Underwater Target Detection and Localization with Feature Map and CNN-Based Classification

Tiantian Guo[1,2] Yunze Song[1] Zejian Kong[1] Enggee Lim[1] Miguel López-Benítez[2,3] Fei Ma[4] Limin Yu[1]

[1]Dept. of Communications and Networking,
School of Advanced Technology,
Xi'an Jiaotong-Liverpool University (XJTLU),
Suzhou, China
Email: Tiantian.guo19@student.xjtlu.edu.cn;
Yunze.song19@student.xjtlu.edu.cn;
Zejian.kong19@student.xjtlu.edu.cn
Enggee.lim@xjtlu.edu.cn;
Limin.yu@xjtlu.edu.cn
[3]ARIES Research Centre,
Antonio de Nebrija University,
28040 Madrid, Spain

[2]Dept. of Electrical Engineering and Electronics,
University of Liverpool,
Liverpool, Merseyside, UK
Email: pstguo2@liverpool.ac.uk;
M.Lopez-Benitez@liverpool.ac.uk
[4] Dept. Applied Mathematics,
School of Science,
Xi'an Jiaotong-Liverpool University (XJTLU),
Suzhou, China
Email: Fei.ma@xjtlu.edu.cn

***Abstract*—The purpose of this paper is to apply the acoustic features, Mel Frequency Cepstral Coefficient (MFCC) and Gammatone Frequency Cepstral Coefficient (GFCC), to underwater signal classification. Underwater acoustic signals are vibration signals, and their characteristics are similar to speech signals. The auditory feature extraction method in speech recognition can also be applied to the underwater environment. For underwater communication, we simulate two models designed for underwater target detection and localization. One is the deterministic model, which is considered as basic model; the other is to combine the deterministic model and statistic model, which is called combined model. The geometric channel model facilitates the generation of the database for different geometric settings. The database is generated by adjusting the parameters of the underwater environment. The classifier adopts a convolutional neural network (CNN). The input to the CNN is the feature maps after feature extraction. We choose continuous wavelet transform (CWT) and short-time Fourier transform (STFT) for comparison. Experiments show the effectiveness of the system architecture and superiority of the proposed algorithm in underwater signal classification and target localization.**

***Index Terms*—Underwater communication, CNN, Mel Frequency Cepstrum Coefficient (MFCC), Gammatone Frequency Cepstral Coefficient (GFCC)**

I. Introduction

Acoustic sensing in the marine environment is affected by marine noise, sound velocity characteristics, seabed acoustic characteristics and other distorting factors like multipath fading. The feature extraction of underwater acoustic signals has developed rapidly in recent years, and the most prominent one is the auditory feature extraction method based on speech recognition. However, the performances of traditional methods are not very suitable for various complex environments [1].

Inspired by the cepstral, Steven B. Davis found that the human ear has a different sensitivity to different frequency bands through the study of the characteristics of human auditory perception and first proposed the Mel Frequency Cepstral Coefficient (MFCC) theory [2]. MFCC has good recognition ability and anti-noise characteristics in speech recognition and has become the mainstream in speaker recognition applications for a while. The Gammatone filter was first proposed by Johanesma in [3]. It has achieved a satisfactory effect on physiological data in auditory experiments, so it has been widely used in many fields. The Gammatone filter has a simple time-domain impulse response, from which a simple transfer function can be derived, which is convenient for analyzing the performance of the filter [4].

Underwater acoustic signals are vibration signals, and the characteristics of it are similar to speech signals. The auditory feature extraction method in speech signal processing can also be applied to underwater acoustic signal processing. Brown et al. theoretically demonstrated the feasibility of underwater signal analysis using auditory perception [5]. Lu and his team combined the research on auditory psychology-related theories, first used MFCC for ship target recognition, and achieved good results [6]. Compared with the MFCC algorithm, the GFCC algorithm used the Gammatone filter bank to have a simple time-domain impulse. In response to this characteristic, the signal is directly filtered in the time domain, so the GFCC algorithm avoids the error caused by the spectrum estimation in the MFCC algorithm [7]. Luo and Feng investigate an underwater acoustic target recognition method based on target radiated noise using GFCC and achieve better performance than the traditional method [8].

In this paper, based on the ray theoretical model, two used underwater acoustic channel models are established: deterministic model and combined with statistic model. The

classifier adopts a convolutional neural network which has 3 convolutional layers with feature maps as the input. Feature extraction uses continuous wavelet transform (CWT) , STFT, MFCC and GFCC. The rest of this paper is organized as follows: II introduce the database setting, feature extraction methods and CNN settings. In III, classification results and analysis are illustrated in charts and tables. IV is the conclusion and future work.

