Yao *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:48
https://doi.org/10.1186/s13634-023-01014-1

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Seal call recognition based on general regression neural network using Mel-frequency cepstrum coefficient features

Qihai Yao[1,2], Yong Wang[1,2*], Yixin Yang[1,2] and Yang Shi[1,2]

*Correspondence:
yongwang@nwpu.edu.cn

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China
[2] Shaanxi Key Laboratory of Underwater Information Technology, Xi'an 710072, China

Abstract

In this paper, general regression neural network (GRNN) with the input feature of Mel-frequency cepstrum coefficient (MFCC) is employed to automatically recognize the calls of leopard, ross, and weddell seals with widely overlapping living areas. As a feedforward network, GRNN has only one network parameter, i.e., spread factor. The recognition performance can be greatly improved by determining the spread factor based on the cross-validation method. This paper selects the audio data of the calls of the above three kinds of seals and compares the recognition performance of three machine learning models for inputting MFCC features and low-frequency analyzer and recorder (LOFAR) spectrum. The results show that at the same signal-to-noise ratio (SNR), the recognition result of the MFCC feature is better than that of the LOFAR spectrum, which is verified by statistical histogram. Compared with other models, GRNN for inputting MFCC features has better recognition performance and can still achieve effective recognition at low SNRs. Specifically, the accuracy is 97.36%, 93.44%, 92.00% and 88.38% for cases with an infinite SNR and SNR of 10, 5 and 0 dB, respectively. In particular, GRNN has the least training and testing time. Therefore, all results show that the proposed method has excellent performance for the seal call recognition.

Keywords: Seal call recognition, MFCC feature, Underwater acoustic signal, GRNN, Machine learning

1 Introduction

Marine mammal acoustics is a science that studies the acoustic behavior of marine mammals. Different from the land–air environment, the energy of light, heat, electromagnetic wave, and other forms will decay rapidly in the marine water environment, but the acoustic signal can spread effectively over a long distance. Therefore, most marine mammals evolved the ability of underwater voice to achieve the purposes of communication, individual recognition, navigation and positioning, foraging for food, and so on. Thus, best method to study marine mammals is to use passive acoustic monitoring (PAM) technology to record and analyze their underwater acoustic signals.

Almost all seals can make sounds underwater, and the vocal behavior is closely related to the habitat and activities of seals during the whole life cycle. Seals spend

most of their time underwater to complete breeding, predation, and other behaviors. The traditional visual monitoring of seal behavior depends on weather, sea conditions, light, and other conditions, so it is difficult to track and monitor seals for a long time and in a wide range. However, PAM technology can get rid of most of the limitations of traditional visual monitoring, which can be effectively used for the investigation and research of seals. Watkins et al. analyzed the air and underwater vocalization of the Ross seal and found that its underwater vocalization includes pulses similar to that in the air. The difference is that the center frequency and bandwidth of underwater vocalization are large [1]. Mossbridge et al. [2] found through spectrum analysis that when leopard seals and killer whales in Antarctica appear in similar sea areas, the two animals will use frequency modulation to avoid the crossover of their vocal frequencies. Rogers et al. [3] studied the seals and Ross seals in Antarctica and found that the vocalization of the two seals was related to factors (e.g., day, night, and seasons). Risch et al. [4] studied the sounds of bearded seals in different Arctic waters and found significant differences in the vocal characteristics of different groups of bearded seals. Frouin-Mouy et al. [5] found that bearded seals vocalize more at night than during the day and more in winter than in summer. Halliday et al. [6] studied the seasonal vocal patterns of marine mammals (e.g., bearded seals, spotted seals, and whales underwater) and obtain the importance of sea ice to seal call. Cziko et al. [7] found that seals can emit ultrasonic waves close to 50 kHz through the broadband data collected underwater for a long time.

Underwater sound is collected by PAM technology, and acoustic features are extracted and input into a machine learning model, which has been widely used in the detection and recognition of marine mammals. Potter et al. used an artificial neural network to detect bowhead whales. The results show that the performance of this method is better than the time series matched filter and hidden Markov model [8]. Mouy et al. [9] used image processing methods for spectrum noise reduction and realized the detection and classification of blue whales and baleen whales through time–frequency analysis technology. Halkias et al. [10] established the restricted Boltzmann machine (RBM) model to realize the classification of baleen whales under a low signal-to-noise ratio (SNR). Parisi et al. [11] distinguished the underwater calls of bearded seals following different center frequencies and bandwidths. Frouin-Mouy et al. [12] used a spectrum map to realize the automatic detection of spotted seals and studied the distribution pattern of spotted seals in the Bering Sea, Chukchi Sea, and Beaufort Sea. Luo et al. proposed the local energy normalization method for inputting underwater sound data spectrum of different lengths. The convolutional neural network (CNN) model was applied for the effective detection of the echolocation sound of odontocetes [13]. Zhong et al. [14] established a transfer learning model and input the extracted small sample underwater sound spectrum for the detection of beluga whales. Shiu et al. [15] built various machine learning models to detect the phonation of North Atlantic right whales. The results show that the CNN model can significantly improve accuracy. Mishachandar et al. [16] applied the CNN model to marine noise recognition, effectively distinguishing different human, natural, and marine animal sounds. Escobar-Amado et al. [17] realized the effective classification of the sound of bearded seals by extracting the region of interest of bearded seals in the spectrum and using the CNN model.

Few studies related to seals, especially studies on the automatical recognition of seal calls, are noted in the field of underwater sound recognition of marine mammals. When different kinds of seals live in a similar sea area, the recognition of seal calls is of great significance to the study of abundance estimation and population structure. Therefore, a method of seal call effective recognition at different SNRs is essential. Machine learning is a data-driven method that has emerged in recent years. It can automatically extract the most representative abstract features of the category as the basis for recognition through a series of deep extractions. It is often of high accuracy and good noise resistance, realizing the intellectualization of underwater acoustic recognition. Most of the existing acoustic signal recognition methods based on machine learning models are complex and require training a large number of network parameters. This makes it difficult to find the optimal structure to adapt to the recognition task, and it is easy to overfit. Besides, the training time is often long.

In this study, a general regression neural network (GRNN) method using Mel-frequency cepstrum coefficient (MFCC) features is proposed to accurately recognizing seal calls. The GRNN method has only one network parameter. The optimal network architecture can be obtained by determining the spread factor, and the training time is short. The MFCC features can extract the feature information of the low-frequency part more effectively, and the frequency of the seal calls is mainly concentrated in the low-frequency part. Based on the above considerations, the GRNN model for inputting MFCC features combines the feature advantages of MFCC and the nonlinear fitting ability of the GRNN in order to improve the recognition performance for seal calls. Figure 1 shows the whole structure of our framework.

2 MFCC

MFCC is a widely used feature in speech recognition. It is proposed based on the characteristics of the human ear. Due to the particularity of human ear structure, the listener can automatically separate the low- and high-frequency part of speech, in which the low-frequency part recognizes speech characteristics. Based on this, the characteristics of the human ear can be simulated and effective spectrum features extracted (i.e., convert the spectrum into Mel spectrum) by setting denser filters in the low-frequency part and fewer filters in the high-frequency part. Cepstrum is used in log functions to transform multiplicative signals into additive signals to reflect the low-frequency envelope spectrum characteristics and high-frequency detail characteristics of the signal. Through cepstrum analysis of the Mel spectrum, MFCC can be obtained and widely used in underwater target recognition [18].

