electronics

MDPI

Article

Marine Mammal Conflict Avoidance Method Design and Spectrum Allocation Strategy

Han Wang, Jiawei Liu, Bingqi Liu and Yihu Xu *

Department of Electronic & Communication Engineering, Yanbian University, Yanji 133002, China
* Correspondence: xuyh@ybu.edu.cn

Abstract: Underwater wireless sensor networks play an important role in underwater communication systems. Communication through collaborative communication is an effective way to solve critical problems in underwater communication systems. Underwater sensors are often deployed in spaces that overlap with those of marine mammals, which can adversely affect them. For this reason, in this paper, a marine mammal conflict avoidance method that can be dynamically adjusted according to the channel idle time duration and sensor node demand is designed, and the derivation of the maximum occupancy time duration is performed. Meanwhile, in addition, combining the potential of reinforcement learning in adaptive management, efficient resource optimization, and solving complex problems, this study also proposes a reinforcement learning-based relay-assisted spectrum switching method (R2S), which aims to achieve a reasonable allocation of spectrum resources in relay collaborative communication systems. The experimental results show that the method proposed in this study can effectively reduce the disturbance to marine mammals while performing well in terms of conflict probability, interruption probability, and quality of service.

Keywords: underwater sensor networks; bio-friendly; spectrum switching; reinforcement learning

check for updates

Citation: Wang, H.; Liu, J.; Liu, B.; Xu, Y.. Marine Mammal Conflict Avoidance Method Design and Spectrum Allocation Strategy. *Electronics* **2024**, *13*, 1994. https://doi.org/10.3390/electronics13101994

Academic Editor: Paul Mitchell

Received: 15 April 2024
Revised: 12 May 2024
Accepted: 17 May 2024
Published: 20 May 2024

1. Introduction

With the increasing development of marine resources, the development of the marine economy is a strong impetus to social progress [1]. In the process of exploring the oceans, underwater wireless communication plays an important role and can be applied to remote control of the offshore oil industry, pollution monitoring of environmental systems, collection of scientific data from submarine stations, disaster detection and early warning, national security, intrusion detection and defense, and the discovery of new resources in a variety of fields [2]. Underwater wireless sensor networks are capable of penetrating underwater environments that are difficult for humans to access, collecting and transmitting data to guide the development and utilization of marine resources. In view of the special characteristics of the underwater environment, the sensor nodes mainly communicate with the base station or buoy through acoustic signals, and the underwater sensor network composed of such sensor nodes is called a hydroacoustic sensor network [3]. An in-depth study of UWSNs will not only help to improve the reliability and accuracy of data in underwater communications but also help to enhance the ability to explore the underwater environment.

Underwater communication systems on which underwater wireless sensor networks rely differ from terrestrial radio systems in that they are subject to performance constraints that include, but are not limited to, lower propagation speeds, narrow available bandwidths, significant multipath delays, and limitations on transmission range and energy [4]. Solving the difficult problems in underwater communication systems by means of collaborative communication has gradually become the focus of research in various countries.

The communication methodology of collaborative communication can effectively improve the performance and reliability of the communication system in the face of different working environments, which gives it a wide range of application scenarios. The application of collaborative communication in free-space optical (FSO) communication systems has been systematically introduced in the literature [5], and in recent years, the research in this field has become more and more comprehensive and sophisticated. The literature [6] has investigated the space–air–ground integrated network (SAGIN) in free-space optical communication systems. In the study, shadow Rice fading is used to characterize the shadowing effect of RF signals from satellite–UAV links, and the effect of atmospheric turbulence on the optical signals of UAV–ground user links is modeled using the Málaga distribution fading model. The performance of the relay system is also improved with the help of outlier detection techniques and intensity modulation and direct detection techniques. The literature [7] has investigated the performance of a UAV-assisted relay system in a SAGIN for vehicular networking based on an AF-based variable gain relay protocol and establishes and investigates a UAV-assisted RF/free-space optical communication system based on variable gain forwarding protocol. It assumes that the RF link follows κ-μ fading and the FSO link undergoes Málaga fading to characterize atmospheric turbulence under weak-to-strong conditions, which provides an important reference for further improving the performance of the SAGIN system. In underwater relay collaborative communication systems, the characterization of the underwater channel is a key consideration for researchers. It includes transmission loss in the underwater channel due to a combination of absorption and propagation loss of the acoustic signal in the underwater environment, communication delay due to the lower propagation speed of the acoustic signal, and multiple sources of noise such as turbulence, shipping, waves, and thermal noise. This tests the reliability of underwater communication systems and raises the performance requirements for collaborative underwater communication systems [8].

