^2 is the eigenvalue associated with the mth mode, $\boldsymbol{N} \in \mathbb{C}^K$ denotes the additive noise, and $\boldsymbol{H} \in \mathbb{C}^{K\times M}$ with $H_{m,k} = \Psi_m(z_k)$. $\boldsymbol{P} = [P_1, \ldots, P_K]^T \in \mathbb{C}^K$, where $(\cdot)^T$ denotes the transpose operation and the pressure field at depth z_k due to a source at (r_s, z_s) is represented as

$$P_k = \sum_{m=1}^{M} H_{m,k} S_m(r_s, z_s). \tag{2}$$

3. PROPOSED METHOD

The proposed method comprises two modules: 1) feature extraction and 2) DNN analysis. The feature extraction module extracts the eigenvectors from the observed acoustic data. The DNN analysis module constructs the one-to-one mapping between the eigenvectors and the source locations. Feature extraction and DNN analysis are mutually independent.

3.1. Feature extraction

Based on the aforementioned signal model, the covariance matrix at a single frequency over D snapshots is expressed as

$$\begin{aligned} \boldsymbol{R}(f) &= \frac{1}{D}\sum_{d=1}^{D} \boldsymbol{P}_d(f)\boldsymbol{P}_d^{+}(f), \\ &= \boldsymbol{H}\boldsymbol{R}_S(f)\boldsymbol{H}^{+} + \boldsymbol{R}_N(f), \end{aligned} \tag{3}$$

where f denotes the frequency, $(\cdot)^+$ denotes the Hermitian transpose, and $\boldsymbol{R}_S(f)$ and $\boldsymbol{R}_N(f)$ are the covariance matrices of the signal and noise. Applying eigenvalue decomposition (EVD) to $\boldsymbol{R}(f)$,

$$\begin{aligned} \boldsymbol{R}(f) &= \boldsymbol{\Lambda}_f \boldsymbol{\Sigma}_f \boldsymbol{\Lambda}_f^{+} \\ &= \boldsymbol{\Lambda}_f^S \boldsymbol{\Sigma}_f^S \boldsymbol{\Lambda}_f^{S+} + \boldsymbol{\Lambda}_f^N \boldsymbol{\Sigma}_f^N \boldsymbol{\Lambda}_f^{N+}, \end{aligned} \tag{4}$$

where the eigenvectors and eigenvalues are obtained as $\boldsymbol{\Lambda}_f = [\mathbf{e}_{f,1}, \ldots, \mathbf{e}_{f,K}] \in \mathbb{C}^{K\times K}$ and $\boldsymbol{\Sigma}_f = diag[\lambda_{f,1}, \ldots, \lambda_{f,K}]$, where the eigenvalues are sorted in descending order. $\boldsymbol{\Sigma}_f^S$ and $\boldsymbol{\Lambda}_f^S = [\mathbf{e}_{f_i,1}, \ldots, \mathbf{e}_{f_i,M'}] \in \mathbb{C}^{K\times M'}$ are eigenvalues and eigenvectors corresponding to the modal signal space and $\boldsymbol{\Sigma}_f^N$ and $\boldsymbol{\Lambda}_f^N$ corresponds to the modal noise space.

Comparing (3) with (4), it can be seen that the M dominant eigenvectors of the covariance matrix span the same space as the columns of $\boldsymbol{H}$ if the modes are sampled sufficiently. The eigenvectors associated with larger eigenvalues span the modal signal space while the remaining eigenvectors span the modal noise space. Note that the eigenvectors of the modal signal space may not correspond to the lowest-order normal modes exactly ($M' \leq M$), if some mode amplitude functions are not activated. The eigenvectors with relative larger eigenvalues ($\boldsymbol{\Lambda}_f^S$), which are considered to be the main feature of the propagating modes of an assumed source location, are used as the input feature of DNN. The remaining eigenvectors are disregarded to suppress the noise.

3.2. DNN analysis

TDNN [22] is selected to construct the functional transformation between the eigenvectors and the source locations, because it has the capability of modeling the temporal dynamics of the feature representation. The architecture of TDNN with one hidden layer is shown in Fig. 1(a). The figure shows the current estimate is not only determined by current input feature but also its adjacent features. The temporal context information is collected by each TDNN unit and the higher layers have the ability to learn wider temporal relationship. The dependencies across layers are localized in time. For the specific TDNN shown in Fig. 1, units from $t-1$ to $t+1$ are spliced at the input layer and the hidden layer. The output of the tth moment depends on $t-2$ to $t+2$ frames in terms of the whole framework. The output of each unit at all layers is obtained by computing the weighted sum of its inputs and passing this sum through a nonlinear function. Such connection is unfolded as Fig. 1(b), which can be viewed as a fully-connected architecture.