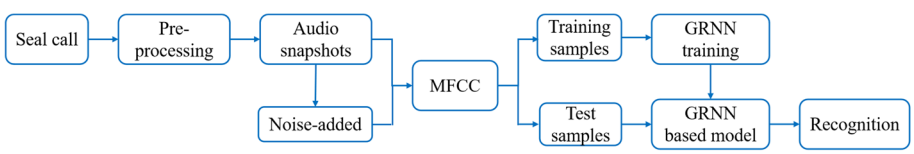


Fig. 1 The whole structure of the proposed method

Figure 2 shows the MFCC feature extraction process, which is mainly composed of preprocessing, fast Fourier transform (FFT), Mel filtering, and discrete cosine transform (DCT).

The preprocessing process includes pre-emphasis, framing, and windowing. Pre-emphasis enables the spectrum of the signal to be gentler by raising the spectrum of the high-frequency part of the signal. Framing can divide the signal into several short-term signals in which the signal can be regarded as a stationary process. In the process of framing, the method of overlapping segmentation is generally adopted to make the frame to frame excessively smooth. Windowing reduces the truncation effect of the signal. Thus, the signal and the window function are set as $s(n)$ and $w(n)$, respectively. The signal obtained after windowing is as follows:

$$s'(n) = s(n)w(n),\ 0 \le n \le N-1 \tag{1}$$

where N is the number of samples and $w(n)$ is the Hamming window.

After preprocessing, FFT is performed on each frame signal to obtain the spectrum. The discrete spectrum $S_a'(k)$ of the signal can be expressed as

$$S_a'(k) = \sum_{n=0}^{N-1} s'(n)e^{-j2\pi k/N}, 0 \le k \le N \tag{2}$$

The spectrum is then filtered through a group of triangular band-pass filters to obtain Mel filters. Moreover, M filters exist, and the center frequency is $f(m)$, of which $m = 1, 2, \cdots, M$. The formula of triangular filter is [19]:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m-1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{3}$$

The logarithmic energy output by each filter is calculated, which is expressed as follows:

$$S^*(m) = \ln\left(\sum_{k=0}^{N-1} \left|S_a'(k)\right| H_m(k)\right), 0 \le m \le M \tag{4}$$

Perform DCT on the M logarithm energies calculated by Formula (4) to obtain the MFCC of order L ($L = 12$–16), where the DCT is [20]:

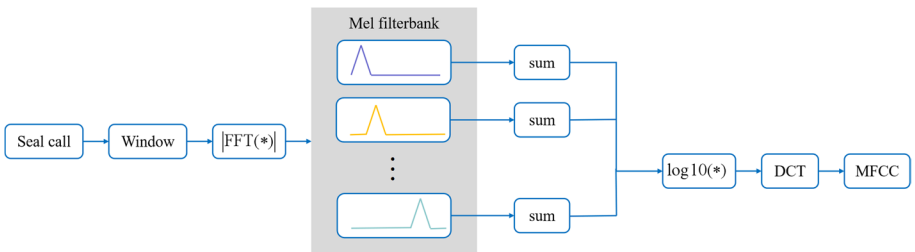


Fig. 2 MFCC feature extraction flow chart

$$C_n = \sum_{m=0}^{N-1} S_m^* \cos \frac{n*(m-0.5)*\pi}{M}, n = 1, 2, \ldots, L \tag{5}$$

In practical application, the cepstrum difference parameter (delta cepstrum) is calculated following the value of L MFCC cepstrum coefficients, which is expressed as [18]:

$$d_n = \begin{cases} C_{n+1} - C_n, & n < K \\ \frac{\sum_{k=1}^{K} k(C_{n+k} - C_{n-k})}{\sqrt{2\sum_{k=1}^{K} k^2}}, & else \\ C_n - C_{n-1}, & n \geq L - K \end{cases} \tag{6}$$

where d represents the nth first-order difference result, C_n represents the nth cepstrum coefficient calculated by Formula (5), L represents the order when calculating MFCC, and K represents the time difference of the first-order derivative, which can be taken as 1 or 2. The second-order difference result can be obtained when the calculation result is brought into Formula (6).

Generally, MFCC and the first- and second-order cepstrum difference parameters are combined as the eigenvector of the signal.

3 GRNN

GRNN is a feedforward neural network. This neural network is based on kernel regression analysis [21] and has a good nonlinear mapping ability. The input characteristic is one-dimensional sequence. GRNN obtains the conditional probability density function by calculating the input and output of the training data and the input of the test data to further obtain the output of the test data.

3.1 Theoretical basis

The theoretical basis of the GRNN model is kernel regression analysis. Kernel regression analysis is a kind of nonlinear regression analysis. $f(x, y)$ is assumed to be the joint probability density function of random variables x and y. The regression of non-independent variable y to independent variable x (conditional mean) calculates y with the maximum probability value. Then, the regression of y to x is as follows:

$$E(y|x) = \frac{\int_{-\infty}^{\infty} yf(x, y)dy}{\int_{-\infty}^{\infty} f(x, y)dy} \tag{7}$$

GRNN estimates $f(x, y)$ according to the training data. Let $X_i (i = 1, 2, \cdots, N)$ and Y_i $(i = 1, 2, \cdots, N)$ be the observed values of the training samples of x and y, respectively, and X and Y are the observed values of the test samples of x and y, respectively. Then, the estimated density function can be provided according to the Parzen nonparametric estimation [22]:

$$\hat{f}(X, Y) = \frac{1}{N(2\pi)^{(M+1)/2}\sigma^{M+1}} \sum_{i=1}^{N} \exp\left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}\right] \exp\left[-\frac{(Y - Y_i)^2}{2\sigma^2}\right] \tag{8}$$

where N is the number of training samples, M is the dimension of random variable x, and σ is the spread factor.

By substituting $\hat{f}(X,Y)$ into $f(x,y)$ in Formula (7) and exchanging the order of integration and summation, $\hat{Y}(X)$ can be calculated as follows:

$$\hat{Y}(X)=\frac{\sum_{i=1}^{N}\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]\int_{-\infty}^{\infty}y\exp\left[-\frac{(Y-Y_i)^2}{2\sigma^2}\right]dy}{\sum_{i=1}^{N}\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]\int_{-\infty}^{\infty}\exp\left[-\frac{(Y-Y_i)^2}{2\sigma^2}\right]dy} \tag{9}$$

Given that $\int_{-\infty}^{\infty}\lambda\exp(-\lambda^2)d\lambda=0$, Formula (9) can be simplified as [23]

$$\hat{Y}(X)=\frac{\sum_{i=1}^{N}Y_i\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]}{\sum_{i=1}^{N}\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]} \tag{10}$$

According to Formula (10), $\hat{Y}(X)$ can be regarded as the weighted average of the observed values (Y_i) of all the training samples, and the weight of each observed value (Y_i) is the exponent of the square of the Euclidean distance between the observed value of the corresponding training sample (X_i) and the observed value (X) of the test sample. When σ approaches zero, $\hat{Y}(X)$ is approximate to observation value Y_i of the training sample corresponding to the observation value (X_i) nearest observation value X. When σ is very large, $\hat{Y}(X)$ is approximate to the mean value of all training sample observations Y_i. σ controls the width of the RBF and determines the fitting degree of the network. It is the only parameter to be optimized. Only when the spread factor is appropriate can we obtain a satisfactory fitting result.