II. Methodology

A. Datasets

For underwater communication, we simulate two models for underwater target detection and localization. One is the deterministic model, which is considered as basic model; the other is to combine the deterministic model and statistic model, which is called combined model. The geometric channel model facilitates the generation of the database for different geometric settings. To generate underwater signal database, we consider the following characteristics for each model:

(1) Ocean simulation: we use a ray-tracing model to simulate underwater acoustic channels. From Fig. 1, we consider 4 types of propagation paths from source S and receiver R with different reflection groups: Top-Bottom (TB), Top-Bottom-Top (TBT), Bottom-Top (BT) and Bottom-Top-Bottom (BTB). If the propagation could not form a group, we set the number of reflection groups as zero. Assume the source signal is $s(t) = Acos(\omega * t)$, the received signal at time can be calculated as:

$$r(T) = \sum_{j=1}^{4}\sum_{i=1}^{N}(-1)^i * A * a_j(i) * *cos(\omega * (t - d_j(i)) + p_j(i)) \quad (1)$$

Where T is the first time the receiver received the signal. N denotes the number of the received signal of one reflection group type, a_j is the attenuation coefficients of th reflection type with different numbers of reflection group, d_j is the time delay of th reflection type, and p_j is the phase shift of th reflection type. There are two underwater communication models in our experiment: the deterministic model, which is considered a basic model; the other is to combine the deterministic and statistic models, which is called the combined model.

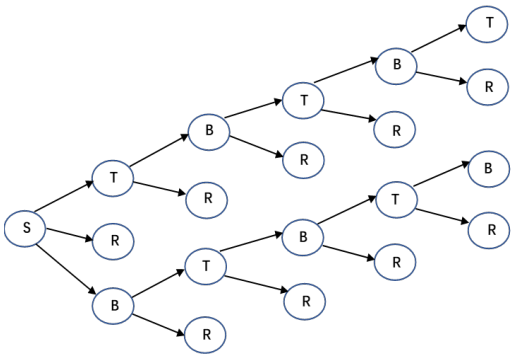


Fig. 1. The refection chain of propagation paths

In an example reflection group in Fig. 2, H denotes the depth of the ocean, Zs denotes the depth of source, Zr denotes the depth of receiver, and L denotes the horizontal distance between sound source and receiver. For simplicity, assume that the sound beam propagates in an underwater area along a straight line with the constant sound speed v. Also, we consider that the sea bottom is a plane and ignore the scattering.

$$Length = \sqrt{L^2 + (Zs + H + (H - Zr)^2)} \quad (2)$$

$$sin\theta = \frac{L}{Length} \quad (3)$$

$$\text{delay time}(d_j) = \frac{Length}{v} \quad (4)$$

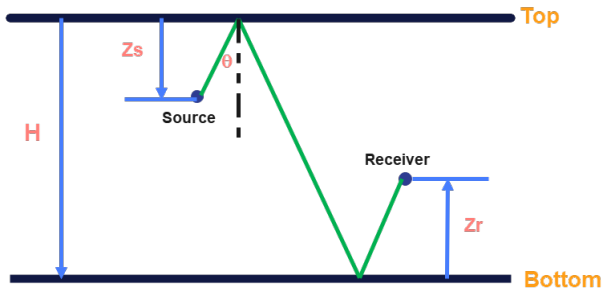


Fig. 2. Example of one reflection group of TB

(2) Noise type: All the noise data are from [9], we choose Cargo Ship, Croaker, Drum Fish, Humpback Whale, Ice Flow, Rain Squall, Snorkeling Submarine, White Whale and AWGN nine noise recordings with different SNR to generate the database. One example of noise and its frequency response is shown in Fig. 3.

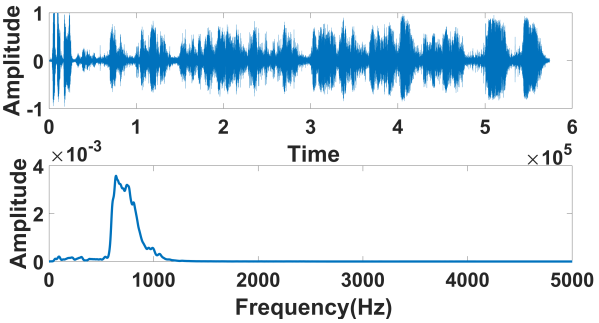


Fig. 3. Time and frequency response of White Whale noise

(3) Classification settings: There are six signal classifications for signal frequency, ocean depth, source depth, receiver depth, the horizontal distance between source and receiver, and ocean bottom type, each with 5 classifications. Through $SNR = -5 : 5 : 40$, 10 SNRs and 9 kinds of noise are generated, and each situation generates 10*9 signals. The database has 450 data, 315 for training and 135 for testing. For signal frequency, which is abbreviated as f, we set the carrier frequency is from 400 Hz to 2000 Hz with a step of 400 Hz.

