Article:

Jiang, J.-J., Bu, L.-R., Duan, F.-J. et al. (4 more authors) (2019) Whistle detection and classification for whales based on convolutional neural networks. Applied Acoustics, 150. pp. 169-178. ISSN 0003-682X

https://doi.org/10.1016/j.apacoust.2019.02.007

Whistle Detection and Classification for Whales Based on Convolutional Neural Networks

Jia-jia Jiang [#, a, 1)], Ling-ran Bu [#, a)], Fa-jie Duan[a)], Xian-quan Wang [a)], Wei Liu [b)], Zhong-bo Sun [a)] and Chun-yue Li [a)]

[a] The State Key Lab of Precision Measuring Technology and Instruments, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, China

[b] The Department of Electronic and Electrical Engineering, University of Sheffield, United Kingdom[a]

[#] Jia-jia Jiang and Ling-ran Bu contributed equally to this work and should be considered co-first authors.

[1] The correspondence author is Jia-jia Jiang (E-mail: jiajiajiang@tju.edu.cn.)

Abstract

Passive acoustic observation of whales is an increasingly important tool for whale research. Accurately detecting whale sounds and correctly classifying them into corresponding whale species are essential tasks, especially in the case when two species of whales vocalize in the same observed area. Whistles are vital vocalizations of toothed whales, such as killer whales and long-finned pilot whales. In this paper, based on deep convolutional neural networks (CNNs), a novel method is proposed to detect and classify whistles of both killer whales and long-finned pilot whales. Compared with traditional methods, the proposed one can automatically learn the sound characteristics from the training data, without specifying the sound features for classification and detection, and thus shows better adaptability to complex sound signals. First, the denoised sound to be analyzed is sent to the trained detection model to estimate the number and positions of the target whistles. The detected whistles are then sent to the trained classification model, which determines the corresponding whale species. A GUI interface is developed to assist with the detection and classification process. Experimental results show that the proposed method can achieve 97% correct detection rate and 95% correct classification rate on the testing set. In the future, the presented method can be further applied to passive acoustic observation applications for some other whale or dolphin species.

1. Introduction

Passive acoustic observation has been used increasingly widely in the field of whale species research. Several countries, such as the USA [1], Australia [2] and a few European countries [3], have established underwater passive acoustic monitoring (PAM) systems to detect and monitor cetacean species such as whales or dolphins. Compared with visual observation methods, PAM has a better monitoring performance. In addition, it can continue at night, in poor weather, and under other conditions in which visual observation is not feasible. These PAM systems can be used to measure the range and seasonal occurrence of whales [4], estimate the quantity of a species in a given area [5], and determine the population structure, etc. [6,7]. For the above applications, an important condition is to detect and identify the target whale signals from the sounds recorded by PAM systems. Accurately detecting whale sounds and correctly classifying various whale sounds into their corresponding whale species can assist observers to monitor the occurrence (appearance) of whales and confirm their species, and so it is a fundamental and primary task in the PAM of whales [8]. Most of the current tasks of whale sound detection and classification still need to be implemented manually. On the one hand, due to the different levels of experience and different sensitivities to sounds, the performance of manual methods varies with operators; on the other hand, the commonly stated range for human hearing is 20Hz to 20kHz [9], and information of sound outside this range cannot be effectively acquired by human ear. Furthermore, it is difficult for manual methods to process the large amount of sound data generated by the large-scale PAM networks such as *Listen to the Deep Ocean Environment* (LIDO) program [3,10]. Automatic methods for whale sound detection and classification is highly desired in this context.

However, due to the unknown statistical signal properties, as well as the use of different recording equipment and low signal to noise ratio (SNR) conditions, automatic detection and classification of marine mammal sounds is still a challenging task in the field of animal bioacoustics.

Several whale sound detection and classification methods have been proposed in the past. Typically, these methods follow the following steps: sound preprocessing, whale sound detection, feature extraction of detected sounds, and feature classification. Among them, Short Time Fourier Transform (STFT) [11-15], Wavelet Transform (WT) [16] and Hilbert Huang Transform (HHT) [17] were used to extract features of whale sounds. Artificial Neural Network (ANN) [13,16], Support Vector Machine (SVM) [11,17] and Sparse Representation-based Classifier (SRC) [18] were used for classifying the extracted features. However, the features extracted by the above methods are generally fixed specific features which are commonly used in sound processing, such as Mel-scale Frequency Cepstral Coefficients (MFCC), STFT Coefficients, Wavelet Coefficients, and Energy Spectrums.

On the one hand, these common features may make it difficult to effectively characterize differences between different types of sound signals to be classified, resulting in low classification performance. On the other hand, these simple features may not be able to adequately characterize the complex and varying time-frequency characteristics of sound signals (such as whale whistles with varied contours or harmonics), leading to a poor classification performance for complex whale signals. Further, with the ongoing upgrade of sound recording equipment and the change of the recording environment, these methods may be difficult to adapt to the large amount of newly recorded data. Besides, there are some low-energy whale sounds, such as whale whistles, that are easily submerged in noise. Traditional methods based on energy or zero-crossing rate cannot provide high detection performance. Therefore, it is necessary to develop an automatic detection and classification method with good adaptability and high performance.

Generally, whale sounds can be categorized as whistles, clicks, and pulsed calls, etc.[19-21]. Whale whistles are vital vocalizations that are widespread in a variety of whales such as killer whales (*Orcinus orca*) and long-finned pilot whales (*Globicephala melas*) [19-21]. Killer whales and long-finned pilot whales are two typical toothed whale species that can produce a wide variety of whistles, clicks and pulsed calls for echolocation and social signaling. Whistles, which are an important vocalization for both whale species, are considered to be used as contact calls between individuals or to maintain group contact during foraging and traveling [3,19-21]. In some monitoring areas, long-finned pilot whales are believed to produce whistles similar in frequency and structure to killer whales, especially in the ultrasonic range [22].

Furthermore, killer whales and long-finned pilot whales are abundant in quantity and widespread in distribution. There is a wide range of overlapped distribution areas between killer whales and long-finned pilot whales. Previous evidence has shown that both whale species may be present in the same area [21, 22]. In passive acoustic monitoring of the two whale species, it is necessary and important to first distinguish and identify their individual whistles from their all kinds of mixtures. In this paper, based on deep convolutional neural networks (CNNs), we propose a novel whistle detection and classification method for both killer whales and long-finned pilot whales. First of all, the method can adaptively learn to extract features that can effectively characterize the sounds to be detected and classified through training data, and implement detection and classification of whale whistles based on these features. Secondly, the whale sounds detected and classified by the trained CNNs model can be sent to the CNNs model for further training and optimization after initial simple screening, which provides the possibility to improve the accuracy of detection and classification further.

This paper is organized as follows. Section 2 describes the details of the sounds used in this paper and the preprocessing steps. Section 3 introduces the algorithms used for denoising, detection, feature extraction and classification and Section 4 presents the experimental process and the results. Finally, the conclusions are drawn in Section 5.

