


RESEARCH

Open Access



Deep reinforcement learning-based adaptive modulation for OFDM underwater acoustic communication system

Xuerong Cui¹, Peihao Yan^{2*} , Juan Li², Shibao Li¹ and Jianhang Liu²

*Correspondence:
s20070011@s.upc.edu.cn

¹ College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China

² College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

Abstract

Due to the time-varying and space-varying characteristics of the underwater acoustic channel, the communication process may be seriously disturbed. Thus, the underwater acoustic communication system is facing the challenges of alleviating interference and improving communication quality and communication efficiency through adaptive modulation. In order to select the optimal modulation mode adaptively and maximize the system throughput ensuring that the bit error rate (BER) meets the transmission requirements, this paper introduces deep reinforcement learning (DRL) into orthogonal frequency division multiplexing acoustic communication system. The adaptive modulation is mapped into a Markov decision process with unknown state transition probability. Thereby, the underwater communication channel environment is regarded as the state of DRL, and the modulation mode is regarded as action. The system returns channel state information (CSI) and signal–noise ratio in every time slot through the feedback link. Because the Deep Q-Network optimizes in the changing state space of each time slot, it is suitable for a variety of different CSI. Finally, simulations in different underwater environments (SWellEx-96) show that the proposed adaptive modulation scheme can obtain lower BER and improve the system throughput effectively.

Keywords: Underwater acoustic communication, Orthogonal frequency division multiplexing, Deep reinforcement learning, Channel estimation and feedback, Channel state information

1 Introduction

Underwater acoustic (UWA) channels are generally recognized as one of the most challenging communication channels [1]. Considering the complexity of underwater acoustic media and the low propagation speed of sound in water, and in order to combat its characteristics of large time delay spread and large-scale fading, researchers usually set up underwater acoustic communication (UWAC) systems based on the channel's most undesirable state before using adaptive transmission technology [2, 3]. By improving the transmitting power of the transmitter, using low-order modulation technology and inserting more redundant error correction coding, we can ensure that the transmission bit error rate (BER) meets the system requirements and the correct information can be

successfully demodulated at the receiver. However, that leads to the low spectral efficiency of the underwater acoustic channel, along with insufficient utilization of channel capacity and low communication efficiency. Adaptive modulation technology (AMT) is always a powerful method for efficient transmission. Therefore, we proposed a deep reinforcement learning-based adaptive modulation for OFDM underwater acoustic communication system to solve the problem. Through real-time estimation and feedback of underwater acoustic channel state, the modulation mode, constellation size, bit rate per symbol, transmit power, and so on are automatically changed.

1.1 Related works

Orthogonal frequency division multiplexing (OFDM) has recently emerged as a more effective solution for underwater acoustic communications because of its robustness to channels that exhibit long delay spreads and frequency selectivity [4, 5]. Radosevic et al. [6] discussed the design of UWA communication adaptive modulation based on OFDM. They proposed two adaptive modulation schemes to maximize the system throughput under the target average BER as the design criterion. The first scheme adjusted only the modulation level and evenly distributed power among subcarriers; the second scheme adaptively adjusted the modulation level and power and then gave the effectiveness of UWA link adaptive modulation results through real-time marine experiments for the first time. Mangione [7] and others designed and implemented a software-defined modem, which can dynamically estimate the acoustic channel conditions, adjust the parameters of the OFDM modulator according to the environment, or switch to a more robust JANUS/FSK modulator under harsh propagation conditions.

In order to overcome the influence of complex and changeable marine environment on underwater acoustic communication signals, artificial intelligence technology has been introduced into the field of underwater acoustic communication applications in recent years. The application of artificial intelligence technology in underwater acoustic communication mainly focuses on the dynamic changes in the marine environment and the physical characteristics of underwater acoustic channels [8, 9]. Mahmutoglu et al. [10] proposed the particle swarm optimization (PSO) algorithm-based adaptive decision feedback equalizer (DFE) for UWAC, in which PSO is independent from channel characteristics and has faster convergence. Although PSO has the highest computational complexity, our simulation results show that the PSO-DFE outperforms other algorithms. Chen Yougan et al. [11] proposed a machine learning-based environment-aware communication channel quality prediction (ML-ECQP) method for underwater acoustic communication networks (UACNs). In ML-ECQP, the logistic regression (LR) algorithm is used to predict the communication channel quality (which is measured according to the bit error rate) between a transmitter and a receiver based under the perceived underwater acoustic channel environmental parameters (such as signal-to-noise ratio, underwater temperature and wind speed). In addition, based on adaptive modulation and coding, Alamgir et al. [12] used support vector machine, k-nearest neighbor algorithm, pseudo-linear discriminant method and the enhanced regression tree method to study the classification of modulation and coding, which further improved the effect of underwater acoustic adaptive modulation and coding. Huang et al. [13] proposed an adaptive approach to pre-set

the modulation scheme for long-range underwater acoustic communication (LR-UWAC). They avoided the direct approach of making the decision based on a simulation over the predicted channel and instead added an abstract layer that classifies the channel or predicts the channel's performance using machine learning tools—support vector machines (SVMs). Upon capturing the important features from the channel, a machine learning-based classifier has better resilience to mismatches in channel prediction.

AI-related algorithms and the Markov decision processes have recently attracted some attention to research in underwater acoustic communication networking. Jin et al. [14] proposed a congestion-avoiding routing protocol for Underwater Acoustic Sensor Networks (UASNs) based on reinforcement learning, which provides an effective way for the node to choose the next forwarder. Su et al. [15] applied cooperative communications to internet of underwater things (IoUT) networks to expand the communication range and alleviate power shortages. They investigated the cooperative communication problem in a power-limited cooperative IoUT system and proposed a reinforcement learning-based underwater relay selection strategy. They formulate the underwater cooperative relaying process as a Markov process and applied reinforcement learning to obtain an effective underwater relay selection strategy. The simulation results have revealed that the DQN-based scheme improved the mutual information and reduced the outage probability. Recently, Q-learning-based AM and coding scheme have been proposed. The performance of an adaptive system depends on the transmitter's knowledge of the channel which is provided via feedback from the receiver. Wang et al. [16] developed an online algorithm based on the reinforcement learning framework for the long-term running regular point-to-point underwater acoustic communication system. They estimated the underwater acoustic channel model parameters recursively and tracked the underwater acoustic channel dynamics and then realize the optimal transmission parameter setting to minimize the long-term cost of the system. The test results obtained from a lake showed that the proposed method can perform better than the benchmark method of ideal non-causal CSI. Song et al. [17] proposed an underwater acoustic adaptive modulation communication strategy based on the reinforcement learning Dyna-q algorithm. The algorithm took the effective signal-to-noise ratio (SNR) as the underwater acoustic channel state parameter, predicted the channel state and communication throughput based on the actual situation and simulation experience of data communication and then used the result to select the modulation parameters combined with the channel state returned by the receiver to maximize the communication throughput. Simulation results showed that the Dyna-q algorithm can achieve higher communication throughput than the direct feedback effective SNR scheme. Su et al. [18] proposed an adaptive modulation and coding scheme for underwater acoustic communication based on reinforcement learning (RLMC). The hot-booting Q-learning algorithm is used to solve the optimization problem under variant quality of services (QoS) requirements. The performance bound of this optimization problem is calculated and analyzed. The scheme dynamically selected the modulation and coding strategy of underwater acoustic communication systems based on network perceived state information such as information service quality requirements, previous transmission quality and energy consumption. Pool and sea trial data showed that it improved the throughput and reduced BER with

less energy consumption compared with the benchmark scheme. However, none of the above methods can deal with continuous channel states.

1.2 Contributions

The inherent Doppler double spread effect delay of the underwater acoustic channel has space–time uncertainty, and hence, there is no unified standard model of the underwater acoustic channel at present. The model uses reinforcement learning to adaptively learn the underwater acoustic channel and realize the adaptive modulation scheme, which is convenient for the parameter setting of the underwater acoustic communication system. It is the key for artificial intelligence technology to break through the bottleneck of underwater acoustic communication. The basic assumptions for our design are: (1) For many applications, the underwater channel is in good enough condition to allow the setup of an OFDM-link, and (2) calibration of the OFDM modulation parameters is possible in scenarios with temporal variability of environmental parameters.