In most application scenarios, UWSNs will fully occupy a specific band resource, but there is a natural acoustic system in the underwater environment, dominated by marine mammals, with which the activity of artificial acoustic systems may cause harmful interference. In order to assess such impacts, the United States divided impacts into two levels, A and B, in the Marine Mammal Protection Act (MMPA) [9]. Class A impacts primarily include direct harm to marine mammals, while Class B impacts primarily include disturbance of their behavioral patterns. In fact, the effects of artificial acoustic systems on marine mammals are not limited to A and B. Therefore, as researchers, it is of great significance for the conservation of marine ocean biodiversity and the sustainable utilization of marine resources to study the method of allocating the scarce and unevenly distributed spectral resources in the underwater environment while avoiding harmful disturbance to the marine animals.

Aiming at the problem of relative scarcity of underwater spectrum resources, the dynamic spectrum sharing technology based on vertical sharing in cognitive radio technology has unique advantages, which include two spectrum sharing modes, overlay and underlay [10], and realizes the reasonable allocation of the spectrum through prioritization. Collaborative communication technology in hydroacoustic communication networks can optimize the underwater spectrum allocation strategy, improve spectrum resource utilization, reduce interference, and enhance system reliability and network performance through cooperation and information sharing among nodes. It also facilitates the development of dynamic spectrum access and management schemes that provide greater flexibility and adaptability in spectrum allocation for hydroacoustic communication networks. The literature [11] has proposed a joint power control and spectrum allocation algorithm called PCCA, which can increase the channel capacity and reduce the interruption probability by analyzing multiple user interference scenarios and establishing the relationship between the interruption probability and the cumulative interference of the channel, adjusting the transmitter power in order to reduce the interruption probability and, at the same time, constructing a channel allocation model by using the Lagrangian optimization theory.

Inspired by research on the allocation of non-exclusive spectra, it has become a hot topic to study how to utilize the spectrum in a way that avoids interference with marine mammals. The literature [12] has proposed a cognitive acoustic transmission scheme in multi-hop aquatic acoustic networks, called dolphin-aware data transmission (DAD-Tx), which utilizes statistical laws to achieve overlapping spectrum sharing through power control by means of statistical information about the animal's location. In [13], the researchers proposed a mammal-friendly channel power allocation algorithm for hydroacoustic sensor networks. The interference level of sensor nodes with marine animals was fully considered, and a game model with a unique Nash equilibrium point was established based on the network interference level and node residual energy. Finally, a bio-friendly underwater communication mechanism was designed by improving the network service quality under the premise of achieving a bio-friendly purpose. The literature [14] developed an efficient deep multiuser reinforcement learning algorithm based on deep neural networks, Q-learning, and collaborative multi-agent systems for CR access policy, which is a user selection algorithm that sets the order of agent activation to select the optimal communication frequency over the appropriate bandwidth. The experimental results show that the method can make the channel allocation always appropriate, but the performance of the proposed algorithm is closely related to the algorithm-related parameters and network environment parameters. The literature [15] introduces the concept of reinforcement learning into underwater acoustic communication networks and constructs a multi-agent network by considering nodes as intelligent agents. By dividing the state space and action space, formulating the reward function and search strategy, a distributed resource allocation algorithm based on cooperative Q-Learning is proposed and its convergence is verified. The experimental results prove that the algorithm can effectively enhance the network transmission capacity and reduce the resource allocation overhead.

In this paper, a marine mammal conflict avoidance method is firstly proposed to address the problem of underwater sensor networks interfering with marine mammals during the communication process, and the bio-friendly goal is realized by designing the stopping occupancy rule and the maximum occupancy duration and predictively deoccupying the active spectrum for marine animals. Second, the design of conflict avoidance actions based on spectrum switching is carried out, and a reinforcement learning-based relay-assisted spectrum switching algorithm (R2S), which includes three steps, biologically friendly shared spectrum matching, reinforcement learning decision making, and dynamic spectrum access, is proposed. The proposed method realize the reasonable allocation of spectrum resources in the relay collaborative communication system, the number of switches and the interruption probability of the system is reduced, the quality of service of the sensor nodes is ensured, and harmonious coexistence between the underwater sensor network and the marine mammals is realized.

