IET Radar, Sonar & Navigation

Research Article

IET Journals
The Institution of Engineering and Technology

Underwater target classification at greater depths using deep neural network with joint multiple-domain feature

ISSN 1751-8784
Received on 28th February 2018
Revised 8th October 2018
Accepted on 5th November 2018
E-First on 11th December 2018
doi: 10.1049/iet-rsn.2018.5279
www.ietdl.org

Xu Cao[1] ✉, Xiaomin Zhang[1], Roberto Togneri[2], Yang Yu[1]

[1]*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China*
[2]*School of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth, WA 6009, Australia*
✉ *E-mail: caoxu@mail.nwpu.edu.cn*

Abstract: For underwater target classification which is supposed to recognise different ships with the radiated acoustic signal, it is the most challenging task to provide excellent classification accuracy in a variety of environments. However, most of the existing systems are optimised to get the best performance on the data set from certain situations which they are trained in, which may lead to generalisation risks when applied to new environments. Here, the authors introduce an underwater target classification framework using a deep neural network to learn deep features from a large joint multiple-domain input. The authors propose to incorporate spectral and wavelet domain information with different resolutions to grasp the 'global' structure and the 'local' transient variation of the raw radiated signals. In contrast to shallow models, a stacked sparse autoencoder (SSAE) model, which is composed of multiple hidden layers and a softmax classifier, is adopted to learn more discriminating features for classification. In the authors' experiments, the proposed SSAE model is evaluated on the data set consisting of underwater acoustic signal received at different ocean depths. The authors' results show that the proposed SSAE model with joint input features achieved a 5% improvement in classification accuracy compared to the state-of-the-art DBN approach.

1 Introduction

Underwater target classification using passive sonar is one of the most important topics for underwater acoustic signal processing, which is widely used in ocean engineering and underwater detection. The goal of underwater target classification is to determine whether the received underwater acoustic signal belongs to a certain kind of vessel or not. Usually, an underwater target classification system is based on three stages, including feature extraction of raw radiated acoustic signal of vessels, feature selection, and object classification. However, solutions for underwater target classification are far from being practically useful, especially in real-ocean environment. This is because of the lack of any prior knowledge about the acoustic information of the target, the unexpected variations of the ocean environment, and the reverberation influence of the spreading channel. These factors contribute to make the classification process a complex problem. For ocean detection, the classification system is not only required to recognise vessels more correctly, but also is supposed to be adaptive enough to different conditions. The current paper is mainly concerned with the generalisation ability of underwater target classification. The request is that, when applied to unknown conditions, a trained system is capable of correctly identifying the vessel class of the radiated noise.

Several classification methods have been proposed to deal with the vessel classification task with passive sonar, using a variety of feature extraction and pattern recognition methods. In [1], a neural network classifier is adopted to identify the radiated noise received by a hydrophone that was far from the ship using the averaged spectral information. In [2], the wavelet transform (WT) is used to extract tonal features from the average power spectral density (PSD) while two neural network classifiers are used to evaluate the classification results. In [3], a new configuration of probabilistic neural network (PNN) called multi-spread PNN (MSPNN) is proposed for marine vessel classification with the features of autoregressive (AR) model. In [4, 5], different feature extraction approaches using wavelet packet demonstration (WPT) are developed. In [6], the cepstral features and the average cepstral features are extracted, which is able to reduce the multipath distortion effects of shallow underwater channel. In [7], the multiple binary MLP classifiers are adopted to construct the class-modular multi-layer perceptron network for classification.

In the past decade, SVMs have been shown to achieve on par, and in some literature, competitive performance in practical situations. In [8], the feature extraction method of line spectrum and classification algorithm of SVM are adopted for recognition of ship radiated noise. The algorithm is proved to reduce the computational complexity and effectively identify the objects. In [9], a support vector machine (SVM) classifier is optimised using the BAT algorithm to work on the acoustic signature captured by the hydrophone. In [10], Sherin and Supriya propose to use a genetic algorithm (GA) to explore the optimal parameters of an SVM classifier with mel-frequency cepstral coefficients (MFCC) features. The performance of the multiclass SVM classifier is improved by incorporating an automatic classifier parameter selection using the GA. The advantage of using an SVM is that the SVM has outstanding model capacity matching the training data complexity, especially when trained using a small number of samples. However, the SVM is essentially a shallow architecture which depends mostly on the prior knowledge of the data set. The shallow architecture has a limited representation and generalisation ability, which may cause performance degradation when applied in the complex real-ocean environment, such as target classification using samples at a different ocean depth to that trained on.

Recently, we have seen the huge potential of the deep learning architecture to hierarchically learn deep representations with models consisting of multi-hidden layers [11]. In contrast to shallow architectures, deep learning can generate high-level features learned from low-level ones, which are more abstract and invariant. For classification tasks, high layer of features leads to enhancement of the discriminant parts of the input [12]. Before deep learning hand-crafted features had to be extracted, these features are derived from or are transformations applied to the signal which capture the discriminative information important to the task. Alternatively, the deep belief network (DBN) is utilised to learn deep features directly from the base short-time Fourier transform (STFT) feature for underwater target classification, and the results confirm the outstanding performance of the deep learning approach in complex ocean environments [13]. In [14], the convolution neural network (CNN) and the DBN method have

been combined with the SVM for classification, which can achieve higher recognition accuracies compared with traditional methods. In [15], a method for feature extraction and identification of underwater noise data based on CNN and extreme learning machine is proposed. These works have showed the great capability of deep networks to model a complex function with high-dimensional input. In this paper, we focus on the stacked sparse autoencoder (SSAE) model, which achieves the deep structure by stacking multiple autoencoders (AEs), to learn the high-level features of the shallow representation.

This work is an extension of our earlier work in [16] where we propose to use the SSAE model for underwater target classification, and consider a high-dimensional log-power spectrum (LPS) feature as the input for the SSAE model. However, the deep models which just rely on the spectrum-based features such as the STFT feature [13] and the LPS feature [16] may dismiss some significant properties of the radiated noise from the underwater target. Note that the single-domain feature captures only one aspect of the characteristics for the radiated acoustic signals, an integration of multiple-domain feature can be more appropriate for the deep models, especially when applied to new conditions. In this work, we propose to construct a joint multiple-domain feature based on the spectral and wavelet domain information. The joint feature group is generated by extracting the shaft frequency feature, the LPS feature, and the wavelet packet component energy (WPCE) feature in various subbands. Specifically, in contrast to [17], which use the maximum-likelihood to estimate the frequency points related to the propeller speed from the PSD of the squared radiated signal, we propose to remove the continuous spectrum frequency components in the PSD of the raw radiated signal to obtain the line spectrum components. Moreover, unlike [18], which sums up the wavelet packet power energy of specific subbands and computes the ratios indices, we compute the ratio of the wavelet packet energy of each subband on the whole band to capture the global energy distribution of the high frequencies.