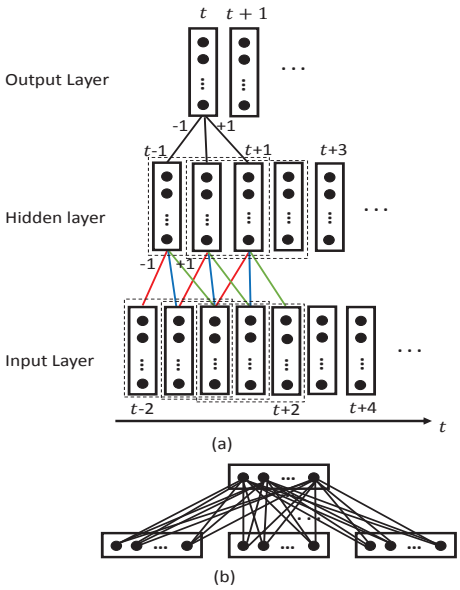

Fig. 1. Architecture of TDNN. (a) TDNN with one hidden layer. (b) A fully-connected architecture. Connections with the same color (red, blue, or green) share the same weights.

The network parameters are updated by minimizing the mean square error (MSE) objective function, given by

$$E = \frac{1}{L}\sum_{l=1}^{L}\left[(r_l - r_l^{'})^2 + (z_l - z_l^{'})^2\right]. \tag{5}$$

where $(r_l^{'}, z_l^{'})$ and (r_l, z_l) denote the reference and estimated range and depth respectively and L denotes the sample number of each batch. Note that the source locations are expressed by range in kilometer and depth in meter.

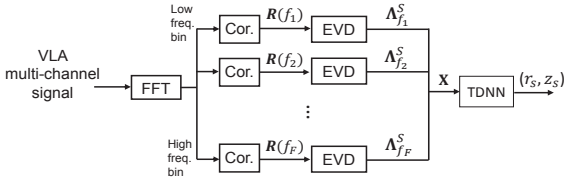

Fig. 2. Block diagram of the proposed method.

3.3. Implementation

Since the extracted eigenvectors are complex values, they cannot be directly addressed by real-valued neural network. Here, the complex values are considered to be two-dimensional real values. The real and imaginary part of the eigenvectors at different frequencies are concatenated as the input vector **x**,

$$\mathbf{x} \triangleq \bigcup_i \bigcup_m \Big[\mathcal{R}(\mathbf{e}_{f_i,m}), \Im(\mathbf{e}_{f_i,m})\Big],\ i = 1, \ldots, F,\ m = 1, \ldots, M^{'}, \tag{6}$$

where i denotes the frequency index. The block diagram of the proposed method is shown in Fig. 2.

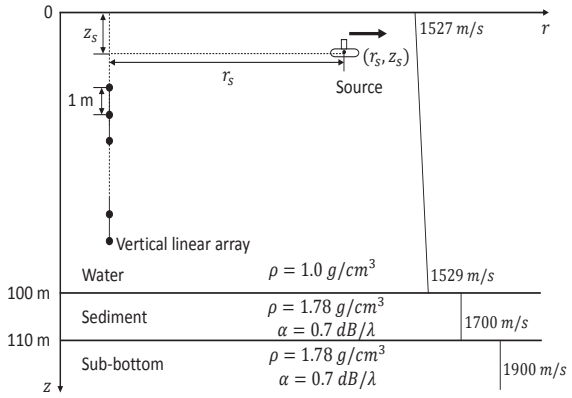

Fig. 3. Schematic diagram of the simulated acoustic environmental model.

4. EVALUATIONS

4.1. Simulation setup

4.1.1. Acoustic environmental model

The simulations were conducted to evaluate the performance of the proposed method. The schematic diagram of the simulated environment was illustrated in Fig. 3. The VLA consisted of 30 hydrophones spanning $30 - 60$ m depth with uniform inter-sensor spacing 1 m. The depths of water and the sediment layer were 100 m and 10m. The sound speed increased from 1527 m/s at the top of the water column to 1529 m/s at the bottom.