2. Design of Marine Mammal Conflict Avoidance Method

2.1. System Model

As shown, the design of this paper is based on a multi-hop cognitive hydroacoustic network model, as shown in Figure 1, with a buoy or ship as the control center on the sea surface; multiple sensor nodes arranged underwater, each of which is capable of collaborative communication; and the presence of multiple types of marine mammals in the sea.

Marine mammal activities often have no specific mathematical model compared with real-time spectrum sensing to control spectrum switching, statistical information-based control and prediction through probabilistic models can reduce the computational overhead of the control center and reduce the energy consumption caused by the exchange of information between nodes, which is very suitable for the design of marine mammal retreat methods.

The Markov ON–OFF channel model shown in Figure 2 is a mathematical model used to describe the stochastic behavior of a communication system, which is often used to describe the switching of the system between the "ON" and "OFF" states, and it assumes that the state transfer of the system is based on a stochastic process.

Where a is the probability that the ON state is transferred to the OFF state, 1-a is the probability of keeping the ON state, b is the probability that the OFF state is transferred to the ON state, and 1-b is the probability of keeping the OFF state. The use of the channel in continuous time is shown in Figure 3, which treats the user's use of the channel as an alternating busy–idle switching process.

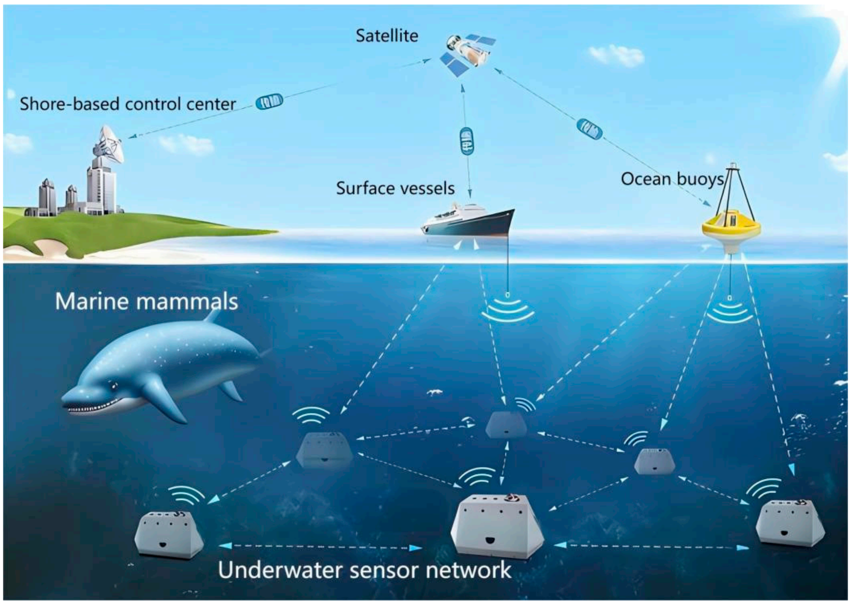


Figure 1. Cognitive underwater acoustic network.

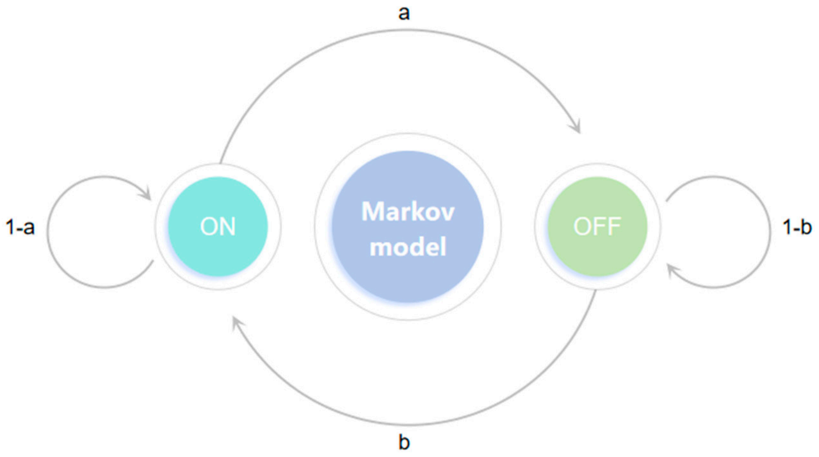


Figure 2. Markov ON–OFF model.

