remote sensing

MDPI

Article

Reinforcement Learning Based Relay Selection for Underwater Acoustic Cooperative Networks

Yuzhi Zhang [1,*], Yue Su [1], Xiaohong Shen [2], Anyi Wang [1], Bin Wang [1], Yang Liu [1] and Weigang Bai [3]

[1] School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710054, China; suyue@stu.xust.edu.cn (Y.S.); wanganyi@xust.edu.cn (A.W.); wangbin@mail.xidian.edu.cn (B.W.); lyang@xust.edu.cn (Y.L.)
[2] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; xhshen@nwpu.edu.cn
[3] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China; wgbai@xidian.edu.cn
* Correspondence: yuzhizhang@xust.edu.cn

Abstract: In the complex and dynamically varying underwater acoustic (UWA) channel, cooperative communication can improve throughput for UWA sensor networks. In this paper, we design a reasonable relay selection strategy for efficient cooperation with reinforcement learning (RL), considering the characteristics of UWA channel variation and long transmission delay. The proposed scheme establishes effective state and reward expression to better reveal the relationship between RL and UWA environment. Meanwhile, simulated annealing (SA) algorithm is integrated with RL to improve the performance of relay selection, where exploration rate of RL is dynamically adapted by SA optimization through the temperature decline rate. Furthermore, the fast reinforcement learning (FRL) strategy with pre-training process is proposed for practical UWA network implementation. The whole proposed SA-FRL scheme has been evaluated by both simulation and experimental data. The simulation and experimental results show that the proposed relay selection scheme can converge more quickly than classical RL and random selection with the increase of the number of iterations. The reward, access delay and data rate of SA-FRL can converge at the highest value and are close to the ideal optimum value. All in all, the proposed SA-FRL relay selection scheme can improve the communication efficiency through the selection of the relay nodes with high link quality and low access delay.

Keywords: cooperative communication; relay selection; reinforcement learning; underwater acoustic networks

Citation: Zhang, Y.; Su, Y.; Shen, X.; Wang, A.; Wang, B.; Liu, Y.; Bai, W. Reinforcement Learning Based Relay Selection for Underwater Acoustic Cooperative Networks. *Remote Sens.* **2022**, *14*, 1417. https://doi.org/10.3390/rs14061417

Academic Editor: Andrzej Stateczny

Received: 31 January 2022
Accepted: 11 March 2022
Published: 15 March 2022

1. Introduction

Over the last two decades, the development of underwater acoustic (UWA) systems for ocean monitoring has grown sharply. As a matter of fact, the demand of UWA data collection has become more and more important for numerous research and industries fields, such as renewable energies, underwater mining, offshore oil and gas [1–3]. However, the ocean environment presents unprecedented challenges for UWA data collection including, but not restricted to, limited communication distance, limited bandwidth, limited service energy and dynamic channel conditions. In the complex and dynamic ocean environment, the platforms or sensing nodes with intelligent learning capabilities have the potential to learning optimum strategy for performance improvement.

The unique characteristics of UWA channel bring many limitations to data transmissions. For example, the large frequency and distance dependent transmission attenuation [4,5] will lead to the limited communication distances. At the same time, the temporal and spatial variation characteristics [6–8] will make the data link lose the connectivity. During the long-term deployment of UWA network, the link disconnection is highly likely

to occur which results in transmission failure. Cooperative communication is a key technology that can achieve reliable transmission, where the relay node can assist the failure nodes to forward data to destination. One of the crucial issues in cooperative communication is effective relay selection. This paper will explore how intelligent reinforcement learning helps to enhance the performance of relay selection in UWA communication networks.

1.1. Related Work

Basic UWA communication networks establish two-way acoustic links between underwater devices such as various sensors and autonomous underwater vehicles. The acoustic network is then connected to a surface station which can be further connected to offshore data center. The network design principles for each network layer have been discussed in [9,10]. The authors explore potential applications and research directions for underwater sensor networking [11], such as monitoring, navigation and localization [12]. In the early-stage research, the design of UWA network is commonly performed in a layered structure, typically including the physical layer, the data link layer and the network layer [9–11]. For higher layer protocol design and network performance evaluation, the physical layer parameters are usually assumed to be ideal, even though the errors and special features of UWA channel will affect network performance. In recent years, the imperfect feature in physical layer has drawn more and more attention for practical applications. For example, in the localization scenario, how the error sources influence the estimated localization result has been analyzed in [13], where the error sources include the length of baseline, the error in ranging, etc. In UWA adaptive communication, how the imperfect channel state information affects the performance of multi-carrier adaptive modulation has been analyzed by experimental data [14]. In this paper, the proposed cooperative relay selection scheme will model the imperfect outdated channel state information to combine with the reinforcement learning process, and will be evaluated by both simulation and experiment.

The UWA cooperative communication has the advantage that it is easily to be integrated with the existing system including both static [15] and mobile networks [16]. For integration with data collection network, Zhang et al. [17] proposed selective relay cooperation protocol and dynamic node cooperative protocol for a practical underwater data collection network, where UWA sensor nodes can cooperate with each other without adding additional relay nodes. Several lake experiments with full protocol implementation have demonstrated that the proposed protocols can significantly improve the network performance. For dynamic UWA network with joined new nodes, Liao et al. [18] proposed a network access mechanism, which allows nodes that have joined the network to help new nodes access the network through multi-hop relay forwarding. In the relay path determination phase, they presented a relay path selection algorithm based on lifetime and energy efficiency. This algorithm can significantly extend the network lifetime and improve energy efficiency. For integration with physical layer modulation scheme, Doosti-Aref et al. [19] investigated both optimal relay selection and power loading issues for UWA orthogonal frequency division multiplexing (OFDM) cooperative transmission. They considered amplify-and-forward (AF) relaying with perfect channel state information (CSI) in the frequency selective Rician fading channel. The optimal relay selection and power allocation problems are solved in two phase to minimize the bit error rate (BER) and maximize the system capacity. For relay cooperation in linear UWA networks, Li et al. [20] researched the relay deployment problem in two aspects: which conditions a relay should be deployed and where to deploy it for the optimal performance. A closed-form expression for the open distance and optimal placing position have been derived, and simulation results demonstrate that properly introducing a relay can reduce the network energy consumption almost without increasing the end-to-end delay. All the above UWA cooperation protocols assume theoretical channel model or perfect channel state information for analysis.

The intelligent algorithms have the potential to learn optimum strategy for performance improvement in complex UWA channels. Some research studies focus on UWA relay selection by exploiting machine learning algorithms. For example, Li et al. [21] modeled

relay selection as a contextual bandit problem, where the relay is selected based on one bit of contextual communication environment information rather than a large amounts of instantaneous or statistical CSI information. The proposed scheme has a stable performance in complex UWA environment and simplifies the relay selection process for efficient cooperative transmission. Zhao et al. [22] proposed a hierarchical adversarial multi-armed bandit learning framework to improve adaptive relay decision. The proposed framework exploits heuristic interactions between hierarchical frameworks to integrate reward estimation, information prediction, adaptive exploration and decision making within a holistic algorithm to maximize learning efficiency. It enables cooperative communication networks with greater stability and lower communication costs in complex and varying underwater environments. The above two studies have shown the potential of artificial intelligence learning for underwater systems. Different from bandits learning problem in [21,22], reinforcement learning (RL) provides a general and powerful computational framework for sequential decision making problem. RL can satisfy the optimization in actual sequential time varying UWA channels, where the channel usually cannot be modeled into a linear pattern.

Reinforcement learning (RL) has been used to address the problem of maximizing rewards or achieving specific goals by learning strategies during the interaction of an agent with its environment. In UWA communication systems, some researchers have employed RL to solve adaptive problems, such as adaptive data and image transmission [23,24], and adaptive routing [25–27]. In terms of relay selection, Jadoon et al. [28] firstly proposed QL-based relay selection algorithm (QL-RSA) in wireless sensor networks. QL-RSA defines the relay nodes as the set of states and whether to select a new relay as the set of actions, which requires less feedback information and provides each source a self-learning capability. QL-RSA receives the feedback reward by learning AF cooperative environment in time-varying Rayleigh fading channels to maximize the network throughput. In wireless sensor networks, Su et al. [29] modeled the process of relay selection for cooperative communications as a Markov decision process (MDP) and proposed a deep reinforcement learning based relay selection scheme, where the channel state and relay nodes are defined as the state and actions of RL, respectively. The best relay is selected from multiple relay nodes for better performance of outage probability and system capacity. The above researches model channel theoretically and assume the feedback is perfect for wireless cooperative sensor networks [28,29]. For implementation RL based relay selection in UWA sensor networks, Su et al. extended the scheme [29] to internet of underwater things (IoUT) network [30] with additional power adjust strategy to maximize the end-to-end signal to noise ratio (SNR) of the system. Han et al. [31] investigated reinforcement learning based joint relay selection and power allocation in energy harvesting UWA cooperative networks. In the proposed model, a joint state expression is presented to better reveal the relationship between learning and environment, and a reward function that consists of channel capacity and energy consumption is proposed for adjusting power strategy to balance instantaneous capacity and long-term quality of service (QoS). Wang et al. [32] investigated the power allocation problem in energy harvesting full duplex UWA cooperative relay network by reinforcement learning. The relay adjusts transmission power by the battery level, harvested energy and CSI to maximize the cumulative rate of the UWA network. In [30–32], the UWA channel is simulated by integrated empirical formula and statistic model, and the relay selection criteria is similar to wireless sensor networks, mainly by CSI. This paper will systemically analyze how the specific phenomena of UWA affect relay selection criteria in Section 1.2, will design an effective state and reward expression for UWA RL based relay selection with the consideration of UWA propagation delay in Section 3.1.

In the above RL based UWA cooperation systems [28–32], the learning related parameters are static. We consider that the learning process is a dynamic process to explore the possible actions for maximizing the reward. Accordingly, in the early stage of learning, the intelligent agent of RL should explore the possible random actions with high possibility. With the number of iterations increasing, the learning will turn to convergence gradually, then the agent should explore random action with low possibility and exploit the previous

learning results for relay selection. Herein, we employ simulated annealing (SA) algorithm to adapt the exploration rate and greedy factor dynamically for better balance between exploration and exploitation. The SA algorithm is one of best-known metaheuristic methods for addressing global optimization problems [33]. The SA algorithm is powerful to be applied in diverse areas, and still attracts much attention in recent research, such as multi-objective optimization problem [34] and federated learning problem [35]. The SA algorithm simulates the cooling down process of solid annealing whose temperature decreases from a high value. With the decrease of temperature, the global optimal solution is explored randomly in the solution space according to the Metropolis criterion to avoid falling into local optimum. For the routing problem, SA is combined into Q-routing algorithm which has self-adaptive learning rate for dynamic exploration [36,37]. We utilize SA algorithm to dynamically adjust the exploration rate and dynamic greedy factor for RL through the temperature decline rate. The proposed algorithm can better balance the exploration and exploitation process of reinforcement learning, and achieve better long-term performance as the agent can learn the best action in both exploration and exploitation stage.

1.2. Motivations and Contributions

Principally, the signal attenuation from the source node to the destination node is distance-dependent in both terrestrial electromagnetic and underwater acoustic networks. Correspondingly, the SNR criteria based optimal relay usually has a shorter distance to the source or destination node than other potential relays. However, the SNR based criteria may not be effective without taking specific characteristics of UWA channel into consideration. Low UWA speed leads to long propagation delay of UWA signals [38,39]. Considering the long propagation delay together with the distance independent characteristics of UWA channel, we will discuss how these phenomena affect the SNR based cooperative criteria.