4.1.2. Data description

Acoustic data was simulated using KRAKEN. The bandwidth of simulation signal was $[50, 1000]$ Hz and the sampling rate was 6000 Hz. The sources incorporated near-surface vessels and underwater targets with source level (SL) 120 dB (at 1000 Hz). All sources moved away from the array ranging from 10 to 28.5 km, while the near-surface vessels with depth from 1.5 to 8.5 m and underwater targets form 28 to 35 m. The white Gaussian noise was used as the ambient noise, which was artificially added to the source signal. Both SL and noise level (NL) were attenuated by -6 dB/Oct. The signal-to-noise rate (SNR) of a single hydrophone at different ranges can be approximately calculated by SL and NL as

$$SNR(f)\ (dB) = SL(f) - 60 - 10\log_{10}(\frac{r}{r_0}) - NL(f), \tag{7}$$

where r is the current source range and $r_0 = 1000$ m is the reference range ($r \geq r_0$). The transmission loss decreases with depth going deeper.

4.1.3. Parameters for feature extraction

The signal transformed to frequency domain by operating fast Fourier transformation (FFT). The frame length was about 0.6827 s. The bandwidth used for feature extraction was set to $[100, 300]$ Hz. Ten eigenvectors at sixteen frequency bins were extracted as the input feature, thus the feature per frame included 9600 ($30 \times 10 \times 16 \times 2$) dimension.

4.1.4. Parameters for TDNN

The configuration of TDNN was that of 8 layers (one input layer + six hidden layers + one output layer) with 1024 hidden nodes. The units were spliced from $t - 1$ to $t + 1$ at the input layer and the second hidden layer. No frame was spliced at the first, and third to fifth hidden layer. Current output was determined by the inputs from $t - 2$ to $t + 2$ (totally 5 frames) in terms of the whole framework.

The simulation data were divided into two parts: training set and test set. There were totally $1,582,200$ training samples and $158,220$ test samples, which were mutually different. The rectified linear units (ReLU), $f(x) = \max(0, x)$, was used as the activation function [23]. The Kaldi [24] toolkit was adopted for TDNN training. TDNN was optimized using the back propagation (BP) algorithm [25] with stochastic gradient descent (SGD) in a mini-batch mode [26]. The initial learning rate was 0.001 and the batch for SGD was 512.

4.1.5. Parameters for MFP

The conventional MFP [7] was taken as the competing method. The grid resolution for calculating the modeled field was chosen to 10 m in range and 0.5 m in depth, which was set the same for search grid. For the sake of fairness, MFP made use of 5 frames (0.6827 s per frame) to calculate the final estimator output and the bandwidth used was $[100, 300]$ Hz. Global maximum in the ambiguity surface indicated the best estimate of source location.

4.2. Simulation results

The simulation investigated the effectiveness of the proposed method under various NLs. The NL were set to 25, 45, and 65 dB (at 1000 Hz). There were three models respectively trained using near-surface vessels and underwater targets with NL=25, 45, and 65 dB, then the tested sources with different NLs were decoded by the corresponding model trained for the coincident NLs. For example, tested source under NL=25 dB was decoded by the model trained for NL=25 dB. The objective evaluation metrics used were the mean absolute error (MAE) and the mean relative error (MRE),

$$MAE = \frac{1}{Q}\sum_{q=1}^{Q}|x_q - x_q^{'}|, \tag{8}$$

$$MRE = \frac{1}{Q}\sum_{q=1}^{Q}\Big|\frac{x_q - x_q^{'}}{x_q^{'}}\Big| \times 100\%, \tag{9}$$

where x represents the estimation value and $x^{'}$ represents the reference value. Q is the sample number.

Table 1. MAE and MRE comparison under different NLs.