2. Sound Data and Preprocessing

We selected 15 sound samples containing either killer whale sound or long-finned pilot whale sound as raw data for generating the data set for the detection and classification model. The recording date of these sounds varies from 1967 to 2002. The total duration of these sounds is about 120 minutes with a sampling rate of 44100 Hz. The sound recording locations include the waters near Antarctica, Canada, Norway, Mexico, and the United States. These sounds mainly contain killer whale sounds (whistles), long-finned pilot whale sounds (whistles and clicks), background noise and other non-target sounds (ship noise and pulse interference). All these sounds are preprocessed as follows.

2.1 Denoising

Firstly, the raw sound data is denoised using the spectral subtraction method [23] to reduce background noise. This method is based on spectral averaging and residual noise reduction, widely used for enhancement of noisy speech signals and can remove the stationary noise included in the sound. The incoming sound signal is buffered and divided into blocks of 256 samples with 128 samples overlapping adjacent blocks. Each block is *Hamming* windowed and then transformed by Discrete Fourier Transform (DFT) to the frequency domain. The over-subtraction factor α is set to 10, and the magnitude estimate factor β is set to 0.02. After spectral subtraction, the magnitude spectrum is combined with the phase of the noisy signal, and transformed back to the time domain. Each signal block is then overlapped and added to the preceding and succeeding blocks to form the final denoised sound signal. Figs. 1(a) and 1(b), as well as Figs. 2(a) and 2(b), show a comparison of the original whale sound and the denoised one.

2.2 Frame Spectrogram

All the denoised sounds in the data set are sequentially cut into sound frames with a duration of t_d (no overlapping between adjacent frames). The sound frame with a length of less than t_d at the end of the sound file is discarded. The Short Time Fourier Transform (STFT), with *Hamming* window, a segment length of $t_d/40$, segment shift of t_d /80 and FFT length of 1024 samples, is computed for each sound frame. In order to show more details in the spectrogram, the STFT coefficients are logarithmized by Eq. (1).

$$Z = \log_{10}\left(|Z|\right) \tag{1}$$

where Z is the STFT coefficients matrix for each sound frame.

If the value of t_d is too small, some short-term pulse interference may also be misdetected; if the value of t_d is too large, the signal detection accuracy is lowered. Based on the durations of the whistles from both whale species, t_d is set to 250ms. In addition, the time interval between most adjacent whistles is greater than t_d, so the paper does not discuss the case where two whistles are falsely detected as a whole whistle due to the short signal interval($<t_d$).

Further, for each sound frame, based on the preprocessed STFT coefficients Z, a frame spectrogram (grayscale) of 180*120 pixels is obtained by the *pcolormesh* method in matplotlib [24] to visualize the STFT result. Fig. 1(b) and Fig. 2(b) show the start and end positions of the frames for the denoised sound, and Fig. 1(c) and Fig. 2(c) show the corresponding frame spectrograms. As can be seen, the contours of whistles have been enhanced.

By viewing the corresponding waveforms and spectrograms, we manually mark the sound frames containing whistles and their corresponding spectrograms as label A (whistles of killer whale) or label B (whistles of long-finned pilot whale). As shown in Fig. 1(c) and Fig. 2(c), these sound frames and spectrograms may only contain part of a complete whistle. Other non-target sounds are marked as label C. These labeled frame spectrograms are used to train and test the whistle detection model.

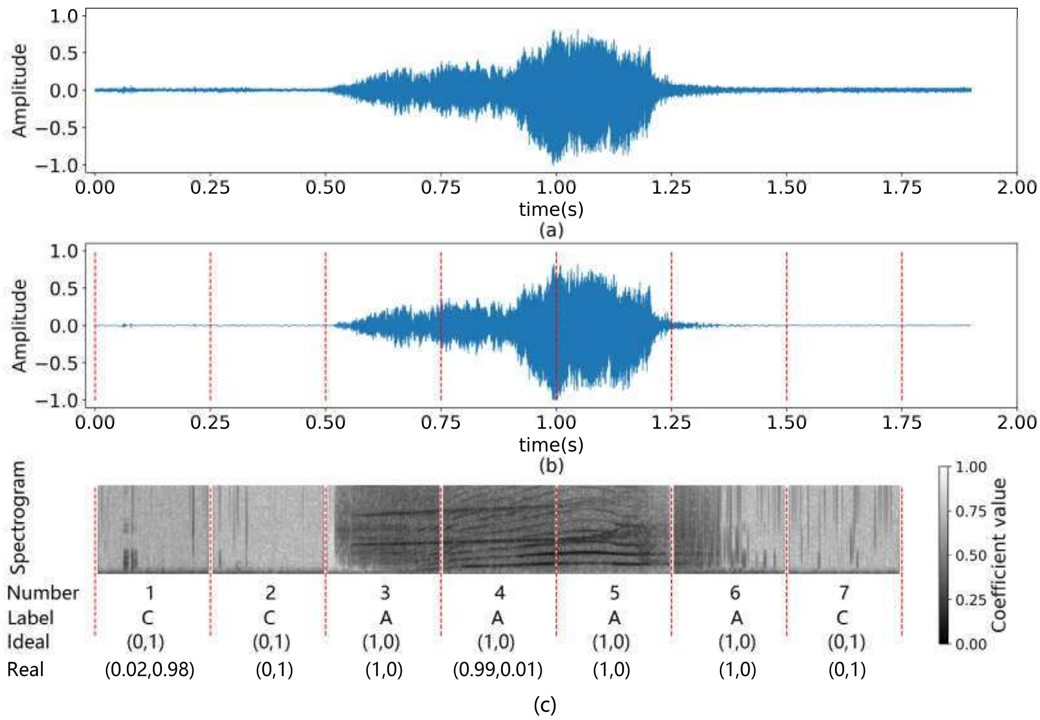

Fig. 1. The preprocessing steps for a whistle signal of the killer whale.

(a) The original whistle waveforms. (b) The denoised whistle waveforms; the red dotted lines are the dividing lines between adjacent frames with the frame duration t_d=25ms; the signal above 1.75ms is deleted because its length is less than t_d. (c) The frame spectrograms corresponding to the frames in (b); their labels (A or C) and the ideal outputs ((1,0) or (0,1)) in the whistle detection model are listed under the spectrograms; their real outputs in the trained detection model (obtained in Section 4.1) are listed under the ideal outputs.