Our approach and contributions are the following:

1. We proposed a metric for channel environment. Firstly, CSI information is stored in a sparse matrix and combined with SNR, including environmental noise, the residual ICI and the noise due to the channel estimation error. The throughput loss due to quantization can be reduced by adjusting states not directly dependent on the complexity of the algorithm. Meanwhile, a DRL-based adaptive modulation scheme combining neural network and RL is used. It can effectively deal with continuous state space problems with fast convergence speed. The reward mechanism includes BER, spectral efficiency, maximum throughput and time consumption. When the transmission does not meet the accuracy requirements, the defined penalty is reset to zero, and non-transmission mode is turned on in addition to the four modulation modes.
2. We investigated the performance of the proposed DRL-based AM scheme under the BELLHOP simulation environment of the SWellEx-96 experiment. A time-varying underwater acoustic channel was established using the temperature, salinity, depth, sound velocity and corresponding time data in the SWELLEX-96 experiment. This channel modeling method is more in line with the real-world scenarios. Compared with the adaptive modulation algorithm people usually use in practice (tabular method with setting fixed threshold) and adaptive modulation method based on improved Q-learning, it demonstrates stable transmission accuracy performance and maximizes the use of channel capacity.

The paper is organized as follows: The second section introduces the experimental model of the OFDM underwater acoustic communication system, the time-varying underwater acoustic channel model and the feedback link; the third section describes the adaptive modulation scheme based on improved DRL, including environmental change setting and system state feedback as well as the reward mechanism and the algorithm process; in fourth section, MATLAB and BELLHOP simulation is used to analyze the anti-environmental interference, BER performance, maximum throughput performance and the defined reward function value of the proposed method in underwater

acoustic channel; and the fifth section summarizes and looks forward to perfecting this paper.

2 System model

2.1 Time-varying underwater acoustic channel model

In most cases, the underwater acoustic channel can be regarded as a slow time-varying coherent multi-path channel [19]. If the observation or processing time is not too long, we can describe the underwater acoustic channel as a time-invariant filter. However, in a continuously operating underwater acoustic communication system, the sound source and receiver's distance and position follow the hydrological changes, and the channel environment varies as well. According to this principle, the assumptions of the multi-path channel simulation model are as follows:

- (a) The sound velocity does not change with the horizontal direction, but only with the depth of the seawater;
- (b) The surface and the bottom of the sea are flat interfaces;
- (c) The position of sound source and receiving point does not change with time;
- (d) The eigen-rays determine the sound field.

The ray from the sound source reaches the receiving point through multiple routes, and the received sound field is the superposition result of all arriving rays (eigen-rays). Figure 1 shows the physical model of three simple propagation paths. We did not consider the bending of the rays caused by different sound velocity profiles, but we simulated the delay. We assume that each response amplitude is not equal, so τ_{21} and τ_{31} represent the delay difference between the second path and the first path and between the third path and the first path, respectively. The received signal is the signal superposition of the three paths.

The transmitted waveform is convoluted with the impulse response, and then, the output is correlated with simulate the multi-path effect, and each arriving sound line is superimposed. Generally, in one transmission τ pulse signal, $h(\tau, t)$ (in which t

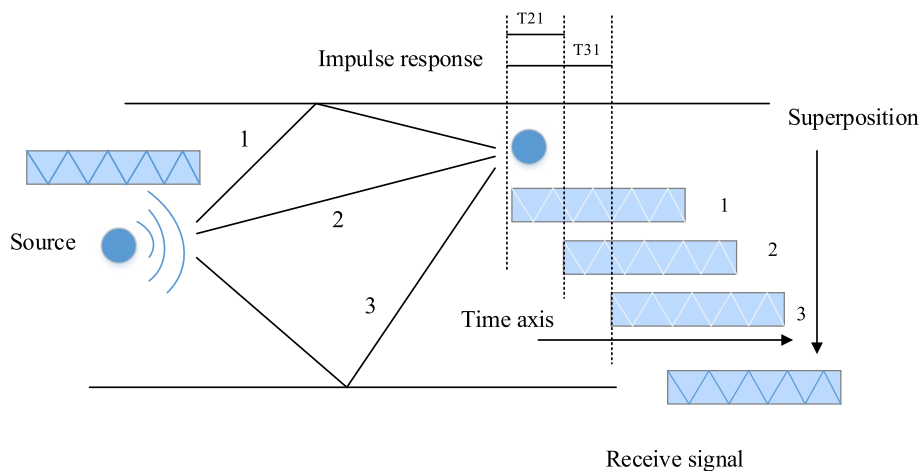


Fig. 1 Physical model of sound line propagation

represents the time) represents the response obtained at a specific time of the signal in the time-varying channel, and the following formula can express h :

$$h(\tau, t) = \sum_{i=1}^{N(t)} A_i(t) \delta(\tau - \tau_i(t)), \quad (1)$$

where i represents the i th arrival sound line, and N is the number of all rays from the transmitting end to the receiving end, or the number of eigen-rays. The underwater acoustic multi-path channel corresponding to this time has n paths. A is the amplitude, δ is the receiving phase, and τ is the arrival delay difference of each path. A_i and τ_i is the propagation attenuation coefficient and relative delay corresponding to different paths.

Another important acoustic property of underwater acoustic channels is marine environmental noise. The marine turbulence, wind noise and thermal noise in the underwater acoustic channel are added through empirical function, which are calculated as follows:

$$AN_{turb}_{dB} = 17 - 30 \log_{10} \frac{f_c}{1000}, \quad (2)$$

$$AN_{wind}_{dB} = 50 + 7.5 \sqrt{s_w} + 20 \log_{10} \frac{f_c}{1000} - 40 \log_{10} \left(\frac{f_c}{1000} + 0.4 \right), \quad (3)$$

$$AN_{thermo}_{dB} = -15 + 20 \log_{10} \frac{f_c}{1000}, \quad (4)$$

where s_w is wind speed for ambient noise level calculation, and f_c is the center frequency of the acoustic band.

2.2 OFDM communication model

We consider a point-to-point underwater acoustic communication system. The transmitter can adaptively adjust the modulation mode. There is a feedback channel between the transmitter and the receiver. The receiver feeds back the CSI for each fixed time slot through the feedback channel.

Figure 2 only shows the process related to the adaptive modulation scheme in the OFDM Underwater acoustic communication system. It is assumed that the transmitted CSI will not be affected by the instability of underwater acoustic channel and system hardware equipment, which is to say that a noiseless transmission is assumed. Our goal is for the actual BER to be less than 10^{-2} , and the system BER after error correction coding to be than 10^{-5} . In order to simplify the environment model, this paper does not consider the error correction coding scheme but only the modulation scheme. The optimal modulation level is determined by CSI [20] and BER in our system to find an optimal transmission strategy for the transmitting transducer. Therefore, to realize demodulation, the sender notifies the receiver of the modulation level before transmitting data in each time slot, for the purpose of maximizing the system throughput under the specified BER requirements in a limited time range.

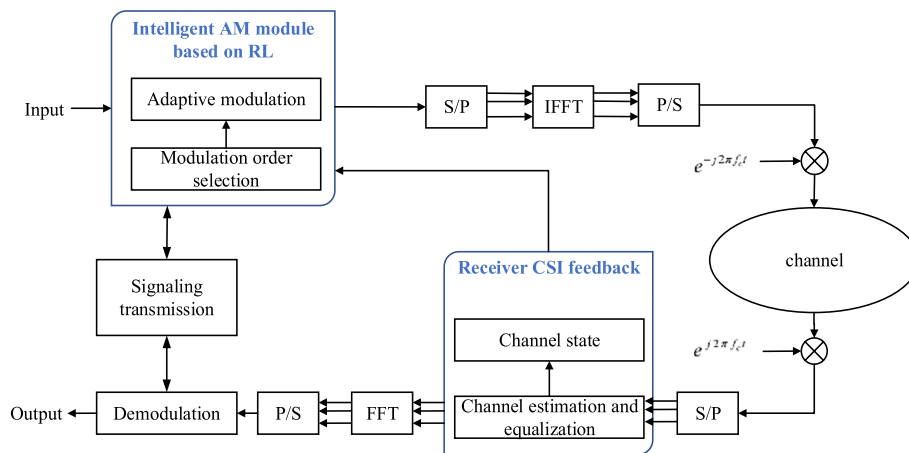


Fig. 2 OFDM system model

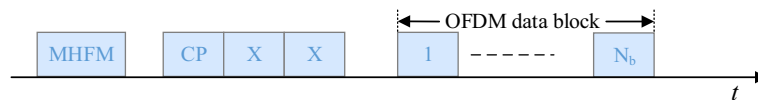


Fig. 3 The frame structure of OFDM system model

1. Frame structure

We adopt the multiple hyperbolic frequency-modulated (MHFM) signals [21] to jointly estimate the arrival time and use cyclic prefixes (CP)-OFDM leader codes with self-repetition for Doppler extension estimation. The frame structure of OFDM system model is shown in Fig. 3.