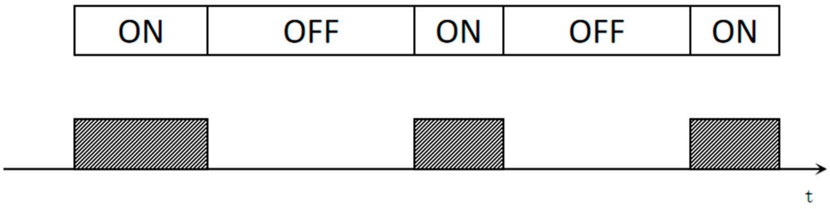


Figure 3. Channel occupancy state.

When the underwater sensor node generating the communication demand is ON with the marine mammal at the same time, the communication behavior of the sensor node will cause interference to the marine mammal, and when the node is ON and the marine mammal is OFF, it will not cause interference. In the ON–OFF model, the ON and OFF phases obey an exponential distribution with parameters λ_{ON} and λ_{OFF}, respectively, and their durations, t, can be represented by the probability density functions shown in Equations (1) and (2), respectively.

$$f(t_{ON}) = \begin{cases} \lambda_{\mathrm{ON}} e^{-\lambda_{ON} t_{ON}} & t_{\mathrm{ON}} \geq 0 \\ 0 & t_{\mathrm{ON}} < 0 \end{cases} \tag{1}$$

$$f(t_{OFF}) = \begin{cases} \lambda_{OFF} e^{-\lambda_{OFF} t_{OFF}} & t_{OFF} \geq 0 \\ 0 & t_{OFF} < 0 \end{cases} \tag{2}$$

There is a close relationship between marine mammal activity and the exponential distribution, which, as a continuous probability distribution, is commonly used to describe the time intervals at which independent random events occur. Marine mammals, such as whales and dolphins, on the other hand, tend to show a certain pattern of time intervals in their activity patterns, such as feeding, migrating, and resting, i.e., multiple activities occurring over a short period of time followed by a longer period of rest. When feeding, marine mammals may strike frequently for a certain period of time and then rest for a period of time before resuming, a pattern of activity intervals that fits with the characteristics of an exponential distribution. Similarly, there are differences in the time intervals between their migrations for a variety of reasons, and such differences fit the description of an exponential distribution. In addition, the exponential distribution is also suitable for describing the social behavior of marine mammals. So, in this paper, this channel model is used for the study.

The communication behavior of the sensor nodes may interfere with marine mammals. However, when the node is in the ON state and the marine mammal is in the OFF state, it will not cause interference. T_h is defined as the time that marine mammals need to be in the OFF state for the required spectrum to be used, and T_r is the time that the node is in the ON state. The value of T_h can be obtained through analysis and prediction methods, and T_r can be obtained by Equation (3).

$$T_r \approx \frac{L}{Blog(1 + SINR)} \tag{3}$$

B represents the bandwidth of the frequency spectrum used, $SINR$ represents the signal-to-interference-plus-noise ratio during communication of the underwater sensor node, and L represents the total amount of data for each communication of each underwater sensor node. When T_h is less than T_r, communication of the underwater sensor node i in the frequency spectrum j (assuming that the time coefficient of frequency spectrum j follows an exponential distribution with parameter λ_j) will cause interference to marine mammals, and the sensor node needs to take evasive actions. The avoidance probability can be expressed by Equation (4).

$$\begin{aligned} P_{ij} &= p\left(T_{h,j} < T_{r,i}\right) \\ &= \int_0^{T_{r,j}} \lambda_j e^{-\lambda_j t} dt \\ &= 1 - e^{-\lambda_j T_{r,j}} \end{aligned} \tag{4}$$

The meaning is that the probability of sensor node i exhibiting avoidance behavior on spectrum j is equal to the probability of the OFF state time of spectrum j being less than the ON state time of the sensor node. Next, based on the relationship between T_r and T_h, we will design a method for avoiding conflicts with marine mammals. According to the characteristics of the underwater channel, avoidance actions will be designed. We propose a relay-assisted spectrum switching method based on reinforcement learning (R2S) to achieve low switching frequency, low interruption probability, and high service quality spectrum switching.