Ocean depth, abbreviated as H, takes a value every 10 m from 40 m to 80 m. source depth and receiver depth are from 10 m to 50 m with a step of 10 m, abbreviated as Zs and Zr. The horizontal distance, which denotes L, takes a value from 1500 m to 3500 m. There are 5 types of ocean bottom: Clay-silt, Sand-silt-clay, Silt, Sand-silt, and Coarse sand, which denote as $B1, B2, B3, B4$, and $B5$. We consider the [0, 3s] window for further processing for each signal. An example of the signal is shown in Fig. 4. Neural Network structure is the same, but each method is run 10 times to average accuracy for each method. To investigate the relationship between accuracy and SNR, we randomly chose several noise types and generated 1200 data for one SNR; 840 for training and 360 for testing. SNR takes a value every 10dB from -10dB to 40dB, totally 6 values.

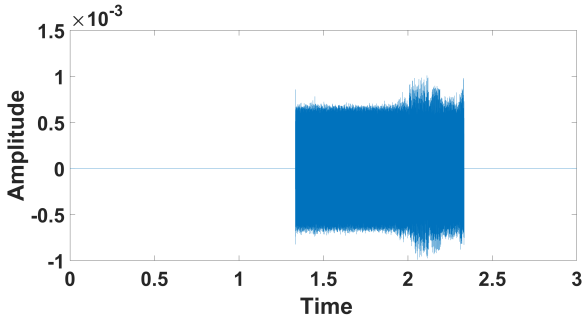


Fig. 4. An example of [0, 3s] observation window

B. Preprocessing Method

We consider several feature extraction methods to process the one-dimension signals to two-dimensions feature maps as the input of Neural Network. Traditional signal processing techniques, continuous wavelet transform (CWT) and short-time Fourier transform (STFT) are included. In addition, two auditory-based feature extraction methods, Mel frequency cepstral coefficient (MFCC) and its derivative methods, and Gammatone frequency cepstral coefficient (GFCC) and its derivative methods, which have been widely used in speech classification and underwater acoustic target recognition, are also investigated in our experiment.

1) MFCC: MFCC feature extraction consists of two key steps: conversion to Mel frequencies, followed by cepstral analysis. The Mel frequency is proposed based on the acoustic characteristics of the human ear, and it has a nonlinear correspondence with the Hz frequency. MFCC uses this relationship between them to calculate the Hz spectral characteristics, and MFCC has been widely used in speech recognition. Equation (5) is the conversion calculation from frequency to Mel frequencies [10]. Fig. 5 is the progress flowchart of MFCC. Fig. 6 is an example of a Mel filter bank.

$$Mel = 2595 * log_{10}(1 + \frac{f}{700}) \tag{5}$$

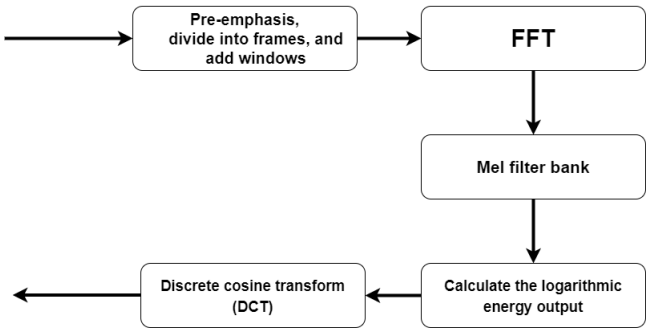


Fig. 5. MFCC flowchart

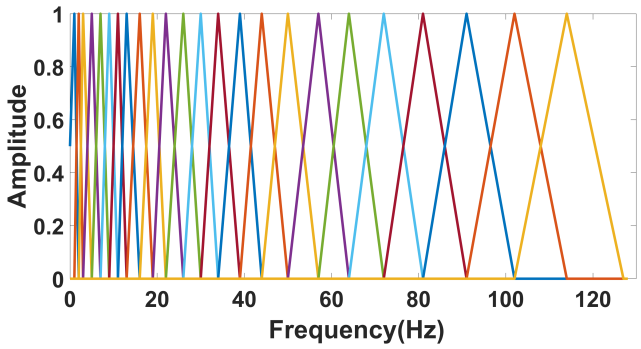
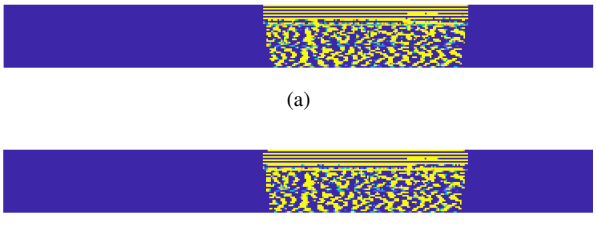


Fig. 6. Mel filter bank

Our experiment set 13 MFCC features and its first derivation and second derivation, a totally 39 features in one frame. So feature map size is $39 * F$, where F is the number of frames. Fig. 7 is examples of MFCC feature map as the input of Neural Network with different ocean depth H. It is impossible to classify which ocean depth classification it belongs to directly from the picture. It can only be classified by inputing it into the CNN.