2. Channel frequency-domain estimation

The input/output relationship between discrete symbol sampling and emitted symbols can be expressed as follows [22]:

$$z = Hs + w \tag{5}$$

where the noise term W contains the ambient noise, the residual ICI and the noise caused by the channel estimation error. Based on the above input/output relationship, with the help of frequency measurements on pilot subcarriers, the path parameters of the channel can be estimated by the minimal military method or the compressed sensing method based on sparse channel estimation [23]. In order to reduce the complexity of the algorithm, we adopted the least square (LS) method which is widely used in practice.

$$\hat{H}_{LS} = \underset{\xi}{\operatorname{argmin}} \|z - Hs\|^2, \tag{6}$$

3. SNR estimation

In the environmental assessment of the underwater acoustic channel, the SNR is an essential parameter that can effectively reflect the magnitude of noise and the environmental quality of the ring channel [24]. According to the received signal of each pilot subcarrier in each OFDM symbol in the system, it is expressed as:

$$y(i, j) = \sqrt{S} \bullet h(i, j) \bullet a(i, j) + \sqrt{N} \bullet n(i, j), \quad (7)$$

where S is the signal power factor, N is the noise power factor, h is the channel coefficient, $a(i, j)$ is the i th pilot subcarrier, the modulation signal at the j th OFDM symbol, and n is the AWGN signal with zero mean added. Then, the SNR is estimated as:

$$\widehat{SNR}_{ML} = \frac{\widehat{S}_{ML}}{\widehat{N}_{ML}} = \frac{\left[\frac{1}{J} \sum_j^{J-1} \text{Re} \left\{ y(i, j) \bullet \widehat{h}^*(i, j) \bullet a^*(i, j) \right\} \right]^2}{\frac{1}{J} \sum_j^{J-1} |y(i, j)|^2 - \left[\frac{1}{J} \sum_j^{J-1} \text{Re} \left\{ y(i, j) \bullet \widehat{h}^*(i, j) \bullet a^*(i, j) \right\} \right]^2}, \quad (8)$$

where J is the number of OFDM symbols used for SNR estimation, $\text{Re}\{\bullet\}$ represents the real part of the complex number, $*$ represents the conjugate of the complex number, and \widehat{h} is the channel time-domain impulse response estimated in the previous part.

3 Our adaptive modulation scheme

3.1 Multi modulation system

We map the adaptive modulation scheme to a finite Markov decision process. Based on this discrete and finite-state theoretical framework, agents and the environment achieve their goals through interactive learning. In finite MDP, function p defines the dynamic characteristics of MDP and specifies a probability distribution for the selection of each state and action.

In order to improve the bandwidth efficiency of the system, an underwater acoustic communication system usually adopts a multi-band modulation scheme. A set of signal constellation points can represent the modulation level of each data symbol M_t . We select modulation scheme {BPSK, QPSK, 8-QAM and 16-QAM.}, where $M_1 = \{2, 4\}$ in circular constellation multiphase shift keying (MPSK) system, $M_2 = \{8, 16\}$ in the square constellation multi-level quadrature amplitude modulation (MQAM) system. It is assumed that the transmitter uses constant symbol period T_s and ideal value is combined to obtain $M = \{2, 4, 8, 16\}$, and the length of each time slot $T_s = \frac{1}{B}$, where B is the bandwidth of the received signal.

(1) State space: since the receiver obtains the CSI information of the feedback link, including channel gain, multi-path, noise and other information, we define the state of each time slot as $S_t = \{s_1, s_2, s_3, \dots\}$.

(2) Action space: in the OFDM underwater acoustic communication system model, since the transmitter automatically adjusts and selects the modulation scheme according to only the current feedback state in each time slot, the action of each time slot is defined as $A_t = \{a_1, a_2, a_3, \dots\}$, where the modulation mode a_i adopted for the i th time slot is selected from the given constellation M system.

(3) Immediate reward function: each time slot obtains the timely reward function $R_t = \{r_1, r_2, r_3, \dots\}$ based upon the feedback, and it includes reward and punishment. The reward function is directly proportional to the number of data bits successfully transmitted, and related to the system throughput. The punishment function is only related to the BER.

A complete MDP consists of four tuples. Given the values of the initial states and actions, the probability of the occurrence of $s' \in S$ and $r \in R$ at time t can be obtained $p(s', r|s, a) = \Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}$. However, in our adaptive modulation scheme in an underwater channel environment, the selected probability distribution of each s and a cannot be obtained, so it is defined as MDP triple $\langle S, A, R \rangle$.

3.2 Value calculation

We are considering designing an adaptive transmission system based on SNR γ_t . If the average data rate is maximized only under the fixed target BER, then $k(\gamma_t)$ can be set to $\text{equallog}_2 M_t$, which can meet the general adaptive M-nary modulation. The accurate BER is obtained through the actual transmission of the system. We send specific data and feedback on the bit error in each time slot and compare the proportion of the number of bits incorrectly accepted by the receiver in the total transmission bits.

Considering various expenditure loads in the communication system, we calculate the total data rate so that the coding rate r is constant. Calculated according to the number of bits per second per *Hertz*, the spectral efficiency is:

$$\psi = (r \times \sigma) \times \frac{T}{T_{b1}} \times \frac{K_d}{K}, \tag{9}$$

where σ is the bit rate, $\sigma = \log_2 M$, and M is the modulation order. T_{b1} is the OFDM symbol length, and $T_{b1} = T + T_{cp}$, T_{cp} is the length of the cyclic prefix, T is the basic OFDM symbol interval, K is the number of subcarriers, and the size of the symbol after FFT transformation and K_d is the number of data subcarriers.

Different from Shannon capacity, in the multi-level modulation system, we use the number of bits sent per unit time as the system throughput:

$$T = \psi \times \max [P_{cf} \times (r \times \sigma)], \tag{10}$$

where P_{cf} is related to the system BER ρ , and $P_{cf} = 1 - \rho$.

The penalty Θ sets the reward to zero. If the BER requirements are not met, all reward values are set to 0. The calculation of the value function is defined as:

$$\begin{cases} r(s, a) = -c_1 \cdot \rho + c_2 \cdot \psi + c_3 \cdot T, \Theta = 0, \rho \leq P_b \\ r(s, a) = 0, \Theta = 1, \rho > P_b \end{cases}, \tag{11}$$

where c_1, c_2, c_3 are constants, representing the weight of each parameter in value calculation. No signal will be transmitted if the penalty bit Θ of each action is set to zero in the t th time slot. It happens when the current environment state is not ideal. According to our automatic modulation strategy, even the modulation mode with the lowest order cannot meet the system requirements, so the optimal strategy is not to transmit.

3.3 Optimization problems

According to the feedback information of the feedback line, we set the optimization problem as:

$$\begin{aligned} & \max_{M_t} r_t(s_t, a_t) \\ \text{s.t.} & \begin{cases} M_t \in M = \{1, 2, 4, 8, 16\}, \forall t = 1, 2, \dots, N \\ \rho_t \leq P_{b,t}, \forall t = 1, 2, \dots, N \end{cases} \end{aligned} \quad (12)$$

where $M_t = 1$ is the action of not transmitting, and the transmitter remains in a static waiting state.