3.2 Network structure

GRNN includes the input, pattern, summation, and output layers. GRNN only needs one network parameter, while other neural network models generally need to select multiple parameters. Therefore, GRNN has obvious advantages in network construction. Figure 3 shows the structure of the GRNN model. The purpose of each layer of GRNN is as follows:

The input layer passes the input vector to the pattern layer.

The number of neurons in the pattern layer is consistent with the number of training samples (N), and each training sample has a corresponding neuron. Neuron G_i is the exponent of the square of the Euclidean distance between training data input X_i and test data input X [23]:

$$G_i=\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right],\ i=1,2,\ldots,N \tag{11}$$

where σ is the spread factor to be selected.

The summation layer includes two types of neurons: the D and S neurons. The D neuron is the arithmetic summation of all the neurons in the pattern layer, and the S neuron is the weighted summation of all the neurons in the pattern layer (the weight between

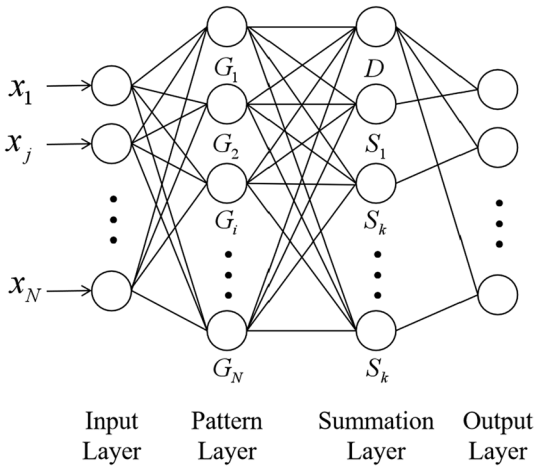


Fig. 3 Structure diagram of GRNN model

the ith neuron in the pattern layer and the kth neuron in the summation layer is y_{ik}). D and S neurons can be expressed as [21]

$$D = \sum_{i=1}^{N} G_i. \tag{12}$$

$$S_k = \sum_{i=1}^{N} y_{ik} G_i \tag{13}$$

The number of neurons in the output layer is dimension K of the output vector, and the output of the corresponding neurons is calculated by dividing the corresponding pattern layer (S_k) by D:

$$\hat{y}_k(X) = \frac{S_k}{D},\ k = 1, 2, \ldots, K \tag{14}$$

3.3 Spread factor determination

The GRNN model has only one network parameter, i.e., spread factor. Only by optimizing the spread factor can the training performance of the network be improved. Generally, the smaller the spread factor is, the better the fitting degree is, but an extremely small spread factor can easily lead to overfitting. In this study, the input feature of the GRNN model is the extracted one-dimensional MFCC feature, and the k-fold cross-validation is used to determine the optimal spread factor. The steps are as follows:

(1) The value range of the spread factor is determined, e.g., from 0.1 to 1.0 with an increment of 0.1.

(2) Recognition accuracy is taken as the measurement index of the recognition result, and the formula is as follows:

$$\text{Accuracy} = \frac{\overline{N}}{N} \times 100\% \tag{15}$$

where N is the number of samples and $\overline{N}$ is the number of correctly recognized samples in N samples.

(3) For k-fold cross-validation, randomly dividing the training samples into k folds with onefold used as validation set and the other k-1 folds used as training set.

(4) For each spread factor, building the network with training set using this spread factor and applying the resulting network for validation set and recording the accuracy; repeating this process k times so that all the k folds are used as the validation set once and computing the average of all the k accuracies (namely, as the averaged accuracies).

(5) The procedure (4) is repeated for all the spread factors, and the spread factor corresponding to the minimum value of the averaged accuracies is taken as the optimal spread factor.

3.4 Computer configuration

The proposed method was developed on a workstation with 11th Gen Intel(R) Core(TM) i7-1165G7 CPU*8. The code was written using MATLAB R2020b (https://www.mathworks.com/).

4 Results and discussion

4.1 Dataset

In this study, calls of leopard, ross, and weddell seals are conducted. Figure 4 shows the distribution maps of leopard, ross, and weddell seals. It shows that the above three kinds of seals have extensive overlap living areas in Antarctica. Therefore, it is of great significance to realize the recognition of seal calls automatically.

The data used in this study are the audio signals of leopard, ross, and weddell seals in Watkins marine mammal sound database [25]. This database provides a wide range of sounds from the 1940's to the 2000's and contains approximately 2000 unique recordings of over 60 species of marine mammals, which can be used as a reference dataset for the detection of marine mammals in PAM data collected worldwide. Figure 5 shows the geographic regions collected by the Watkins database, and the seal audio used in this study is from the elliptical area. Figure 6 shows the waveforms of the calls of different kinds of seals.

4.2 Extraction of input features

The data used in this study are from 'Best of Cuts' in Watkins marine mammal sound database. The sound clips of 'Best of Cuts' are considered to be of higher quality and lower noise. The sampling rates of leopard, ross and weddell seal data are 5000 Hz, 20000 Hz and 20480 Hz, respectively. On the one hand, the sampling rate needs to be unified. On the other hand, the sound characteristics of seal calls are concentrated in the low-frequency part. Therefore, all sounds are resampled at 5000 Hz. For the MFCC feature and LOFAR spectrum extracted from the resampled audios, data is divided into frames with a window length of 25 ms for 125 samples with a step size of 10 ms for 50