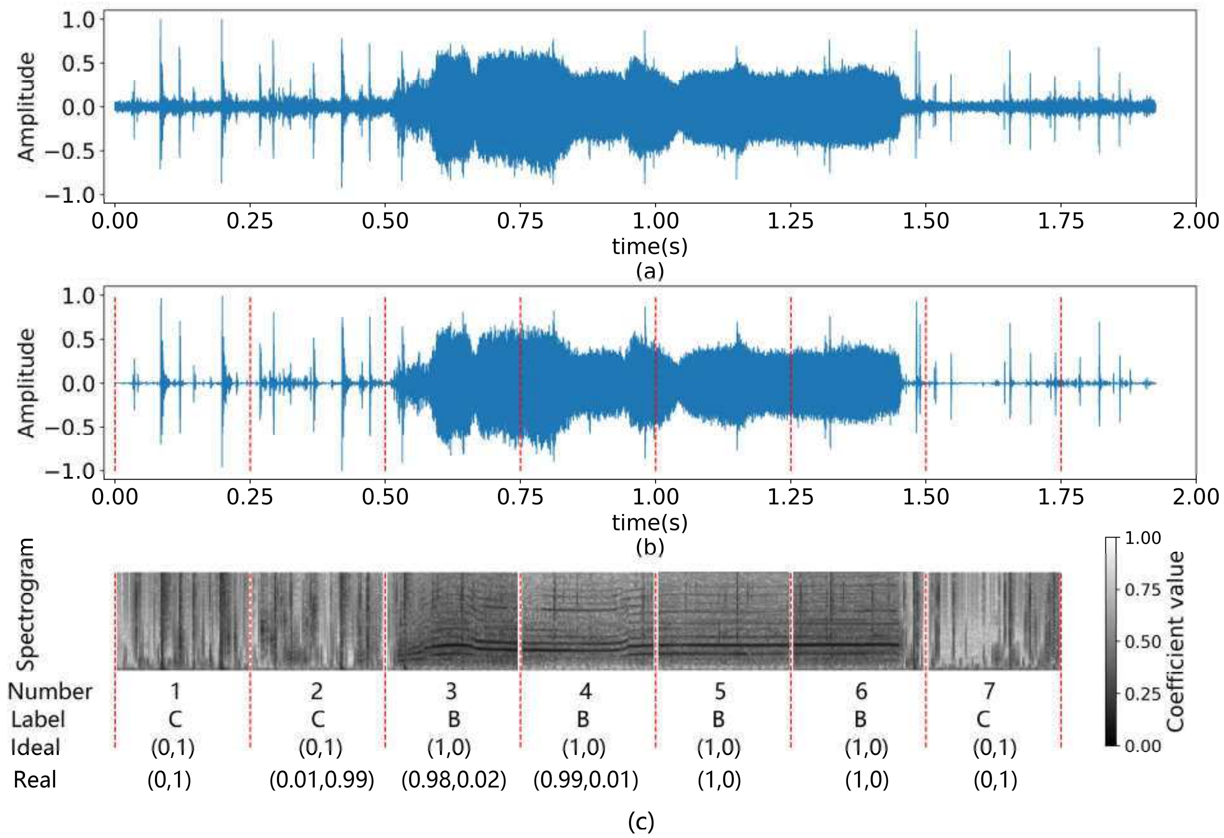

Fig. 2. The preprocessing steps for a whistle signal of long-finned pilot whale.

(a) The original whistle waveforms. (b) The denoised whistle waveforms; the red dotted lines are the dividing lines between adjacent frames with the frame duration t_d=25ms; the signal above 1.75ms and 2ms is deleted because its length is less than t_d. (c) The frame spectrograms corresponding to the frames in (b); their labels (B or C) and the ideal outputs((1,0) or (0,1)) in the whale sound detection model are listed under the spectrograms; their real outputs in the trained detection model (obtained in Section 4.2) are listed under the ideal outputs.

2.3 Whistle Spectrogram

In addition to cutting all sound data into fixed-length frames (t ms) in Section 2.2, we also manually extract the complete killer whale and long-finned pilot whale whistle signals from the denoised sound data. As shown in Fig. 3, the extracted sound contains the complete whistle signal, and their length is variable. According to the spectrogram calculation method described in Section 2.2, we calculate the STFT coefficients for each extracted whistle signal and visualize the results. The parameters used in this process are the same as those used in Section 2.2. Thus corresponding to each complete whistle signal, a spectrogram (grayscale) of 180*120 pixels can be obtained. Based on the whale species corresponding to the whistle, we manually mark these whistle signals and their corresponding spectrograms as label A (whistles of killer whale) or label B (whistles of long-finned pilot whale). These labeled whistle spectrograms are used to train and test the classification model.

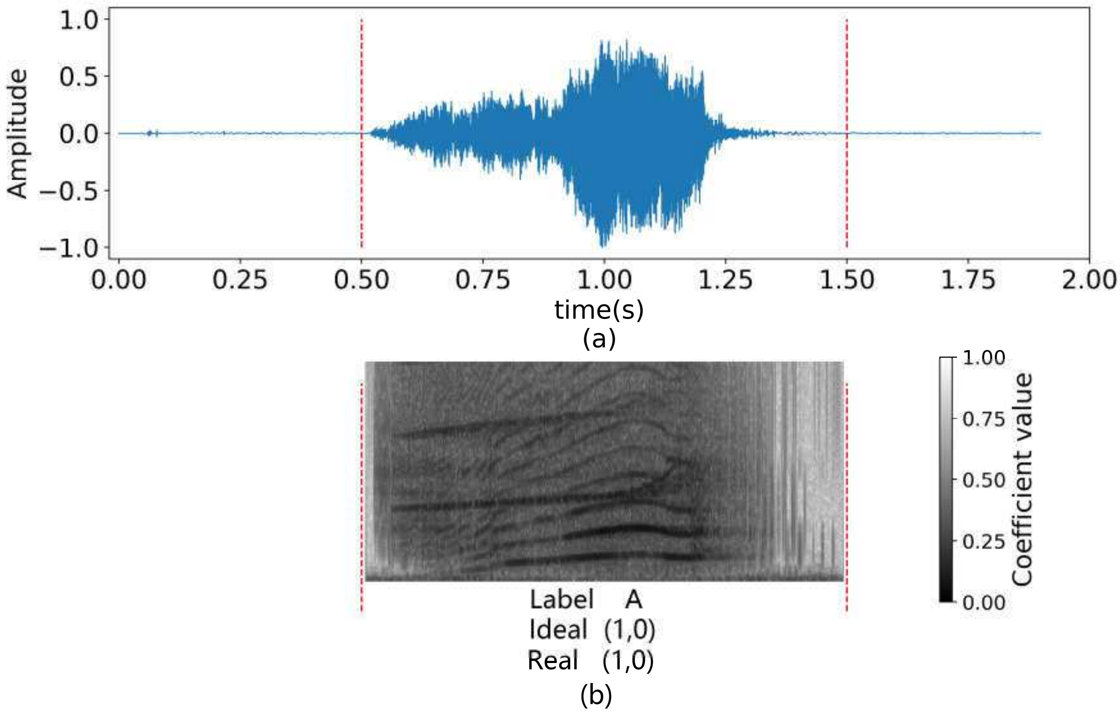

Fig. 3. The detected whistle of killer whale (a) and its whistle spectrogram (b).