3.4 The proposed adaptive modulation scheme based on DRL

In the traditional Q-learning algorithm, the Q value is stored in a table. The horizontal axis of the two-dimensional table is the state, the vertical axis is the action, and the median value is the Q value of the action corresponding to each state. For the low-dimensional state space, the Q table can be stored in all states, and the optimal action can be selected by directly querying the Q table. However, there is a large-scale and continuously changing state space for the time-varying underwater acoustic channel [25]. Deep Q-Network (DQN) is model-free, aiming to find the mapping relationship between the action state and the Q value. The temporal difference (TD) method combines the Monte Carlo sampling method and the bootstrapping of the dynamic programming method (using the value function of the subsequent state to estimate the current value function) so that it can be applied to the model-free algorithm and is updated in one step with faster speed. For the Q-learning method in which action is a discrete variable, $Q^*(s, a)$ is approximated by a deep neural network. We still consider transforming the continuously changing state into the discrete state of each time slot. Because the neural network can automatically extract complex features, we do not quantify the CSI but keep the feature vector as the input.

3.4.1 Model input

Usually, people take the combination of the channel frequency-domain response estimated by the receiver and the SNR of the equalized subcarrier as the input vector of the network. However, due to the sparsity of the underwater acoustic channel, we consider transforming the channel frequency-domain response estimated by the receiver into the time-domain impulse response. Through the storage method of a sparse matrix, we can effectively reduce the amount of data and denote the network input signal as x :

$$x = [h_{\text{sparse}} \text{ SNR}], \quad (13)$$

In order to reduce the dimension of the input data, n peaks of the time-domain impulse response are extracted in advance to keep the data input size consistent.

$$h_{\text{sparse}} = \begin{bmatrix} A_1 & A_2 & \dots & A_n \\ \tau_1 & \tau_2 & \dots & \tau_n \end{bmatrix}. \quad (14)$$

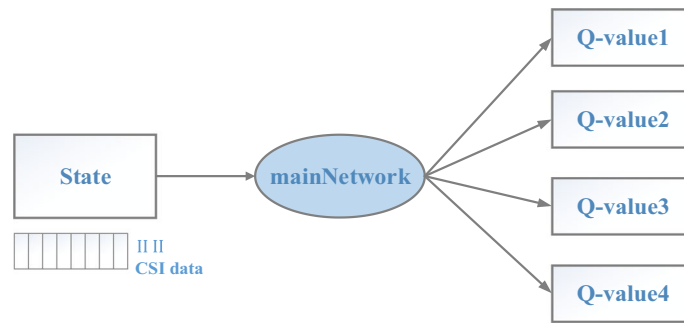


Fig. 4 Through main network output, the Q values of all actions corresponding to a CSI state

As in Eq. (1), A_i and τ_i is the amplitude and relative delay corresponding to different paths. The network input and output relationship is shown in Fig. 4.

3.4.2 Adaptive modulation algorithm based on deep reinforcement learning

The modification of Q-learning by DRL is mainly reflected in three aspects:

- (1) DRL uses a depth neural network to approximate the value function;
- (2) DRL uses experience replay to train the learning process of reinforcement learning;
- (3) DRL independently sets up the target network to deal with the TD deviation in the time difference algorithm.

In DRL, the enhanced learning Q-learning algorithm and the SGD deep learning training are carried out synchronously. We used the powerful fitting ability of the neural network to approximate the action-value function in Q learning, so $Q(s, a) \approx Q(s, a; \theta)$. The updated method is calculated as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right], \tag{15}$$

where $\text{Target}Q = r + \gamma \max_{a'} Q(s', a')$, the loss function in the algorithm uses the mean square loss to update the parameters in the iteration: $L(\theta) = E \left[(\text{Target}Q - Q(s, a))^2 \right]$, with θ parameters representing the network, α as the learning rate, s' and a' as the state and action in the next iteration, respectively, and γ being the discount factor in the TD method.

In the learning phase of each time slot, we consider that passing ε – greedy strategy traverses all possible operations in each channel state. The greedy algorithm generates an optimal global solution through optimal local strategy, which is usually set ε as a constant in a system. A number between 0 and 1 is randomly generated ξ :

$$\begin{cases} a_t = \arg \max_{a_t} Q(s_t, a_t), \xi < \varepsilon \\ a_t = \text{rand}(A_t), \xi \geq \varepsilon \end{cases}, \tag{16}$$

The actions that maximize the Q value have a higher probability of occurrence, and the Q values corresponding to all possible actions can be learned. Therefore, we set the training state:

$$\varepsilon = \max \left(0.01, 0.2 - 0.1 \times \left(\frac{N_{\text{episode}}}{N_0} \right) \right), \tag{17}$$

where N_{episode} is the number of episodes that the system continues to cycle and N_0 is a constant about episodes. ε will gradually decrease with the continuous training of the current event. If the environment changes rapidly, the value of ε needs to be increased to make the system more likely to train Q values corresponding to other actions in this state, which can maximize the reward of the selected action.

Each episode is carried out in each time slot for a period of time. In a certain period of time, the channel environment changes slowly. The agent judges whether the performance meets the requirements and calculates the reward value and punishment by observing the average maximum throughput and the average BER corresponding of each modulation order. The penalty value Θ represents the negative value of reward. In the same network, the weight after learning a task may completely change when learning a new task because the optimization target values are different in the time-varying underwater acoustic channel environment with significant differences. The objective function is the same, but the data sets are different. The old weights are easily damaged, so the batch random sampling method is adopted. We store 10,000 groups of data. Each set of data contains the current environmental state information, modulation mode and value score obtained. The agent samples 100 groups of data from the experience replay, learns all samples of the whole batch, calculates the average gradient, and then updates them. During the update, only the Q value corresponding to the current modulation is updated, and others remain unchanged.

We consider the sample buffer with a sliding window mechanism. The sequence stored in the experience replay is $[S_t, A_t, R_t, S_{t+1}]$, the cache length is $L = 1000$, and the initialization is empty. The learned state transition sequence is added every time. When it is completed, we delete the old sample at the top layer and then store the new sample.

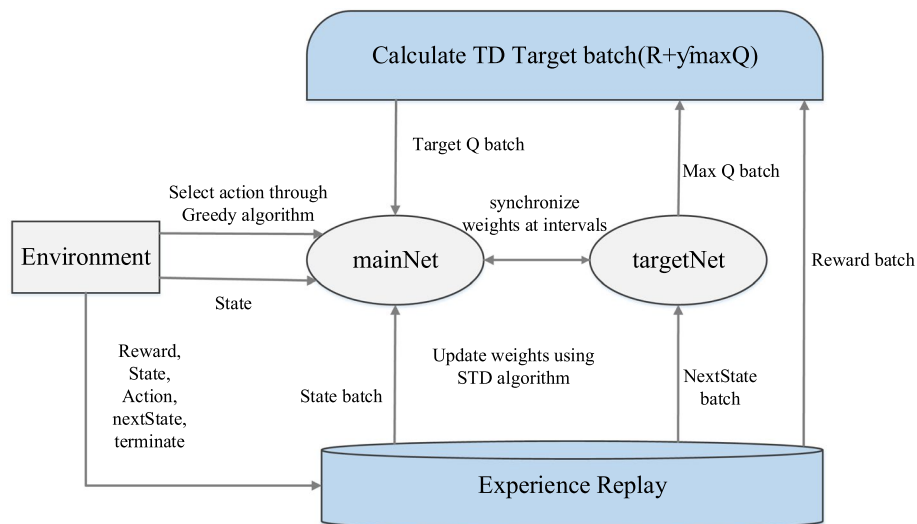


Fig. 5 Algorithm flowchart of DQN

There are two networks with precisely the same structure but different parameters in DQN. As shown in Fig. 5, the prediction of Q estimation network mainNet is the virtual training network. Each step updates parameter θ according to the samples collected by mini-Bach, while the prediction of Q reality network targetNet parameters was used some time ago. The weight of mainNet is copied to targetNet every certain number of iterations C . $Q(s, a; \theta_i)$ represents the output of the current network mainNet, which is used to evaluate the value function of the current state action pair; $Q(s, a; \theta_i^-)$ indicates the output of targetNet. It mainly provides maxQ, which can solve the target. Therefore, when the agent acts on the environment, it can calculate Q according to the formula and update the parameters of mainNet according to the loss function. In order to prevent overestimation and make the Q value closer to its real value, our optimal action selection is based on the parameters of the Q network currently being updated. It completes an episode workout.

Specifically, the pseudo-code of using the DRL algorithm to find the best transmission strategy is shown in Algorithm 1.