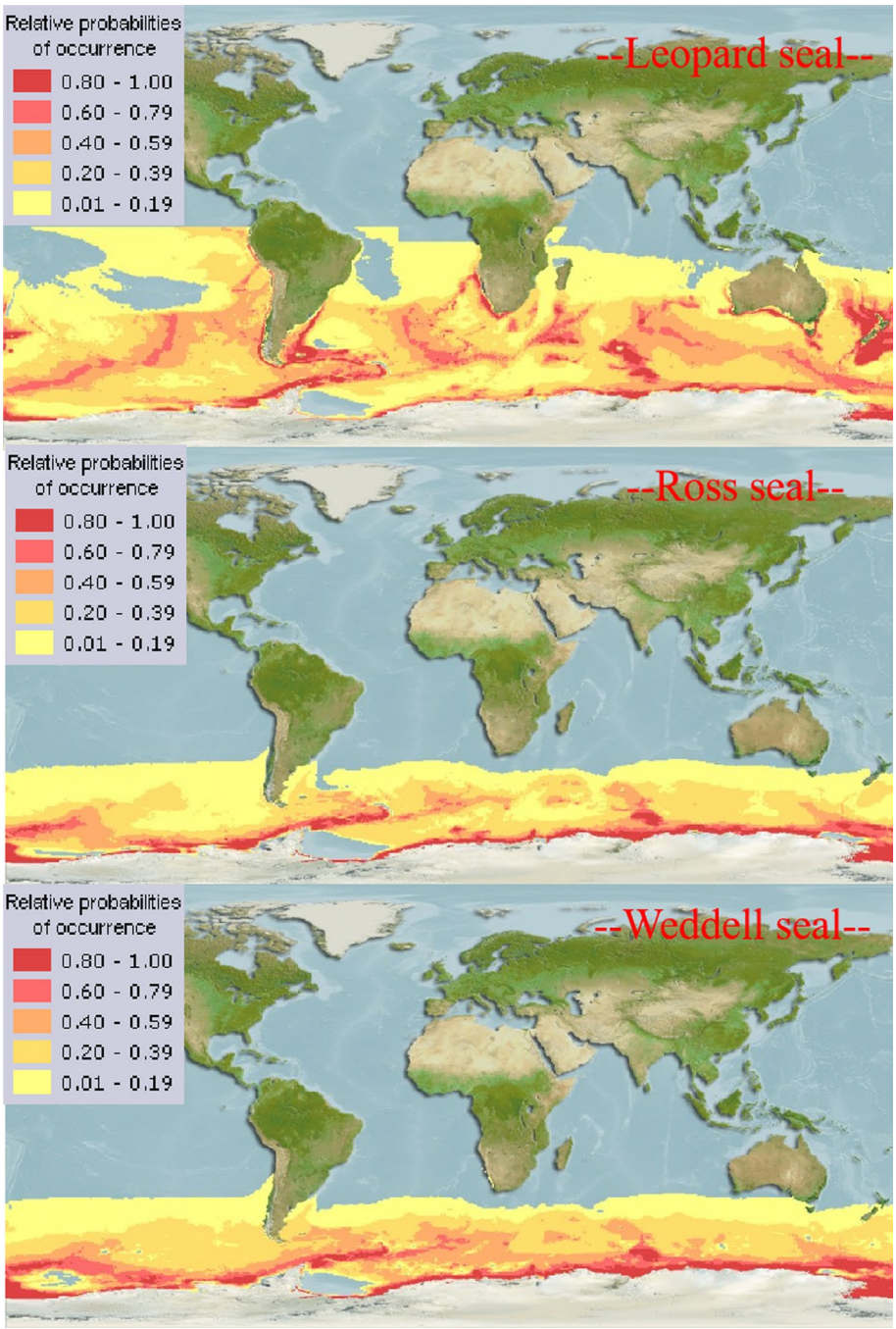


Fig. 4 Distribution maps of leopard, ross, and weddell seals [24]

samples. Table 1 shows the information of raw recordings of seal calls and number of samples after framing.

In the process of extracting MFCC features of each segment of data, 24 groups of filters are set, and their first- and second-order difference coefficients are calculated, so that each segment of data can obtain a 1×36 feature vector. This is used

Fig. 5 Geographic regions collected by the Watkins database [25]

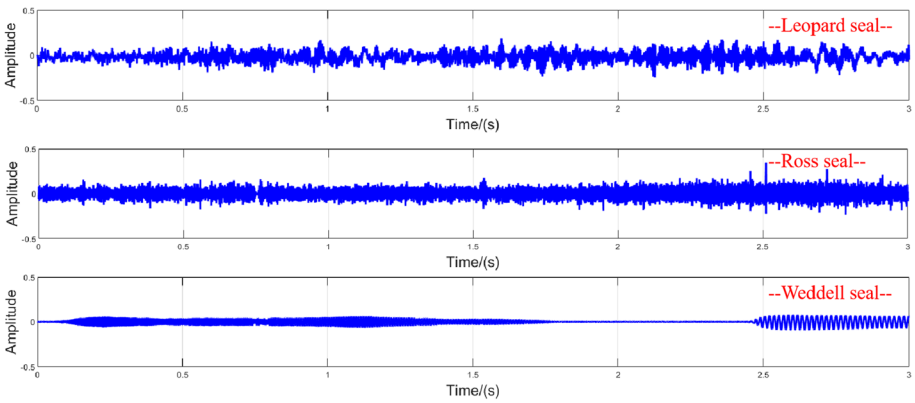
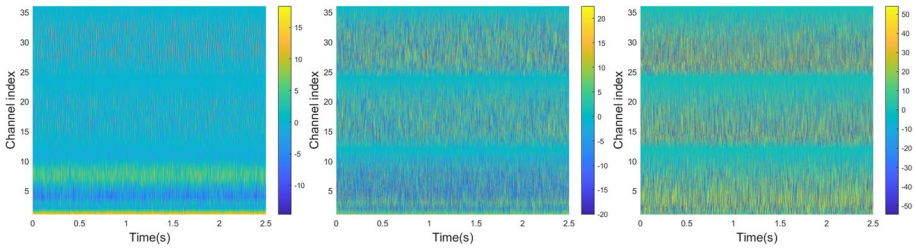


Fig. 6 Sound waveforms of different kinds of seal calls

Table 1 The information of raw recordings of seal calls

Name	Recording time and location	Total duration (s)	Number of samples
Leopard	Jan-1966. Cape Jones, Antarctica	52	5218
Ross	Jan-1966. Cape Hallett, Antarctica	72	7194
Weddell	Nov-1964. McMurdo Sound, Antarctica	43	4296

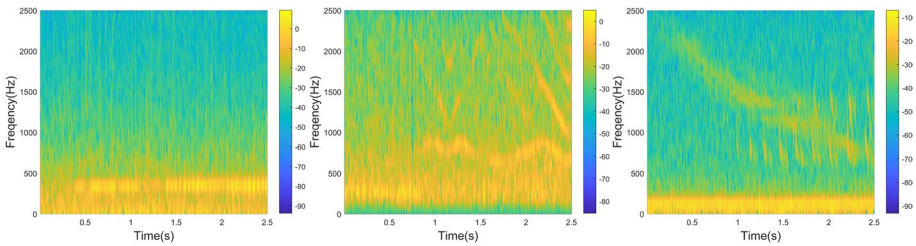
as the feature input of a single sample. Leopard, ross, and weddell seals all take the MFCC feature of the 2.5-s sound data as an example as shown in Fig. 7a–c. In this study, 5218, 7194 and 4296 MFCC features are extracted from the calls of leopard seal, ross seal and weddell seal, respectively, with a total of 16,708 MFCC features. We randomly select 75% of each category for training and the rest for testing. A total of 12,531 training samples and 4177 test samples are obtained.



(a) Leopard seal (b) Ross seal (c) Weddell seal

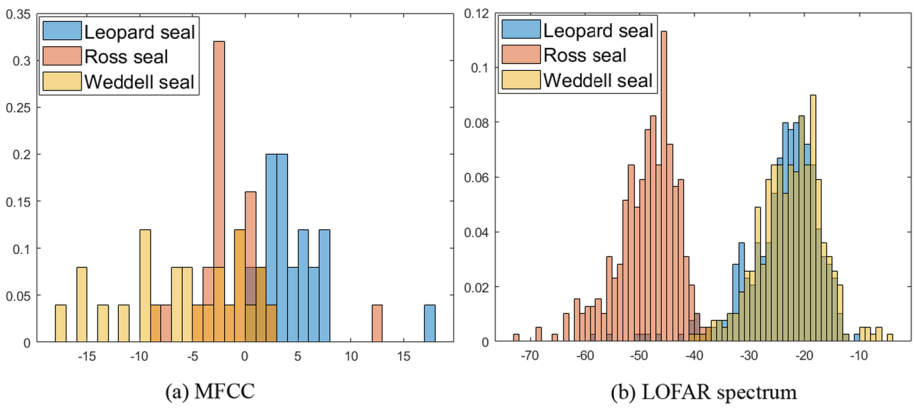
Fig. 7 MFCC features of different kinds of seal calls

In this study, low-frequency analyzer and recorder (LOFAR) spectrum of signal energy variation characteristics in the time–frequency domain can be obtained by continuous sampling of signal and short-time Fourier transform (STFT) of continuous signal samples. For STFT, the frequency interval of STFT is set to 10 Hz. The LOFAR spectrum of the three kinds of seal calls are shown in Fig. 8. It can be seen that the spectrum characteristics change significantly with time. For LOFAR spectrum, the input feature is the amplitude of all frequency points under each time frame, and its dimension is 1×251. Similarly, a total of 16,708 LOFAR spectrum are extracted. When allocating training and test samples, ensure that the time frame corresponding to the training and test data of LOFAR spectrum is consistent with the MFCC features, so that the subsequent recognition performance comparison between the MFCC characteristics and the LOFAR spectrum is meaningful.