(a)

(b)

Fig. 7. MFCC feature map: (a) H=80m, SNR=-10dB; (b) H=80, SNR=10dB

2) GFCC: The Gammatone filter was first proposed by Johanesma in [3]. It is a signal processing structure that mimics the human cochlea, and its performance theoretically approaches the resolution of the best sonar, but discards frequency components outside the hearing range [11]. The impulse response in time domain of a Gammatone filter is defined as [12]:

$$g(f,t) = \begin{cases} at^{n-1}e^{-2\pi bt}cos(2\pi ft + \phi), & t \geq 0 \\ 0, & else \end{cases} \quad (6)$$

where f is the centre frequency of the filter, t refers to time, the constant a defines the output gain, n is the order of the filter, b is related to filter bandwidth, and ϕ is the phase that is usually set to zero. The bandwidth of each Gammatone filter is the value of Equivalent Rectangular Bandwidth (ERB) and each band are related to the human ear's critical band. A Gammatone filter bank involves a set of Gammatone filters with different center frequencies fc, and these center frequencies are equally distributed on the ERB scale. The relation between ERB scale and Hz is defined as:

$$E = 21.4log_{10}(1 + \frac{4.37f}{1000}) \quad (7)$$

Fig. 8 is the frequency response of a Gammatone filter bank. Fig. 9 is the flowchart of GFCC progress.

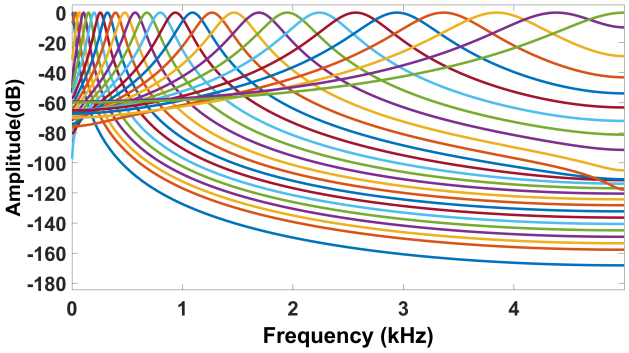


Fig. 8. Gammatone filter bank

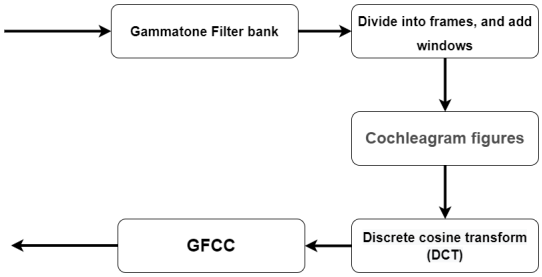
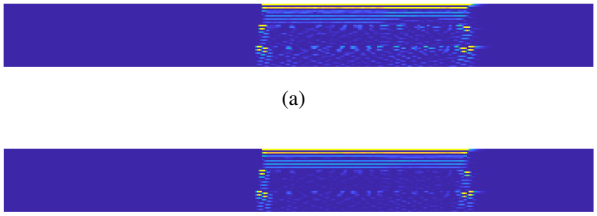


Fig. 9. GFCC flowchart

Our experiment also set 13 GFCC features and their first derivation and second derivation, a totally 39 features in one frame. So feature map size is $39 * F$, where F is the number of frames. Fig. 10 is one example of a GFCC feature map as the input of Neural Network.

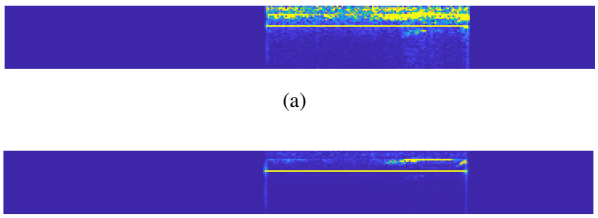


(a)

(b)

Fig. 10. GFCC feature map: (a) H=80m, SNR=-10dB; (b) H=80, SNR=10dB

3) STFT: STFT use the same window size, overlap samples and FFT numbers as MFCC and GFCC to obtain the same feature map size. Fig. 11 is the feature map of STFT with different SNR. Since now it is a narrow band signal, SFTF features have obvious characteristics at the position of carrier frequency set before. Scatters will happen around the distinctive characteristics when adding noises with different SNR, especially when SNR = -10dB in Fig. 11(a).

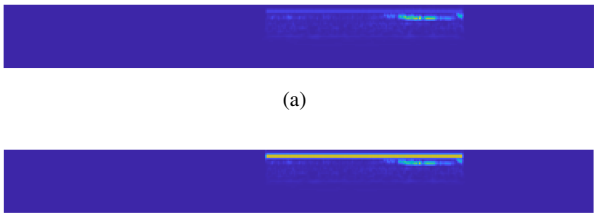


(a)

(b)

Fig. 11. STFT feature map: (a) H=80m, SNR=-10dB; (b) H=80, SNR=10dB

4) CWT: Currently, we use simple CWT with MATLAB function and apply analytic Morse wavelet. To compare with other processing methods, we will resize the CWT feature map to $39 * F$. Fig. 12 is the feature map of CWT. Like STFT features extraction, CWT features also have obvious characteristics at the carrier frequency position especially when SNR = 10dB in Fig. 12(b). In Fig. 12(a), the signal characteristic is hard to find. Because we resize the feature map, the position is a little different from the STFT feature map in Fig. 11.