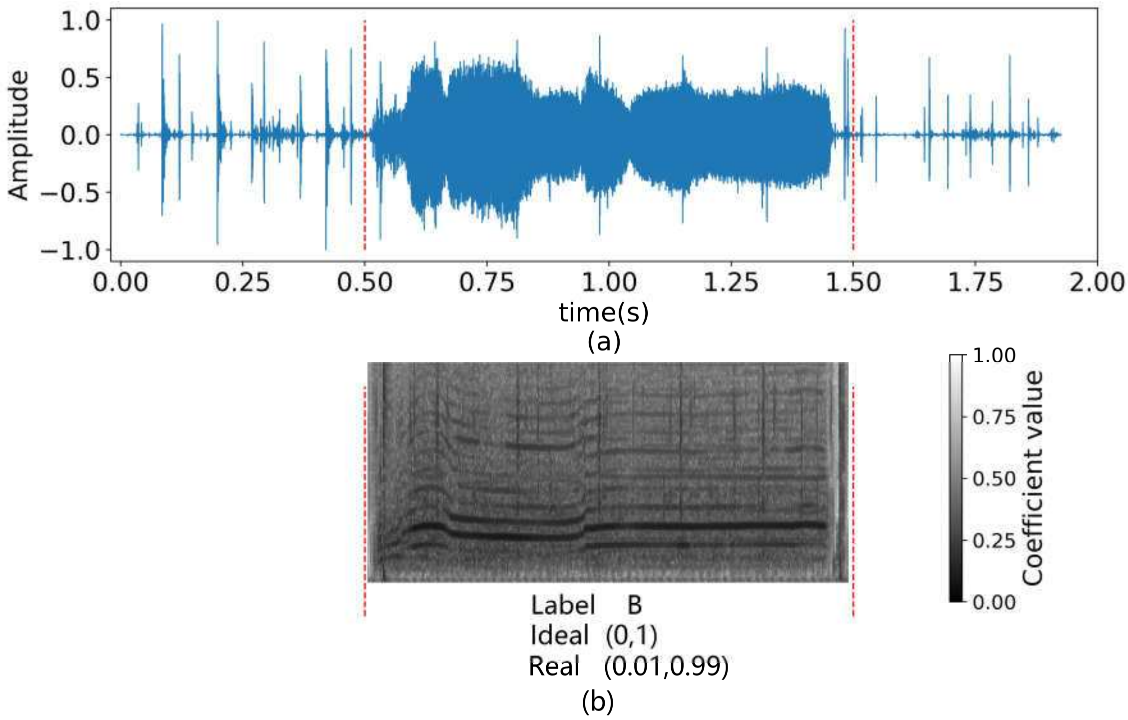

Fig. 4. The detected whistle of long-finned pilot whale (a) and its whistle spectrogram (b).

3. Description of Algorithms

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [25] are a class of deep feed-forward artificial neural networks, which are most commonly employed to analyze images. CNNs have been tremendously successful in practical applications, and already demonstrated good performance in many speech-related [26] and music-related tasks [27]. There are three important characteristics of CNNs: sparse interactions, parameter sharing, and equivariant representations [28]. Based on the above three characteristics, CNNs can well perceive the 2D structural features of the input images.

In this paper, based on CNNs, a whistle detection model is designed together with a whistle classification model. Firstly, the detection model and the classification model are trained respectively by labeled frame spectrograms and labeled whistle spectrograms data set obtained in Sections 2.1 and 2.2. Then, in the process of detecting and classifying the target whistles in unknown sound, the trained detection model takes the frame spectrograms of the unknown sound as inputs, and only judges whether the corresponding frame spectrograms contains whistles or not. Furthermore, based on outputs of the detection model, the number and positions of whistles in the input sound can be estimated, and then the detected complete whistle signal is extracted from the sound. Next, spectrograms of the detected whistles are calculated and sent to the trained classification model in turn. Finally, the classification model predicts the whale species to which the input spectrograms belong (killer whale or long-finned pilot whale). Through the above processes, the whistles in the input sound can be detected and classified into the corresponding whale species.

3.2 Whale Whistles Detection Model

The LeNet5 [29] model can achieve a high recognition accuracy of 99.2% on the MNIST handwritten digit set, and it is relatively simple compared to other CNN structures. As can be seen from Fig. 3(c) and Fig. 4(c), there are contours similar to handwritten numbers in the time-frequency spectrograms of whistles from both whale species. Therefore, this paper draws on the structure of LeNet5 to design the detection model and the classification model. The structure of the detection model is shown in Fig. 5. The hyperparameters of each layer of the detection model are as follows:

(1) C1 is a convolutional layer containing 32 convolution kernels of size 5*5. The convolution step is 1 (stride) with padding, and the ReLU function is used as the activation function of output.

(2) S2 is a pooling layer, and the pooling strategy is average pooling with pooling size 2*2, pooling step 2, and full 0 padding.

(3) C3 is a convolutional layer containing 64 convolution kernels of size 5*5, the convolution step is 1 (stride) with padding, and the ReLU function is used as the activation function of output.

(4) S4 is a pooling layer, and the pooling strategy is the average pooling with pooling size 2*2, pooling step 2, and full 0 padding.

(5) F5 is a fully connected layer containing 64 neurons, and each neuron is fully connected with all output units of layer S4, and the ReLU function is used as the activation function.

(6) D6 is a dropout layer with dropout rate 0.2.

(7) F7 is a fully connected layer containing 2 neurons (corresponding to the final output layer),

and each neuron is fully connected with all output units of layer D6 with no activation function.

(8) S8 is a softmax layer that converts the output into a pair of probabilities (P1, P2), 0⩽P1, P2⩽1, and P1+P2 =1. If P1>P2, the model predicts that the input signal contains a whistle signal; otherwise the model judges that the input time-frequency diagram does not contain a whistle signal. Therefore, for a frame spectrogram labeled A or B, the ideal output y of the model is (1,0), and for a frame spectrogram with label C, the ideal output y of the model is (0,1).

The cross entropy, which is widely used in softmax output classification, is adopted as the loss function of the detection model:

$$L = \sum_{i=1}^{2} y * \log(\hat{y}) \tag{2}$$

where y is the ideal output of the model and $\hat{y}$ is the predicted output. The Adam optimization method [30] is applied to model optimization in order to adapt the learning rates of model parameters.

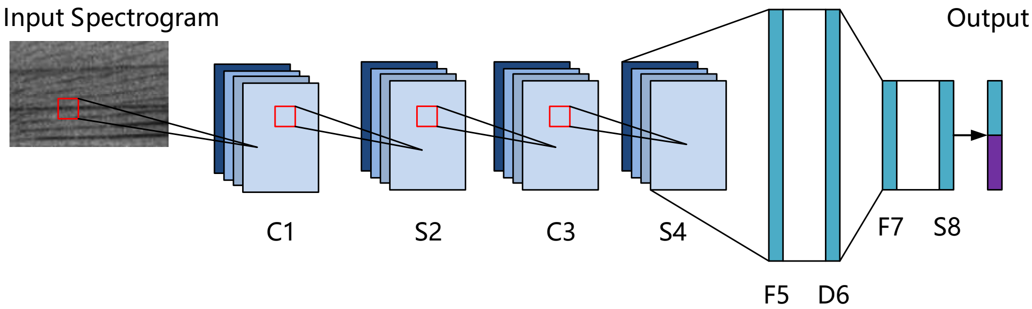

Fig. 5. Structure of the detection model and the classification model.