- **Shadow zones:** In radio communications, the path of the radio wave is modeled as straight line. In large scale underwater communication networks, the speed of sound in seawater is not a constant, so the actual UWA propagation path is a curve where the sound ray bends to the lower velocity layer of the water and leads to nonlinear sound propagation [40]. As shown in Figure 1, the nonlinear UWA propagation will form communication shadow zones and convergence zones. The sound rays will converge to convergence zones, and the node in shadow zones can only receive few sound rays. Assuming that potential relay node r_1 and r_2 are located among the source S and destination D, and r_1 is located in the convergence zone and r_2 is located in the shadow zone. By SNR based cooperative criteria, r_1 will be chosen as optimal relay. However, due to the nonlinear propagation of sound, the sound ray of r_1 might transmit a longer way which consumes more time for one-trip data transmission. The channel capacity obtained from higher SNR may not compensate the longer propagation delay, and finally leading to lower throughput.
- **Underwater obstacles:** Underwater random obstacles will interfere the signals between potential relay nodes and source/destination. For example, as shown in Figure 2, potential relay node r_2 is close to the destination node but there are fish schools as obstacle, which results in worse channel condition than r_1. While relay node r_1 is far away but there are no obstacles. By SNR only cooperative criteria, r_1 might be chosen as optimal relay, while it is far away from destination. In this case, there is no guarantee which relay node is better in the view of network throughput. So we should take both channel quality and propagation delay into consideration.

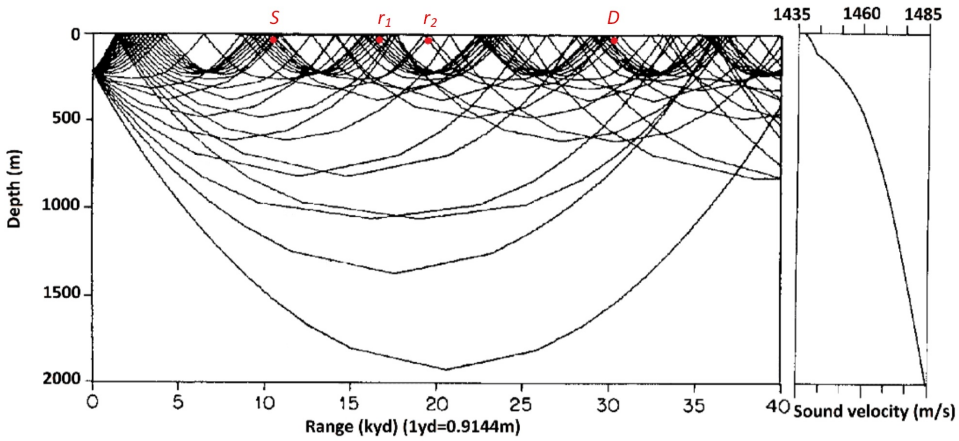

Figure 1. Test results of nonlinear sound ray propagation in the Arctic region, (Urick [40]).

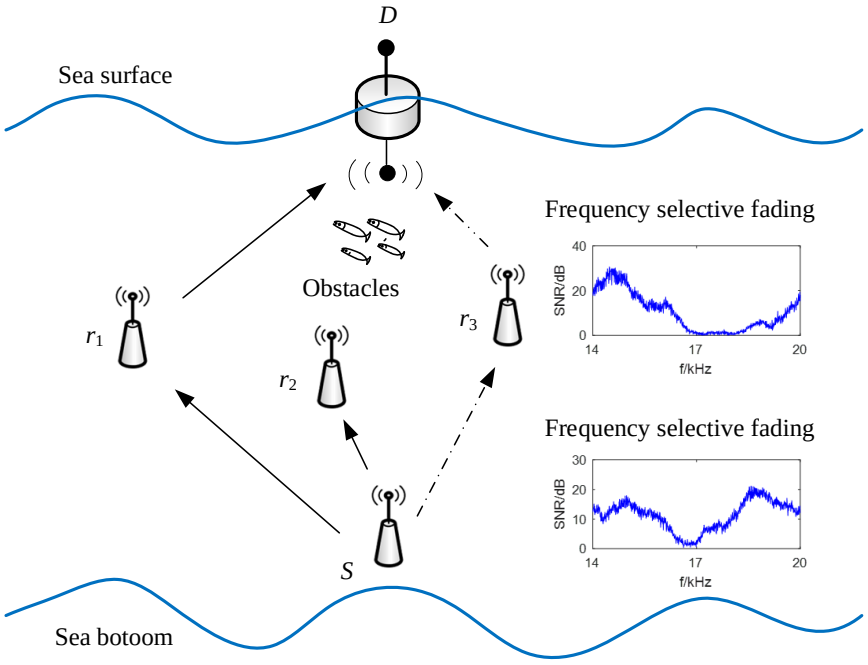

Figure 2. The impact of obstacles and frequency selective fading on relay selection.

- **Frequency selective fading:** UWA channel is a typically frequency selective fading channel with severe multi-path effect, where multi-path propagation can result from signal reflections from surfaces, bottoms and water objects or bending along the axis of the lowest sound speed. The signals from the source node propagate through different paths to the destination node may result in the signal elimination on specific frequency. The signal attenuation due to frequency-selective fading is independent of distance, and it only depends on the signal frequency and the multi-path delay difference between the multiple paths. For example, as shown in Figure 2, suppose that the channels between S to r_3 and r_2 to D are in frequency selective fading, and r_1 will be chosen as relay by SNR based criteria. Nevertheless, with longer propagation delay, it is uncertain that r_1 will have higher throughput than r_3.

Different from radio network solutions, if above typical phenomena of the UWA channel are not considered, the best relay selection may not be effective in practical underwater scenario. In this paper, we propose a RL based relay selection scheme for UWA networks that takes both channel quality and transmission delay into consideration, which is more

effective and applicable for UWA networks. Furthermore, the parameters of the learning process for relay selection strategy will be dynamically optimized for less convergence time and better convergence value. The main contributions of this paper are as follows:

- We propose a RL based relay selection strategy for UWA cooperative communication with the knowledge of feedback CSI and transmission delays of relay nodes. Effective state, action and reward expression have been formulated to better reveal the relationship between agent and environment. The combination of outdated CSI and the system mutual information (MI) is defined as the state of RL. The selection of different relay nodes is set as the action. The joint function of system MI and access delay is established as the reward. The proposed scheme selects relay nodes with good link quality and low access delay.
- We exploit the simulated annealing (SA) algorithm to improve the convergence performance of UWA relay selection strategy. In the RL process, exploration rate is dynamically adapted by SA optimization through the temperature decline rate. In turn, the temperate decline rate will regulate the probability to accept the new solution of Q-values. The proposed strategy can improve the convergence speed and value in comparison with RL with constant exploration rate.
- We propose the fast reinforcement learning (FRL) scheme for implementation of the proposed relay selection scheme in practical UWA networks. Before experimental implementation, the similar simulated channel model will be utilized for pre-training of SA-RL scheme. So the possible action in different channel state will be explored. And then, the on-line RL stage will be implemented with the pre-trained parameters. Evaluated by experimental data, it shows that the proposed FRL scheme can accelerate the convergence and can improve data rate UWA network.

The remainder of this paper is organized as follows. Section 2 formulates the system model and channel model in details. Section 3 models the UWA dynamic cooperative problem by RL, and has proposed SA-FRL method for performance improvement. Section 4 presents the simulation and experimental results in comparison with other RL based UWA relay selection methods. Section 5 provides a discussions. Finally, Section 6 concludes the research work.

2. System Model

2.1. UWA Cooperative Communication System Model

In this paper, we consider an UWA sensors network as shown in Figure 3. The sensor nodes are deployed underwater for sensing and transmitting important data to the sea-surface destination node. The sensor nodes are equipped with UWA modem for data transmission. As the sensor nodes are equipped with power limited battery, the transmission power can not be too high for the consideration of network lifetime. At the same time, the attenuation of UWA channel is severe and the loss of connectivity usually happens. On this account, the relay nodes which are randomly distributed between source and destination can help to forward the collected data to improve transmission quality of service of whole system. In this paper, we consider an application scenario where a source node S sends data to a destination node D and selects an optimal relay node r_i for cooperation, as shown in Figure 3.

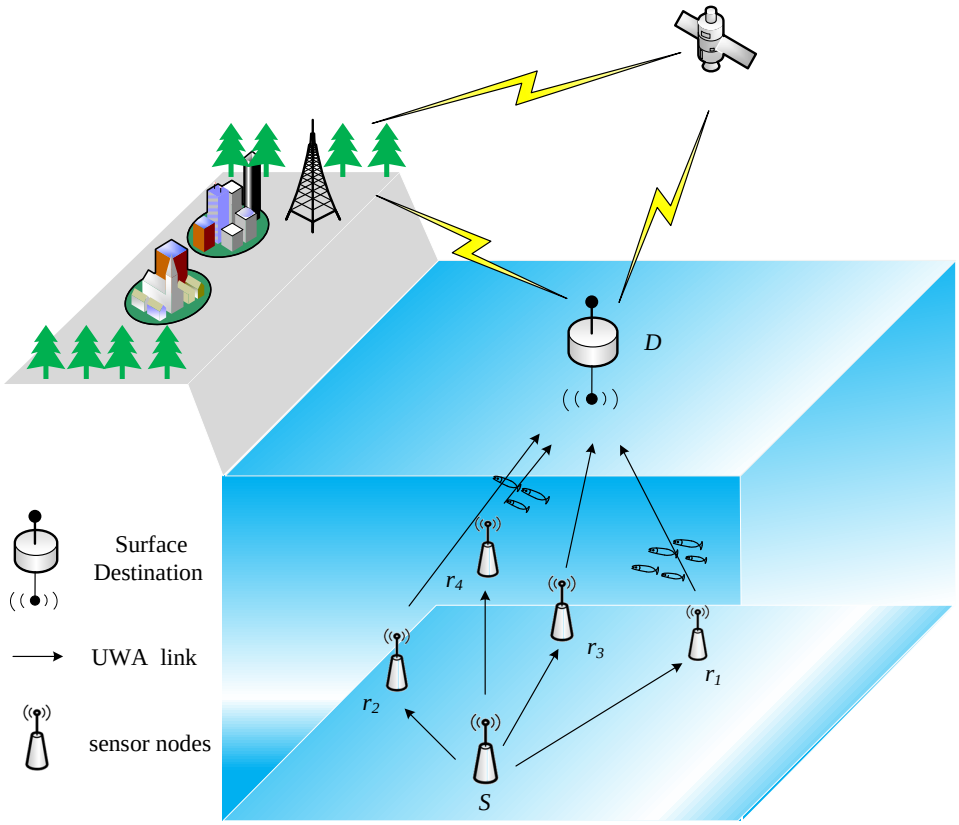

Figure 3. An UWA cooperative network system.

In this paper, time division multiple access (TDMA) protocol is employed to avoid interference among the nodes. When the relay node forwards the information, it consumes time or frequency resources, which will downgrade the spectrum efficiency. For example, when the system selects M relay nodes for cooperation, only one efficient data packet has been transmitted in $M+1$ time slots. The idea of "opportunistic relay" [41] has proved that the "best relay" cooperative strategy can also achieve full set gain compared to the strategy of selecting multiple relays. The best relay strategy greatly reduces the system cooperative overhead and still can realize relatively high throughput. Therefore in this paper, the data is sent in time round, and one best relay is selected in one round. In the initialization round, the source node broadcasts the data and one random relay node forwards the data. In the normal round of data transmission, node S sends the data in the first time slot, and the selected best relay forwards the data in the second time slot. In this way, the destination node only needs to select and notify an optimal underwater relay node without calculating and feeding back channel information of all relay nodes, which greatly scale down the complexity of the system.