(a) Leopard seal (b) Ross seal (c) Weddell seal

Fig. 8 LOFAR spectrum of different kinds of seal calls



(a) MFCC (b) LOFAR spectrum

Fig. 9 Statistical histograms of MFCC features and LOFAR spectrum

The statistical histograms of MFCC features and LOFAR spectrum are shown in Fig. 9. The different colors represent three types of seals; Fig. 9a shows that the three types of seal calls have obvious differences in terms of MFCC, which are specifically expressed in the location, shape, skewness and kurtosis of the distributions. Figure 9b shows that the calls of leopard and weddell seals are very similar in terms of LOFAR spectrum, which greatly reduces the recognition performance.

4.3 Design of recognition model

In practical application, the ocean often has different degrees of environmental noise. This paper analyzes the recognition performance of this method at different SNRs. Gaussian white noise is added to the original audio signals. Taking the calls of leopard seals as an example, Fig. 10 shows a comparison of the waveform of the original data and the corresponding samples at different SNRs. It can be obtained that the lower the SNR, the more burrs appear in the signal.

In this study, three machine learning models, namely GRNN, support vector machine (SVM) and CNN, are designed to study the recognition performance with the input feature of MFCC and LOFAR spectrum at different SNRs.

For the GRNN model, the spread factor is determined using tenfold cross-validation. The spread factor range is set to be [0.01 : 0.01 : 0.10.2 : 0.1 : 1.0]. Figure 11a–b, respectively, shows the relationship between the spread factor and the recognition performance of the GRNN method using MFCC features and LOFAR spectrum at different SNRs. The results show that the recognition performance of the MFCC features and LOFAR spectrum decreases with the increase in the spread factor, and the optimal spread factor is concentrated between 0 and 0.1. Compared with LOFAR spectrum, GRNN method with MFCC feature input feature has higher average accuracy after determining the optimal spread factor.

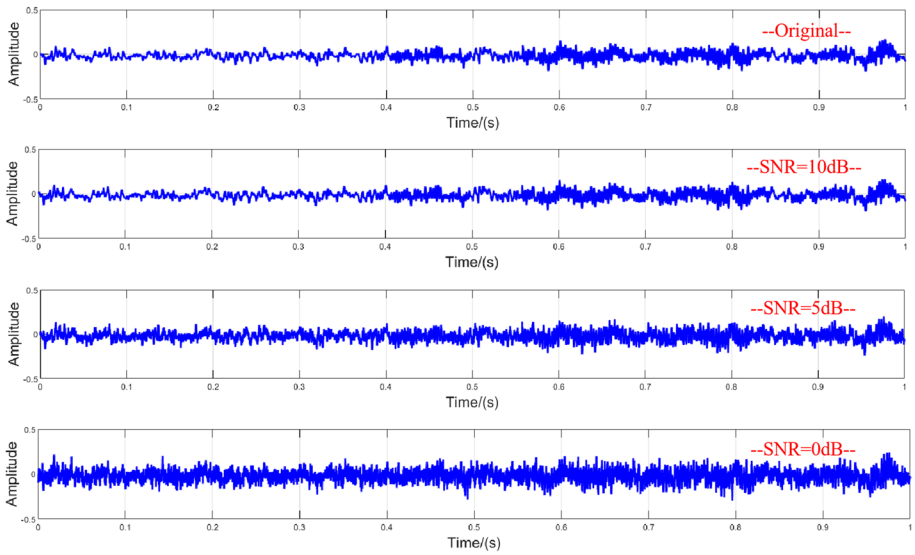


Fig. 10 Waveform at different SNRs

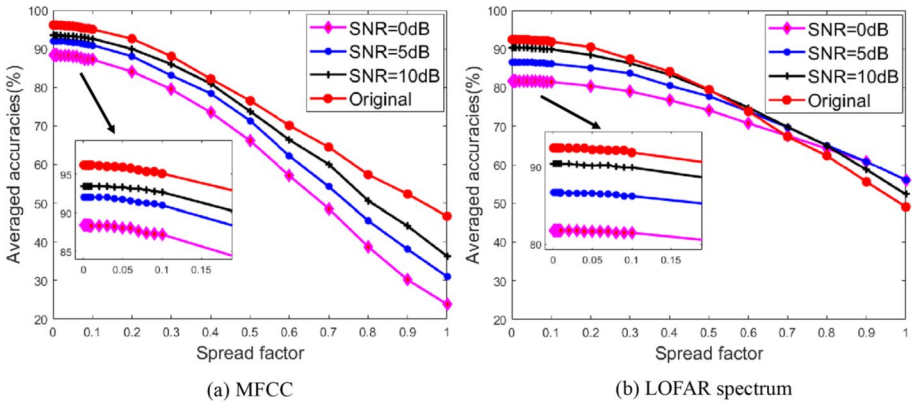


Fig. 11 Determination of the optimal spread factor of GRNN at different SNRs

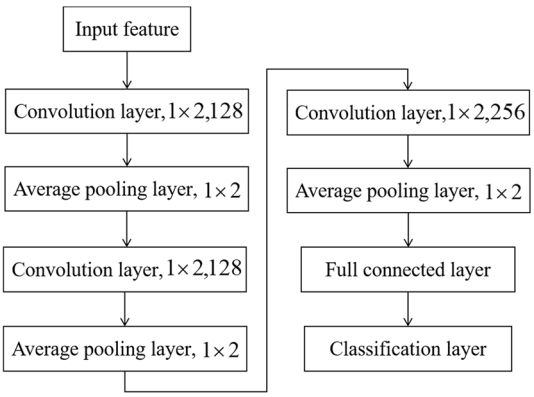


Fig. 12 CNN structure

For the SVM model, it solves the inner product operation in high-dimensional space by introducing kernel function [26]. In this study, the radial basis function (RBF) is used as the kernel function of the SVM model.

RBM is a kind of randomly generated neural network that can obtain the probability distribution by training the data set. RBM contains a visible layer and a hidden layer. In this paper, RBM is combined with the softmax activation function, and the number of hidden units is set to 200 [10].

CNN is a special neural network. The convolution layer extracts the local features of the data through the convolution kernel while reducing the impact of unrelated factors [27]. Most of the CNN models used in the field of underwater acoustic recognition of marine mammals have a relatively few number of layers, including 2–4 convolution layers [13, 15, 17, 28]. The CNN classification model built in this paper has three convolution layers. Convolution kernel sizes are all set to 1×2, and number of convolution kernels is set of 128, 128 and 256, respectively. After each convolution layer, the rectified linear unit (ReLU) is used as the activation function. The pooling layer reduces the dimensionality based on the local correlation of the feature data using the average-pooling method while keeping the feature scale constant. An

average-pooling layer is set behind each ReLU, and the size of the pooling filter is set of 1×2. The features extracted through the convolution and pooling operations are passed through the fully connected layers, which are connected to the output layer. The CNN structure is shown in Fig. 12. The obtained model is trained by Stochastic Gradient Descent with Momentum (SGDM) algorithm with the learning rate of 0.0001. The minibatch size is set of 10. The CNN model is trained until the loss function converges. The cross-entropy is used as the loss function.