(a)

(b)

Fig. 12. CWT feature map: (a) H=80m, SNR=-10dB; (b) H=80, SNR=10dB

C. Convolutional Neural Network

In our experiment, we consider a 3-convolution-layer CNN structure. Fig. 13 is the structure of CNN. *Conv* is the convolution layer, and *Fc* is the fully connected layer. The input

is $39 * F * 1$ feature map. The last layer gives a classification score to 5 categories for each underwater signals settings.

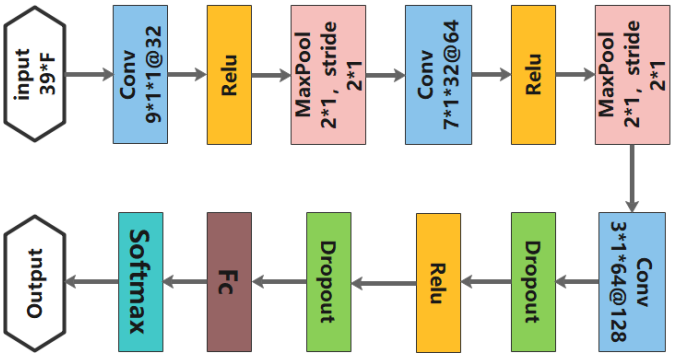


Fig. 13. CNN archtecture

III. Underwater signal classification

A. Basic model

The final classification results of the basic underwater model are shown in Table I. For frequency and horizontal length, each processing method all achieve good results. For the depth of ocean, source and receiver, GFCC achieves the best accuracy, and CWT achieves the lowest accuracy. Its because wavelet transform needs to choose a suitable wavelet and proper decomposition level. Fig. 14 is the bar chart of accuracies. Frequency and horizontal length classifications have the best performance with each feature extraction method. The accuracy of feature maps with GFCC features is highest so that it is more suitable for underwater target classification.

TABLE I
CNN accuracy for underwater basic model

Accuracy(%)	CWT	STFT	MFCC	GFCC
Frequency	97.484	99.038	99.186	97.854
Ocean depth	36.445	79.408	84.517	87.186
Source depth	44.074	74.962	78.815	82.812
Recerver depth	53.48	77.408	68.741	85.037
Horizontal length	100	100	100	100
Bottom type	67.483	69.777	63.481	79.481

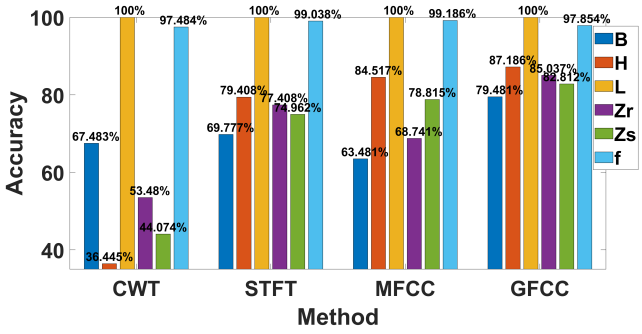


Fig. 14. Bar chart of accuracies of bascic model

TABLE II
CNN accuracy for underwater basic model with different SNR

Acc(%)	SNR	f	H	Zs	Zr	L	B
CWT	**-10**	100	43.667	44.137	39.918	100	41.917
	0	99.748	43.192	47.055	51.084	100	70.332
	10	99.553	54.473	63.362	68.527	100	80.724
	20	99.888	91.611	97.778	97.277	100	80.721
	30	100	98.387	99.331	99.916	100	95.5
	40	100	98.749	100	100	100	97.722
STFT	**-10**	99.055	44.694	46.141	42.583	100	52.333
	0	100	61.304	62	54.722	100	71.89
	10	99.972	94.333	96.501	95.39	100	76.611
	20	100	99.916	100	100	100	75.999
	30	100	100	99.972	100	100	87.75
	40	100	100	100	100	100	91.055
MFCC	**-10**	99.666	58.306	77.25	66.472	100	30.277
	0	99.917	71.332	72.639	71.196	100	45.636
	10	100	90.889	81.111	94.334	100	68.001
	20	100	99.944	94.583	99.972	100	96.499
	30	100	100	96.416	100	100	100
	40	100	100	98.001	100	100	100
GFCC	**-10**	92.862	79.222	74.83	75.722	100	42.971
	0	98.806	88.861	74.20	79.305	100	40.917
	10	99.832	84.502	88.807	86.806	100	49.111
	20	100	96.831	97.722	91.831	100	52.862
	30	100	99.861	96.222	99.806	100	81.006
	40	100	99.167	97.194	98.25	100	91.083