The relay modes for cooperative communication are usually divided into amplify-and-forward (AF) and decode-and-forward (DF) according to how the relay nodes process the overheard received signals. The AF mode amplifies and forwards the received signal directly to the destination. While the DF mode decodes the received signal, re-encodes the signal and then forwards it to destination [42,43]. The AF is simple to implement, but the noise will also been amplified. The DF has no such drawback due to the decoding and encoding operations at the relay node. Especially in UWA channel, the sources of noise are complex, such as the wind waves noise, turbulence noise, shipping activity noise, thermal noise, etc. The noise level is high, which is usually several tens of dB. When AF

is employed, it will amplify the noise when amplify the whole received signal. Previous studies in [17,44] have shown how DF can efficiently work in UWA communication and network with experimental evaluation. So different from the previous UWA QL based cooperation [30–32], the research focus of this paper is on the scenario that the relay nodes employ the DF protocol for cooperation.

For DF system, the mutual information of the entire system when selecting relay i at time round t can be expressed as

$$I^{(t)} = \frac{1}{2}\log_2\left(1 + {\gamma_{SD}}^{(t)} + {\gamma_{r_iD}}^{(t)}\right) \tag{1}$$

where γ_{SD} and γ_{r_iD} are the received SNRs to measure the channel quality.

γ_{SD} is the SNR of signals received by node D sent from node S in tth round of data transmission, which can be expressed as

$${\gamma_{SD}}^{(t)} = \frac{P_S {G_{SD}}^{(t)}}{\sigma^2} \tag{2}$$

where P_S is the transmission power of the source node, and ${G_{SD}}^{(t)}$ represents the channel gain of S-D communication link.

γ_{r_iD} is the SNR of signals received by D sent from node relay node r_i in tth round, which can be represented as

$${\gamma_{r_iD}}^{(t)} = \frac{P_r {G_{r_iD}}^{(t)}}{\sigma^2} \tag{3}$$

where P_r is the transmission power of the relay node and ${G_{r_iD}}^{(t)}$ represents the channel gain of r_i-D communication link.

2.2. UWA Channel Model

UWA systems have to be designed to operate in the complex underwater environment, where the system geometry and channel conditions are varying during the deployment. Especially for resources allocation system, it is crucial to have a relatively accurate channel model for evaluation of the algorithms and protocols before experiment. Beam tracing tools which use ray theory can provide deterministic channel impulsive response (CIR), but they do not consider random channel variations, such as BELLHOP [45,46]. Some studies have been conducted to statistically model the UWA channel. The relay selection schemes in [30,31] have utilized the path loss function to model channel gain, Wenz noise model to calculate the noise power level (NL), and fading coefficient that follows specific fading distribution to simulate small scale fading. This type of method can simulate the received SNR statistically.

In this paper, the time-varying UWA channel has been simulated in the view of physical layer with the consideration of both acoustic propagation loss and inevitable random channel variations. The random UWA channel variations include location uncertainty of transmitter/receiver, the motion of sea surface, and small-scale random scattering, etc. The variations will affect not only the locally averaged received power, but also the instantaneous fast varying channel response. The channel transfer function will be modeled with acoustic propagation loss and channel variations, and then the gain contained in the channel can be calculated accordingly [45].

The propagation path loss of UWA signal is related to carrier frequency f and transmission distance d. The UWA propagation loss $A(d, f)$ can be calculated by the empirical formula in reference [4] as

$$A(d, f) = d^k \alpha(f)^d \tag{4}$$

where d^k stands for spreading loss and $\alpha(f)^d$ stands for absorption loss in water. k is the spreading factor. $\alpha(f)$ is absorption coefficient which can be calculated by the Thorp's empirical formula in dB per kilometer as

$$10\lg[\alpha(f)] = 0.11\frac{f^2}{1+f^2} + 44\frac{f^2}{4100+f^2} + 2.72\times10^{-4}f^2 + 0.003 \tag{5}$$

where f is in KHz.

Due to the displacement of the transmitter/receiver in underwater environment, as well as the changes in the sea surface height or the shape of seafloor, the system geometry is varying. These factors cause the uncertain change of multi-path. Suppose that the pth path has an average communication distance $\bar{d}_p$, and d_p can be modeled as a variable that follows Gaussian distribution. The channel transfer function is proportional to reflection and inversely proportional to path loss. For pth path, the corresponding average channel transfer function $\bar{H}_p$ of pth path with average propagation distance $\bar{d}_p$ can be expressed as

$$\bar{H}_p(f) = \frac{\Gamma_p}{\sqrt{A(\bar{d}_p, f)}} = \frac{\Gamma_p}{\sqrt{\left(\frac{\bar{d}_p}{\bar{d}_0}\right)^k \alpha(f)^{\bar{d}_p - \bar{d}_0}}}\bar{H}_0(f) = \bar{h}_p\bar{H}_0(f) \tag{6}$$

where Γ_p is the cumulative reflection coefficient of the seabed and the sea surface. For normalized expression to add all multi-paths together, normalized multi-path propagation distance $\bar{d}_0$ is introduced. $\bar{H}_0(f)$ is the normalized reference channel transfer function referred to $\bar{d}_0$. $\bar{h}_p$ is the average gain of the normalized path.

Modeling the pth multi-path as a path with gain h_p and delay τ_p, then the total channel transfer function can be represented as

$$H(f) = \bar{H}_0(f)\sum_p h_p e^{-j2\pi f\tau_p} \tag{7}$$

where τ_p can be expressed as a function of propagation speed c as $\tau_p = d_p/c - t_0$, and t_0 is the normalized reference delay. The changes in path gain h_p and path delay τ_p caused by the temporal change of the position of the UWA node can be expressed as a function of the path length which is expressed as $d_p = \bar{d}_p + \Delta d_p$, where Δd_p is a random offset distance. The path gain h_p in Equation (7) can be characterized as a function of $\bar{h}_p$ in Equation (6) as

$$h_p = \bar{h}_p\frac{1}{\sqrt{\left(1+\frac{\Delta d_p}{\bar{d}_p}\right)^k \alpha_0^{\Delta d_p}}} \tag{8}$$

where α_0 is the normalized absorption coefficient.

For the further step, we consider the widely existed scattering in UWA channel, which is caused by placement of scattering points within a scattering field. If scattering occurs along a path, the path will be split to a number of micro-paths. Then the channel transfer function can be expressed as

$$H(f) = \bar{H}_0(f)\sum_p h_p\beta_p(f)e^{-j2\pi f\tau_p} \tag{9}$$

where β_p is a small-scale fading coefficient, which can be expressed as

$$\beta_p(f) = \frac{1}{h_p}\sum_{i\geq0} h_{p,i}e^{-j2\pi f\delta_{\tau_{p,i}}} \tag{10}$$

where $h_{p,i}$ is the random gain within the multi-path cluster, $\tau_{p,i} = \tau_p + \delta_{\tau_{p,i}}$ is the random delay within the multi-path cluster, $h_{p,i}$ and $\delta_{p,i}$ characterize the random scattering in the UWA channel.

Finally, the instantaneous channel gain $\tilde{G}(t)$ of time-varying UWA channel is expressed as

$$\tilde{G}(t) = \frac{1}{B}\int_{f_0}^{f_0+B} |H(f,t)|^2 df \tag{11}$$

where $H(f,t)$ is the total channel transfer function including both large and small scale channel variations, which is related to the large-scale path propagation loss, interface reflection, small-scale scattering and other factors.

2.3. UWA Channel with Long Transmission Delays

The speed of UWA propagation is only about 1500 m/s, which leads to large delay of UWA signal during its transmission. The transmission delay of UWA communication is five orders of magnitude higher than that of land-based radio communication. The impact on UWA communication and network should not be ignored.

On one hand, the location of the relay nodes are different, so that their transmission delays to the destination node are different either. According to the analysis in Section 1, the nodes with better channel quality may not be able to reach the destination node with the shortest access delay. From the perspective of system throughput, the channel quality and access delay should be comprehensively considered. Define access delay of one round in an UWA cooperative communication system as the sum of the transmission delay and the packet duration. The total access delay T_{Sr_iD} from source node to the destination node can be expressed as

$$T_{Sr_iD} = (T_{Sr_i} + T_p) + (T_{r_iD} + T_p) \tag{12}$$

where T_{Sr_i} is the transmission delay from the source node to the relay node r_i, T_{r_iD} is the transmission delay from the relay node r_i to the destination node, and T_p is the duration of the data packet. The source node can know the channel status and transmission delay information of each potential relay through the feedback of the destination node [47]. In RL based UWA relay selection, the influence of access delay will be designed in the expression of system reward.

On another hand, the relay selection is decided according to the UWA feedback CSI. When the selected relay transmits data, it experiences a certain transmission delay after the time of feedback. That is to say, the actual transmission is decided by the feedback CSI one transmission round ago. The phenomenon of the outdated feedback CSI will be considered in the state design of RL based UWA relay selection.

3. RL Based Relay Selection Method for UWA Cooperative Networks

With the description of system model, the system actions are discrete, which are corresponding to the selection of different relay nodes. For a given system model, the UWA relay selection process is comparable to the state transition process, where the agent selects best relay based on the current system state and then obtains the next system state. The system state in the next round is only related to current state and action, and is independent of the other previous states and actions. Therefore, the relay selection process can be modeled as Markov Decision Process.

Reinforcement learning can derive optimal strategy for MDP system through the interactions with the environment by "attempt and failure" mechanism. The basic principle is that if the chosen action receives a larger reward from the environment, the probability that the agent adopting this action strategy in the future will increase. On the contrary, when the less reward is obtained, the probability that the agent choosing the action will be weakened.

Theoretical relay selection strategies often assume that the channel state information is ideal. However, in the dynamically changed UWA channel, the channel state is time-

and frequency- varying and with outdated property. Reinforcement learning can train the best action strategy a according to the known channel state s and the corresponding reward r, and finally obtain the best relay selection. RL relay selection can work without prior channel knowledge and system model. And even with delayed imperfect CSI, it can still train the action under specific available states, which reveals the relationship between channel and actions.

3.1. RL Based Relay Selection for UWA Cooperative Communication

The objective of the proposed RL based relay selection for UWA cooperative communication is to select the potential relay node with high mutual information and low access delay for the improvement of long-term network throughput. A framework of the RL based UWA cooperative communication is shown in Figure 4, where the environment is the UWA channel, the agent is the source node. The specific design of the states, actions and rewards are as follows.

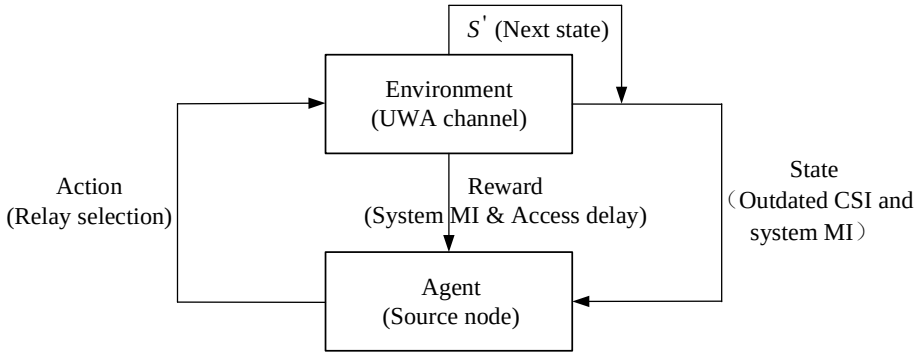

Figure 4. Framework of RL based UWA cooperative communication.