In this paper, we further investigate the advantages of the SSAE architecture to model the complex underwater acoustic signals by using the joint multiple-domain information as the feature input. We do this by exploring three different characterisations from the training data set of the radiated noise: the propeller shaft frequency information, the spectral structure, and the wavelet packet energy distribution. The SSAE model is trained with the joint multiple-domain input of these low-level features. In the testing stage, to evaluate the generalisation ability to unseen conditions, we evaluate the trained deep model with the testing samples recorded at new depths. Our experimental results show that the proposed SSAE model performs significantly better than the DBN model [13] and the sparse coding (SC) approach [19] in terms of the classification accuracy and the confusion matrices. The contributions of this paper are summarised as follows:

i. The SSAE model is used to automatically learn the high-level representation of the radiated acoustic signals in an unsupervised manner, which is more discriminative compared to previously hand-crafted features.
ii. We construct a joint feature group based on the spectral and wavelet domain information for the deep model to take full advantage of the deep architecture.
iii. Our method is tested on the underwater acoustic signals recorded at new depths and achieves outstanding performance compared with the state-of-the-art DBN approach [13].

The rest of this paper is organised as follows. Section 2 introduces the related work of deep learning in acoustic signal recognition. In Section 3, we outline the system framework of the proposed method. Section 4 details the extraction of joint feature input from multiple domain information. In Section 5, the deep network for high-level feature learning and classification is discussed. The experiments using the radiated noise recorded at different depths are reported in Section 6. Section 7 draws the conclusion of the research.

2 Related work

A lot of recent works have made progress in acoustic signal recognition by using deep networks to directly learn the inside structure of raw input rather than the shallow models. In the Kaggle competition on classification of whale vocalisations, many participants applied a deep learning approach (in particular, a convolutional network) and achieved high scores. In their deep learning approach, the spectrogram of a right whale call is treated as an image in much the same way as a handwritten digit [20]. Deep learning has also been proven to compare very well to other published techniques on classification tasks. The work in [21] outlines a sound event classification framework that compares auditory image front-end features with spectrogram image-based front-end features, using SVM and deep neural network (DNN) classifiers. The experimental results on the standard database show that given richer input features, the discriminative abilities of the DNN appear more able to extract meaningful classification relationships than SVM classifiers. In [22], Zhang *et al.* extend the DBM-based approach further by proposing novel features derived from spectrogram energy triggering, firstly allied with the powerful classification capabilities of a convolutional neural network (CNN). In [23], Espi *et al.* introduce the use of a CNN in acoustic event detection with multiple resolution spectrograms as input, to model 'local' properties of acoustic events, which provided better results in the evaluation task compared with log spectral patches. The approach taken in [24] is similar to [21], that is, the classification performance of the DNN classifier is compared against the results using an SVM at various signal-to-noise ratios (SNRs) with a number of individual features. The sparse AEs are adopted to learn the non-linear representation in an unsupervised manner [25], which are learned layer-by-layer through the deep model constructed by stacking the sparse AEs and the softmax classifier.

Classification of the underwater target is based on the radiated noises of marine vessels, which contain machinery noise, propeller noise, and hydrodynamic noise. The work in [13] first introduces a deep learning architecture to underwater acoustic signal classification. Similar to [21], a daisy chain of restricted Boltzmann machines (RBMs) is formed by using features learned by one of the layers of the RBM as the input to the next layer and so on for generating a probabilistic model. Unlike [1], the input of the proposed DBN model is just the T–F representation without resorting to hand-crafted features (e.g. MFCC), and the derived output features for classification are learned by the deep network in an unsupervised manner. The results indicate that the deep learning approach is capable of capturing several layers of intermediate representations of the underlying signal that are more abstracted at the higher layers. In this paper, we use the DBN method in [13] as the baseline system.

3 System overview

The framework of the proposed underwater target classification system is described in Fig. 1. In the first stage, we need to extract the low-level features in different domains and organise the multiple-domain feature vector into the joint feature input which is suitable for training. The joint feature group is composed of the shaft frequency feature, the LPS feature, and the WPCE feature. In the training stage, we stack multiple AEs to construct the SSAE model, which can be trained layer by layer in an unsupervised manner. The proposed SSAE model is trained from a collection of radiated noise signals recorded at two different depths represented by the joint feature input. We fine-tune the whole network to further improve the performance. In the classification stage, the well-trained SSAE model is fed with the joint feature input of signals at other depths to classify the targets with the softmax classifier. The detailed description of the joint feature extraction and the training of the SSAE model can be found in next sections.