2.2. Avoidance Algorithm Design

Methods incorporating spectrum sensing have been widely used in bio-friendly spectrum sharing research. However, this method needs to sense the occupation of the spectrum by marine animals first and then perform spectrum switching or pause the occupation operation, which will lead to back-off delay and spectrum overlapping, forming harmful interference. Predictive decommissioning methods can utilize statistical information to decommission the spectrum in advance, eliminate decommissioning delays, and protect marine mammals in the event of node damage.

Here, we define the probability of communication events occurring for marine mammals at each independent moment: $\delta \in (0,1]$, which can be calculated based on factors such as time, season, species, or location. We define the event of communication occurrence for marine mammals as A, with time denoted as *t*. As time t increases, the probability of event A occurring, $P(A \mid t)$, gradually increases. This can be expressed using a conditional probability distribution as $P(A \mid t_1) < P(A \mid t_2) < P(A \mid t_3) < \ldots < P(A \mid t_\mathrm{n})$, where $t_1 < t_2 < t_3 < \ldots < t_\mathrm{n}$, indicating that as time progresses, the probability of communication events occurring also increases gradually. In continuous time, the probability of marine mammals engaging in communication events will certainly become greater. Therefore, we can define an increasing Function (5) with respect to *t* to represent the influence of time on the occurrence of communication events A for marine mammals.

$$f(t) = 1 - (1 - \delta_0)^{t+1} \tag{5}$$

Its range is $(\delta, 1)$, and when time *t* is sufficient, marine mammals will definitely engage in communication activities on the occupied spectrum. In other words, $\lim_{t \to \infty} f(t) = 1$ always holds true.

When it is necessary to communicate on the frequency spectrum where marine mammals are active, the T_h value of the spectrum must be greater than the T_r value of the node. However, the time difference $\Delta T = T_r - T_h$ does not decrease over time. The node will continue to occupy the spectrum within $\mathrm{T_r}$, so a time discount factor $\gamma = 1 - (1 - \delta_0)^{t+1}$ is designed to avoid prolonged occupation of the marine mammal activity spectrum by underwater sensor networks. T_γ is defined as the discount time where $T_\gamma = T_h - \gamma \cdot \Delta T$, and the difference between T_γ and T_r decreases over time but eventually converges to and is not equal to zero. Therefore, in order to avoid occupying a single spectrum for a long time, as shown in Equation (6), the stop-occupancy rule is designed based on the characteristic that the ratio difference can reflect the relative magnitude between T_h and T_r. Defining Γ_γ as the threshold for the time discount factor when $\gamma = \Gamma_\gamma$, stops occupying the spectrum at that time.

$$\begin{gathered} T_r = \frac{\Delta T}{T_r + T_h} \times T_\gamma \\ T_\gamma = \frac{T_h - T_r}{T_h + T_r} \times [T_h - \gamma(T_h - T_r)] \\ T_r^2 + T_r T_h = (1-\gamma)T_h^2 + (2\gamma - 1)T_r T_h - \gamma T_r^2 \\ \Gamma_\gamma = 1 - \tfrac{2{T_r}^2}{(T_h - T_r)^2} \end{gathered} \tag{6}$$

The design of the above equation allows for dynamic adjustment of the threshold value γ. When there is a significant difference between T_h and T_r (T_h >> T_r), the threshold value is higher, resulting in a longer occupation time. When the difference is relatively small ($T_h \approx T_r$), the threshold value is lower, leading to a corresponding reduction in occupation time. Based on the definition of γ, the derivation of the maximum occupation duration M is conducted. During the request process, T_h and T_r are already determined. For the sake of derivation convenience, let $\frac{2T_r^2}{(T_h - T_r)^2} = b, b \in (0,1)$, As a constant in the derivation process, b can also reflect the relative relationship between T_h and T_r. That is, with the increase of

b, ΔT will shrink accordingly, Γ_γ becoming increasingly smaller and lower, leading to a shorter maximum available time for the spectrum. The derivation of M is shown in (7), which means that $\gamma = \Gamma_\gamma$ for the value of t at the time.