In 2015, He et al. [29] studied the CNN model with residual connection in depth, which can be used to build a deeper network architecture. Residual CNN is a widely used and state-of-the-art model. For residual CNN, this paper uses the typical ResNet18 as the recognition model. The network parameter configuration is consistent with the above common CNN.

Tables 2 and 3 show the training and testing time of different models with the input feature of MFCC feature and LOFAR spectrum, respectively. The results show that since GRNN has only one network parameter, and compared with LOFAR spectrum, the MFCC feature has fewer parameters, the GRNN method for inputting MFCC feature has the least training and testing time, while the ResNet18 model has the longest time due to the most network parameters.

Table 2 The training time (s) of different models

Feature	GRNN	SVM[26]	CNN	ResNet[29]	RBM[10]
MFCC	**15.5**	32.8	80.6	125.8	38.2
LOFAR	20.8	42.3	105.1	151.2	50.9

Table 3 The testing time (s) of different models

Feature	GRNN	SVM[26]	CNN	ResNet[29]	RBM[10]
MFCC	**1.9**	3.8	7.5	8.1	4.3
LOFAR	2.5	5.4	8.9	9.2	6.1

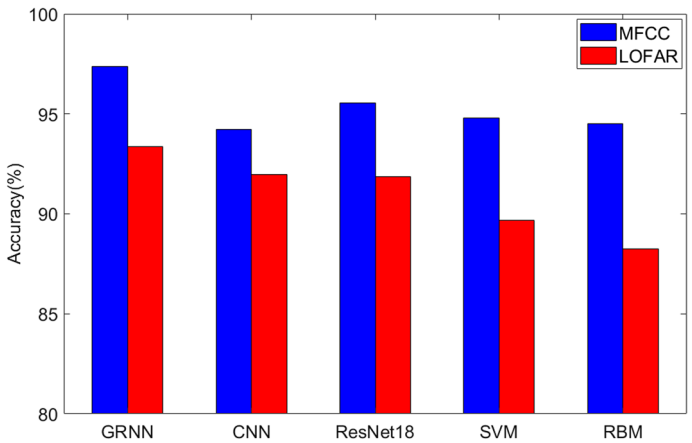


Fig. 13 Comparison of the recognition results using MFCC features and LOFAR spectrum of the original data

4.4 Recognition results

Figure 13 shows the comparison of the three machine learning methods with the input feature of MFCC and LOFAR spectrum of the original data. The results show that in the scene without added noise, GRNN method has better recognition performance than other methods for the same input features, and the accuracy of GRNN method with the input features of MFCC reach 97.36%. When MFCC features are given as input, the recognition performance of RBM, CNN and SVM are similar, and that of ResNet18 is slightly better. When inputting LOFAR spectrum, the recognition performance of CNN and ResNet18 is obviously better than that of SVM. The recognition accuracy of RBM is only 88.53%, while that of GRNN, CNN, ResNet18 and SVM models are 93.41%, 92.39%, 92.08% and 90.12%, respectively, that is, RBM has the worst recognition performance. The MFCC features are obtained according to the human ear auditory mechanism, which can extract the feature information of the low-frequency part more effectively, and the frequency of the seal calls is mainly concentrated in the low-frequency part, so the recognition performance of MFCC features is obviously better than that of the LOFAR spectrum for the same method.

Table 4 Accuracy (%) of seal calls recognition by the GRNN model

SNR/dB	MFCC				LOFAR spectrum			
	A	B	C	Total	A	B	C	Total
20	99.12	97.58	94.01	**96.90**	95.25	92.32	91.05	**92.87**
15	97.83	96.02	93.91	**95.92**	91.71	91.35	90.69	**91.25**
10	93.55	95.53	91.24	**93.44**	92.96	89.23	89.28	**90.49**
5	91.96	94.67	89.37	**92.00**	90.17	85.82	83.21	**86.40**
0	88.01	90.65	86.47	**88.38**	84.42	81.71	79.81	**81.98**

Table 5 Accuracy (%) of seal calls recognition by the CNN model

SNR/dB	MFCC				LOFAR spectrum			
	A	B	C	Total	A	B	C	Total
20	98.85	97.13	91.71	**94.23**	94.51	91.25	90.15	**91.97**
15	93.61	96.73	90.19	**93.51**	93.47	90.02	88.37	**90.62**
10	92.52	95.24	90.61	**92.79**	89.90	89.88	86.32	**88.70**
5	90.48	93.46	89.18	**91.04**	86.75	85.05	80.71	**84.17**
0	85.91	85.21	85.44	**85.52**	78.52	75.38	73.89	**75.93**

Table 6 Accuracy (%) of seal calls recognition by the ResNet18 [29] model

SNR/dB	MFCC				LOFAR spectrum			
	A	B	C	Total	A	B	C	Total
20	98.32	93.80	92.82	**94.98**	94.25	90.19	89.52	**91.32**
15	96.92	93.76	91.38	**94.02**	93.85	89.09	89.61	**90.85**
10	94.21	93.22	90.31	**92.58**	88.03	85.31	83.71	**85.69**
5	91.55	90.15	85.06	**88.92**	85.19	83.22	80.38	**82.93**
0	85.59	82.92	80.28	**82.56**	74.52	74.41	71.21	**73.38**

Table 7 Accuracy (%) of seal calls recognition by the SVM [26] model

SNR/dB	MFCC				LOFAR spectrum			
	A	B	C	Total	A	B	C	Total
20	94.51	97.31	92.52	**94.78**	91.85	88.02	89.17	**89.68**
15	92.10	95.25	90.33	**92.56**	89.36	86.71	85.31	**87.12**
10	91.10	93.35	89.66	**91.37**	84.64	86.47	81.37	**84.16**
5	88.05	85.39	83.02	**85.48**	81.66	81.85	79.91	**81.14**
0	81.65	82.00	81.63	**81.76**	74.02	70.92	68.03	**70.99**

Table 8 Accuracy (%) of seal calls recognition by the RBM [10] model

SNR/dB	MFCC				LOFAR spectrum			
	A	B	C	Total	A	B	C	Total
20	94.06	93.68	91.83	**93.19**	90.36	87.38	85.69	**87.81**
15	92.42	93.08	90.65	**92.05**	87.80	86.71	83.25	**85.92**
10	91.02	86.23	87.32	**88.19**	82.53	79.50	79.62	**80.55**
5	82.38	79.43	74.95	**78.92**	77.08	74.29	73.72	**75.03**
0	75.97	69.27	65.03	**70.09**	64.05	65.94	60.57	**63.52**

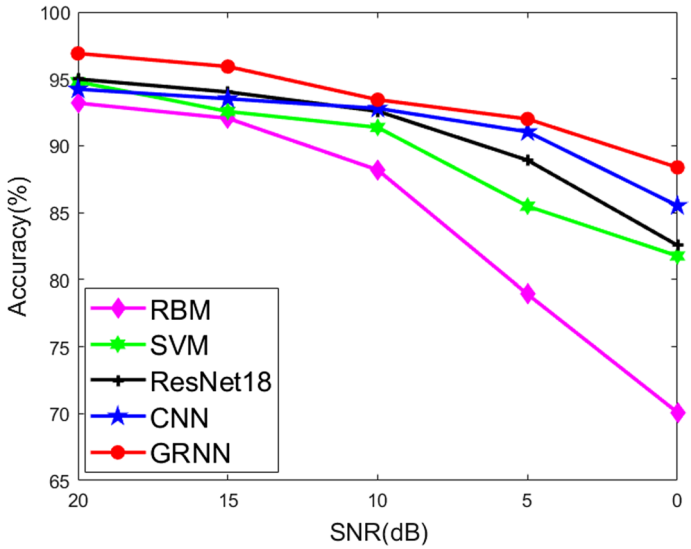


Fig. 14 Comparison of the five models with the input feature of MFCC

Tables 4, 5, 6, 7 and 8 show the recognition accuracy of GRNN, CNN, ResNet18, SVM and RBM models, respectively, inputting MFCC features and LOFAR spectrum at different SNRs. A, B, and C represent the leopard, ross, and weddell seals, respectively.