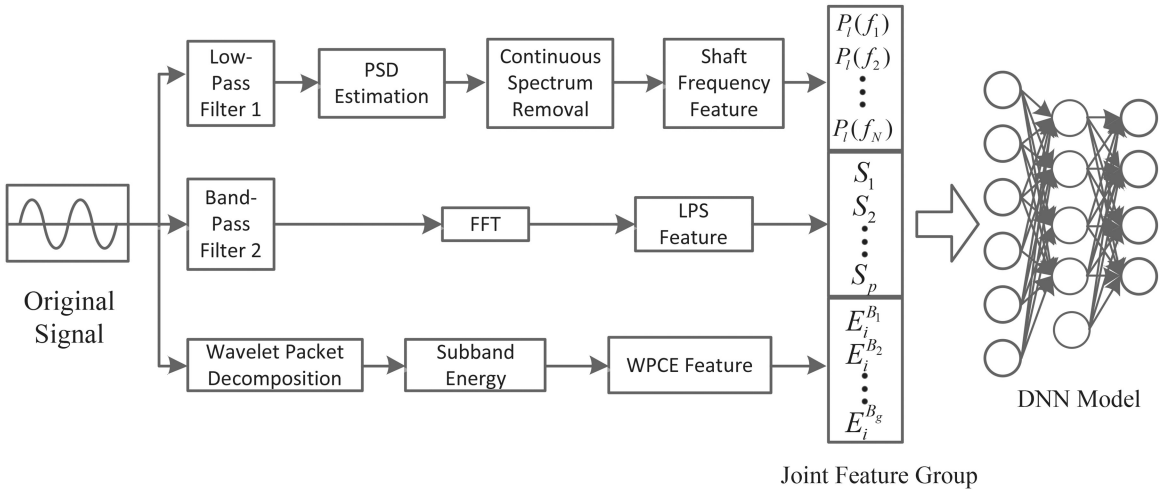

Fig. 1 *Framework of the proposed underwater target classification system*

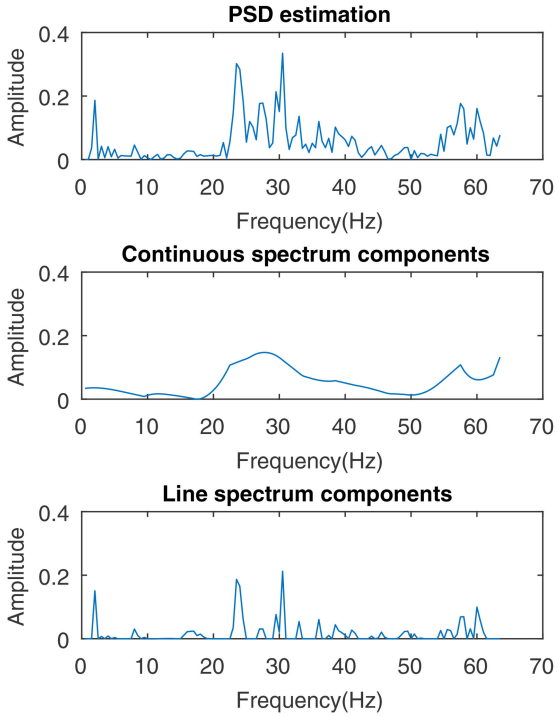

Fig. 2 *Continuous spectrum components and the line spectrum components of a specific vessel*

4 Joint feature group with multiple domain information

In this section, we integrate the spectral and wavelet packet features together to construct a joint spectral–wavelet feature input for the deep learning network.

4.1 Shaft frequency feature

In underwater target classification systems, a variety of features have been studied to fit the characteristics of the radiated noise. There is no doubt that the spectrum features are most widely used since they are easy to derive and the frequency-domain representation contains much useful information of different kinds of objects. However, there are several periodic low-frequency components of the underwater acoustic signal, which are caused by the turning of the propeller and are corresponding to the speed of the propeller [17]. This low-frequency information is a significant contributor to classifying ships since the low-frequency components are more robust, especially over long distances. In this paper, we propose to capture the shaft frequency feature directly from the power spectral density (PSD) of the raw underwater acoustic signal.

The PSD estimation contains line spectrum components and continuous spectrum components. If we extract the shaft frequency from the PSD estimation directly, then the trend of continuous spectrum components may lead to a high false rejection and detection rate. Thus, we need to remove the continuous spectrum components (see Fig. 2).

The PSD of the raw acoustic signal is defined as $P(f_i)$ where $f_i = f_1, f_2, \ldots, f_N$ denotes the frequency points of the PSD. Since a continuous spectrum components represent the spectral trending of $P(f_i)$, in this paper, we propose to use the least-squares polynomial fit (LSPF) to find the polynomial $P_n(f_i)$ of degree at most n to fit $P(f_i)$

$$P_n(f_i) = \sum_{j=1}^{n} b_j f_i^j \tag{1}$$

where $b_j\{j = 1, 2, \ldots, n\}$ represents the fitting parameters of P_n. The goal of the LSPF approach is to minimise the cost function $J(P_n)$ to get the best setting of each b_j. $J(P_n)$ is defined as

$$J(P_n) = \sum_{i=1}^{N} (P(f_i) - P_n(f_i))^2 \tag{2}$$

The details for computing b_j can be found in [18].

The shaft frequency feature can be treated as the raw PSD with the continuous spectrum components removed, which can be written as

$$P_l(f_i) = P(f_i) - P_n(f_i) \tag{3}$$

where $P_l(f_i)\{f_i = f_1, f_2, \ldots, f_N\}$ denotes the shaft frequency feature.

In this paper, unlike [17], we do not define any signal models in advance to estimate the propeller shaft speed; instead, we feed the shaft frequency components to the SSAE model to learn the discriminating structure in an unsupervised manner.

4.2 LPS feature

There is a lot of useful information for classification in the PSD from the radiated noise of underwater target. Since the continuous spectrum of vessels is mainly caused by the propeller cavitation, different objects often have various continuous spectrum profiles, which depend on the propeller structure and the power system of the objects. However, due to the complex underwater environment and the reverberation of the acoustic signal, it is very important to design a feature extractor which is adaptive enough. The deep network has shown great potential representing such complex

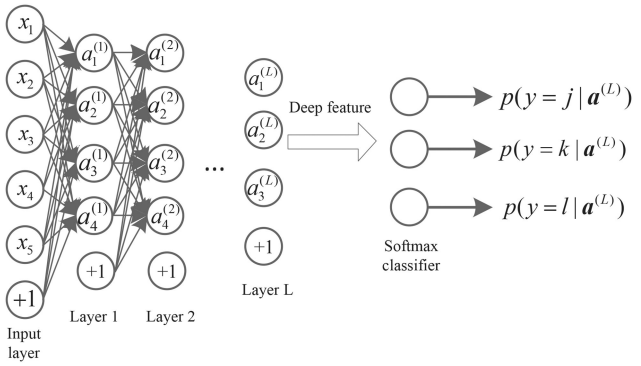

Fig. 3 *Simplified instance of the proposed SSAE model*

models, and deep learning can contribute to extract progressively more effective features which are invariant to the local change of the input. We still choose the LPS feature for the SSAE model, but we just focus on the subband from 100 Hz to 1 kHz to compute the LPS feature, which contains the main spectral structure. It is defined as

$$\boldsymbol{S} = \{S_1, S_2, \ldots, S_p\} \tag{4}$$

where p denotes the dimension number of the LPS feature.

4.3 WPCE feature

Compared with the STFT and WT methods, the wavelet packet transform (WPT) generates the full decomposition tree to provide the low-pass and high-pass analysis for the non-stationary underwater acoustic signal. The WPCE is an effective approach for classifying the specific characteristics of an underwater acoustic signal in the time–frequency domain. In our work, we propose to compute the wavelet packet global energy distribution to model the high-frequency instantaneous components of the radiated acoustic signal.