Table II is the accuracies in different SNR. Signal frequency is abbreviated as f. Ocean depth is abbreviated as H. Source depth and receiver depth are abbreviated as Zs and Zr. Horizontal distance denotes as L. Ocean bottom type denotes as B. Obviously, higher SNR has better classification accuracy. Fig. 15 are the line charts of each feature extraction method. GFCC and MFCC achieve better classification performance in lower SNR situations than other methods; GFCC is better than MFCC. When SNR reaches about 30dB, the classification performances are similar. When the signals are significantly disturbed by noise (SNR=-10dB and 0dB), the four feature extraction methods have no significant difference in the basic model for the signal frequency and the horizontal distance between the source and the receiver. However, for ocean depth, source depth, receiver depth and bottom type, MFCC and GFCC can still achieve good classification results, about 80%. Severely corrupted signals are best to use GFCC and MFCC feature extraction methods to classify.

B. Combined model

Tabel III is the accuracy for the underwater combined model. All of them achieve good classification performances. Fig. 17 is the bar chart of accuracies. The classification results are better than the basic model. Every methods' accuracy achieves more than 90%. GFCC achieves the best classification performance as the basic model. However, STFT features have excellent performance when classifying ocean bottom type and ocean depth (H).

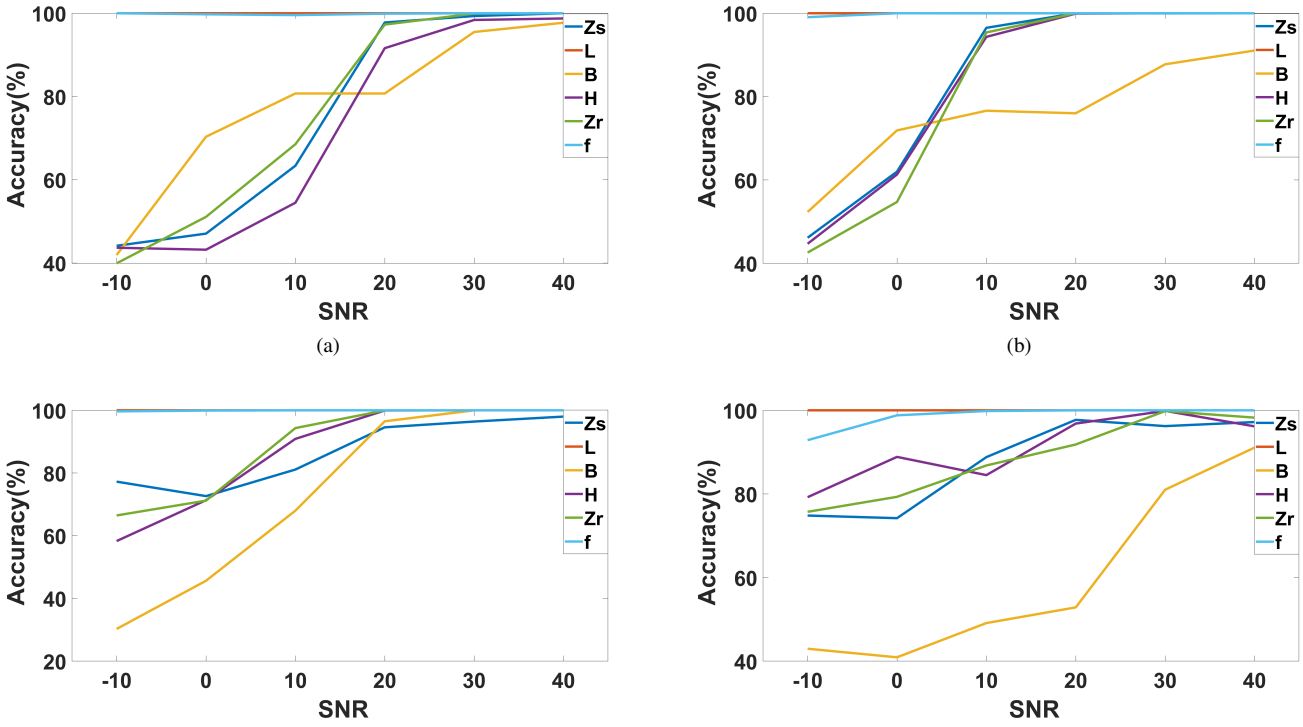


Fig. 15. Line chart of each method with different SNR: (a) CWT; (b) STFT (c) MFCC; (d) GFCC