States: The state for reinforcement learning is combination of received feedback SNR and system mutual information:

$$\mathbf{D}^{(t)} = [\gamma_{SD}^{(t-1)}, \gamma_{r_iD}^{(t-1)}, I^{(t-1)}] \tag{13}$$

where $\gamma_{SD}^{(t-1)}$ and $\gamma_{r_iD}^{(t-1)}$ represent the channel states of S-D and r_i-D link in time round $t-1$, respectively. $\gamma_{SD}^{(t)}$ and $\gamma_{r_iD}^{(t)}$ are defined in Equations (2) and (3), and the channel gain can be calculated as Section 2. $I^{(t-1)}$ represents the whole system state at times round $t-1$ in mutual information. It should be noted that when the agent performs an action at time t according to the state $s^{(t)}$, the SNR and mutual information are actually the feedback at time $t-1$. UWA system has to select the relay with the available outdated CSI.

Actions: The action is to select one of the N nodes as a relay

$$\mathbf{A} = \{a^{(t)}\} \tag{14}$$

where $\{a^{(t)}\} \in \mathbf{A} = \{1, 2, \ldots, i, \ldots, N\}$. $a^{(k)} = i$ indicates that node i is selected as a relay for cooperation in time round t.

Rewards: According to the analysis in Sections 1.2 and 2.3, We define the reward as a function of system mutual information and total access delay when selecting a relay

$$r^{(t)} = \beta_1 \times \ln^{(I^{(t)})} - \beta_2 \times T_{Dr_iD} - \rho \times u \tag{15}$$

where β_1 is the proportional coefficients of mutual information, and concavity view of logarithmic function can well capture the system utility for data analysis. β_2 is the proportional coefficients of access delay, and ρ is the scale coefficient of energy consumption factor u.

In the UWA cooperative network, when the selected relay node has the higher mutual information, the system will achieve higher rewards and throughput. The cumulative reward of the system with RL can be expressed as

$$r^{(0)} + \zeta r^{(1)} + \zeta r^{(2)} + \cdots = \sum_{t=0}^{M} \zeta r^{(t)} \tag{16}$$

where $r^{(t)}$ is the immediate reward of tth round, ζ is the discount factor and M is the total number of iterations of RL. The smaller the ζ value is, the more immediate reward is considered. The greater the ζ value is, the more future reward is considered. The objective of RL based dynamic cooperation is to select the optimal relay node to reach the maximum amount of data during the long-term network deployment.

State-action Q function: The Q-table is used in the RL algorithm to store the maximum future expectation reward that can be obtained by taking different actions in each state. The agent will use the ε-greedy algorithm to select the relay node with the maximum Q-value for cooperative communicate at the current moment. The Q-value update function for iteration is

$$Q\left(s^{(t)}, a^{(t)}\right) \leftarrow (1-\sigma) Q\left(s^{(t)}, a^{(t)}\right) + \sigma[r^{(t)} + \zeta V(s^{(t+1)})] \tag{17}$$

$$V(s^{(t)}) = \max_{a \in \{1,2,\cdots,N\}} Q(s^{(t)}, a^{(t)}) \tag{18}$$

where $\sigma \in (0, 1]$ is the learning rate, which represents the speed of updating Q-value. V-table stores the maximum Q-value in each state.

3.2. SA Optimized RL Relay Selection Strategy

According to the values in Q-table, optimal relay can be selected by ε-greedy algorithm to avoid being trapped in a local maximum. The ε-greedy selection strategy can be expressed as

$$\pi(s|a') = \begin{cases} \arg\max Q\left(s^{(t)}, a'\right), \text{ probability } 1-\varepsilon \\ \text{random action, probability } \varepsilon \end{cases} \tag{19}$$

where ε is the greedy factor. In state s, the ε-greedy algorithm selects the action corresponding to the maximum Q-value with probability $1-\varepsilon$, and selects random action with probability ε. ε is an invariant constant in classical ε-greedy algorithm.

For invariant constant ε, even when the rewards turn to be stable, the selection strategy will still explore other possible actions with probability ε, which will cause the reward to fluctuate. Therefore, this paper exploits the thought of SA algorithm to optimize the strategy of relay selection dynamically. The agent will fully explore all of the actions at the beginning of RL with a high value ε. With the number of iteration rounds increasing and reward converges to be stabilized, the value of ε will decrease, which indicates that the agent will explore random action with lower probability.

In this paper, the action selection strategy is optimized by Metropolis criterion based SA algorithm. The greedy factor ε is related to the exploration rate, and it can be dynamically adapted by SA optimization through the temperature decline rate. Meanwhile, the temperate decline rate will regulate the probability to accept the new solution of Q-values. In each round, the new solutions of Q-value will be generated by RL, and whether to accept new solution is determined by Metropolis criterion under certain temperature. The greedy factor ε of RL is corresponding to the temperate in SA algorithm, and ε is gradually reduced to the lowest temperature until the end of iteration.

The flowchart of SA-RL algorithm is shown as Figure 5, and the specific implementation steps are described as follows.

Step 1: Initialize the solution Q_0, V and temperate T_0. The elements of initial Q-table are set to 0. For each state, V is defined as Equation (19). The initial temperature for SA is set to 1.

Step 2: In t time round, calculate a new solution Q_{new} by RL algorithm.

Step 3: Implement Metropolis criterion to decide whether to accept the new solution under temperature T_t of SA. Define $\Delta E = V - Q_{new}$, which represents the difference between the maximum value V in the current state and the generated new Q-value Q_{new}. If $\Delta E \leq 0$, then the new solution will be accepted. If $\Delta E \geq 0$, the new solution will be accepted with a probability given by $Pr(\Delta E) = \exp(-\Delta E / KT_t)$, where K is Boltzmann constant, and usually can be set to 1.

Step 4: Cooling schedule. The cooling schedule represents the procedure to reduce the temperature as the convergence is reached, which means the agent will explore the random action with lower probability. In this paper, the temperature reduces according to $T_{t+1} = \lambda T_t$, where λ is the cooling rate which decides the decreasing speed of the temperature. The value of cooling rate is less than 1. In SA-RL, the temperature is set to be greedy factor.

Step 5: If the stop criteria is met, stop the algorithm, or else go to Step 2. The termination condition is set to stop the algorithm when the temperature reaches the minimum value or the algorithm reaches the number of iterations. In this paper, the algorithm will be terminated when the number of iterations reaches the scheduled number.

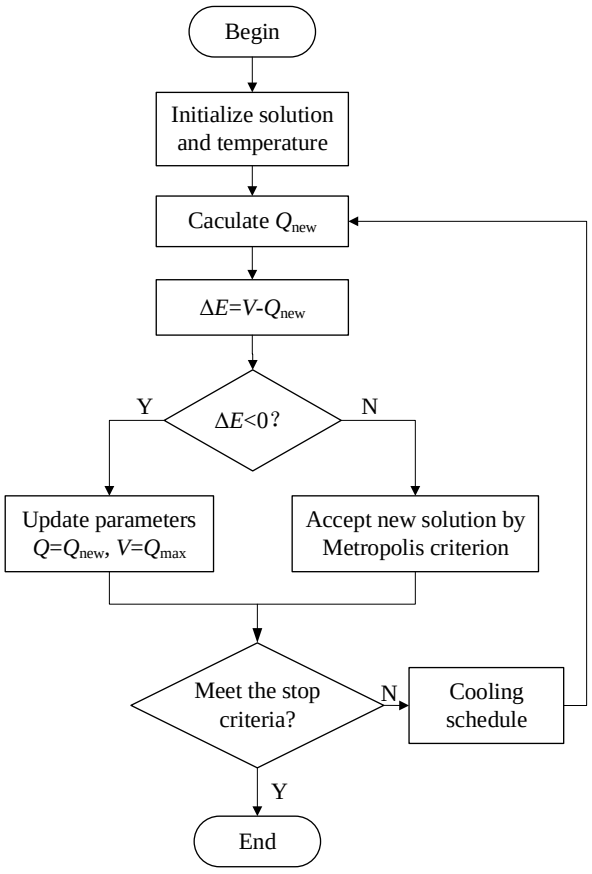

Figure 5. Flowchart of the SA-RL algorithm.

3.3. Fast Reinforcement Learning Scheme

The RL algorithms used in literature [30–32] initialize the Q-table to a full zero matrix during parameter initialization for UWA relay selection. For implementation of the proposed relay selection scheme in practical UWA networks, it is reasonable to pre-train the

SA-RL scheme before experimental implementation. The simulated or channel with similar settings, or the pretest data can be utilized for training. It helps to explore possible actions in certain state.

Herein, a fast reinforcement learning scheme is proposed with the consideration of actual field implementations. It includes two stages: pre-training stage and on-line learning stage. In the parameter pre-training stage, the training data is obtained by simulation in a similar UWA channel scenario. Through the training, a pre-trained Q^* table is obtained. In the on-line learning stage, the Q-table matrix is initialized to Q^*. In each iteration round, the agent will update Q-table by SA-RL algorithm. As the proposed SA-FRL scheme has explored possible actions in the first stage, it can better exploit the exploration results in the second stage for performance improvement.

The pseudocode of the SA-FRL algorithm for UWA cooperative communication networks is shown as Algorithm 1.

Algorithm 1: SA-FRL algorithm.

STAGE1: Pre-training stage
Initialize $\sigma, \zeta, \varepsilon, \lambda, K, s^{(0)}, Q^*(s,a) = 0, V^*(s) = 0$.
for episode=1:Max_episode **do**
  for $t = 1, 2, 3, \ldots, M$, **do**
  Take action $a^{(t)}$ according the latest Q-table.
  The destination node obtain $\gamma_{sd}^{(t)}, \gamma_{rd}^{(t)}, I^{(t)}$.
  Calculate $r^{(t)}$ via(16).
  Calculate $Q_{new}(s^{(t)}, a^{(t)})$ via(18).
  Calculate the difference $V(s^{(t)}) - Q_{new}(s^{(t)}, a^{(t)})$.
  if $V(s^{(t)}) < Q_{new}(s^{(t)}, a^{(t)})$
    $Q(s^{(t)}, a^{(t)}) = Q_{new}(s^{(t)}, a^{(t)})$.
  else
    $p = \exp(-(V(s^{(t)}) - Q_{new}(s^{(t)}, a^{(t)}))/K\varepsilon)$.
    n = rand.
    if $p <$ n
    $Q(s^{(t)}, a^{(t)}) = Q(s^{(t)}, a^{(t)})$.
    else
    $Q(s^{(t)}, a^{(t)}) = Q_{new}(s^{(t)}, a^{(t)})$.
    end if
  end if
  Update $V(s^{(t)})$ via(19).
  $s^{(t+1)} = [\gamma_{sd}^{(t)}, \gamma_{rd}^{(t)}, I^{(t)}]$.
  if $t \geq M$
    go to 28.
  else
    $\varepsilon_{t+1} = \lambda \times \varepsilon_t$.
  end if
  end for
end for
Save Q^*;
STAGE2: On-line RL stage
Initialize $Q(s,a) = Q^*, V(s) = 0$.
for episode=1:Max_episode **do**
  for $t = 1, 2, 3, \ldots, T$, **do**
    Select actions $a^{(t)}$ by ε strategies.
    Repeat 7-27.
  end for
end for

4. Numerical Results

4.1. Simulation Results

4.1.1. Simulation Setup

Simulations have been conducted to evaluate the performance of the proposed relay selection scheme in UWA data collection networks. The simulated UWA network includes

one source node, one destination node and four relay nodes, which are located in an area with 3km diameter.