Method	NL (dB)	Depth (m)	Range (km)
TDNN	25	**0.34**	**0.04 (0.2%)**
	45	**0.39**	**0.14 (0.7%)**
	65	**0.55**	**0.18 (1.0%)**
MFP	25	1.13	0.07 (0.3%)
	45	1.31	0.17 (0.7%)
	65	11.3	8.02 (43.7%)

The MAE of range and depth and MRE of range (in the brackets) are averaged over all tested sources (including near-surface vessels and underwater targets). The results of the proposed method (called TDNN) and MFP are shown in Table 1, one finds that MAE and MRE of two methods consistently achieve a relative low error when NL=25 and 45 dB. Nevertheless, the performance degrades when NL increases, e.g. NL=65 dB, particularly serious for MFP. It indicates that, for the proposed method, the eigenvectors are disrupted by environmental noise, so the features can not well represent the propagating modes of the source. For MFP, it can estimate depth and range of sources accurately under favorable environments, however, the modeled field fails to be matched with the experimental field when SNR is low. Overall, TDNN outperforms MFP in all tested conditions, especially when noise becomes serious, which reveals that the proposed method can achieve a better performance under adverse environment. Besides, it should be noted that the proposed method can give the reliable estimates as long as the range and depth of the test data are within that of the training data.

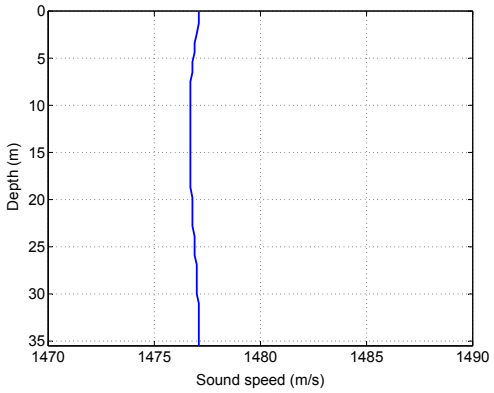

Fig. 4. Sound speed profile for the real environment.

4.3. Experimental results

The simulations showed that the proposed method and MFP can estimate source location when environmental parameters were certain. In fact, the environmental information was not always precise. The real environmental data collected in March 1999 in the Yellow Sea were used for testing. The data were recorded by a VLA with 16-element hydrophones spanning $0.5-30.5$ m depth with inter-sensor spacing 2 m. The surface vessel traveled toward the sensors from 11.5 km away from the sensor and lasted about 15 minutes. The sound speed profile (SSP) measured in the experiment is shown in Fig. 4, where the water depth was about 35.5 m. The hydrophone sampling rate was 12 kHz. The bandwidth used for feature extraction and MFP was $[100, 150]$ Hz. The configuration of TDNN is identical to that described in Sec 4.1.4.

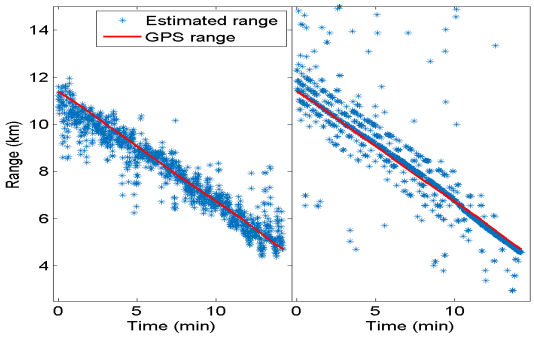

Fig. 5. Source ranging using experimental data. The left figure shows the results of TDNN and the right shows the results of MFP.

Table 2. MAE and MRE of range estimation for experimental data.

Method	MAE	MRE
TDNN	**0.41**	**5.3%**
MFP	0.52	6.4%

For DNN method, since we did not have adequate data that recorded in the same environment as training set, the data were simulated as training data. Simulation data under various NLs were generated in order to cover the real case because NL was unknown for real data. The training set included $1,107,000$ samples. As only range varies with time, the range estimates are shown in Fig. 5. From this figure, we can observe that both TDNN and MFP can localize the source accurately in general. The MAE and MRE between the estimated range and GPS range are summarized in Table 2, the results show that TDNN achieves a better accuracy than MFP. It should be mentioned that the proposed method may achieve a better performance if there are another few experimental data in the same environment for training.

From this experiment, the results demonstrate that simulation data is helpful when training data are insufficient. The model trained by simulation data can also achieve a fairly good performance on experimental data.

5. CONCLUSIONS

This paper proposes a novel approach to localize source in a shallow water environment by utilizing DNN. In summary, our contributions are two-fold: (i) We applied TDNN to source localization task, and the experimental results show the effectiveness of the proposed method for source localization. (ii) Simulation data are available for source localization when laking of real-environment training data. Simulation data in close environments can be merged to train a general model. The general model can still achieve a fairly good performance on experimental data, as long as the tested condition is covered by that of simulation data. They promotes the proposed method to be deployed in a wider range of situations in real-world environment.