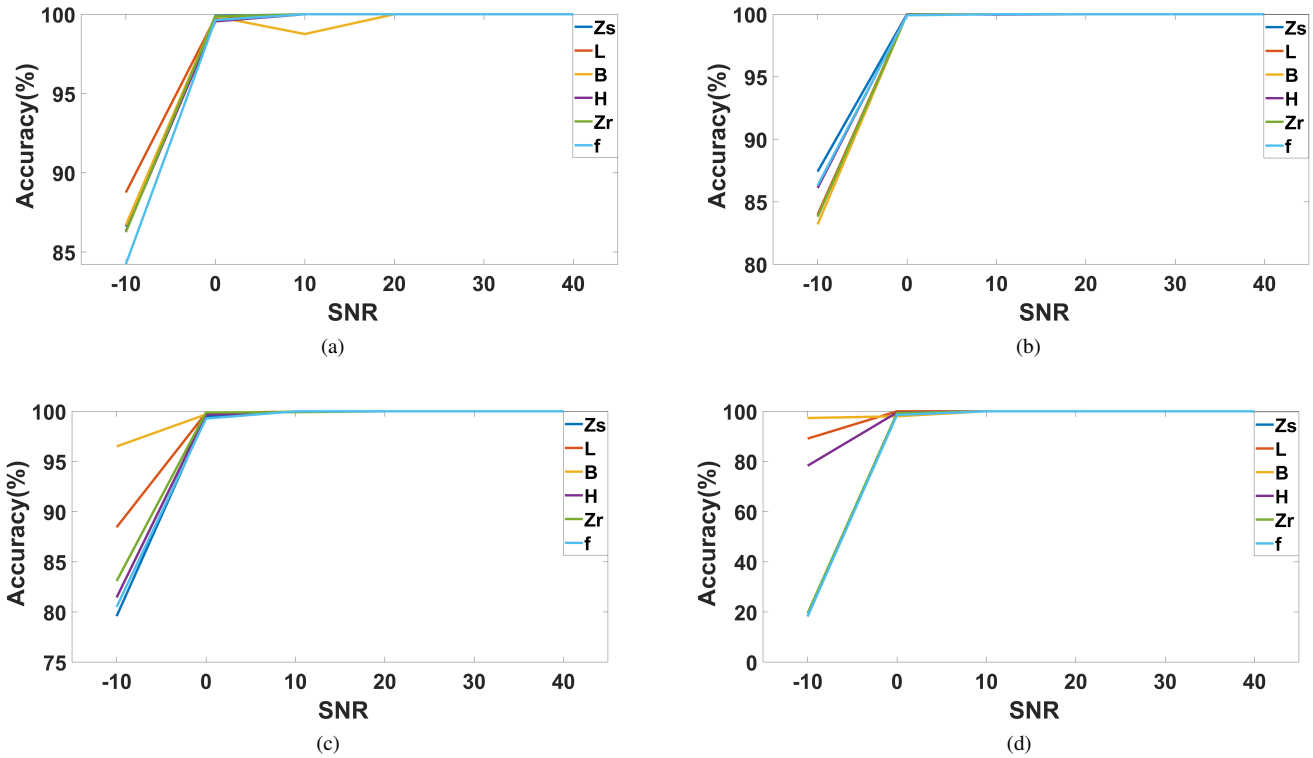


Fig. 16. Line chart of each method with different SNR: (a) CWT; (b) STFT (c) MFCC; (d) GFCC

TABLE III
CNN ACCURACY FOR UNDERWATER COMBINED MODEL

Accuracy(%)	CWT	STFT	MFCC	GFCC
Frequency	96.744	99.408	97.928	98.964
Ocean depth	98.742	99.852	98.742	99.186
Source depth	99.89	100	99.408	98.742
Recerver depth	97.41	98.89	99.408	99.038
Horizontal length	100	98.594	100	100
Bottom type	98.816	99.556	98.964	99.26