The trained detection model based on the labeled frame spectrograms can be used to detect the target whistles (the whistles of killer whale or long-finned pilot whale) in the input sound and determine the number and positions of target whistles in the input sound. The detection process is achieved by the following steps: firstly, as described in Sections 2.1 and 2.2, the input sound is denoised and cut into fixed length frames (t_d) in order. Frames are numbered sequentially with 1, 2, 3, …, n-1, n, where n is the total number of frames and the frame spectrogram is obtained for each numbered frame. Then, the frame spectrogram for each frame is fed into the trained model in turn, and the model outputs the probabilities (P_i1,P_i2). The *frame number-model output* sequence [i, (P_i1, P_i2)], where i ($1⩽i⩽n$) is the frame number, can be obtained through the above process. In the sequence, the position s ($1⩽s⩽n$) where the sequence changes from $P_s1<P_s2$ to $P_{s+1}1>P_{s+1}2$ can be regarded as the position where a whistle starts, and the position e ($1⩽e⩽n$) where the result changes from $P_e1<P_e2$ to $P_{e+1}1>P_{e+1}2$ can be regarded as the position where the whistle ends. The estimated start positions t_s and end positions t_e can be calculated by Eq. (3):

$$t_s = s * t_d$$
$$t_e = e * t_d \tag{3}$$

For example, in the *frame number-model output* sequence shown in Fig.1, according to the above calculation rule, s=2, and e=6. Therefore, it can be obtained by Eq. (3) that t_s=50ms and t_e=150ms. The detected complete whistle of Fig. 1 is shown in Fig. 3. Similarly, in Fig. 2, we can obtain that s=2 and e=6, and then t_s=50ms and t_e= 150ms. The detected complete whistle of Fig. 2 is shown in Fig. 4.

Through the above process, the number and positions of detected whistles in the input sound can be estimated respectively.

3.3 Whale Whistle Classification Model

The structure and hyperparameters of the classification model are the same as those of the detection model presented in Section 3.2; however, the input data, output results and training processes of the two models are different. The classification model takes the whistle spectrograms of the detected whistles as inputs to determine the corresponding whale species. More specifically, the input of classification model is whistle spectrograms (grayscale) of 180*120 pixels and the output is a pair of probabilities (R1,R2), $0 \leqslant R1,R2 \leqslant 1$, and R1+R2=1. If R1>R2, the model predicts that the input whistle is produced by a killer whale; otherwise it is produced by a long-finned pilot whale. Therefore, for the whistle spectrograms labeled A, the ideal output is (1,0), and for the whistle spectrograms labeled B, the ideal output is (0,1).

The classification model is first trained using the whistle spectrogram data set generated in Section 2.3, and then the trained model can be applied for classifying the two types of whale whistles. In the classification process, all the detected whistles are first cut from the sound according to the estimated start positions t_s and end positions t_e. For each whistle, the whistle spectrogram is obtained through the visualization method described in Section 2.3. All the whistle spectrograms are fed into the classification model in turn and classified into their corresponding whale species.

The whole detection and classification process are shown in Fig. 6. Through the two steps of detection and classification, the two types of whistles in the unknown sound are automatically positioned and classified into their whale species. For both the detection model and the classification model, there is no process of extracting time-frequency features directly from whistles. The inputs to both models are time-frequency spectrograms that characterize the overall information of whistles, rather than the specified features extracted by the specified algorithms. The feature extraction pattern and the calculated features of the two models are learned from the training data and its ideal output. Compared with the traditional detection and classification methods, the detection and classification algorithms proposed in this paper are more robust. Firstly, by optimizing the loss function, both models can learn and adjust CNN parameters, such as values of convolution kernels and weights of fully connected layers. Through this process, CNNs can adaptively learn from the input time-frequency spectrograms and extract deep features that are more suitable for detection or classification. Secondly, when new sound data is collected and filtered, these data can be used as raw data to train CNNs, so that CNNs can learn new features in new data. In the paper, these techniques are implemented with MATLAB R2014 and Python 3.6.

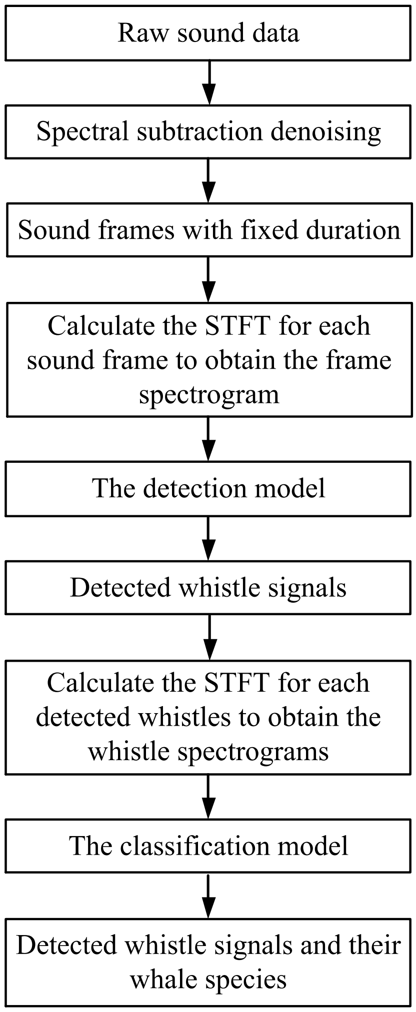

Fig. 6. The overall process of whistle detection and classification

4. Experiments

4.1 Detection Performance

The frame spectrogram data set calculated in Section 2.2 are used to train and test the detection model. The dataset contains 4028 frame spectrograms. Among them, the data set size corresponding to the ideal output (1,0) is 2054 (1298 for label A and 756 for label B), and the data set size of output (0,1) is 1974 (label C). The number of samples in the two data subsets is approximately balanced. We randomly extracted 200 images from each of the two data subsets as the testing set, and all the remaining spectrograms are used as the training set for detection model training.

The model is developed on a PC with Intel(R) Core(TM) i5-8400 CPU and NVIDIA GeForce GTX 1080 GPU. The code is written using TensorFlow 1.4.0, which is an open-source python library for dataflow programming across a range of tasks such as machine learning.

The weight parameters of each layer in the detection model are randomly initialized with zero mean and standard deviation of 0.1. The initial value of the learning rate is an empirical value of 0.01. The sum of cross entropy L_t in a batch is calculated and recorded as the loss in each epoch by Eq. (4):

$$L_t = \sum_{j=1}^{S}\sum_{i=1}^{2} y_j * \log(\hat{y}_j) \tag{4}$$

where S is the number of spectrograms in the training batch and testing batch (batch size), and in our paper, S=50, which means that 73 iterations can complete the traversal of the training data set. The model is trained for 25 epochs, and the accuracy on the testing set in each epoch is calculated and recorded too. Fig. 7(a) shows the average value (marked as L_m) of loss L_t in each epoch as it is being minimized during training. The loss goes as low as around 0 at the end of the training. After each epoch, the testing set is sent to the model to calculate and record the detection correct rate τ [13,31] on the model test set by Eq. (5).

$$\tau = N_c / N_s \tag{5}$$

where N_s is the amount of testing data (N_s=400 in our paper), and N_c is the amount of correctly classified data. Fig. 7(b) shows the curve for the detection correct rate τ. As can be seen, τ is stable at around 97% in the last eight epochs, which means most of the whistles in the testing data can be accurately detected. The detection model demonstrates good adaptability to the input frame spectrograms.