Algorithm 1: DRL-Based AM Algorithm With TD Strategy

Input : CSI and SNR parameters

Output: optimal action of each time slot at a_t^*

Initialize replay memory D to capacity N

Initialize state-value buffer B to capacity P

Initialize state-value Q -function with random weights θ

Initialize target state-value Q -function with weights $\theta^- = \theta$

Initialize sequence S , A , and R

For episode = 1, M **do**

 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence

For step = 1, T **do**

 With probability, ϵ select a random action a_t

 Otherwise, choose optimal action at $a_t = \max_a Q^*(s_t, a; \theta)$ each time slot.

 Evaluate BER and θ after passing through the system of each rate region

 Evaluate spectral efficiency and maximum throughput

 Obtain reward r_t

 Send the information to the receiver through the feedback link in every time slot

 Store transition $\{s_t, a_t, r_t, s_{t+1}\}$ in P

 Sample random mini-batch of transition from P

 Set $y_t = \begin{cases} r_j, & \text{if episode terminate at step } t + 1 \\ r_j + d\hat{Q}(s', \arg\max_a \hat{Q}(s', a, \theta); \theta^-), & \text{otherwise} \end{cases}$

 Perform a gradient descent step on $(y_t - Q(s_t, a_t; \theta))^2$ concerning θ

 Every C steps reset $\theta^- = \theta$

End for

End for

4 Simulation results and discussion

In this section, we present numerical and experiment results to evaluate the performance of our proposed DRL-based adaptive transmission modulation scheme. We compare its performance with the most commonly used fixed threshold method and RLMS [18] based on Q-learning in many different hydrological environment changes, in terms of BER, system throughput and a value function defined by this paper. Fast-changing channel state estimation and feedback are considered. The improvement of the primary look-up table method is based on quantization, and the channel environment has different SNR restrictions. The adaptive modulation schemes are BPSK, QPSK, 8-QAM and 16-QAM.

As shown in Fig. 6, the transmission system is divided into fixed time slots. In each time slot, dynamic changes will occur between the transmitter and receiver, including the change of hydrological environment and the random movement of equipment. The transmitter sends data information for channel estimation (including pilots and data packets). In each time slot, pilots are used for estimation and feedback, and the underlying data packets are transmitted continuously.

4.1 Simulation environment settings

In this paper, the channel simulation uses BELLHOP software to simulate the hydrological environment. The simulation uses the sound velocity profile of the SWellEx-96 experiment about 12 km away from the tip of Loma angle near San Diego, California, considering the acoustic characteristics of the sea surface, sound attenuation and absorption, and seafloor reflection loss. The sound velocity profile is shown in Fig. 7.

Firstly, we select the underwater acoustic channel scenario. Figure 8 shows the different channel responses obtained during the gradual change of the distance between the transmitter and the receiver in one channel environment. S_d is the depth of transmitter, and R_d is the depth of receiver. The water depth is 200 m, the hydrophone is 50 m away from the sea surface, the transmitting transducer is 24 m away from the sea surface, and the distances between the five transmitting and receiving are 5.01 km, 5.02 km, 5.03 km, 5.04 km and 5.05 km, respectively.

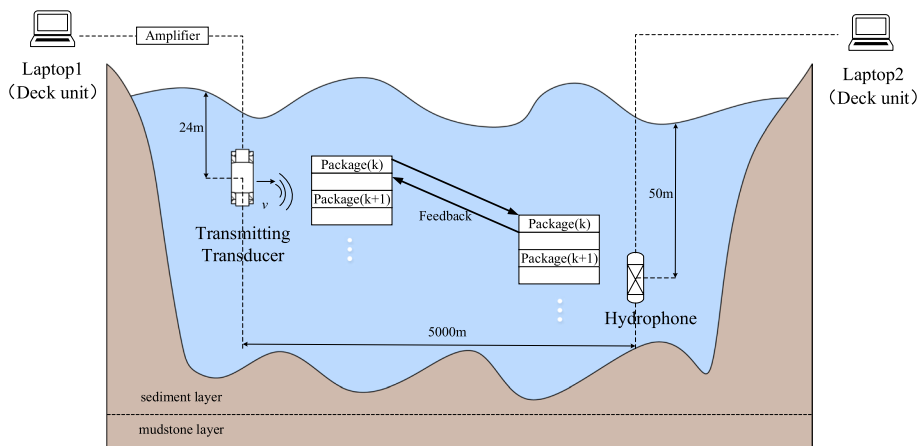


Fig. 6 Transmission process of each simulation

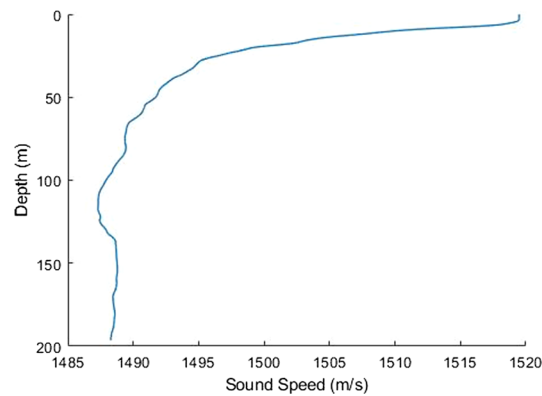


Fig. 7 Simulation of sound velocity profile

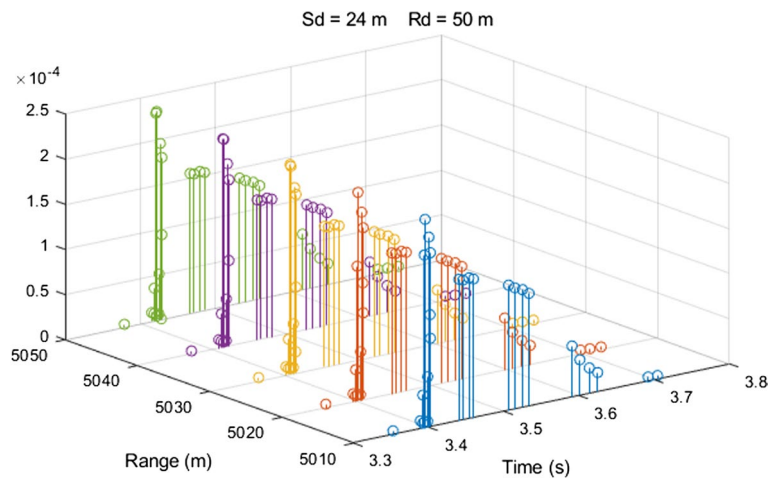


Fig. 8 The impulse response of hydrophone at different receiving points (time-varying)

The sound wave emitted by the sound source is disturbed by the background environment in the propagation process of the underwater acoustic channel, including marine dynamic noise, biological noise, traffic noise, industrial noise, seismic noise and underwater noise. In addition, the location of sound source and receiver also leads to the complexity and variability of marine environmental noise.

Since each time slot of the system feeds back a set of CSI information, we assume that the channel environment is unchanged during this time slot. We consider processing the channel impulse response information of each group of time slots. After estimation, we take 16 multi-path information with higher amplitude and store them in the sparse matrix. Based on this channel information, the estimated time-domain channel h and the estimated SNR are calculated to provide the current state for the adaptive modulation scheme at the transmitter.

4.2 BER analysis of fixed threshold algorithms

Table 1 shows the simulation parameters of OFDM system.