To generate more realistic CSI for reinforcement learning, the time-varying UWA channels are simulated as Section 2. The channels can be simulated by setting the geography and environmental parameters, such as communication distance and depth, communication frequency band, speed of sound, the variation of distance and height of transmitter/receiver, sea surface/bottom, et al. This simulation method fully considers the actual signal transmission characteristics of physical layer, including large-scale path variation and small-scale scattering [45]. According to this method, the time-varying channel impulsive response will be generated, and the channel gain of the corresponding frequency band can be obtained accordingly. The SNRs can be calculated by Equations (2) and (3), where the sound source level (SL) can be set to a certain value, and the noise power level (NL) can be calculated by Wenz model. To incorporate these parameters with the time-varying channel gain together, the received SNR sequence can be calculated finally.

To mitigate the effect of frequency selective fading, the multi-carrier modulation has been used for physical layer communication scheme. Specifically, OFDM is employed where the number of subcarriers is set to 1024 and the frequency is 14–20 kHz. The settings are matched with AquaSeNT OFDM modem, which are accordance with the experimental setup which we will further used for evaluation.

The parameters of UWA environment for simulation are listed as Table 1. It can be seen that parameters including the sound speed, the sea surface and bottom, the location of transmitter/receiver, the scattering intra-paths are all time varying within defined variation range. The topology in Figure 3 is used as an example for simulation in this paper. The distance between the nodes can be generated randomly.

Table 1. Simulation parameters of UWA environment.

UWA Environment Parameters	Value
Sound speed	1200–1500 m/s
Spreading factor	1.7
Surface height	100 m
Variation of surface height	±1 m
Transmitter/receiver height	25 m
Variation of transmitter/receiver height	±1 m
Channel distance	[3, 2.5, 2.25, 1.75, 1.45] km
Variation of channel distance	±2 m
Number of intra-paths	±1 m
Mean of intra-path amplitudes	0.025
Variance of intra-path amplitudes	0.000001

Figure 6 has shown examples of the CIR with certain communication distance from sensor nodes to destination node D for illustration. It can be seen that the multi-path delay and amplitude keep varying during the simulation. From the sub-figures of Figure 6, it can be seen that the longer communication distance leads to lower amplitude of CIR. The multi-path structures of the five nodes are different. With the generated CSI, the corresponding channel gain can be calculated. According to the generated time-varying CIRs, the range of channel gains of S-D and r_i-D are [−30, −28.5], [−28.8, −27.4], [−28.3, −26.7], [−27, −25.4], [−26, −24.8], respectively.

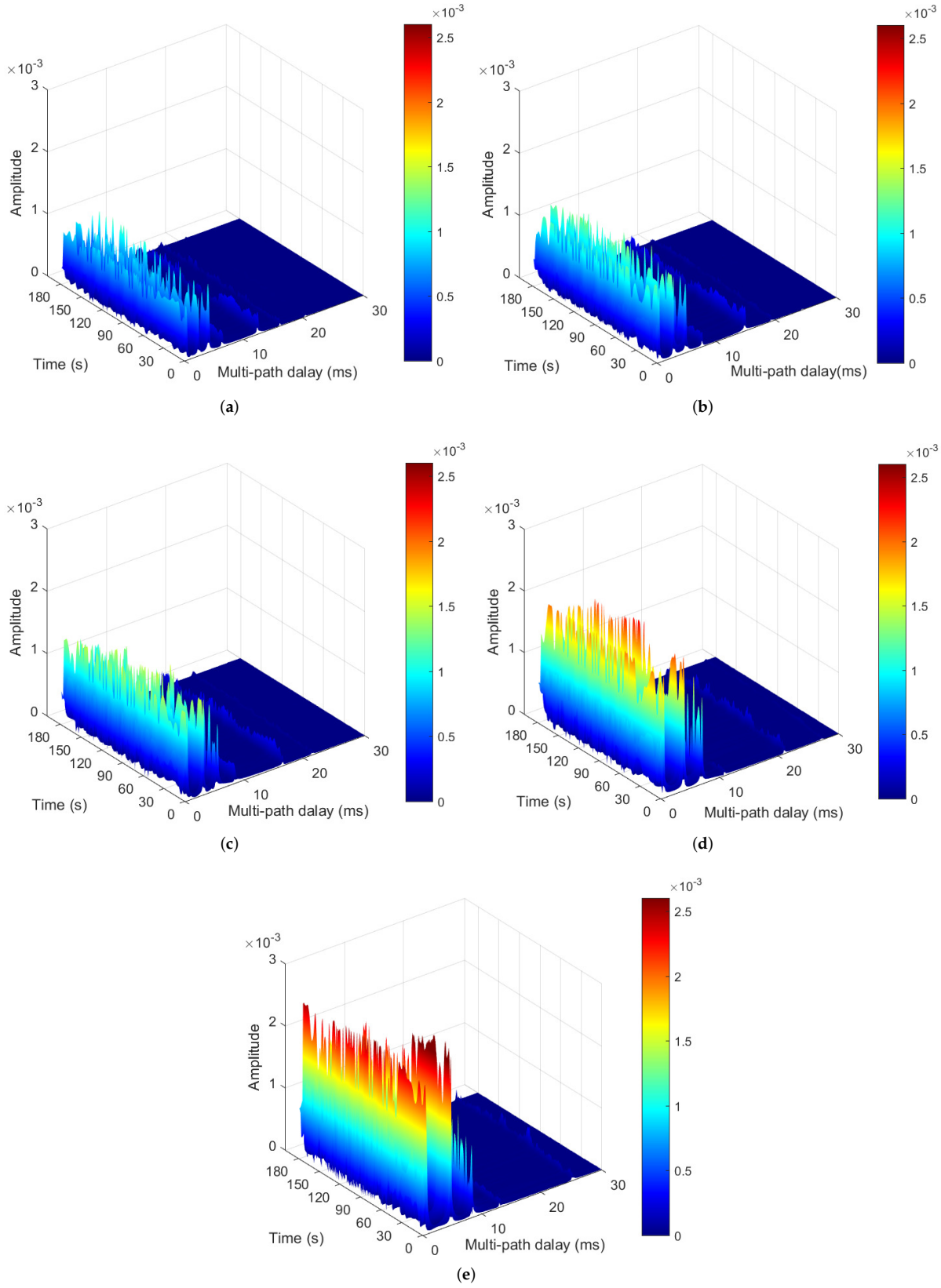

Figure 6. Channel impulse responses from sensors nodes to destination node. (**a**) Received CIR by D sent from S, d = 3 km. (**b**) Received CIR by D sent from r_1, d = 2.5 km. (**c**) Received CIR by D sent from r_2, d = 2.25 km. (**d**) Received CIR by D sent from r_3, d = 1.75 km. (**e**) Received CIR by D sent from r_4, d = 1.45 km.

For reinforcement learning, the CSI is quantized according to the CSI sequence to construct a Q-table. The channel states of *S-D* and r_i-*D* are uniformly quantized to 6 levels, and the corresponding mutual information is also 6 levels. Thus the total number of states is $6^3 = 216$. There are 4 relay nodes in the simulated UWA network, so the size of Q-table is 216 × 4. The propagation delay is calculated by distance and speed of sound in the simulation. In the practical UWA network implementation, the delay can be obtained by the message exchanges of sensor nodes. The simulation parameters for SA-RL algorithm are set as Table 2.

Table 2. Simulation parameters of SA-RL.

Parameters of SA-RL		
Max_episode	Max episode of learning	10
M	Rounds of learning per episode	2000
σ	Learning rate	0.9
ζ	Discount factor	0.9
λ	Cooling rate	0.996
K	Constant	1
$\varepsilon_{\max}$	Initial maximum greedy factor	1
$\varepsilon_{\min}$	minimum greedy factor	[0.05, 0]

4.1.2. Performance of SA-RL Relay Selection Scheme

The performance of the SA-RL UWA relay selection scheme with dynamic ε is simulated and compared with RL scheme with invariant ε strategy. For SA-RL relay selection strategy, the initial value of annealing temperature is set to $\varepsilon_{\max} = 1$, and then approaches to minimum value. Two parameter settings, $\varepsilon_{\min} = 0.05$ and $\varepsilon_{\min} = 0$ have been simulated for evaluation. For invariant ε comparison, the settings are ε = 0.2 and ε = 0.1. The random relay selection is compared as a lower bound.

As shown in Figure 7, the proposed SA-RL relay selection can achieve optimal reward after convergence, as the greedy factor is dynamically adapted. The fluctuation of reward of SA-RL tends to be stabler than invariant ε strategy after convergence, because the greedy factor is very small after convergence, which means the agent explores random action with low probability.

In the first 800 rounds, the reward of SA-RL is lower than invariant ε. As in the early stage of SA-RL, the numerical value of ε is high, which indicates that the agent will explore the relay more randomly with higher probability ε as Equation (19).

With the iteration rounds increases, ε gets lower, and the probability of randomly selection decreases. The reward keeps increasing and finally converges to the optimal value. After convergence, the reward of SA-RL scheme with $\varepsilon_{\min} = 0$ is about 43% higher than invariant ε = 0.1, 60% higher than invariant ε = 0.2, and 88% higher than random selection. The reward of SA-RL with $\varepsilon_{\min} = 0$ is still increased by 21% in comparison with $\varepsilon_{\min} = 0.05$.

Figure 8 has shown the access delay performance of the proposed SA-RL algorithm. The normalized access delay is defined as $T_{Sr_iD} / \min\{T_{Sr_iD}\}$. It can be seen that normalized access delay of SA-RL scheme is close to the minimum and is most stable after convergence. Compared with other invariant ε value strategies, the SA-RL dynamic ε strategy reduces the access delay of the system by 2%, 4%, respectively. SA-RL reduces the access delay by 15% in comparison with random relay selection. From Figures 7 and 8, it is approved that the proposed SA-RL has the highest reward and lowest access delay compared to classic instant ε strategy. It should be noticed that, for classical RL with invariant ε, if ε is very small at the beginning of training, it is easily to be trapped in local maximum.

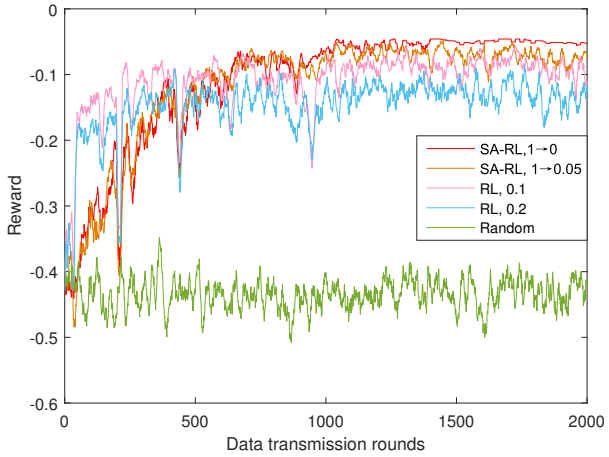

Figure 7. Comparison of rewards of RL with/without SA strategy.