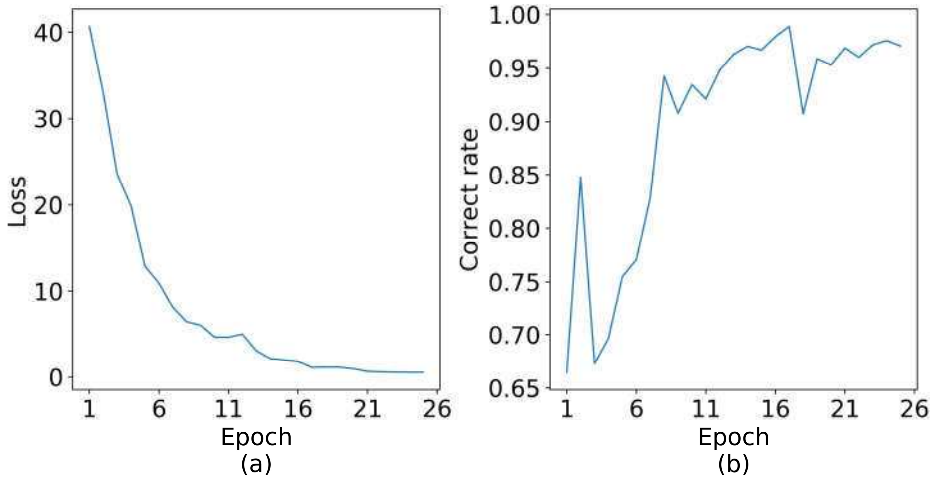

Fig. 7. The loss curve (a) and detection correct rate curve (b) of the detection model. For each epoch, we calculate L_t on each batch, and the average values of loss L_t are shown in (a). In each epoch, after the training is completed, we send the testing data set to the model, and the detection correct rates on the testing set of each trained model are shown in (b).

4.2 Classification Performance

The classification model is trained and tested using the whistle spectrogram data set obtained in Section 2.3. The dataset contains 980 whistle spectrograms (530 for label A, output (1,0) and 450 for label B, output (0,1)). 100 and 80 spectrograms are randomly extracted from the label A data set and label B data set respectively as the testing set, and all the remaining spectrograms are used as the training set for classification model training. The classification model is developed under the same software and hardware conditions as the detection model. The loss L_m and the classification correct rate τ on the testing set in each epoch are also calculated and recorded. As shown in Fig. 8(a), the mean loss decreases from 45 to 0.5 at the end of the training. Fig. 8(b) shows the classification correct rate curve on the testing set. At the beginning of the training (epoch 1), the model shows a poor classification performance. Then, the classification correct rate starts to improve gradually. At epoch 11, the model shows a correct rate higher than 0.9, which goes around 0.95 at the end of the training, meaning most of the whistles in the testing data can be correctly classified into their corresponding whale species.

The trained detection model and classification model are saved in the checkpoint file of TensorFlow.

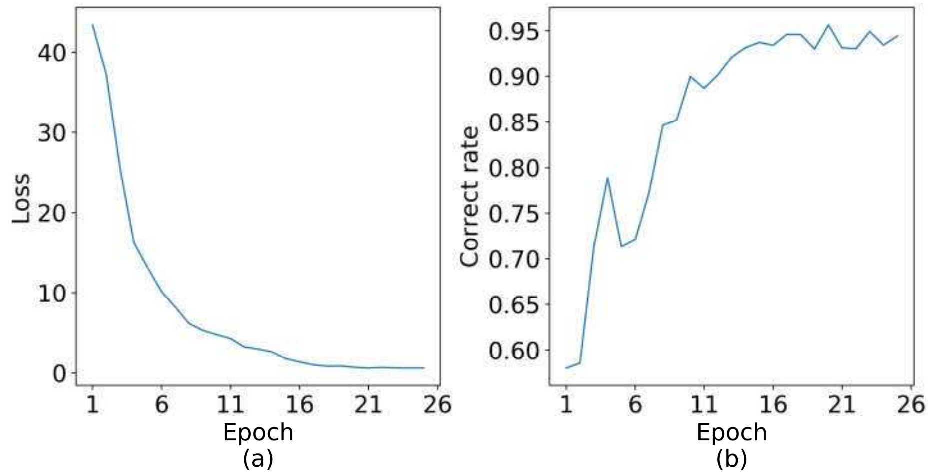

Fig. 8. The loss curve(a) and the classification correct rate curve(b) of the classification model. For each epoch, we calculate L_t on each batch, and the average values of loss L_t are shown in (a). In each epoch, after the training is completed, we send the testing data set to the model, and the classification correct rates on the testing set of each trained model are shown in (b).

4.3 Application

As shown in Fig. 9, using pyqt5, we have developed a GUI (graphical user interface) software to visualize both the detection and classification processes for the sound to be analyzed. First, the operator imports the sound file (*.wav* format) to be analyzed, and then imports the TensorFlow checkpoint files, including the trained detection model and the trained classification model. Further, the whistle detection and classification process can be performed automatically by the software. The log of the analysis processes will be displayed in the text box at the right side of the interface. The analyzing results, including the estimated start positions t_s and the end positions t_e, the whale species and their probabilities (the larger value of R1 and R2), will be saved in an Excel file. At the same time, the waveforms, spectrograms and classification results of the detected whistles can be viewed through the GUI.

Through the GUI, a sound containing killer whale whistles is utilized to test the proposed detection model and classification model. The total length of the sound is 264.65s with sampling rate of 44100Hz, and the sound contains 56 whistles of killer whale and some pulse interference.

The checkpoint files obtained in Sections 4.1 and 4.2, as well as the sound, are sent to the GUI respectively, and then the GUI performs the whistle signal detection and classification operation. The whole process takes 43.10s in total.

53 whistles are detected in the detection process, 3 whistles are missed and no signal is falsely detected. Therefore a detection correct rate of 0.947 is achieved. Compared to the real positions, the errors of the output positions (t_s and t_e) calculated by the detection model are within the range of ±350ms. The classification model has correctly classified all 53 detected whistles with a minimum classification probability of 0.97. Fig. 10 shows a number of whistles detected and correctly classified by the detection model. It can be seen that the model can completely detect and extract most of the whistles and the classification model then accurately identifies and classifies the detected whistles with a variety of contours.