Under the given average channel environments, Fig. 9 gives four different modulation modes transmission BER under the different SNRs. The modulation schemes

Table 1 Simulation parameters of the OFDM packet

Parameter	Value
Bandwidth: B/Hz	8000
Center frequency /Hz	12,000
Number of subcarriers	1024
Pilot form	Block
pilot length	256
Signal length	256 × 256 × 8
Cyclic prefix (CP) length/protection interval	400

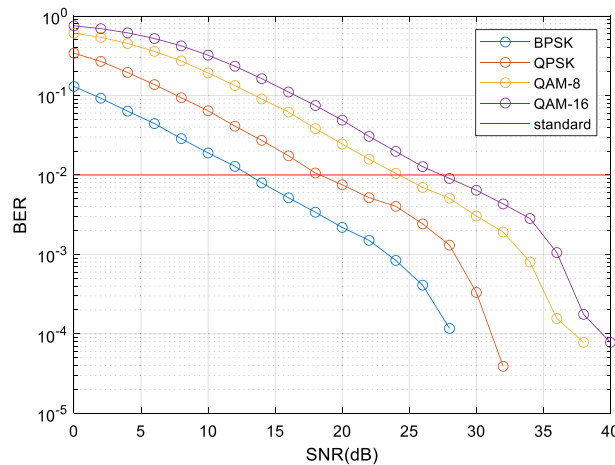


Fig. 9 Simulation of different modulation modes of OFDM system in multi-path fading

Table 2 Modulation scheme in UNDERWATER acoustic channel

Transmission	SNR(dB)
Model 1 (BPSK)	13 dB < SNR ≤ 18 dB
Model 2 (QPSK)	18 dB < SNR ≤ 24 dB
Model 3 (QAM-8)	24 dB < SNR ≤ 27.5 dB
Model 4 (QAM-16)	SNR > 27.5 dB

are BPSK, QPSK, 8-QAM and 16-QAM. We define that the system requires the BER to be less than 0.01. Hence, the target SNR P_b is fixed. According to the transmission criteria we defined, the change of each of the modulation scheme is shown in Table 2. When the SNR is quite low ($SNR \leq 13$ dB), the lowest-order modulation mode BPSK cannot meet the BER of 0.01; therefore, it chooses not to transmit. When the SNR is not so low ($13 \text{ dB} < SNR \leq 18 \text{ dB}$), the system selects QPSK as the modulation scheme, and the bit rate is twice that of the BPSK system. When the SNR is not so high ($18 \text{ dB} < SNR \leq 27.5 \text{ dB}$), the system selects 8-QAM as the modulation scheme, and the bit rate is twice higher than that of the BPSK system. When the SNR is high ($SNR > 27.5 \text{ dB}$), the system selects 16-QAM as the modulation scheme, the data rate is four times higher than BPSK, and the BER performance remains < 0.01 .

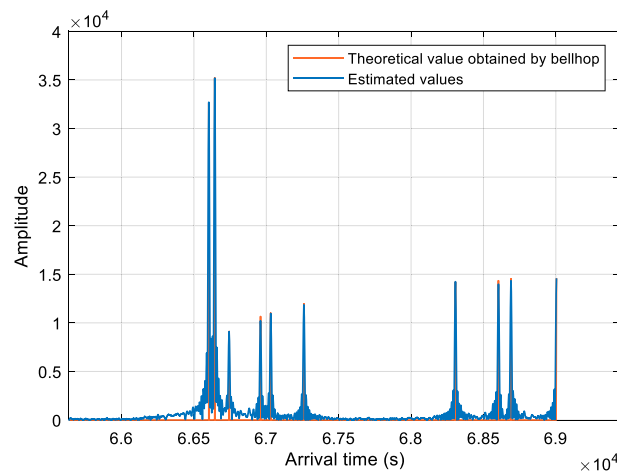


Fig. 10 Impulse response: theoretical and estimated values

Table 3 System network parameters

Parameter	Value
SNR	0–35 dB
Loss function	MSE
Epoch	$30 \times 92 \times 10$
Train parameter epochs	100
Initial learning rate	0.01
Network training function	traingdx
Optimizer	Adam

4.3 Channel estimation and feedback

HFM signal can still obtain good energy accumulation under significant Doppler frequency shift, and it has obvious ambiguity function. The ambiguity function of the HFM signal is the output of the matched filter, which has good autocorrelation and cross-correlation [26]. We superimpose HFM signal as the training sequence on the data signal to estimate the channel's frequency response, equalize the channel and eliminate the interference caused by signal superposition. As shown in Fig. 10, this method can correctly estimate the multi-path information of the channel with small side lobes.

4.4 Training and learning based on DRL

We consider using 30 different channel environments, adding different noise changes, and each episode is trained $k = 30 \times 92 \times 10$ times. Table 3 shows the neural network parameters of our deep reinforcement learning method. During the test, we use different environments of SWellEx-96 experiment with changing the position and noise of the transmitter and receiver and then compare the performance of the proposed adaptive modulation system based on DRL algorithm.

Figure 11 shows the performance of the DRL method when using the changed ϵ -greedy algorithm to select actions. The abscissa is several episodes, and the ordinate is the sum of earnings per episode. At the beginning of network training, in the first 30 episodes,

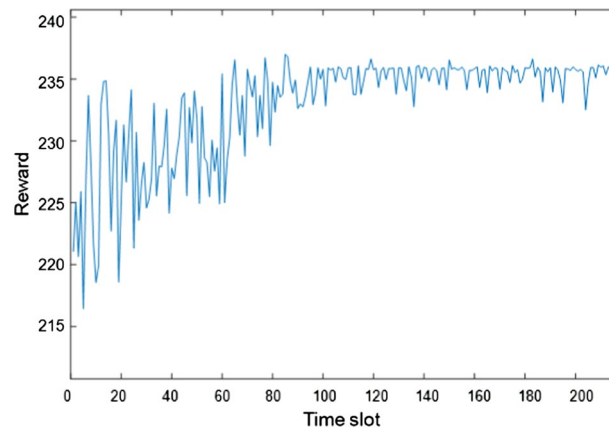


Fig. 11 Learning and training process of DRL

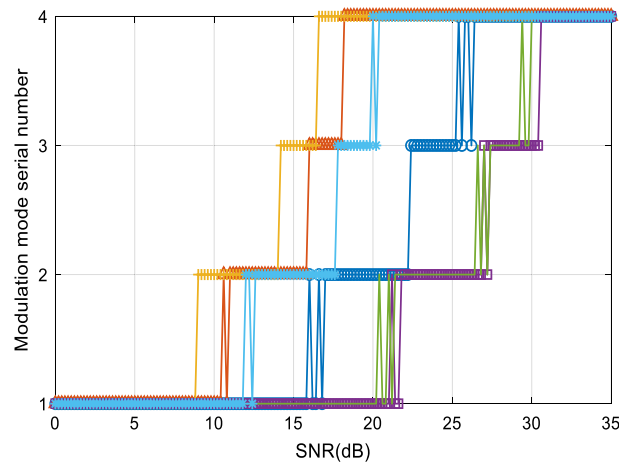


Fig. 12 The modulation mode serial number of SNR in some simulated environments

the income of each episode changes significantly. Because the value of ϵ in the greedy algorithm is relatively large, the network traverses all possible actions in the current state to prevent local optimization. Due to the time-varying channel, the state space of each episode process is not totally the same. However, DRL can quickly learn the relationship between state and $Q(s, a)$ to update network parameter θ without taking up additional storage. After training for a while, the system converges quickly, proving the fitting ability of a neural network, which proves DRL has learned the optimal strategy. After the output converges to 235, because the greedy algorithm selects the action, it still exists after 100 episodes, though the value of ϵ is smaller than before. Therefore, there will be suboptimal situations when implementing this strategy, resulting in small fluctuations in return.

4.5 The performance of our proposed adaptive scheme is compared

Figure 12 shows the 6 different environment modulation-order selection of the data set. When the SNR is quite low ($\text{SNR} < 7$ dB), the scheme selects modulation mode BPSK. When the SNR is high ($\text{SNR} > 31$ dB), the scheme selects 16-QAM as the modulation

scheme. In other cases, proposed adaptive scheme automatically adjusts the threshold to select the modulation mode.

As shown in Fig. 13, the BER of the proposed DRL-based AM scheme is compared in different channel environments. According to different SNRs, experiments show that the tabular method with fixed threshold and RLMC based on Q-learning cannot meet the requirements of $BER < 0.01$, and there is the possibility of non-compliance in some non-ideal conditions as high as 0.02. Under the low signal-to-noise ratio ($SNR < 7$ dB), even if the BPSK with the lowest modulation order is adopted, it still causes dramatical bit error, and thus, continuous transmission is not considered. When $SNR = 17$ dB, 23 dB and 27 dB, the change of a certain state does not comply with the quantitative CSI of the table, resulting in the sudden change of BER, making the average BER close to 0.02, which does not meet the system transmission requirements defined by the paper. The RLMC method also shows great fluctuations, which cannot guarantee that the bit error rate is always less than 0.01 at $SNR > 12$ dB. The proposed method has no mutation of BER in the whole transmissible SNR range. It adapts to various channel states, maintains the average performance, and demonstrates better stability.