After WPT, the raw signal sequence is transformed to the subband signal sequence $\{d_i^j[k] \big| k = 1, 2, 3, \ldots, M\}$, where i denotes the number of decomposition levels and j the number of subbands, $j = 0, 1, 2, \ldots, 2^i - 1$. Also, M denotes the length of the subband signal sequence. Then the total signal energy can be expressed by the 2^i wavelet packet component energies. The jth subband energy of the ith decomposition level E_i^j is defined as

$$E_i^j = \frac{\sum_{k=1}^{M} \| d_i^j[k] \|^2}{\sum_{j=0}^{2^i-1} \sum_{k=1}^{M} \| d_i^j[k] \|^2} \tag{5}$$

Since the WPT can provide more detailed information in the high-frequency region, we propose to select those subbands over 1 kHz to construct the WPCE feature in this paper. Then, the selected WPCE feature can be written as

$$\boldsymbol{T} = \{E_i^{B_1}, E_i^{B_2}, \ldots, E_i^{B_g}\} \tag{6}$$

where $B_1, B_2, \ldots, B_g$ denote all the subbands over 1 kHz in our analysis band.

5 Deep network architecture

In this paper, the deep model proposed is composed of the SSAE and the softmax classifier. The joint features captured in Section 4 are still low-level ones which are not robust enough. The SSAE is powerful enough to learn the interesting structure of the input in an unsupervised manner. The joint feature group is used as the input feature vector for the SSAE to generate the deep feature, at the last hidden layer L (see Fig. 3). Assume a sample data set $\{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$ of m training samples, where $\boldsymbol{x}^{(i)}$ and $y^{(i)}$ are the input feature vector and the category label for the ith sample, respectively. The softmax classifier is able to estimate the classification probability of each sample, with which we can determine the category.

Before describing the SSAE, let us consider the single sparse AE which only has three layers: the input layer, the hidden layer, and the output layer. The target values of the output layer are expected to be equal to the inputs. Specifically, the input layer (which is set to the input feature vector) is mapped into a hidden representation, and the activity of the hidden layer is then used to reconstruct the output. When the number of hidden layer units is less than the input layer units, the AE is expected to learn the sparse representation of the raw inputs. Thus, a sparse AE can be seen as an effective feature extractor which maps the input to a non-linear representation.

To train the single sparse AE, we need to define the cost function. For the input feature vector of a single sample, $\boldsymbol{x}$, the value of the output layer for the sparse AE, $h_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x})$, is supposed to be close to the raw input $\boldsymbol{x}$, which can be computed with the forward propagation algorithm as follows:

$$h_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x}) = \mathrm{sigm}(\boldsymbol{W}^{(2)}\boldsymbol{a}^{(1)} + \boldsymbol{b}^{(2)}) \tag{7}$$

where

$$\boldsymbol{a}^{(1)} = \mathrm{sigm}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) \tag{8}$$

and $\boldsymbol{W}^{(1)}$, $\boldsymbol{b}^{(1)}$ denote the weight connect parameter matrix and bias parameter vector between the input layer and the hidden layer, while $\boldsymbol{W}^{(2)}$, $\boldsymbol{b}^{(2)}$ are the network parameters between the hidden layer and the output layer. The activation function $\mathrm{sigm}(\cdot)$ is the sigmoid function.

Then, the cost function can be written as $J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}) = (1/2) \| h_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x}) - \boldsymbol{x} \|^2$. For m samples, we impose a sparse constraint on the cost function to limit the size of the hidden layer. To implement the constraint, we will add an extra penalty term to our cost function. Thus, the cost function of sparse AE for m samples $J_{\mathrm{sparse}}(\boldsymbol{W}, \boldsymbol{b})$ is as follows:

$$\begin{aligned} J_{\mathrm{sparse}}(\boldsymbol{W}, \boldsymbol{b}) &= \left[\frac{1}{m}\sum_{i=1}^{m} J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}^{(i)})\right] \\ &\quad + \frac{\lambda}{2} \| \boldsymbol{W} \|^2 + \beta \sum_{j=1}^{s_1} \mathrm{KL}(\rho \parallel \hat{\rho}_j) \\ &= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2} \| h_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x}^{(i)}) - \boldsymbol{x}^{(i)} \|^2\right)\right] \\ &\quad + \frac{\lambda}{2}(\| \boldsymbol{W}^{(1)} \|^2 + \| \boldsymbol{W}^{(2)} \|^2) + \beta \sum_{j=1}^{s_1} \mathrm{KL}(\rho \parallel \hat{\rho}_j) \end{aligned} \tag{9}$$

where the second term is a weight decay term and λ the decay parameter which can be used to limit the value of the weight connect parameters. Besides, the last term is the added sparse penalty term, while β is the weight of the sparse penalty term. Here, s_1 denotes the number of nodes in the hidden layer of the single sparse AE. Also, $\mathrm{KL}(\cdot)$ represents the KL divergence which is defined as follows:

$$\mathrm{KL}(\rho \parallel \hat{\rho}_j) = \rho \lg \frac{\rho}{\hat{\rho}_j} + (1-\rho) \lg \frac{1-\rho}{1-\hat{\rho}_j} \tag{10}$$

where

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m} a_j^{(1)}(\boldsymbol{x}^{(i)}) \tag{11}$$

and $\hat{\rho}_j$ is the average activation of the hidden node $a_j^{(1)}$ for the ith sample $\boldsymbol{x}^{(i)}$. To ensure the sparsity of the hidden layer, we enforce

Table 1 Details about the data set including the number of samples for each vessel at each depth

Depth, m	A	B	C	D	E
50	1440	1320	1440	600	1800
70	1440	2840	460	1440	320
100	2400	2280	1800	1680	2400
150	2760	3240	600	2160	400
200	1320	820	280	840	280

the $\hat{\rho}_j$ to be equal to the sparsity parameter ρ, since the value of $\mathrm{KL}(\rho \parallel \hat{\rho}_j)$ increases unless $\hat{\rho}_j = \rho$.

The goal of training the sparse AE is to estimate the model parameters $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(2)}$, $\boldsymbol{b}^{(1)}$, $\boldsymbol{b}^{(2)}$ by minimising the cost function in (9). In this paper, we use the L-BFGS algorithm to optimise the cost function [26]. In the training stage, the partial derivatives are computed with the backpropagation algorithm.

The sparse AE is capable of learning the sparse representation from the raw input. To learn more discriminating representation of the data, we propose to build the SSAE model with a multiple-layer structure, which is constructed by stacking the input layer and the hidden layer of sparse AE together layer by layer [27]. However, the output layers of the sparse AEs are not needed for constructing the SSAE, so the reconstruction layer of each single sparse AE is removed. After stacking, the output of the hidden layer in the first sparse AE $\boldsymbol{a}^{(1)}$ is treated as the input of the second sparse AE. By repeating these steps, we can achieve a deep structure for the SSAE model with multiple sparse AEs. This can be formulated as

$$\boldsymbol{a}^{(l)} = \mathrm{sigm}(\boldsymbol{W}^{(l-1,1)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l-1,1)}) \tag{12}$$

$$\boldsymbol{a}^{(l+1)} = \mathrm{sigm}(\boldsymbol{W}^{(l,1)}\boldsymbol{a}^{(l)} + \boldsymbol{b}^{(l,1)}) \tag{13}$$

where $\boldsymbol{W}^{(l,1)}$, $\boldsymbol{b}^{(l,1)}$ denote the network parameters of the lth hidden layer in the SSAE model constructed layer by layer. We observe that the activation of the $(l+1)$th layer $\boldsymbol{a}^{(l+1)}$ is the output of the lth layer based on the activation of the lth layer $\boldsymbol{a}^{(l)}$. If we adopt L sparse AEs to build our SSAE model, then the output of the last hidden layer $\boldsymbol{a}^{(L)}$ represents the deep feature we learn from the raw input with the deep structure, which can also be used for classification in the next step.