$$\begin{aligned} \Gamma_\gamma &= 1 - b \\ 1 - (1-\delta)^{t+1} &= 1 - b \\ \left(1-\delta\right)^{t+1} &= b \\ t &= log_{1-\delta} b - 1 \end{aligned} \tag{7}$$

Substituting the equations to find the maximum occupancy duration M is shown in Equation (8):

$$M = log_{1-\delta} \frac{2{T_r}^2}{\left({T_h}^2 - {T_r}^2\right)} - 1 \tag{8}$$

When $M > 0$, the sensor is able to occupy the spectrum based on the value of M. In other cases, the spectrum should not be occupied.

The flow of the final proposed evasion algorithm is shown in Figure 4. After the sensor generates a transmission request, it obtains the idle state of the channel based on the results of spectrum sensing and calculates the maximum occupancy duration, and it takes evasive action when it detects the communication activity of a marine mammal or when it reaches the maximum occupancy duration.

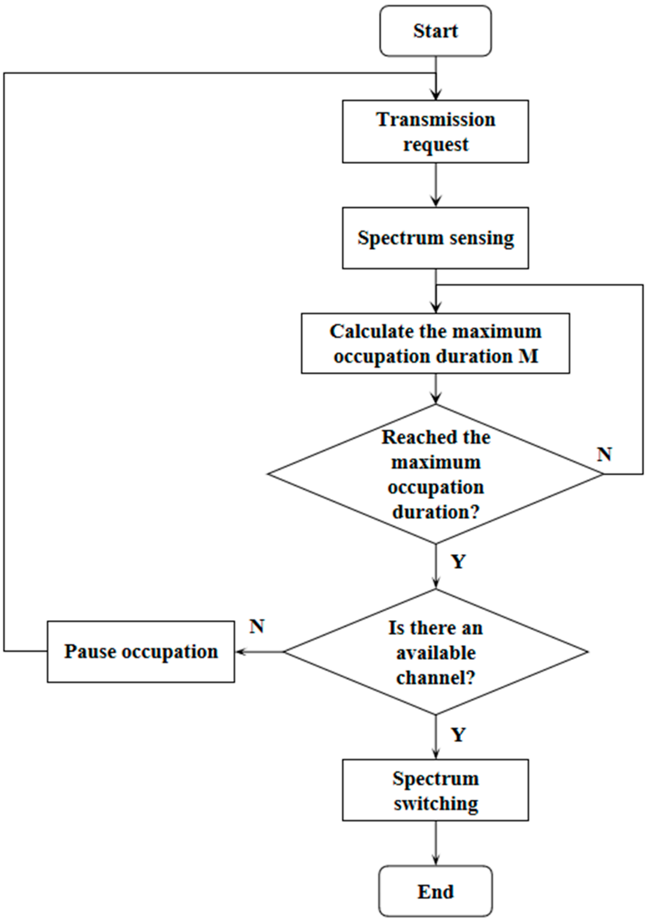

Start
Transmission request
Spectrum sensing
Calculate the maximum occupation duration M
Reached the maximum occupation duration?
N
Y
Is there an available channel?
N
Pause occupation
Y
Spectrum switching
End

Figure 4. Algorithm flow chart.

During the action execution process, the source node will search for other channels that can reach the destination node and perform switching and access operations. If there is no available channel, it will pause the current channel occupation and enter a waiting process. If the waiting time is still within the maximum acceptable delay for each communication, the above process is repeated; if it exceeds the time threshold for each communication, the communication is abandoned directly.

After making the decision to avoid, the sensor nodes need to search for an available spectrum to switch to or temporarily pause usage, which can increase the probability of communication interruption. To address this issue, we have introduced a relay assistance mechanism. By selecting cooperative communication, relay nodes' channels that can reach the vicinity of the destination node are chosen to form single-hop or multi-hop relay links. When there is a direct channel to the destination node, competition with the relay link will be considered, selecting a channel with better service quality and lower switching probability for communication.