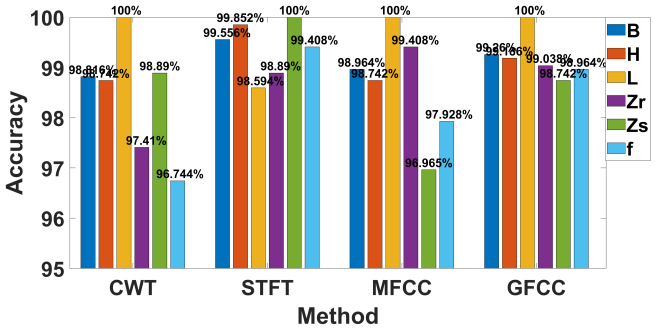


Fig. 17. Bar chart of accuracies of combined model

TABLE IV
CNN ACCURACY FOR UNDERWATER COMBINED MODEL WITH DIFFERENT SNR

Acc(%)	SNR	f	H	Zs	Zr	L	B
CWT	-10	84.222	86.305	87.472	86.25	88.75	87.528
	0	99.637	99.473	99.693	99.86	99.776	99.832
	10	100	100	100	100	100	99.472
	20	100	100	100	100	100	100
	30	100	100	100	100	100	100
	40	100	100	100	100	100	100
STFT	-10	86.306	85.585	87.75	83.805	86.612	85.028
	0	99.916	99.58	100	99.972	99.944	99.72
	10	100	99.972	100	100	100	100
	20	100	100	100	100	100	100
	30	100	100	100	100	100	100
	40	100	100	100	100	100	100
MFCC	-10	80.498	81.444	79.583	83.083	88.446	96.499
	0	99.278	99.666	99.498	99.888	99.832	99.693
	10	100	99.944	100	99.916	99.972	100
	20	100	100	100	100	100	100
	30	100	100	100	100	100	100
	40	100	100	100	100	100	100
GFCC	-10	18.307	78.305	18.22	19.504	89.112	97.332
	0	98.583	99.527	99.22	99.332	100	98.028
	10	100	100	100	100	100	100
	20	100	100	100	100	100	100
	30	100	100	100	100	100	100
	40	100	100	100	100	100	100

Also, Table IV is the accuracies in different SNR. Obviously, higher SNR has better classification accuracy. Fig. 16 is the line chart of each method. If signals are not severely distorted, each feature extraction method could perform well. Even at SNR = -10 or 0dB, GFCC and MFCC have no advantage except bottom type classification. The simulation model combined with the statistic model has better classification performance than the statistic model. MFCC features' performance is more stable in noisy signals than GFCC.

IV. Conclusion and Future Work

In this paper, a deterministic underwater model and the model combined with the statistic model are simulated to generate underwater acoustic signals database. The geometric channel model facilitates the generation of the database for different geometric settings. Then we compared several feature extraction methods to construct the input of CNN. Experiment results verify that GFCC and MFCC feature extraction methods are more suitable for underwater signals classifications, especially severely disturbed signals with lower SNR scenarios. GFCC method could even achieve higher accuracies. When the signals are significantly disturbed by noise (SNR=-10dB and 0dB), the four feature extraction methods have no significant difference in the basic model for the signal frequency and the horizontal distance between the source and the receiver. However, for ocean depth, source depth, receiver depth and bottom type, MFCC and GFCC can still achieve good classification results, about 80%. Severely corrupted signals are best to use GFCC and MFCC feature extraction methods to classify. On the combined model, all four feature extraction methods have good results, and for severly disturbed signals, GFCC and MFCC have no advantage. The simulation model combined with the statistic model has better classification performance than the statistic model. In the future, more suitable feature extraction methods will be explored through different wavelet filters and different decomposition layers. Similar to MFCC and GFCC filter bank, wavelet filter bank could also be applied to extract wavelet features.

Acknowledgment

This research was partially funded by Research Enhancement Fund of XJTLU (REF-19-01-04), National Natural Science Foundation of China (NSFC) (Grant No. 61501380), and by AI University Research Center (AI-URC) and XJTLU Laboratory for Intelligent Computation and Financial Technology through XJTLU Key Programme Special Fund (KSF-P-02), Jiangsu Data Science and Cognitive Computational Engineering Research Centre, and ARIES Research Centre.

References

[1] Z. Lian, K. Xu, J. Wan, G. Li, and Y. Chen, "Underwater acoustic target recognition based on gammatone filterbank and instantaneous frequency," in *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, 2017, pp. 1207–1211.

[2] Davis and B. Steven, "Evaluation of acoustic parameters for monosyllabic word identification," *Journal of the Acoustical Society of America*, vol. 64, no. S1, pp. S180–S181, 1978.

[3] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," 1972.

[4] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 3809–3812.

[5] G. J. Brown, R. W. Mill, and S. Tucker, "Auditory-motivated techniques for detection and classification of passive sonar signals," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3344, 2008.

[6] Z. Lu, X. Zhang, and J. Zhu, "Feature extraction of ship-rediated noise based on mel frequency ceptrum coefficients," *SHIP SCIENCE AND TECHNOLOGY*, vol. 26, no. 2, pp. 51–54, 1 2004.

[7] X. Wang, A. Liu, Y. Zhang, and F. Xue, "Underwater acoustic target recognition: A combination of multi-dimensional fusion features and modified deep neural network," *Remote. Sens.*, vol. 11, p. 1888, 2019.

[8] X. Luo and Y. Feng, "An underwater acoustic target recognition method based on restricted boltzmann machine," *Sensors (Basel, Switzerland)*, vol. 20, 2020.

[9] https://www.hnsa.org/manuals-documents/historic-naval-sound-and-video/sound-in-the-sea/.

[10] A. Chowdhury and A. Ross, "Extracting sub-glottal and supra-glottal features from mfcc using convolutional neural networks for speaker identification in degraded audio signals," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 608–617.

[11] X. Luo and Y. Feng, "An underwater acoustic target recognition method based on restricted boltzmann machine," *Sensors (Basel, Switzerland)*, vol. 20, 2020.

[12] Z. Lian, K. Xu, J. Wan, and G. Li, "Underwater acoustic target classification based on modified gfcc features," in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2017, pp. 258–262.