In this paper, we choose the softmax classifier as the last classification approach, which is popular in multiple-class classification [28]. After the feature learning of the SSAE model, the sample space $\{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$ is transformed to the feature space $\{(\boldsymbol{a}^{(L,1)}, y^{(1)}), (\boldsymbol{a}^{(L,2)}, y^{(2)}), \ldots, (\boldsymbol{a}^{(L,m)}, y^{(m)})\}$, where $\boldsymbol{a}^{(L,i)}$ is the output of the Lth hidden layer for the ith sample, $\boldsymbol{a}^{(L,i)} \in \boldsymbol{R}^{s_L}$, and s_L denotes the number of nodes in the Lth hidden layer of the SSAE. Note that in this paper, we use s_i to represent the number of nodes in the ith hidden layer of the SSAE. For the deep feature $\boldsymbol{a}^{(L,i)}$, the softmax classifier can give a probability that the input belongs to category index j, and this probability can be written as follows:

$$p(y^{(i)} = j \,|\, \boldsymbol{a}^{(L,i)}; \boldsymbol{\theta}) = \frac{\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{a}^{(L,i)}}}{\sum_{t=1}^{K} \mathrm{e}^{\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{a}^{(L,i)}}} \tag{14}$$

where θ_j denotes the model parameter of the jth unit for the softmax classifier, and $\theta_j \in \boldsymbol{R}^{s_L}$, which has the same dimension as $\boldsymbol{a}^{(L,i)}$. Here, K is the total number of categories. Then the cost function of these m samples for the softmax classifier can be written as follows:

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{K} 1\{y^{(i)} = j\}\log\frac{\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{a}^{(L,i)}}}{\sum_{t=1}^{K}\mathrm{e}^{\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{a}^{(L,i)}}}\right] + \frac{\gamma}{2}\sum_{j=1}^{K} \parallel \boldsymbol{\theta}_j \parallel^2 \tag{15}$$

where $1\{\cdot\}$ is the indicator function, $1\{\mathrm{True}\} = 1$ and $1\{\mathrm{False}\} = 0$. The second term is the weight decay term and γ is the controlling parameter. The softmax classifier can also be trained using the L-BFGS method to get the best parameters: $\boldsymbol{\theta}$.

The whole SSAE model consisting of multiple hidden layers and the softmax classifier can be trained by the greedy layer-wise training method [29]. In the unsupervised training phase, the hidden layers are trained hierarchically with the sample input $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\}$, which means that we train the multiple sparse AEs separately. After training, the hidden layer parameters $\boldsymbol{W}^{(l,1)}$, $\boldsymbol{b}^{(l,1)}$ are retained as the initial configuration for the fine-tuning processing. We still use the L-BFGS method to optimise the whole model with the labels $\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$ in the fine-tuning stage.

6 Experiments and results

We mainly tested our proposed approach by running through experiments on the recorded radiated noise data set of underwater targets. The data set provides the radiated signal recorded at different depths in the sea. We conducted experiments on the data set to evaluate how different parameters affect the SSAE approach based on multiple-domain information. We also compared the best scores of these settings with the DBN approach reported recently [13] and the SC method [19].

6.1 Data set description

The radiated acoustic data set used in this paper was collected by the School of Marine Science and Technology of Northwestern Polytechnical University from the South China Sea in 2015. The water depth is ~330–350 m. The data set contains five vessels (A, B, C, D, and E) with different properties such as weight, size, propeller structures, and power system. The original underwater acoustic signals of these objects were collected from the real sea trial with the single-hydrophone placed at various depths of 50, 70, 100, 150, and 200 m. Depth is a significant condition for the received signal since it is related to the multi-path effect of the channel and affects the SNR. Every time the vessels moved across the hydrophone on the sea, only a portion of the recording was kept corresponding to ranges from +500 to −500 m in horizontal distance with respect to the position of the hydrophone. The radiated signals from the vessels were sampled using 32 bits and at the sampling frequency of 50 kHz.

For each recording, the multiple-domain features were computed using 1 s frame to generate the input sample. Details about the whole data set can be found in Table 1. To evaluate the performance of the proposed SSAE model, at first we selected the training set and test set homogenously among samples collected from all depths. Furthermore, to assess the generalisation capability to unseen conditions, we tested the performance of the proposed method at new depths. We separated the whole data set in Table 1 into a training set and testing set for each different depth. The training set used to train the SSAE model was generated by the samples at the depths of 50 and 150 m, while the testing set consisting of the records at the other three depths, 70, 100, and 200 m.

Table 2 Overall classification accuracy for different settings of the number of model layers and the number of nodes on the testing set from all depths

Model layers	Node number				
	50	100	200	300	400
3	0.8812	0.9041	0.9103	0.9024	0.8951
4	0.9083	0.9274	0.9353	0.9231	0.8991
5	0.9114	0.9334	**0.9423**	0.9267	0.9126
6	0.8984	0.9314	0.9123	0.9002	0.8833

Bold values indicate the best overall accuracy of all settings.

6.2 Experimental setting

According to Section 4, there are three subsystems that perform the shaft frequency feature extraction, the LPS feature, and the wavelet packet decomposition in different subbands. The feature vector group was computed with these low-level features.