Figure 14 shows the change in the accuracy curves with SNR using MFCC features as input features. These models can effectively recognize seal calls at high SNRs, but the reduction of SNR has a great impact on the RBM method. The error of the RBM method is large and cannot realize effective recognition when the SNR is 0 dB.

Compared with RBM, SVM and ResNet18, the CNN method is less affected by the reduction of SNR, but the results of CNN method also have a large error at low SNRs, especially at 0 dB. Because the texture, gradient and other features of underwater acoustic signal are not obvious, namely there is less detailed information, too deep network is easy to cause gradient disappearance and overfitting during the training. Therefore, compared with traditional CNN, the estimation performance of ResNet18 is poor at low SNRs. The SNR reduction has the least impact on the GRNN method. The GRNN method can realize accurate recognition when the SNR exceeds 5 dB, and it can still realize approximate recognition when the SNR is 0 dB.

Figure 15 shows the change in the accuracy curves with SNR using LOFAR spectrum as input features. For the GRNN, ResNet18 and CNN models, inputting the LOFAR spectrum can realize effective recognition at high SNRs, but the error is large at low SNRs. For the RBM and SVM method, inputting the LOFAR spectrum can only realize approximate recognition, which is no longer applicable at low SNRs. Especially for RBM, it is difficult to determine the optimal network parameters, which leads to the convergence of the training to the local optimal solution. It is also easy to overfit. In addition, RBM belongs to the traditional shallow neural network, which is difficult to extract deep features. Therefore, compared with other models, its recognition results are poor at low SNRs.

Figure 16 shows the recognition accuracy of different methods with the input feature of MFCC and LOFAR spectrum. Due to the poor recognition performance of RBM and SVM, only GRNN, ResNet18 and CNN are shown here. The results show that the underwater acoustic features have a greater impact on the recognition results than the recognition models. At different SNRs, the recognition results of MFCC features using various models are better than those of LOFAR spectrum.

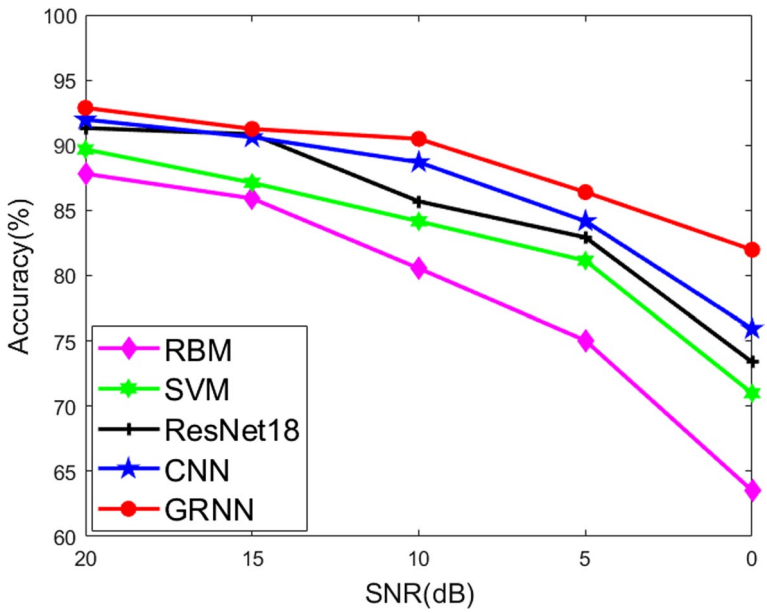


Fig. 15 Comparison of the five models with the input feature of LOFAR spectrum

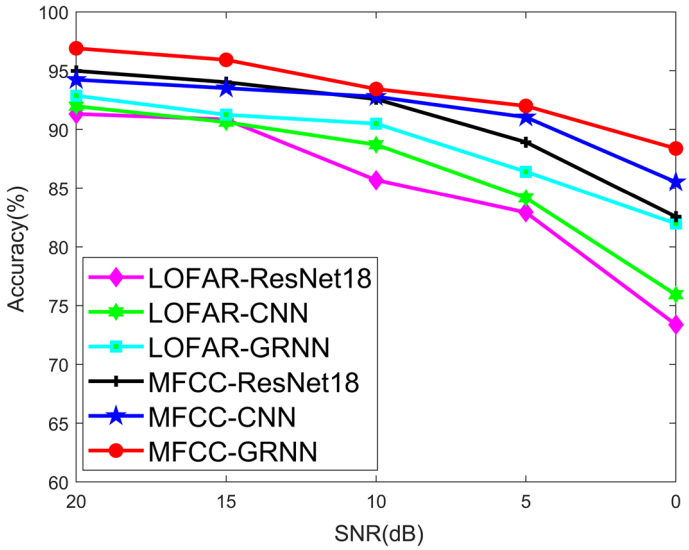


Fig. 16 Comparison of the various models with the input feature of MFCC and LOFAR spectrum

In general, because the GRNN model for inputting MFCC features combines the feature advantages of MFCC and the nonlinear fitting ability of the GRNN, the seal call recognition can be realized stably at various SNRs, and its recognition accuracy of the test samples is better than that of other models at the same SNR. Few studies related to seals, especially studies on the automatical recognition of seal calls, are noted in the field of underwater sound recognition of marine mammals. All results show that the proposed method has excellent performance for the seal call recognition, which fills the blank in this field.

5 Conclusion

This paper presents an accurate seal call recognition method based on the GRNN with the input feature of MFCC. The RBM, SVM, CNN, ResNet18 and GRNN models are compared for inputting MFCC and LOFAR spectrum. The results show that at the same SNR, the recognition result of the MFCC feature is better than that of LOFAR spectrum. The accuracy of GRNN method with the input features of MFCC of the original data reaches 97.36%. Compared with other models, the GRNN model has better recognition performance and can still realize effective recognition at low SNRs. In particular, GRNN has the least training and testing time. In addition, although this method is used for seal call recognition in this paper, it is not limited to this. In the field of passive acoustics, this method is also applicable to the preliminary work of other underwater vocal biological species observation (e.g., distance and seasonal measurement, abundance estimation, and population structure determination). It can also be used in other target sound recognition (e.g., natural sound and ship noise recognition).

Abbreviations

GRNN	General regression neural network
MFCC	Mel-frequency cepstrum coefficient

LOFAR	Low-frequency analyzer and recorder
SNR	Signal-to-noise ratio
PAM	Passive acoustic monitoring
CNN	Convolutional neural network
FFT	Fast Fourier transform
STFT	Short-time Fourier transform
SVM	Support vector machine
RBF	Radial basis function
ReLU	Rectified linear unit
SGDM	Stochastic gradient descent with momentum
RBM	Restricted Boltzmann machine

Acknowledgements
Not applicable.