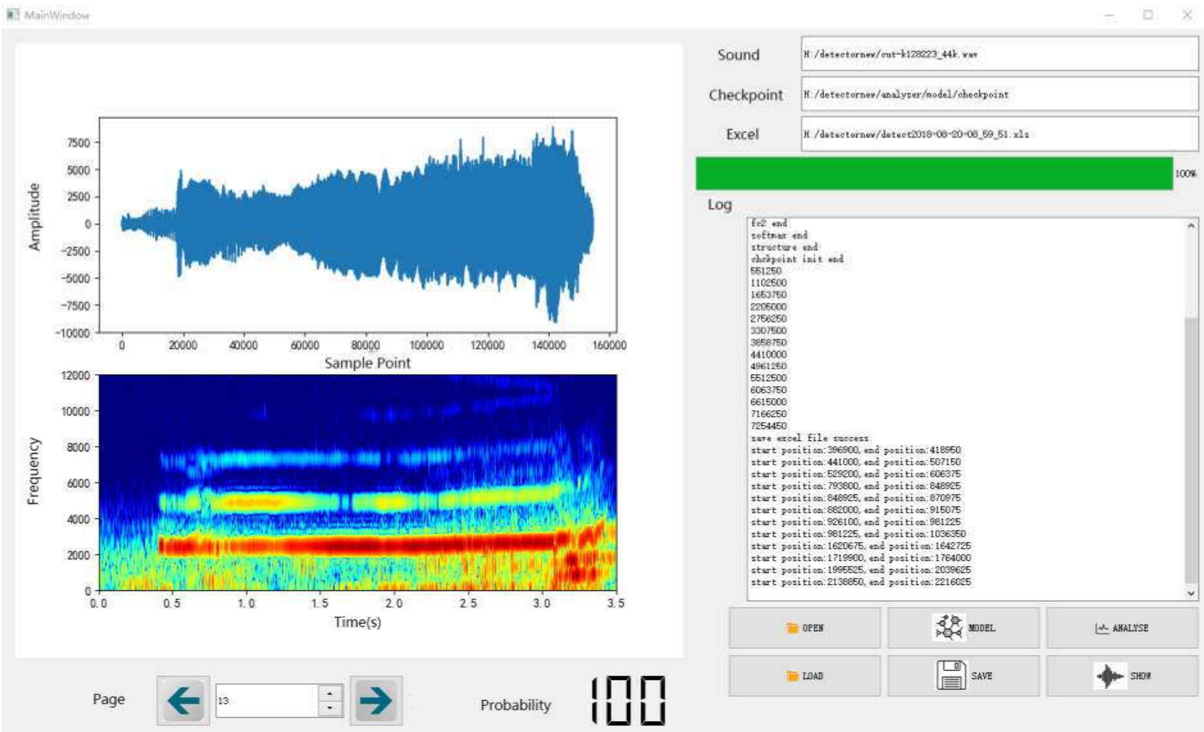

Fig. 9. Display interface of the GUI.

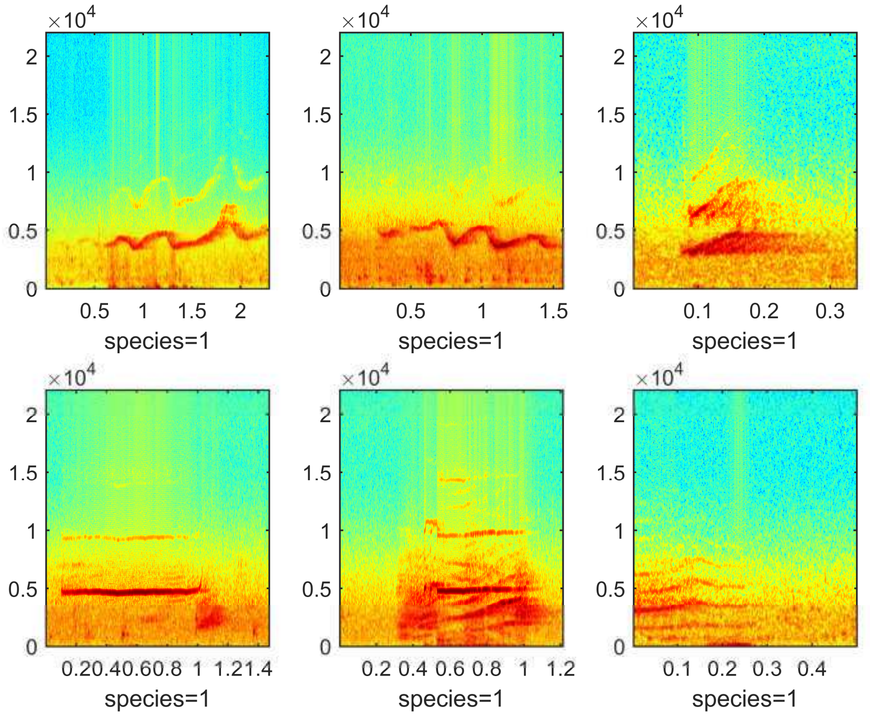

Fig. 10. The detection and classification performance on the testing sound.
Species=1 means the detected whistle is from a killer whale. The *x*-axis represents time in s, while the *y*-axis represents frequency in Hz.

5. Conclusion

In this paper, a CNN-based method has been proposed for accurately detecting and classifying whistles of both killer whales and long-finned pilot whales. The complete process of the proposed method, including denoising, whistle detection, and whistle classification, was presented in detail, together with the corresponding detection model and classification model. The experimental results show that both models can adaptively learn the structural features of the input data and achieve a correction rate of 95% (either detection or classification) on the corresponding testing data set. A GUI interface was developed to assist with the detection and classification processes. Compared with the existing methods presented in Section 1, the proposed method shows a better classification performance for both whale species. Moreover, although the proposed method is used here for whistle detection and classification of only killer whales and long-finned pilot whales, it is not limited to this application and can be easily adapted for other whale or dolphin species that can produce whistles or other sounds; it can also be employed to perform some preliminary work in passive acoustic observation applications for whale or dolphin species, such as range and seasonal occurrence measurement, abundance estimation, and population structure determination, together with some bio-inspired underwater detection or communication systems[32-39].

Acknowledgments

This work was supported in part by the Tianjin Natural Science Foundations of China under Grant No. 17JCQNJC01100, National Natural Science Foundations of China under Grant No. 61501319, 51775377, 61505140, 61501471, National key research and development plan (2017YFF0204800), Young Elite Scientists Sponsorship Program By Cast of China under Grant No. 2016QNRC001, Open Project (MOMST2015-7) of Key Laboratory of Micro Opto-electro Mechanical System Technology, Tianjin University, Ministry of Education, Photoelectric Information and Instrument-Engineering Research Center of Beijing Open Project No.GD2015007. The authors want to thank the Macaulay Library at the Cornell Lab of Ornithology and Watkins Marine Mammal Sound Database for providing the recordings.

References

[1] Fargues, M. P., & Bennett, R. (1995, October). Comparing wavelet transforms and AR modeling as feature extraction tools for underwater signal classification. In *Signals, Systems and Computers, 1995. 1995 Conference Record of the Twenty-Ninth Asilomar Conference on* (Vol. 2, pp. 915-919). IEEE.