The shaft frequency feature was calculated by putting the original signal through the low-pass filter of 0–64 Hz. Owing to the fact that there may exist pseudo-spectral artefacts near the real line-spectrum, we need to improve the analysis resolution. The filtered signal was re-sampled with the sampling rate 1 kHz, while the length of fast Fourier transform (FFT) was set to 2048 (2 s) points with 50% overlap between the adjacent frames. Then the length N of the PSD was set to 1024. In the fitting of the curve, we set the degree of the fitting polynomial n to 3. We then chose the first 128 dimensions of the shaft frequency feature corresponding to the subband of 0–64 Hz. For the LPS feature, the raw acoustic signal was filtered to 64–1024 Hz. Similar to [16], the FFT was computed using 1024 (0.25 s) points with the sampling rate of 4 kHz. To improve the SNR of the LPS feature, we used the average FFT results of four adjacent sections to generate the final LPS feature. Then, the dimension of the LPS feature input was set to 240, corresponding to the subband 64–1024 Hz. For wavelet packet decomposition, the sampling rate was 4 kHz and the length of frame was also 4096 (1 s). For WPT, we adopted the wavelet function *db*5 as the decomposition function, which is widely used in non-stationary signal analysis. The decomposition level and the number of subbands were set to 8 and 256, respectively. However, we just focused on the high-frequency range of 1024–2048 Hz, which is consisted of 128 subbands for the feature input. The three feature input groups were concatenated together used as the input to the SSAE model.

The SSAE model contains several hidden layers, and the softmax classifier was added to the final hidden layer as the output of the model. The node number of the input layer and the output layer were set to 496 and 5, corresponding to the dimensions of joint feature vector and the classes of target, respectively. The number of epochs for each layer of sparse AE pre-training with L-BFGS was 100, while the weight decay parameter was set to 0.003. As for the sparse term, the weight of sparsity penalty term was set to 3. The number of epochs for the last fine-tuning was 200.

6.3 Performance of the testing set generated from all depths

In this section, we evaluated the performance of the SSAE model by selecting the training set and testing set homogenously among the whole data set in Table 1 from all depths. The 5-fold cross-validation was used for the classification experiments. The whole data set was separated into five parts homogenously and each part contains the same number of samples from each depth and each vessel. We evaluated the SSAE model on a single part and use the other four parts as the training set. This process was repeated five times and the final classification accuracy was computed by averaging the results of the five processes.

6.3.1 Evaluation of network settings: For deep networks, the model structure plays an important role in the classification accuracy since it determines the performance of the learned feature in terms of invariance and abstraction. For the proposed SSAE model, we looked into the number of model layers and the number of nodes for hidden layer to find out the best settings of the proposed SSAE model. In Table 2, we tried several setting groups. The number of model layers ranges from 3 to 6, including the input layer and output layers. In the experiment, the number of nodes for all hidden layers was set to be equal, for values of 50, 100, 200, 300, and 400. These different settings were evaluated with the data set homogenously selected with the 5-fold cross-validation. We can see from Table 2 that the number of hidden layers does help to increase the overall classification accuracy until it reaches 5. It can be seen that the increase in the number of nodes may result in low classification accuracy. The best setting for the proposed SSAE model on the testing set from all depths is 496–200–200–200–5.

6.3.2 Comparison with the other classification systems: We compared the SSAE model with the other classification systems on the testing set collected from all depths with the 5-fold cross-validation above. The DBN model has been applied to the T–F representation for underwater target classification in [13], which yields better results in noise environments. It can be treated as the baseline method in this paper. The SC can be seen as a modification of the sparse AE in which we can learn an associated basis to transform the learned features from the data space to the feature space. In [19], the sparse feature is learned by using SC from the spectral patches to train SVMs for acoustic event detection. In this paper, we adopted the SC and SVM (SC–SVM) in [19] for underwater target classification as a comparison with the SSAE model.

We still used the same joint feature group as the input of the DBN model and the SC–SVM method. The number of hidden layers for the DBN model and the learning rate for the RBM training using contrastive divergence were set to 5 and 0.0001, respectively. To apply SC in [19] to underwater target classification, the input dimension and the regularisation parameter were set to 496 and 0.001.

As is clear from Fig. 4, the SSAE model achieves the best performance compared to the DBN model [13] and the SC–SVM method [19] for the testing set from five different depths. Since the SNRs of the radiated acoustic signals at different depths vary a lot, the results demonstrate the robustness to noise variations of the SSAE model. It can be seen that the SSAE model provides a 3% improvement over the DBN model [13]. The best scores of the SSAE model and the DBN model are much greater than the best score of SC–SVM method [19], which shows the advantages of the deep structure.

6.4 Performance of the testing set generated from new depths

To further evaluate the generalisation ability to new conditions of the SSAE model, we evaluated the performance on the testing set generated from new depths (70, 100, and 200 m) with the training set from two other depths (50 and 150 m).

6.4.1 Evaluation of network settings: Table 3 shows the overall classification accuracy for the different settings of the number of model layers and the number of nodes on the testing set from new depths. We can see that the best number of model layers for the SSAE model is still 5. It can be seen that the best setting for the proposed SSAE model is 496–100–100–100–5. We know that the best performing structure of the deep model is not always the most complex structure, especially for a small data set, since the deep network may suffer from over-fitting.

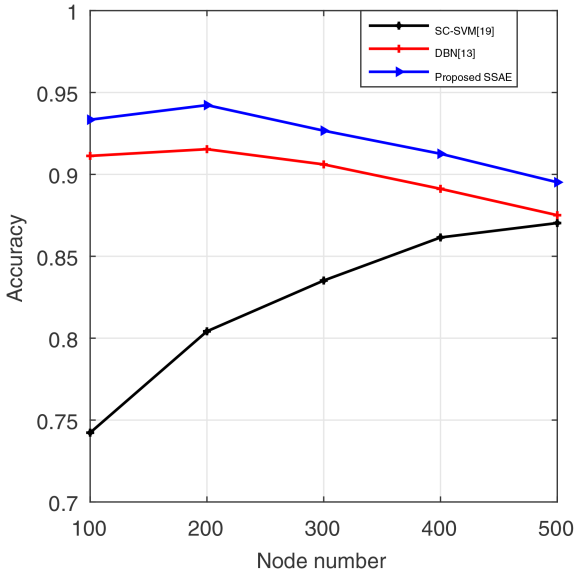

Fig. 4 *Classification results of the SSAE model, the DBN model, and the SC method for the testing set from all depths*

Since the distribution of the samples for each class is not balanced (see Table 1), we adopted the confusion matrix to show the details of the classification results. The confusion matrix generated based on the classification results of the best setting (496–100–100–100–5) is given in Fig. 5.

6.4.2 Comparison with DNN model using single-domain features: To explore the generalisation ability of deep network, the SSAE model with the joint spectral and wavelet information was evaluated on the testing set which consists of recordings at different depths. In this section, the classification performance of joint feature group was compared with the SSAE model using single-domain features, such as the T–F representation and the wavelet feature. Similar to [16], the T–F feature was computed using the STFT with the frame length of 1024 points and sample frequency 4 kHz. The dimension of the feature vector was 512. The WPCE feature was extracted as in Section 6.2, but we used all the 256 subbands to generate the feature vector. All the parameters of the SSAE model used for T–F feature and WPCE feature were as specified in Section 6.4.1. We still chose the best setting as the model structure.