Author contributions
All authors contributed to the conception and design of the experiments and the interpretation of simulation results. Y conceived the idea, prepared the manuscript, and conducted numerical and experimental validations. W substantially revised the manuscript, and Y and S contributed additional revisions of the text. All authors read and approved the final manuscript.

Funding
This work was supported in part by National Key R&D Program of China (2021YFB3203001) and Shaanxi's Young Science and Technology Star Program under Grant 2021KJXX-07.

Availability of data and materials
The dataset supporting the conclusions of this article is available at https://cis.whoi.edu/science/B/whalesounds/index.cfm.

Declarations

Competing interests
The author declares that they have no competing interests.

Received: 29 August 2022 Accepted: 18 April 2023
Published online: 01 May 2023

References

1. A.W. Watkins, In-air and underwater sounds of the Ross seal *Ommatophoca rossi*. J. Acoust. Soc. Am. **77**(4), 1598–1598 (1985). https://doi.org/10.1121/1.392003
2. J.A. Mossbridge, J.A. Thomas, An "acoustic niche" for Antarctic killer whale and leopard seal calls. J. Acoust. Soc. Am. **109**(5), 2390–2390 (2001). https://doi.org/10.1121/1.4744428
3. T.L. Rogers, G.A. Rowney, M.B. Ciaglia, D.H. Cato, Seasonal and diurnal calling patterns of ross and leopards. J. Acoust. Soc. Am. **118**(3), 1938–1938 (2005). https://doi.org/10.1121/1.4780983
4. D. Risch, C.W. Clark, P.J. Corkeron, Vocalizations of male bearded seals, *Erignathus barbatus*: classification and geographical variation. Anim. Behav. **73**(5), 747–762 (2007). https://doi.org/10.1016/j.anbehav.2006.06.012
5. H. Frouin-Mouy, X. Mouy, B. Martin, Underwater acoustic behavior of bearded seals (*Erignathus barbatus*) in the northeastern Chukchi Sea, 2007–2010. Mar. Mammal Sci. **32**(1), 141–160 (2016). https://doi.org/10.1111/mms.12246
6. W.D. Halliday, S.J. Insley, T. Jong, Seasonal patterns in acoustic detections of marine mammals near Sachs Harbour, Northwest Territories. Arct. Sci. **4**, 259–278 (2017)
7. P.A. Cziko, L.M. Munger, N.R. Santos, J.M. Terhune, Weddell seals produce ultrasonic vocalizations. J. Acoust. Soc. Am. **148**(6), 3784–3796 (2020). https://doi.org/10.1121/10.0002867
8. J.R. Potter, D.K. Mellinger, C.W. Clark, Marine mammal call discrimination using artificial neural networks. J. Acoust. Soc. Am. **96**(3), 1255–1262 (1994). https://doi.org/10.1121/1.410274
9. X. Mouy, M. Bahoura, Y. Simard, Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence. J. Acoust. Soc. Am. **126**(6), 2918–2928 (2009). https://doi.org/10.1121/1.3257588
10. X.C. Halkias, S. Paris, H. Glotin, Classification of mysticete sounds using machine learning techniques. J. Acoust. Soc. Am. **134**(5), 3496–3505 (2013). https://doi.org/10.1121/1.4821203
11. I. Parisi, G. Vincenzi, M. Torri, E. Papale, Underwater vocal complexity of Arctic seal *Erignathus barbatus* in Kongsfjorden (Svalbard). J. Acoust. Soc. Am. **142**(5), 3104–3115 (2017). https://doi.org/10.1121/1.5010887
12. H. Frouin-Mouy, X. Mouy, C.L. Berchok, S.B. Blackwell, K.M. Stafford, Acoustic occurrence and behavior of ribbon seals (*Histriophoca fasciata*) in the Bering, Chukchi, and Beaufort seas. Polar Biol. **42**(4), 657–674 (2019). https://doi.org/10.1007/s00300-019-02462-y
13. W. Luo, W. Yang, Z. Yu, Convolutional neural network for detecting odontocete echolocation clicks. J. Acoust. Soc. Am. **145**(1), 7–12 (2019). https://doi.org/10.1121/1.5085647
14. M. Zhong, M. Castellote, R. Dodhia, J.L. Ferres, A. Brewer, Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. **147**(3), 1834 (2020). https://doi.org/10.1121/10.0000921
15. Y. Shiu, K.J. Palmer, M.A. Roch, E. Fleishman, H. Klinck, Deep neural networks for automated detection of marine mammal species. Sci. Rep. **10**(1), 1–12 (2020). https://doi.org/10.1038/s41598-020-57549-y

16. B. Mishachandar, S. Vairamuthu, Diverse ocean noise classification using deep learning. Appl. Acoust. (2021). https://doi.org/10.1016/j.apacoust.2021.108141
17. C.D. Escobar-Amado, M. Badiey, S. Pecknold, Automatic detection and classification of bearded seal vocalizations in the northeastern Chukchi Sea using convolutional neural networks. J. Acoust. Soc. Am (2022). https://doi.org/10.1121/10.0009256
18. K. Yang, X. Zhou, Deep learning classification for improved bicoherence feature based on cyclic modulation and cross-correlation. J. Acoust. Soc. Am **146**(4), 2201–2211 (2019). https://doi.org/10.1121/1.5127166
19. L. Muda, M. Begam, I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques. Ttps, (2010), 2. https://doi.org/10.48550/arXiv.1003.4083.
20. T. Lim, K. Bae , C. Hwang. Classification of underwater transient signals using MFCC feature vector. in *2007 9th international symposium on signal processing and its applications*. https://doi.org/10.1109/ISSPA.2007.4555521
21. D.F. Specht, A general regression neural network. IEEE Tran. Neural. Netw. **2**(6), 568–576 (1991). https://doi.org/10.1109/72.97934
22. E. Parzen, On estimation of a probability density function and mode. Ann. Math. Statist. **33**(3), 1065–1076 (1962). https://doi.org/10.1214/aoms/1177704472
23. T. Masters, W. Land, A new training algorithm for the general regression neural network. IEEE **3**, 1990–1994 (1997). https://doi.org/10.1109/ICSMC.1997.635142
24. K. Kaschner, K.Kesner-Reyes, C. Garilao, J. Segschneider, J. Rius-Barile, T. Rees, R. Froese, AquaMaps: Predicted range maps for aquatic species. World wide web electronic publication (2019). http://www.aquamaps.org.
25. L. Sayigh, M.A. Daher, J. Allen, H. Gordon, K. Joyce, C. Stuhlmann, Fourth international conference on the effects of noise on aquatic life the watkins marine mammal sound database: an online, freely accessible resource 040013 (2019).
26. C. Saunders, M.O. Stitson, J. Weston, Support vector machine. Com. Sci. **1**(4), 1–28 (2002)
27. K. Fukushima, S. Miyake, T. Ito, Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Trans. Syst. Man Cybern. (1983). https://doi.org/10.1109/TSMC.1983.6313076
28. J. Jiang, L. Bu, F. Duan et al., Whistle detection and classification for whales based on convolutional neural networks. Appl. Acou. **150**, 169–178 (2019). https://doi.org/10.1016/j.apacoust.2019.02.007
29. K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition. IEEE Conf. Com. Vis. Pat. Rec. (2016). https://doi.org/10.1109/CVPR.2016.90

Publisher's Note