[2] André, M., Van Der Schaar, M., Zaugg, S., Houégnigan, L., Sánchez, A. M., & Castell, J. V. (2011). Listening to the deep: live monitoring of ocean noise and cetacean acoustic signals. *Marine pollution bulletin*, *63*(1-4), 18-26.

[3] Salgado Kent, C., Gavrilov, A., Recalde-Salas, A., Burton, C., McCauley, R., & Marley, S. (2012). Passive acoustic monitoring of baleen whales in Geographe Bay, Western Australia. *Proceedings of the Acoustical Society of Australia*.

[4] Mellinger, D. K., Nieukirk, S. L., Matsumoto, H., Heimlich, S. L., Dziak, R. P., Haxel, J., ... & Miller, H. V. (2007). Seasonal occurrence of North Atlantic right whale (Eubalaena glacialis) vocalizations at two sites on the Scotian Shelf. *Marine Mammal Science*, *23*(4), 856-867.

[5] Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., & Tyack, P. L. (2009). Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales. *The Journal of the Acoustical Society of America*, *125*(4), 1982-1994.

[6] Deecke, V. B., Ford, J. K., & Spong, P. (1999). Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (Orcinus orca) dialects. *The Journal of the Acoustical Society of America*, *105*(4), 2499-2507.

[7] Zimmer, W. M. (2011). *Passive acoustic monitoring of cetaceans*. Cambridge University Press.

[8] Oswald, J. N., Rankin, S., Barlow, J., & Lammers, M. O. (2007). A tool for real-time acoustic species identification of delphinid whistles. *The Journal of the Acoustical Society of America*, *122*(1), 587-595.

[9] Rosen, S., & Howell, P. (2011). *Signals and syst*ems *for speech and hearing* (Vol. 29). Brill.

[10] Zaugg, S., Van Der Schaar, M., Houégnigan, L., Gervaise, C., & André, M. (2010). Real-time acoustic classification of sperm whale clicks and shipping impulses from deep-sea observatories. *Applied Acoustics*, *71*(11), 1011-1019.

[11] Gillespie, D. (2004). Detection and classification of right whale calls using an'edge'detector operating on a smoothed spectrogram. *Canadian Acoustics*, *32*(2), 39-47.
[12] Xian, Y., Nolte, L., Tantum, S., Liao, X., & Zhang, Y. (2015). On marine mammal acoustic detection performance bounds. *arXiv preprint arXiv:1510.05520*.
[13] Bahoura, M., & Simard, Y. (2010). Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network. *Digital Signal Processing*, *20*(4), 1256-1263.
[14] Dennis, J., Tran, H. D., & Li, H. (2011). Spectrogram image feature for sound event classification in mismatched conditions. *IEEE signal processing letters*, *18*(2), 130-133.
[15] Nanayakkara, S. C. , Chitre, M. , Ong, S. H. , & Taylor, E. . (2007). Automatic classification of whistles produced by indo-pacific humpback dolphins (Sousa chinensis). *Oceans*. IEEE.
[16] Seekings, P., & Potter, J. (2003, April). Classification of marine acoustic signals using wavelets & neural networks. In *Proceeding of 8th Western Pacific Acoustics conference (Wespac8)*.
[17] Adam, O. (2006). Advantages of the Hilbert Huang transform for marine mammals signals analysis. *The Journal of the Acoustical Society of America*, *120*(5), 2965-2973..
[18] Esfahanian, M., Zhuang, H., & Erdol, N. (2014, May). A new approach for classification of dolphin whistles. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 6038-6042). IEEE.
[19] Thomas, J. A., Kastelein, R. A., & Supin, A. Y. (1999). *Marine Mammal Sensory Syste*ms. Springer US.
[20] Richardson, W. J., Greene Jr, C. R., Malme, C. I., & Thomson, D. H. (2013). *Marine mammals and noise*. Academic press.
[21] Perrin, W. F., Würsig, B., & Thewissen, J. G. M. (Eds.). (2009). *Encyclopedia of marine mammals*. Academic Press.
[22] Vester, H. I. (2017). *Vocal repertoires of two matrilineal social whale species Long-finned Pilot whales (Globicephala melas) & Killer whales (Orcinus orca) in northern Norway* (Doctoral dissertation, Georg-August-Universität Göttingen).
[23] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, *27*(2), 113-120.
[24] Devert, A. (2014). *Matplotlib Plotting Cookbook*. Packt Publishing Ltd.
[25] LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.
[26] Abdel-Hamid, O., Deng, L., & Yu, D. (2013, August). Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech* (Vol. 2013, pp. 1173-5).
[27] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017, March). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392-2396). IEEE.

[28] LeCun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., ... & Simard, P. (1995, October). Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks* (Vol. 60, pp. 53-60).

[29] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). Cambridge: MIT press.

[30] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[31] Jiang, J. J., Bu, L. R., Wang, X. Q., Li, C. Y., Sun, Z. B., & Yan, H., et al. (2018). Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network. *Applied Acoustics*, *141*, 26-34..

[32] Jia-jia, J., Xian-quan, W., Fa-jie, D., Xiao, F., Han, Y., & Bo, H. (2018). Bio-Inspired Steganography for Secure Underwater Acoustic Communications. *IEEE Communications Magazine*, *56*(10), 156-162.

[33] Jiang, J., Wang, X., Duan, F., Li, C., Fu, X., & Huang, T., et al. (2018). Bio-Inspired Covert Active Sonar Strategy. *Sensors*, *18*(8), 2436.

[34] Jiang, J., Sun, Z., Duan, F., Liu, W., Wang, X., & Li, C., et al. (2018). Disguised Bionic Sonar Signal Waveform Design With its Possible Camouflage Application Strategy for Underwater Sensor Platforms. *IEEE Sensors Journal*, *18*(20), 8436-8449.

[35] Jiang, J., Wang, X., Duan, F., Fu, X., Huang, T., & Li, C., et al. (2019). A sonar-embedded disguised communication strategy by combining sonar waveforms and whale call pulses for underwater sensor platforms. *Applied Acoustics*, *145*, 255-266.

[36] Jiang, J., Wang, X., Duan, F., Liu, W., Bu, L., & Li, F., et al. (2019). Study of the relationship between pilot whale (Globicephala melas) behaviour and the ambiguity function of its sounds. *Applied Acoustics*, *146*, 31-37.

[37] Liu, S., Ma, T., Qiao, G., Ma, L., & Yin, Y. (2017). Biologically inspired covert underwater acoustic communication by mimicking dolphin whistles. *Applied Acoustics*, *120*, 120-128.

[38] Liu, S., Qiao, G., Ismail, A., Liu, B., & Zhang, L. (2013). Covert underwater acoustic communication using whale noise masking on DSSS signal. In *OCEANS-Bergen, 2013 MTS/IEEE* (pp. 1-6). IEEE.

[39] Qiao, G., Zhao, Y., Liu, S., & Bilal, M. (2017). Dolphin Sounds-Inspired Covert Underwater Acoustic Communication and Micro-Modem. *Sensors*, *17*(11), 2447.