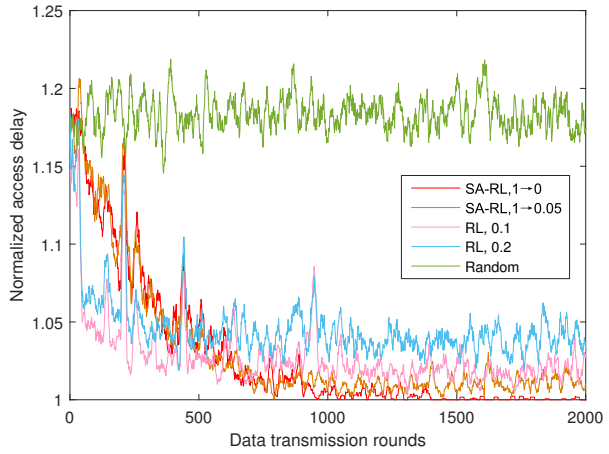

Figure 8. Comparison of access delay of RL with/without SA strategy.

4.1.3. Performance of SA-FRL Relay Selection

The following cases are considered for comparative analysis.

- **Ideal optimum:** as an upper bound, the ideal optimum supposed CSIs are perfectly known, and the relay with maximum reward will be selected accordingly.
- **SA-FRL, 1 → 0:** simulated annealing optimized fast reinforcement learning relay selection with dynamic ε from 1 to 0.
- **SA-RL, 1 → 0:** simulated annealing optimized reinforcement learning relay selection with dynamic ε from 1 to 0.
- **SA-RL, 1 → 0.05:** simulated annealing optimized reinforcement learning relay selection with dynamic ε from 1 to 0.05.
- **Random:** the relay is selected randomly from the available relay set in each time round.

As shown in Figure 9, the rewards of proposed scheme and other schemes have been compared. It can be obviously seen that the proposed SA-FRL and SA-RL schemes have much better reward than random relay selection scheme. After convergence, the rewards of proposed schemes with different parameters are all over 80% better than random relay selection.

Among the SA based learning schemes, SA-FRL with $\varepsilon_{\min} = 0$ has best convergence performance. For convergence speed comparison, SA-FRL with $\varepsilon_{\min} = 0$ reaches convergence after 500 transmissions rounds, while the SA-RL with $\varepsilon_{\min} = 0$ and SA-RL with $\varepsilon_{\min} = 0.05$ schemes begin to converge to a stable strategy after 800 time rounds. By using the fast learning strategy, the convergence speed has increased by 38% in comparison with

the stratergy without pre-training. For convergence value comparison, With $\varepsilon \to 0$, both SA-FRL and SA-RL will achieve the rewards that are near to ideal optimum solution, and SF-FRL is still a little better than SA-RL.

The simulation results have shown that the proposed SA-FRL scheme has the fastest convergence speed and the best learning reward. Even with outdated feedback, our proposed scheme can achieve optimum reward, owing to the dynamic exploration strategy.

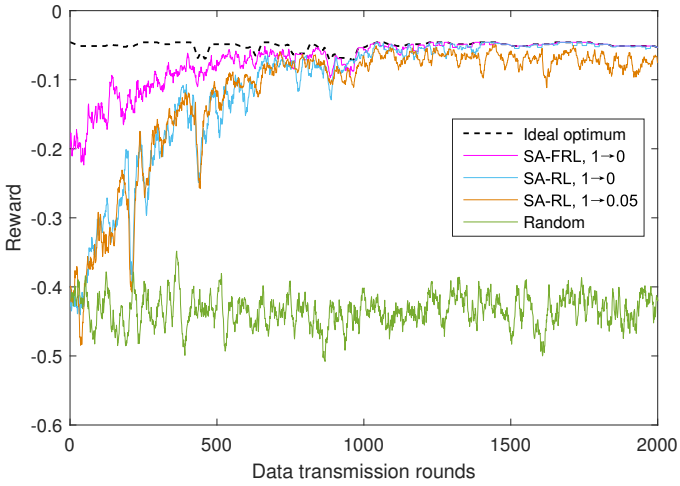

Figure 9. Performance convergence of reward.

Figure 10 shows the convergence of normalized access delay with the iteration rounds of relay selection. Three SA and RL based relay selection schemes can achieve the stable convergence state with the increase of the number of signal transmissions. The normalized access delay of three SA and RL based relay selection schemes are much less than random strategy, and the access delay has been reduced about 15%.

The SA-FRL scheme with $\varepsilon_{\min} = 0$ can achieve the optimal stable state at the fastest speed. With fast learning strategy, SA-FRL scheme with $\varepsilon_{\min} = 0$ begin to converge after about 500 transmissions round, while the SA-RL with $\varepsilon_{\min} = 0$ and SA-RL schemes with $\varepsilon_{\min} = 0.05$ begin to converge to a stable strategy after 800 time rounds. The proposed SA-FRL scheme is pre-trained with similar scenarios, and it selects relay nodes with higher reward and lower access delay in the initial learning process, so it has the fastest learning speed and learning results. When $\varepsilon \to 0$, the minimum access delay can be achieved.

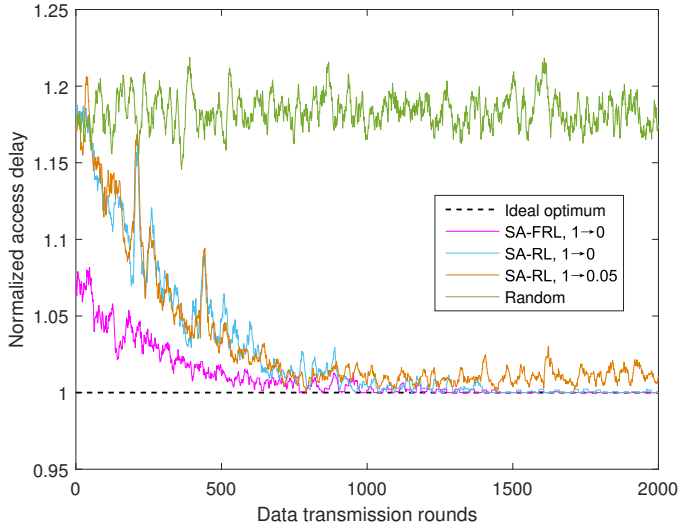

Figure 10. Performance convergence of access delay.

Figure 11 compares the data rates of three schemes, including SA-FRL considering access delay, SA-RL considering access delay, SA-RL without considering access delay (SNR based criteria in [30]), where $\varepsilon_{\min} = 0$. It can be seen from Figure 11 that regardless of the access delay, the data transmission rate is about 2.9 bps/Hz with SNR based criteria. When considering the access delay, the data transmission rate is about 3.45 bps/Hz after convergence, which has 17% improvement. SA-FRL also has the best data rate and convergence speed. In summary, the SA-FRL UWA relay selection scheme proposed in this paper can select the relay node with high channel link quality while considering the access delay of the relay node, and finally achieves the optimum data rate.

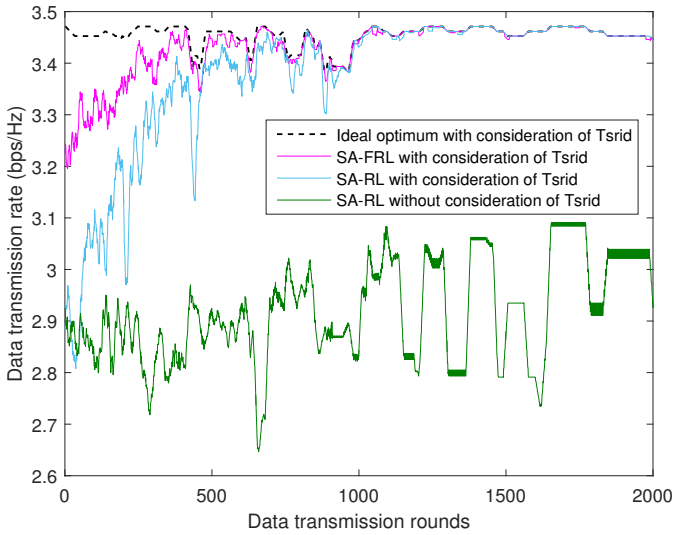

Figure 11. Data transmission rate.

4.2. Experimental Results

To further prove the performance of proposed SA-RL and SA-FRL, our experimental data measured in Mansfield Hollow Lake, Connecticut, USA in October 2014 is used for evaluation [14]. There are four sensor nodes, including one node that works as destination data center, one node works as source node and two nodes work as relay nodes. The source node is 198 m away from the destination node. The relay node 1 and the relay node 2 are 78 m and 77 m away from the destination node, respectively. Relay nodes can overhear the source node, and employ DF mode for cooperative communication. The depth of the lake is about 3–5 m. The AquaSeNT OFDM modem is hanged about 1 m from the water surface. The lake has a muddy bottom full of plants, and a lot of plankton. The sailing of motor boats have brought additional interference during the lake test. It is a complex time-varying UWA multi-path channel, and the CSIs from relay nodes to the destination node are time variant.

With the experimental channel state information, three relay selection schemes have been compared, including of SA-FRL, SA-RL and random relay selection, where $\varepsilon_{\min} = 0$.

Figure 12 has shown the rewards of the three schemes. It can be seen that both SA-FRL and SA-RL schemes can achieve obviously better rewards than random relay selection. The reward of SA-FRL is closer to the ideal optimum than RL after convergence, due to the benefit of per-training. After convergence, the average reward of SA-FRL have improved by 52% and the average reward of SA-RL have improved by 50% in comparion with random relay selection. At the same time, SA-FRL scheme takes fewer rounds to converge to a relatively stable status. Figure 12 has shown that the SA-FRL scheme has faster learning speed and better learning result than the SA-RL and random relay selection scheme.

Figure 13 has shown the performance of the access delay for these three schemes. The SA-FRL scheme will turn to convergence with the least iteration rounds, and is closest to the ideal optimum access delay. Later, the SA-RL scheme begins to converge to be stable

with the continually increase of the number of signal transmissions, and the convergence value is a little bit smaller than SA-FRL.

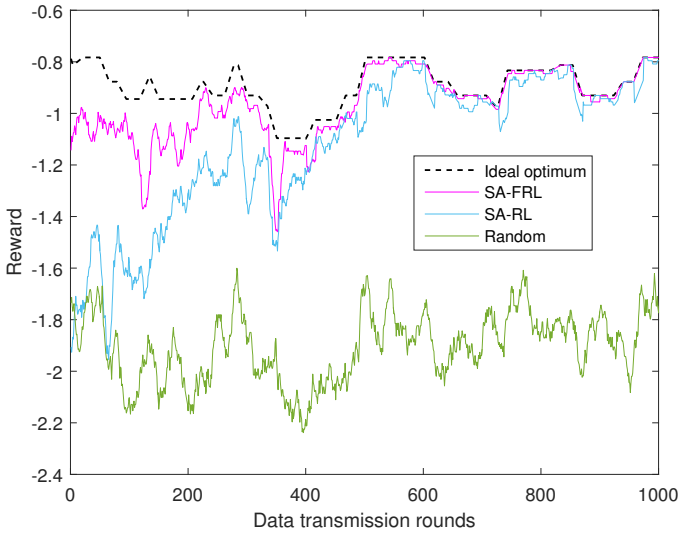

Figure 12. Performance convergence of reward, experimental data.