Figure 14 shows that in a time-varying marine environment, different underwater acoustic environments generated under each signal-to-noise ratio under $0 \text{ dB} < SNR < 35 \text{ dB}$ are tested. By calculating the convergence average BER, spectral efficiency, throughput and penalty value, the appropriate modulation mode is selected according to the system mechanism to display the average maximum system throughput obtained by the time-varying channel.

The proposed method shows better average performance in a variety of marine environments. When the SNR is so low ($SNR < 7$ dB) and very large ($SNR > 28$ dB), it can maintain the same performance as the look-up table method and RLMC based on Q-learning. No matter what CSI, the SNR is very small, BPSK is selected as the modulation mode; the SNR is large, 16-QAM modulation transmission with the highest modulation order is selected, and the same reward can be obtained. In more general cases, as compared in Fig. 14:

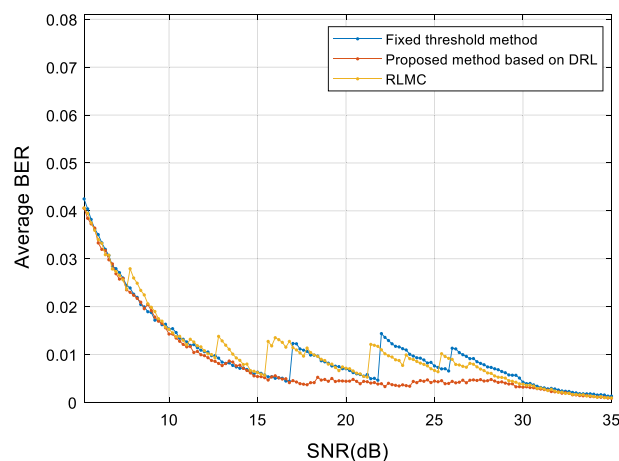


Fig. 13 Average BER in simulated environments

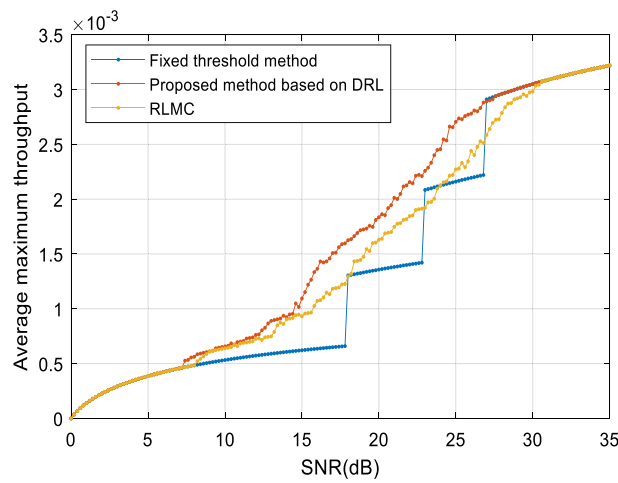


Fig. 14 Average maximum throughput in simulated environments

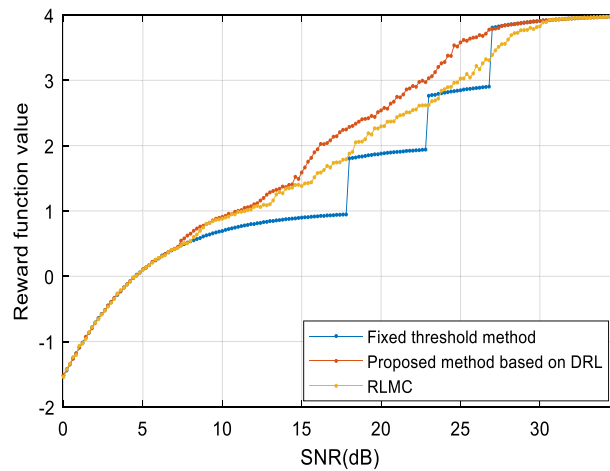


Fig. 15 Average reward function value in simulated environments

$$C = R \cdot P_{cf} \cdot \log_2(1 + \text{SNR}_{\text{dB}}), \tag{18}$$

where $R = \frac{I}{B \cdot T_s \cdot N_f}$ and $P_{cf} = (1 - \text{Error})^I$, $I = \text{bps}$. B is the bandwidth of OFDM system. T_s is OFDM symbol length. N_f is number of OFDM symbols per frame.

Under the condition of signal-to-noise ratio $7 \text{ dB} < \text{SNR} < 28 \text{ dB}$, the performance of the reward function value defined by us in defining the BER interval is shown according to Fig. 15. The reward mechanism of the adaptive modulation scheme proposed by us is calculated by the formula (11).

Comparing Figs. 14 and 15, it is found that in our OFDM underwater acoustic communication system, the transmission conditions are met in a specific SNR range, and the system throughput performance is significantly improved.

Therefore, DRL based on neural networks is more suitable for the environment with underwater state changes. The tabular method quantifies the channel state and cannot maximize the system throughput. RLMC based on Q-learning. RLMC is not suitable

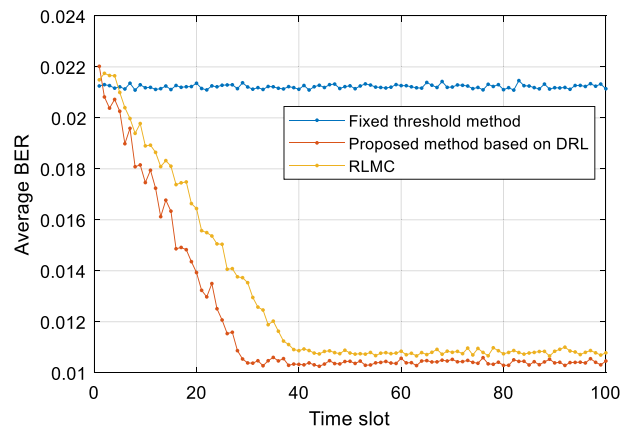


Fig. 16 Average BER of each time slot in simulated experiments

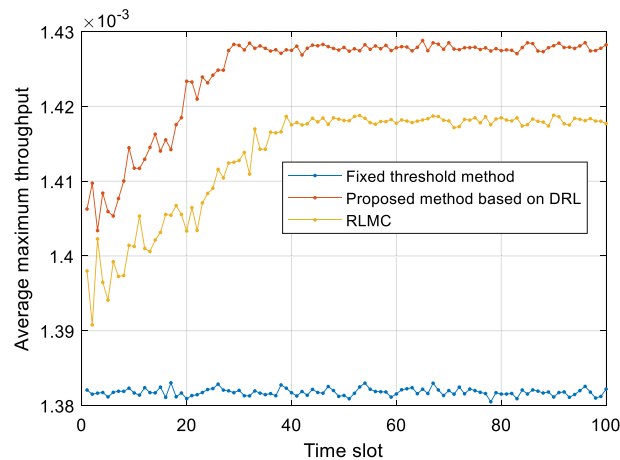


Fig. 17 Average maximum throughput of each time slot in simulated environments

for continuously changing channel environment and cannot minimize the bit error rate only by using SNR as evaluation criterion. However, compared with the tabular method, RLMC has been able to effectively improve the communication throughput while ensuring the bit error rate. However, the DRL method mapping state and Q-value relationships are suitable for both the old and newly changed channel environments, which has stronger robustness.