Fig. 6 reports the classification accuracy of applying the SSAE model to the T–F feature, WPCE feature, and the joint feature group, separately. The incorporation of spectral and wavelet information is shown to improve the overall classification accuracy by 4.11 and 5.79%, compared with WPCE feature and T–F representation, respectively. As far as the performance of different depths for the sensor, the joint feature input is also seen to give the highest classification accuracy. It can also be observed that when the sensor goes deeper, the results get worse. This may result from the attenuation increasing with depth, which degrades the SNR. Our proposed multiple-domain feature achieves 89.68% classification accuracy, but only 84.34 and 84.06% for WPCE feature and T–F representation, respectively.

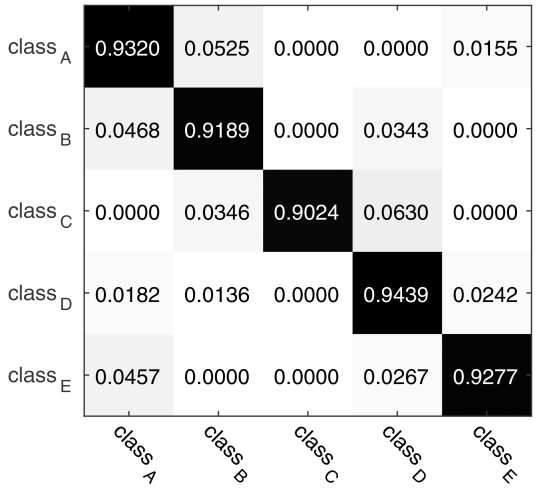

Fig. 5 *Confusion matrix for the classification results of the best setting. X-axis indicates the predicted label and Y-axis indicates the true label*

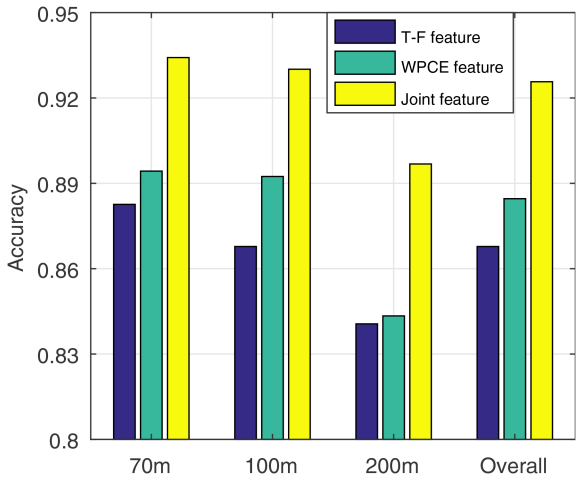

Fig. 6 *Classification accuracy for new depths using T–F feature, WPCE feature, and joint spectral and wavelet feature*

6.4.3 Comparison with the other classification systems: We also compared the performance of the SSAE model with the DBN model [13] and the SC–SVM method [19] for the testing set from new depths. The basic parameters of the DBN model and the SC–SVM method were remained as in Section 6.3.2. Fig. 7 reveals the classification results of the three methods for the data set at different depths (70, 100, 200 m and the overall results).

It can be seen that the proposed SSAE approach provides better classification results compared to the DBN model [13] and the SC–SVM method [19] for all the new depths (70, 100, and 200 m), which proves to be robust for different noises. We also find that the SSAE model offers a 5% improvement over the DBN model for the testing set from new depths in overall classification accuracy, which is greater than the improvement achieved in Section 6.3.2. This can demonstrate that the proposed SSAE model is capable of improved generalisation ability to unseen conditions. Moreover, the best results of the SSAE model are also significantly better than those of the SC–SVM method, which proves that the deep

Table 3 Overall classification accuracy for different settings of the number of model layers and the number of nodes on the testing set from new depths

Model layers	Node number				
	50	100	200	300	400
3	0.8701	0.8924	0.8947	0.8886	0.8782
4	0.8943	0.9104	0.9217	0.9025	0.8821
5	0.8862	**0.9257**	0.9135	0.9067	0.8901
6	0.8720	0.9178	0.9017	0.8833	0.8621

Bold values indicate the best overall accuracy of all settings.

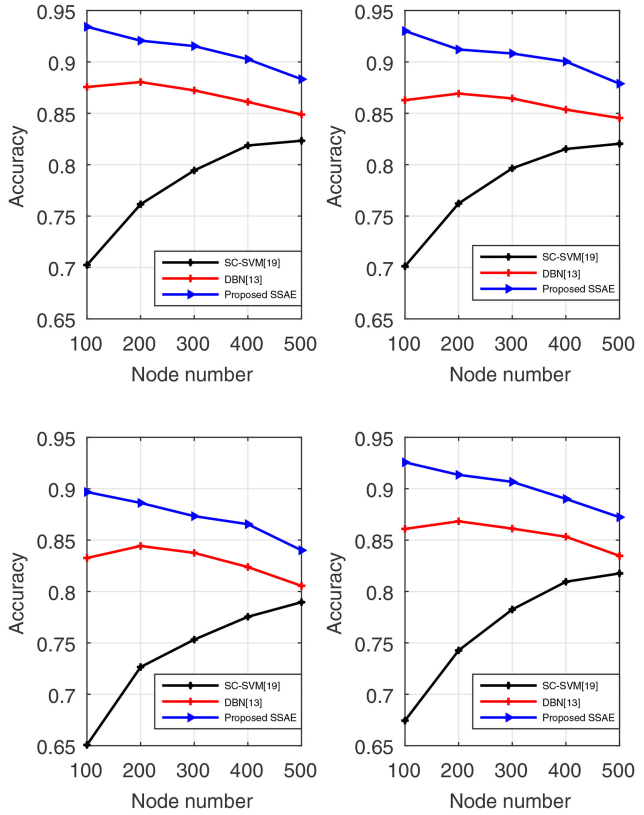

Fig. 7 *Classification results of three methods for the testing set at new depths of 70 m (upper left), 100 m (upper right), 200 m (lower left), and the overall results (lower right)*

architecture can learn a more discriminative representation of the radiated noise.

7 Conclusion

In this paper, we propose an underwater target classification for new depths using an SSAE model with joint multiple-domain feature. The main contribution of this work is combining the deep networks with the joint multiple-domain feature group based on the spectral and wavelet information, to improve the generalisation ability to unseen conditions. We test the proposed approach on the data set consisting of records at new depths and find that the SSAE model with 3 hidden layers and 100 nodes for hidden layers is optimal. In addition, the joint feature group is shown to achieve a significant improvement of 4–5% in classification accuracy over single-domain features. We then compare the classification results of the state-of-the-art DBN model and the SC approach, and showed that the proposed SSAE provided a significant performance improvement over these methods.