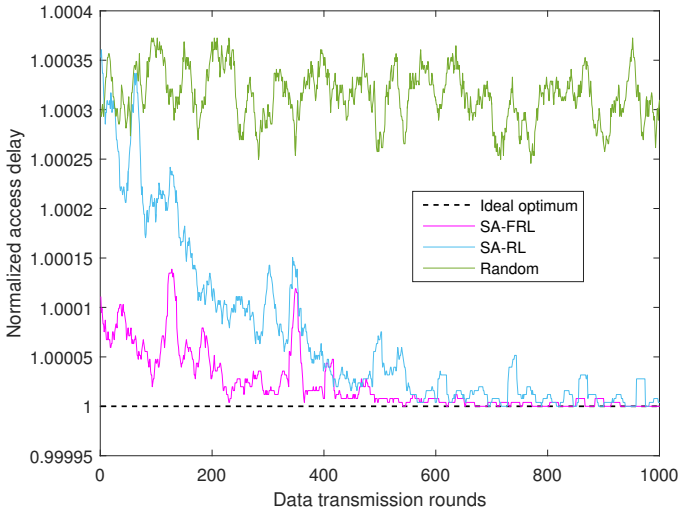

Figure 13. Performance convergence of access delay, experimental data.

Figure 14 has compared the data rate of three schemes including, SA-FRL considering access delay, SA-RL considering access delay and traditional RL without considering access delay (SNR based criteria). In the lake experiment, as the distances between nodes are relatively close, the distinction of access delay is not obvious. Herein, the data rate of SA-FRL and SA-RL have increased about 1%. Although the improvement is not great, the performance curves already can converge to be close to ideal optimal value. This observation is consistent with that from simulation results. In conclusion, the effectiveness of the proposed scheme has been approved by lake experimental data.

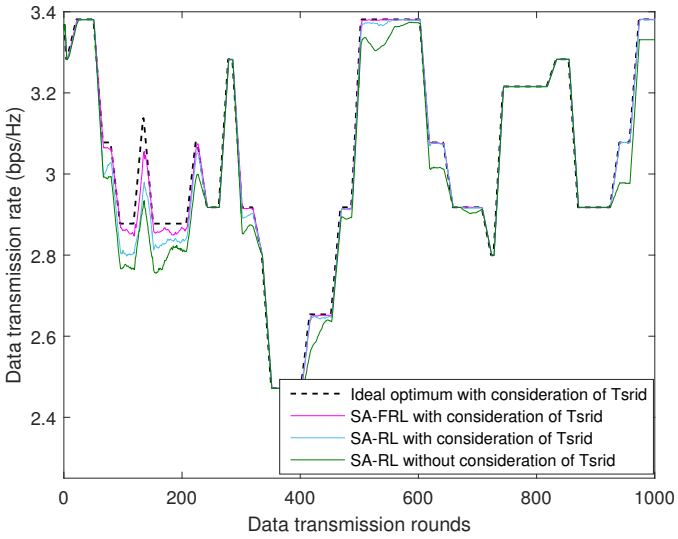

Figure 14. Data transmission rate, experimental data.

5. Discussion

The proposed reinforcement learning UWA cooperative scheme comprehensively considers the physical layer channel characteristics to design the expression of RL for improvement of network throughput. At the same time, the integration of simulated annealing algorithm with reinforcement learning in the relay selection strategy and pre-training of RL can help to accelerate the learning process and improve system performance. The results and findings of this paper and the implications are discussed as follows.

For the design of state, action and reward for RL based relay selection, the long propagation delay characteristic is considered in this paper. The joint reward expression is established with system mutual information, access delay and energy consumption. In research area of reinforcement learning based UWA relay selection, reference [30] firstly considered mutual information as reward of RL. Furthermore, with the consideration of energy consumption, reference [31] set channel capacity and energy consumption as reward, and reference [32] set channel gain and battery level as state, and channel capacity as reward. This paper additionally considers the long access delay characteristic of UWA channel for RL system design. The transmission rate has been improved compared to reference [30] as shown in Figures 11 and 14. As our proposed scheme jointly considers channel quality and access delay, the node with higher channel quality and lower access delay will be selected as relay, which will finally lead to optimal effective throughput. The consideration of long propagation delay also can be employed in related UWA network applications.

With the consideration of dynamic exploration of RL system, simulated annealing algorithm is integrated to the relay selection strategy to better balance exploration and exploitation of RL. In the early stage, the SA-RL relay selection will fully explore the random actions. As shown in Figure 7, the reward is fluctuating and has small values during the first 800 rounds. With the number of learning iteration increasing, the temperate of SA is getting lower, so that the RL will explore random action with lower probability. Specifically, when the minimum ε is set to 0, the exploration rate will approach to 0 after the training turns to convergence. The convergence results will approach to ideal optimum as shown in the figures in Sections 4.1.3 and 4.2. In comparison with constant ε-greedy strategy in [30–32], their curves fluctuate more severely after convergence than our proposed dynamic exploration strategy. When constant ε is a small value, the constant ε-greedy strategy is easily trapped into local optimum in complex environments. As SA-RL can explore with higher temperature, it has the advantage of avoiding to be trapped in local optimum. The integration of simulated annealing algorithm with reinforcement learning in

the relay selection strategy can be adopted in various applications related to the underwater acoustic communication and network applications.

As UWA environment is complex and there is no standard channel model for algorithm evaluation, experiment is a persuasive evaluation method. Reinforcement learning has the ability to learn from the interaction of environment and agent without any prior knowledge, so the Q-table can be initialized with all zeros, such as in [24–32]. Considering practical implementation, this paper has shown how the pre-training helps to improve the system performance. From the simulation results in Figures 9 and 10, it can be observed that the pre-training can accelerate the process of convergence, finally improve the long-term system reward. The experimental results in Figures 12 and 13 also show the same results.

Then we will discuss how much per-trained computational burden is held for performance improvement. The pre-training data set and test data set are 1:1 in Section 4. A number of simulations with different number of training rounds are conducted for comparison. The average values during the whole learning process are calculated. As shown in Tables 3 and 4, the reward and access delay of SA-FRL relay selection with different number of training rounds are compared, respectively. It can be seen that with the number of training rounds increasing, the reward is higher. After some rounds of training, the long-term reward is close to an extremum value. The reward is approximately the extremum value at about 1200, where the reward of SA-RL is converged to a stable value as shown in Figure 9. It indicates that the amount of per-training data do not have to be as large as test data, the computational burden is not heavy. Furthermore, even with limited number of training rounds, e.g., 400, the pre-training still can provide obvious improvement compared to RL without pre-training. Although the number of experimental data is limited, it still can be observed from Tables 5 and 6, that the value is accessing to extremum value with more training rounds. Even with small number of training, the reward still can be improved. So we can conclude that the pre-training is effective for reinforcement learning and necessary before practical system implementation.

Table 3. Reward of SA-FRL with different amount of pre-training data, simulated channel.

Number of Training Rounds	Reward of SA-FRL Relay Selection	Improvement (%) Compared to SA-RL Relay Selection	Improvement (%) Compared to Random Relay Selection
400	−0.088	33%	80%
800	−0.085	35%	81%
1200	−0.078	40%	82%
1600	−0.078	40%	82%
2000	−0.078	40%	82%

Table 4. Normalized access delay of SA-FRL with different amount of pre-training data, simulated channel.

Number of Training Rounds	Normalized Access Delay of SA-FRL Relay Selection	Improvement (%) Compared to SA-RL Relay Selection	Improvement (%) Compared to Random Relay Selection
400	1.017	1.9%	14.1%
800	1.016	2.1%	14.2%
1200	1.014	2.3%	14.4%
1600	1.013	2.4%	14.5%
2000	1.013	2.4%	14.5%

Table 5. Reward of SA-FRL with different amount of pre-training data, experimental channel.

Number of Training Rounds	Reward of SA-FRL Relay Selection	Improvement (%) Compared to SA-RL Relay Selection	Improvement (%) Compared to Random Relay Selection
200	−1.019	9.4%	46.5%
400	−0.996	11.6%	47.7%
600	−0.987	12.2%	48.1%
800	−0.977	13.2%	48.7%
1000	−0.972	13.7%	49%

Table 6. Normalized access delay of SA-FRL with different amount of pre-training data, experimental channel.

Number of Training Rounds	Normalized Access Delay of SA-FRL Relay Selection	Improvement (%) Compared to SA-RL Relay Selection	Improvement (%) Compared to Random Relay Selection
200	1.000038	0.0032%	0.027%
400	1.000031	0.0039%	0.027%
600	1.000029	0.0042%	0.028%
800	1.000025	0.0046%	0.028%
1000	1.000024	0.0047%	0.028%

6. Conclusions

This paper has proposed a reinforcement learning based relay selection scheme for UWA cooperative communication, where the unique characteristics of UWA channel have been considered for the design of state and reward. Moreover, the action exploration strategy is improved by integrating the simulated annealing algorithm with dynamic greedy factor to speed up the learning process and convergence, and the fast reinforcement learning scheme has been proposed for practical implementation. Both simulation and experimental data have been used for system evaluation. Numerical results have revealed that the proposed scheme can select the best cooperative relay with good channel quality and low access delay for optimum data rate in comparison to the scheme without the consideration of access delay. The SA exploration strategy can effectively improve the convergence performance. The SA-RL and SA-FRL can improve the system reward and reduce access delay in comparison with basic RL scheme. With pre-trained Q-table, SA-FRL scheme has the fastest learning speed and the best learning result compared with other schemes. The experimental results have shown the effectiveness of proposed scheme in the actual UWA network.

Author Contributions: Conceptualization, Y.Z. and Y.L.; data curation, Y.Z.; funding acquisition, Y.Z., A.W., B.W., Y.L. and W.B.; investigation, Y.Z. and Y.S.; methodology, Y.Z. and X.S.; resources, B.W. and W.B.; software, Y.S.; supervision, Y.Z. and X.S.; validation, Y.S.; writing—original draft preparation, Y.S.; writing—review and editing, Y.Z. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China under grant No. 61801372, U19B2015, 62001360, 61801371.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://millitsa.coe.neu.edu/projects.html, accessed on 30 November 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UWA	Underwater acoustic
RL	Reinforcement learning
SA	Simulated annealing
FRL	Fast reinforcement learning
OFDM	Orthogonal frequency division multiplexing
DCC	Dynamic coding cooperative
AF	Amplify-and-forward
CSI	Channel state information
BER	Bit error rate
AI	Artificial intelligence
QL-RSA	QL-based relay selection algorithm
IoUT	Internet of underwater things
SNR	Signal to noise ratio
QoS	Quality of service
TDMA	Time division multiple access
DF	Decode-and-forward
CIR	channel impulsive response
MDP	Markov decision process
SL	Source power level
NL	Noise power level