6. REFERENCES

[1] A. Tolstoy, "Matched Field Processing for Underwater Acoustics," Singapore: World Scientific, 1993.

[2] G. R. Wilson, R. A. Koch, and P. J. Vidmar, "Matched mode localization," *J. Acoust. Soc. Am.*, vol. 84, no. 1, 1998, pp. 310–320.

[3] E. K. Westwood, "Broadband matched-field source localization," *J. Acoust. Soc. Am.*, vol. 91, no. 1, 1992, pp. 2777–2789.

[4] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng*. vol. 18, no. 4, 1993, pp. 401–424.

[5] G. B. Smith, C. Feuillade, D. R. Del Balzo, and C. L. Byrne, "A nonlinear matched field processor for detection and localization of a quiet source in a noisy shallow-water environment," *J. Acoust. Soc. Am.* vol. 85, no. 3, 1989, pp. 1158–1166.

[6] Z. -H. Michalopoulou and M. B. Porter, "Matched-field processing for broad-band source localization," *IEEE J. Ocean. Eng.*, vol. 21, no. 4, 1996, pp. 384–392.

[7] R. Zhang, Z. Li, J. Yan, Z. Peng, and F. Li, "Broad-band Matched-Field Source Localization in the East China Sea," *IEEE Journal of Oceanic Eng.*, vol. 29, no. 4, 2004, pp. 1049–1054.

[8] S. E. Dosso and M. J. Wilmut, "Maximum-likelihood and other processors for incoherent and coherent matched-field localization," *J. Acoust. Soc. Am.*, vol. 132, no. 4, 2012, pp. 2273–2285.

[9] C. L. Byrne, R. T. Brent, C. Feuillade, and D. R. DelBalzo, "A stable data-adaptive method for matched-field array processing in acoustic waveguides," *J. Acoust. Soc. Am.*, vol. 87, no. 6, 1990, pp. 2493–2502.

[10] A. M. Thode, "Source ranging with minimal environmental information using a virtual receiver and waveguide invariant theory," *J. Acoust. Soc. Am.*, vol. 108, no. 4, 2000, pp. 1582–1594.

[11] C. Cho, H. C. Song, and W. S. Hodgkiss, "Robust source-range estimation using the array/waveguide invariant and a vertical array," *J. Acoust. Soc. Am.*, vol. 139, no. 1, 2016, pp. 63–69.

[12] H. C. Song and C. Cho, "Array invariant-based source localization in shallow water using a sparse vertical array," *J. Acoust. Soc. Am.*, vol. 141, 2017, pp. 183–188.

[13] S. -C. Chan, K. -C. Lee, T. -N. Lin, and M. -C. Fang, "Underwater positioning by kernel principal component analysis based probabilistic approach," *Appl. Acoust.*, vol. 74, no. 10, 2013, pp. 1153–1159.

[14] R. Lefort, G. Real, and A. Drémeau, "Direct regressions for underwater acoustic source localization in fluctuating oceans," *Appl. Acoust.*, vol. 116, 2017, pp. 303–310.

[15] J. M. Ozard, P. Zakarauskas, and P. Ko, "An artificial neural network for range and depth discrimination in matched field processing," *J. Acoust. Soc. Am.*, vol. 90, no. 5, 1991, pp. 2658–2663.

[16] P. Zakarauskas, J. M. Ozard, and P. Brouwer, "Neural networks for independent range and depth discrimination in passive acoustic localization," *IEEE Transactions on Signal Processing*, vol. 41, no. 3, 1993, pp. 1394–1398.

[17] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," *J. Acoust. Soc. Am.*, vol. 142. no. 3, 2017, pp. 1176–1188.

[18] H. Niu, E. Ozanich, and P. Gerstoft, "Ship localization in Santa Barbara Channel using machine learning classifiers", *J. Acoust. Soc. Am.*, vol. 142, EL455-460, 2017.

[19] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. T. Jin, "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2657–2661.

[20] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, 2006, pp. 1527–1554.

[21] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, 2015, pp. 85–117.

[22] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, 1989, pp. 328–339.

[23] X. Glorot, A. Bordes , and Y. Bengio, "Deep Sparse Rectifier Neural Networks", in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 15, 2011, pp. 315–323.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. M. Qian, P. Schwarz and et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, 1986, pp. 533–536.

[26] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," In *International Conference on Learning Representations: Workshop track*, 2015.