8 Acknowledgments

The research was supported by the National Science Foundation of China (grant no. 61601369) and the Natural Science Basis Research Plan in Shaanxi Province of China (program no. 2017JM6053). This work was completed when the first author was a visiting student in the School of Electrical, Electronic and Computer Engineering, University of Western Australia.

9 References

[1] Filho, W.S., Seixas, J.M., Caloba, L.P.: 'Averaging spectra to improve the classification of the noise radiated by ships using neural networks'. Proc. Sixth Brazilian Symp., Rio de Janeiro, Brazil, November 2000, pp. 156–161

[2] Chen, C., Lee, J., Lin, M.: 'Classification of underwater signals using wavelet transforms and neural networks', *Math. Comput. Model*, 1998, **27**, (2), pp. 47–60

[3] Farrokhrooz, M., Karimi, M.: 'Marine vessels acoustic radiated noise classification in passive sonar using probabilistic neural network and spectral features', *Intell. Autom. Soft. Comput.*, 2011, **17**, (3), pp. 369–383

[4] Azimi-Sadjadi, M.R., Yao, D., Huang, Q.: 'Underwater target classification using wavelet packets and neural networks', *IEEE Trans. Neural Netw.*, 2000, **11**, (3), pp. 784–794

[5] Shi, M., Xu, X.: 'Underwater target recognition based on wavelet packet entropy and probabilistic neural network'. Proc. Int. Conf. on Signal Processing, Communication and Computing, KunMing, China, August 2013, pp. 1–3

[6] Das, A., Kumar, A., Bahl, R.: 'Marine vessel classification based on passive sonar data: the cepstrum-based approach', *IET Radar Sonar Navig.*, 2013, **7**, (1), pp. 87–93

[7] Filho, J.B.O.S., Seixas, J.M.: 'Class-modular multi-layer perceptron networks for supporting passive sonar signal classification', *IET Radar Sonar Navig.*, 2016, **10**, (2), pp. 311–317

[8] Jian, L., Yang, H., Zhong, L.: 'Underwater target recognition based on line spectrum and support vector machine'. Proc. Int. Conf. Mechatronics, Control Electronic Engineering, Shenyang, China, November 2014, pp. 79–84

[9] Sherin, B.M., Supriya, M.H.: 'Selection and parameter optimization of SVM kernel function for underwater target classification'. IEEE Int. Symp. Underwater Technology (UT), Chennai, India, February 2015, pp. 1–5

[10] Sherin, B.M., Supriya, M.H.: 'GA based selection and parameter optimization for an SVM based underwater target classifier'. IEEE Int. Symp. Ocean Electronics (SYMPOL), Kochi, India, November 2015, pp. 1–7

[11] Hinton, G., Deng, L., Yu, D., *et al.*: 'Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups', *IEEE Signal Process. Mag.*, 2012, **29**, (6), pp. 82–97

[12] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444

[13] Kamal, S., Mohammed, S.K., Pillai, P.S., *et al.*: 'Deep learning architectures for underwater target recognition'. IEEE Int. Symp. Ocean Electronics (SYMPOL), Kochi, India, October 2013, pp. 48–54

[14] Yue, H., Zhang, L., Wang, D., *et al.*: 'The classification of underwater acoustic targets based on deep learning methods'. Proc. Control, Automation and Artificial Intelligence, Ohrid, Macedonia, July 2017, pp. 526–529

[15] Hu, G., Wang, K., Peng, Y., *et al.*: 'Deep learning methods for underwater target feature extraction and recognition', *Comput. Intell. Neurosci.*, 2018, **1**, (2018), pp. 1–10

[16] Cao, X., Zhang, X., Yu, Y., *et al.*: 'Deep learning-based recognition of underwater target'. Proc. Digital Signal Processing, Beijing, China, October 2016, pp. 89–93

[17] Lourens, J.G., Du Prcez, J.A.: 'Passive sonar ML estimator for ship propeller speed', *IEEE J. Ocean. Eng.*, 1998, **23**, (4), pp. 448–453

[18] da Silveira, T.L., Kozakevicius, A.J., Rodrigues, C.R: 'Automated drowsiness detection through wavelet packet analysis of a single EEG channel', *Expert Syst. Appl.*, 2016, **55**, (2016), pp. 559–565

[19] Lu, X., Tsao, Y., Shen, P.: 'Spectral patch based sparse coding for acoustic event detection'. Proc. Int. Symp. on Chinese Spoken Language Processing, Singapore, September 2014, pp. 317–320

[20] Smirnov, E.: 'North Atlantic right whale call detection with convolutional neural networks'. Proc. Int. Conf. on Machine Learning, Atlanta, USA, June 2013, pp. 78–79

[21] McLoughlin, I., Zhang, H., Xie, Z., *et al.*: 'Robust sound event classification using deep neural networks', *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2015, **23**, (3), pp. 540–552

[22] Zhang, H., McLoughlin, I., Song, Y.: 'Robust sound event recognition using convolutional neural networks'. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Vancouver, Canada, April 2015, pp. 559–563

[23] Espi, M., Fujimoto, M., Kinoshita, K., *et al.*: 'Exploiting spectro-temporal locality in deep learning based acoustic event detection', *EURASIP J. Adv. Sig. Process.*, 2015, **1**, (2015), pp. 26–35

[24] Sharan, R.V., Moir, T.J.: 'Robust acoustic event classification using deep neural networks', *Inf. Sci.*, 2017, **396**, (2017), pp. 24–32

[25] Deng, J., Zhang, Z., Marchi, E., *et al.*: 'Sparse autoencoder-based feature transfer learning for speech emotion recognition'. Proc. Affective Computing and Intelligent Interaction, Geneva, Switzerland, September 2013, pp. 511–516

[26] Liu, D.C., Nocedal, J., Rodrigues, C.R.: 'On the limited memory BFGS method for large scale optimization', *Math. Program.*, 1989, **45**, (1), pp. 503–528

[27] Xu, J., Xiang, L., Liu, Q., *et al.*: 'Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images', *IEEE Trans. Med. Imag.*, 2016, **35**, (1), pp. 119–130

[28] Nasrabadi, N.M.: '*Pattern recognition and machine learning*' (Academic Press, New York, 2006, 1st edn)

[29] Bengio, Y., Courville, A., Vincent, P.: 'Representation learning: a review and new perspectives', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (8), pp. 1798–1828