References

1. Williamson, B.; Blondel, P.; Armstrong, E.; Bell, P.; Hall, C.; Waggitt, J.; Scott, B. A Self-Contained Subsea Platform for Acoustic Monitoring of the Environment Around Marine Renewable Energy Devices–Field Deployments at Wave and Tidal Energy Sites in Orkney, Scotland. *IEEE J. Ocean. Eng.* **2016**, *41*, 67–81.
2. Baron, V.; Finez, A.; Bouley, S.; Fayet, F.; Mars, J.; Nicolas, B. Hydrophone Array Optimization, Conception, and Validation for Localization of Acoustic Sources in Deep-Sea Mining. *IEEE J. Ocean. Eng.* **2021**, *46*, 555–563. [CrossRef]
3. Hansen, L.; Pedersen, S.; Durdevic, P. Multi-Phase Flow Metering in Offshore Oil and Gas Transportation Pipelines: Trends and Perspectives. *Sensors* **2019**, *19*, 2184. [CrossRef]
4. Stojanovic, M.; Preisig, J. Underwater acoustic communication channels: Propagation models and statistical characterization. *IEEE Commun. Mag.* **2009**, *47*, 84–89. [CrossRef]
5. Van Walree, P. Propagation and Scattering Effects in Underwater Acoustic Communication Channels. *IEEE J. Ocean. Eng.* **2013**, *38*, 614–631. [CrossRef]
6. Xu, B.; Wang, X.; Guo, Y.; Zhang, J.; Razzaqi, A. A Novel Adaptive Filter for Cooperative Localization under Time-Varying Delay and Non-Gaussian Noise. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–15. [CrossRef]
7. Tu, X.; Xu, X.; Song, A. Frequency-Domain Decision Feedback Equalization for Single-Carrier Transmissions in Fast Time-Varying Underwater Acoustic Channels. *IEEE J. Ocean. Eng.* **2021**, *46*, 704–716. [CrossRef]
8. Zhang, Y.; Venkatesan, R.; Dobre, O.; Li, C. Efficient Estimation and Prediction for Sparse Time-Varying Underwater Acoustic Channels. *IEEE J. Ocean. Eng.* **2020**, *45*, 1112–1125. [CrossRef]
9. Sozer, E.M.; Stojanovic, M.; Proakis, J. Underwater acoustic networks. *IEEE J. Ocean. Eng.* **2000**, *25*, 72–83. [CrossRef]
10. Cui, J.; Kong J.; Gerla M.; Zhou S. The Challenges of Building Scalable Mobile Underwater Wireless Sensor Networks for Aquatic Applications. *IEEE Netw.* **2006**, *20*, 12–18.
11. Heidemann, J.; Ye, W.; Wills, J.; Syed, A.; Li, Y. Research challenges and applications for underwater sensor networking. In Proceedings of the IEEE Wireless Communications and Networking Conference, Las Vegas, NV, USA, 3–6 April 2006; pp. 228–235.
12. Ullah, I.; Gao, M.; Kamal, M.; Khan, Z. A survey on underwater localization, localization techniques and its algorithms. In Proceedings of the 3rd Annual International Conference on Electronics, Electrical Engineering and Information Science, Guangdong, China, 8–10 September 2017; pp. 8–10.
13. Cario, G.; Casavola, A.; Gagliardi, G.; Lupia, M.; Severino, U.; Bruno, F. Analysis of error sources in underwater localization systems. In Proceedings of the OCEANS, Marseille, France, 17–20 June 2019; pp. 1–6.
14. Zhang, Y.; Huang, Y.; Wan, L.; Zhou, S.; Shen, X.; Wang, H. Adaptive OFDMA with Partial CSI for Downlink Underwater Acoustic Communications. *J. Commun. Netw.* **2016**, *3*, 387–396. [CrossRef]
15. Yu, W.; Chen, Y.; Wan, L.; Zhang, X.; Zhu P.; Xu, X. An Energy Optimization Clustering Scheme for Multi-Hop Underwater Acoustic Cooperative Sensor Networks. *IEEE Access* **2020**, *8*, 89171–89184. [CrossRef]
16. Villa, J.; Aaltonen, J.; Virta, S.; Koskinen, K.T. A Co-Operative Autonomous Offshore System for Target Detection Using Multi-Sensor Technology. *Remote Sens.* **2020**, *12*, 4106. [CrossRef]

17. Zhang, Y.; Chen, Y.; Zhou, S.; Xu, X.; Shen, X.; Wang, H. Dynamic Node Cooperation in an Underwater Data Collection Network. *IEEE Sens. J.* **2016**, *16*, 4127–4136. [CrossRef]
18. Liao, Z.; Li, D.; Chen, J. A Network Access Mechanism for Multihop Underwater Acoustic Local Area Networks. *IEEE Sens. J.* **2016**, *16*, 3914–3926. [CrossRef]
19. Doosti-Aref, A.; Ebrahimzadeh, A. Adaptive Relay Selection and Power Allocation for OFDM Cooperative Underwater Acoustic Systems. *IEEE Trans. Mob. Comput.* **2017**, *17*, 1–15. [CrossRef]
20. Li, Y.; Zhang, Y.; Zhou, H.; Jiang, T. To Relay or not to Relay: Open Distance and Optimal Deployment for Linear Underwater Acoustic Networks. *IEEE Trans. Commun.* **2018**, *66*, 3797–3808. [CrossRef]
21. Li, X.; Liu, J.; Zhou, H.; Yan, L.; Han, S.; Guan, X. Relay Selection in Underwater Acoustic Cooperative Networks: A Contextual Bandit Approach. *IEEE Commun. Lett.* **2017**, *21*, 382–385. [CrossRef]
22. Zhao, H.; Li, X.; Han, S.; Yan, L.; Yu, J. Adaptive Relay Selection Strategy in Underwater Acoustic Cooperative Networks: A Hierarchical Adversarial Bandit Learning Approach. *IEEE Trans. Mob. Comput.* 2021, *early access*. [CrossRef]
23. Chang, H.; Feng, J.; Duan, C. Reinforcement Learning-Based Data Forwarding in Underwater Wireless Sensor Networks with Passive Mobility. *Sensors* **2019**, *19*, 256. [CrossRef]
24. Su, W.; Tao, J.; Pei, Y.; You, X.; Xiao, L.; Cheng, E. Reinforcement Learning Based Efficient Underwater Image Communication. *IEEE Commun. Lett.* **2021**, *25*, 883–886. [CrossRef]
25. Valerio, V.; Presti, F.; Petrioli, C.; Picari, L.; Spaccini, D.; Basagni, S. CARMA: Channel-Aware Reinforcement Learning-Based Multi-Path Adaptive Routing for Underwater Wireless Sensor Networks. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2634–2647. [CrossRef]
26. Zhang, Y.; Zhang, Z.; Chen, L.; Wang, X. Reinforcement Learning-Based Opportunistic Routing Protocol for Underwater Acoustic Sensor Networks. *IEEE Trans. Veh. Technol.* **2021**, *70*, 2756–2770. [CrossRef]
27. Lu, Y.; He, R.; Chen, X.; Lin, B.; Yu, C. Energy-Efficient Depth-Based Opportunistic Routing with Q-Learning for Underwater Wireless Sensor Networks. *Sensors* **2020**, *20*, 1025. [CrossRef]
28. Jadoon, M.; Kim, S. Relay selection algorithm for wireless cooperative networks: A learning-based approach. *IEEE Trans. Commun.* **2017**, *11*, 1061–1066. [CrossRef]
29. Su, Y.; Lu, X.; Zhao, Y.; Huang, L.; Du, X. Cooperative Communications With Relay Selection Based on Deep Reinforcement Learning in Wireless Sensor Networks. *IEEE Sens. J.* **2019**, *19*, 9561–9569. [CrossRef]
30. Su, Y.; Wang, M.; Gao, Z.; Huang, L.; Du, X.; Guizani, M. Optimal Cooperative Relaying and Power Control for IoUT Networks With Reinforcement Learning. *IEEE Internet Things J.* **2021**, *8*, 791–801. [CrossRef]
31. Han, S.; Li, L.; Li, X. Deep Q-Network-Based Cooperative Transmission Joint Strategy Optimization Algorithm for Energy Harvesting-Powered Underwater Acoustic Sensor Networks. *Sensors* **2020**, *20*, 6519. [CrossRef] [PubMed]
32. Wang, R.; Yadav, A.; Makled, E.; Dobre, O.; Zhao, R.; Varshney, P. Optimal Power Allocation for Full-Duplex Underwater Relay Networks With Energy Harvesting: A Reinforcement Learning Approach. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 223–227. [CrossRef]
33. Gendreau, M.; Potvin, J. *Handbook of Metaheuristics*, 2nd ed.; Springer Publishing Company: New York, NY, USA, 2010.
34. Bandyopadhyay, S.; Saha, S.; Maulik, U.; Deb, K. A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA. *IEEE Trans. Evol. Comput.* **2008**, *12*, 269–283. [CrossRef]
35. Nguyen, L.T.; Kim, J.; Shim, B. Gradual Federated Learning with Simulated Annealing. *IEEE Trans. Signal Process.* **2021**, *69*, 6299–6313. [CrossRef]
36. Lopez, A.; Heisterkamp, D. Simulated Annealing Based Hierarchical Q-Routing: A Dynamic Routing Protocol. In Proceedings of the Eighth International Conference on Information Technology: New Generations, Las Vegas, NV, USA, 11–13 April 2011; pp. 791–796.
37. Rovira-Sugranes, A.; Afghah, F.; Qu, J.; Razi, A. Fully-Echoed Q-Routing With Simulated Annealing Inference for Flying Adhoc Networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 2223–2234. [CrossRef]
38. Anjangi, P.; Chitre, M. Propagation-Delay-Aware Unslotted Schedules With Variable Packet Duration for Underwater Acoustic Networks. *IEEE J. Ocean. Eng.* **2017**, *42*, 977–993. [CrossRef]
39. Noh, Y.; Lee, U.; Han, S.; Wang, P.; Torres, D.; Kim, J.; Gerla, M. DOTS: A Propagation Delay-Aware Opportunistic MAC Protocol for Mobile Underwater Networks. *IEEE Trans. Mob. Comput.* **2014**, *13*, 766–782. [CrossRef]
40. Urick, R.J. *Sound Propagation in the Sea*, 3rd ed.; Peninsula Publishing: Los Altos, CA, USA, 1982.
41. Bletsas, A.; Khisti, A.; Reed, D.; Lippman, A. A simple Cooperative diversity method based on network path selection. *IEEE J. Sel. Areas Commun.* **2006**, *24*, 659–672. [CrossRef]
42. Bansal, A.; Bhatnagar, M.; Hjorungnes, A.; Han, Z. Low-Complexity Decoding in DF MIMO Relaying System. *IEEE Trans. Veh. Technol.* **2013**, *62*, 1123–1137. [CrossRef]
43. Li, Y.; Wang, W.; Kong, J.; Peng, M. Subcarrier pairing for amplify-and-forward and decode-and-forward OFDM relay links. *IEEE Commun. Lett.* **2009**, *13*, 209–211. [CrossRef]
44. Chen, Y.; Wang, Z.; Lei, W.; Hao, Z.; Xu, X. OFDM-Modulated Dynamic Coded Cooperation in Underwater Acoustic Channels. *IEEE J. Ocean. Eng.* **2015**, *40*, 159–168. [CrossRef]
45. Qarabaqi, P.; Stojanovic, M. Statistical Characterization and Computationally Efficient Modeling of a Class of Underwater Acoustic Communication Channels. *IEEE J. Ocean. Eng.* **2013**, *38*, 701–717. [CrossRef]

46. Porter, M.; Bucker, H. Gaussian beam tracing for computing ocean acoustic fields. *J. Acoust. Soc. Am.* **1987**, *82*, 1349–1359. [CrossRef]
47. Peng, Z.; Zhou, Z.; Cui, J.; Shi, Z. Aqua-Net: An underwater sensor network architecture: Design, implementation, and initial testing. In Proceedings of the IEEE OCEANS, Biloxi, MS, USA, 26–29 October 2009; pp. 1–8.