In Figs. 16, 17 and 18, the three graphs compare the performance of the proposed adaptive modulation strategy based on DRL algorithm. As the experience updates every time slot, the BERs, maximum system throughput and reward function value converged to optimal values. In every time slot, we test the environmental changes of all 30 CSI between $\text{SNR} \in [0 \text{ dB}, 35 \text{ dB}]$. In Fig. 16, the target average BER for our OFDM systems is set to 10^{-2} , and the non-adaptive scheme should reduce the overall throughput. The average BER of the tabular method maintains a value of about 0.0213. The proposed adaptive modulation strategy is just a little higher than the target. The BER decreased by 0.015. RLMC based on Q-learning is slightly higher than the method based on DRL. In Fig. 17, the average maximum system throughput of

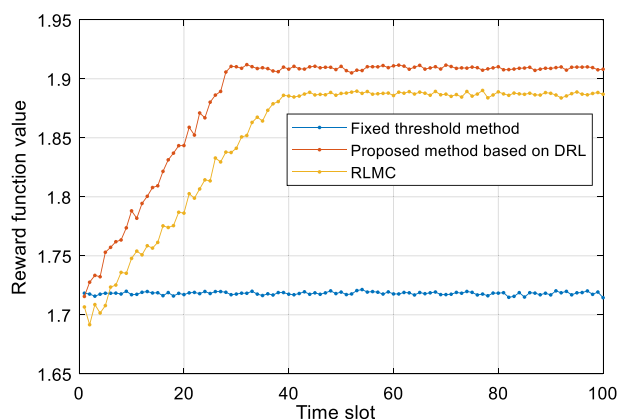


Fig. 18 Average reward function value of each time slot in simulated environments

proposed method based on DRL increased by 4.8×10^{-5} . More specifically, the DRL scheme begins to converge to a stable strategy after 25 time slots, and the convergence speed increases by 37.5% compared to RLMC scheme.

5 Conclusion

The underwater acoustic channel has a profound multi-path effect, Doppler effect and ocean noise compared with the wireless channel. The advantages of OFDM technology are that it can better adapt to the characteristics of apparent multi-path effect, low-frequency band and narrow bandwidth of the underwater acoustic channel. In the large time-varying underwater acoustic channel environment, the neural network is more and more widely used. In this paper, based on the OFDM underwater acoustic communication system, we propose a DRL adaptive modulation scheme. In order to improve the communication service quality, to integrate the system bit error rate, and to maximize the system throughput, we define a reward function. The proposed strategy maps the state, and Q value correspondence through the neural network uses the underwater acoustic Doppler insensitive HFM signal as the pilot to estimate the channel state and automatically selects the modulation scheme according to the real-time feedback of the time slot link. Finally, we tested the model in 30 different channel environment change experiments. The simulation shows that the proposed DRL adaptive modulation scheme has relatively lower and more stable BER performance. It can maximize the system throughput and improve the communication service quality and communication efficiency under the condition of meeting the transmission requirements.

Author contributions

XRC put forward the innovation, designed ideas of the paper and substantively revised it. PHY analyzed underwater acoustic channel data, realized the verification of the algorithm and was a major contributor in writing the manuscript. JL puts forward constructive suggestions to the paper and modifies the grammar of the English paper. SBL and JHL provided data and experimental conditions. All authors read and approved the final manuscript.

Funding

This work was supported by National Natural Science Foundation of China (Nos. 52171341, 61902431, and 61972417), science and technology project of Qingdao west coast new area under Grant No. 2020-84 and the science foundation of Shandong province under Grant No. ZR2020MF005.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 May 2022 Accepted: 13 December 2022

Published online: 03 January 2023

References

1. M. Stojanovic, J. Preisig, Underwater acoustic communication channels: propagation models and statistical characterization. *IEEE Commun. Mag.* **47**(1), 84–89 (2009)
2. M. Sadeghi, M. Elamassie, M. Uysal, Adaptive OFDM-based acoustic underwater transmission: system design and experimental verification, in *Proceedings of the 2017 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, F 5–8 June 2017 (2017)
3. M. Huda, N.B. Putri, T.B. Santoso, OFDM system with adaptive modulation for shallow water acoustic channel environment, in *Proceedings of the 2017 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, F 5–7 Oct. 2017 (2017)
4. L. Wan, H. Zhou, X. Xu et al., Adaptive modulation and coding for underwater acoustic OFDM. *IEEE J. Ocean. Eng.* **40**(2), 327–336 (2015)
5. M.J. Bocus, A. Doufexi, D. Agrafiotis, Performance of OFDM-based massive MIMO OTFS systems for underwater acoustic communication. *IET Commun.* **14**(4), 588–593 (2020)
6. A. Radosevic, R. Ahmed, T.M. Duman et al., Adaptive OFDM modulation for underwater acoustic communications: design considerations and experimental results. *IEEE J. Ocean. Eng.* **39**(2), 357–370 (2014)
7. S. Mangione, G.E. Galioto, D. Croce et al., A channel-aware adaptive modem for underwater acoustic communications. *IEEE Access* **9**, 76340–76353 (2021)
8. P. Jiang, T. Wang, B. Han et al., AI-aided online adaptive OFDM receiver: design and experimental results. *IEEE Trans. Wirel. Commun.* **20**(11), 7655–7668 (2021)
9. H. Wang, Y. Li, J. Qian, Self-adaptive resource allocation in underwater acoustic interference channel: a reinforcement learning approach. *IEEE Internet Things J.* **7**(4), 2816–2827 (2020)
10. Y. Mahmutoglu, K. Turk, E. Tugcu, Particle swarm optimization algorithm based decision feedback equalizer for underwater acoustic communication, in *Proceedings of the 2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, F 27–29 June 2016 (2016)
11. Y. Chen, W. Yu, X. Sun et al., Environment-aware communication channel quality prediction for underwater acoustic transmissions: a machine learning method. *Appl. Acoust.* **181**, 108128 (2021)
12. M.S.M. Alamgir, M.N. Sultana, K. Chang, Link adaptation on an underwater communications network using machine learning algorithms: boosted regression tree approach. *IEEE Access* **8**, 73957–73971 (2020)
13. J. Huang, R. Diamant, Adaptive modulation for long-range underwater acoustic communication. *IEEE Trans. Wirel. Commun.* **19**(10), 6844–6857 (2020)
14. Z. Jin, Q. Zhao, Y. Su, RCAR: a reinforcement-learning-based routing protocol for congestion-avoided underwater acoustic sensor networks. *IEEE Sens. J.* **19**(22), 10881–10891 (2019)
15. Y. Su, M. Liwang, Z. Gao et al., Optimal cooperative relaying and power control for IoUT networks with reinforcement learning. *IEEE Internet Things J.* **8**(2), 791–801 (2021)
16. C. Wang, Z. Wang, W. Sun et al., Reinforcement learning-based adaptive transmission in time-varying underwater acoustic channels. *IEEE Access* **6**, 2541–2558 (2018)
17. Q. Fu, A. Song, Adaptive modulation for underwater acoustic communications based on reinforcement learning, in *Proceedings of the OCEANS 2018 MTS/IEEE Charleston*, F 22–25 Oct. 2018 (2018)
18. W. Su, J. Lin, K. Chen et al., Reinforcement learning-based adaptive modulation and coding for efficient underwater communications. *IEEE Access* **7**, 67539–67550 (2019)
19. P. Qarabaqi, M. Stojanovic, Statistical characterization and computationally efficient modeling of a class of underwater acoustic communication channels. *IEEE J. Ocean. Eng.* **38**(4), 701–717 (2013)
20. Z. Yi-Jia, Z. Lan-Yue, M. Jia-Xin, Modeling and estimation of the space-time varying channels, in *Proceedings of the 2021 OES China Ocean Acoustics (COA)*, F 14–17 July 2021 (2021)
21. X. Cui, P. Yan, J. Li et al., Timing estimation of multiple hyperbolic frequency-modulated signals based on multicarrier underwater acoustic communication. *Trans. Emerg. Telecommun. Technol.* **33**, e4636 (2022)
22. X. Feng, J.F. Wang, X.Y. Kuai et al., Message passing-based impulsive noise mitigation and channel estimation for underwater acoustic OFDM communications. *IEEE Trans. Veh. Technol.* **71**(1), 611–625 (2022)
23. C.R. Berger, S. Zhou, J.C. Preisig et al., Sparse channel estimation for multicarrier underwater acoustic communication: from subspace methods to compressed sensing. *IEEE Trans. Signal Process.* **58**(3), 1708–1721 (2010)
24. P. Chen, Y. Rong, S. Nordholm et al., Joint channel estimation and impulsive noise mitigation in underwater acoustic OFDM communication systems. *IEEE Trans. Wirel. Commun.* **16**(9), 6165–6178 (2017)
25. W. Li, J.C. Preisig, Estimation of rapidly time-varying sparse channels. *IEEE J. Ocean. Eng.* **32**(4), 927–939 (2007)
26. S. Zhao, S. Yan, L. Xu, Doppler estimation based on HFM signal for underwater acoustic time-varying multipath channel, in *Proceedings of the 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, F 20–22 Sept. 2019 (2019)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.