3. Design of Relay-Assisted Spectrum Switching Method Based on Reinforcement Learning

The spectrum switching caused by marine mammal activities implies that sensor nodes can only transmit on the spectrum not occupied by marine mammals. As cognitive radio opportunistically utilizes licensed spectra, if only bio-friendly design objectives are considered throughout the communication process, sensor nodes will frequently switch frequencies, significantly impacting their service quality. Therefore, the ultimate goal of this design is to minimize the number of switches, reduce interruption probability, and ensure high service quality when spectrum switching is needed. The reinforcement learning-based relay-assisted spectrum switching method (R2S) proposed in this paper includes three steps: shared spectrum matching, reinforcement learning decision making, and dynamic spectrum access.

3.1. Shared Spectrum Matching Algorithm

In the shared spectrum matching process, a parameter P_i is defined to represent the minimum tolerable non-switching probability for sensor node i during communication, and $1 - P_i$ represents the maximum allowable switching probability for the sensor node during communication. These parameters serve as the basis for shared spectrum matching. During shared spectrum matching, the sensor node selects a spectrum or relay link that is closest to the maximum switching probability. This process can be represented by Equation (9).

$$\begin{aligned} H_i &= \arg\left\{ \underset{j}{min} \left| P_{ij} - (1 - P_i) \right| \right\} \\ &= \left| 1 - e^{-\lambda_j T_{r,j}} - (1 - P_i) \right| \\ &= \left| P_i - e^{-\lambda_j T_{r,j}} \right| \end{aligned} \tag{9}$$

Here, *Hi* represents the spectrum hole *j* chosen by node *i* with the smallest difference in switching probability from its maximum allowable value. This matching method ensures that the spectrum obtained by the sensor node is closest to its desired spectrum characteristics. However, in practical environments, it is possible that the conflict probabilities on all available spectra are greater than the maximum conflict probability or the conflict probability on the spectrum with the smallest difference from the maximum conflict probability is still greater than the maximum conflict probability. In such cases, the switching and access decisions made will affect the QoS of the sensor node. Therefore, we need to adopt the spectrum reservation scheme as shown in Figure 5, where the reserved frequencies with lower switching probabilities are designated as the targets for switching and access.

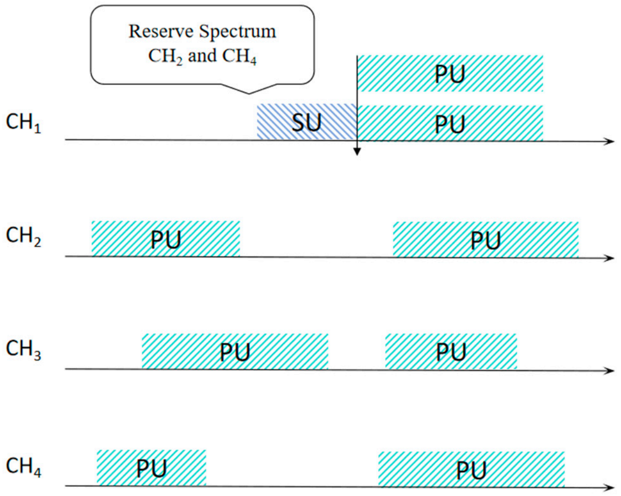


Figure 5. Reserved spectrum switching.

Reserved spectrum switching involves pre-establishing a candidate channel list for spectrum switching before the need arises. When a switch is required, a channel is directly selected from this list for the switch. This approach can avoid wasting time on spectrum sensing in a dynamic underwater environment and effectively maintain communication continuity.

To facilitate the utilization and research of reserved spectrum, the symbol τ_i in the Equation (10) is used to represent the time factor of reserved spectrum holes, which denotes the duration of reserved spectrum holes needed to exactly meet the requirements of the node.

$$\begin{cases} P(T_h > (1-\tau_i)T_{r,i}) > P_i \\ 1 - P(T_h < (1-\tau_i)T_{r,i}) \geq P_i \\ e^{-\lambda(1-\tau_i)T_{r,i}} \geq P_i \\ \tau_i \geq 1 + \dfrac{\ln(P_i)}{\lambda T_{r,i}} \end{cases} \tag{10}$$

The value range of τ_i is between 0 and 1. When τ_i is 0, it indicates that there is no need to access reserved spectrum holes. When τ_i is 1, it means that the sensor node needs to access reserved spectrum holes throughout its communication time $T_{r,i}$. The duration of reserved spectrum holes is closely related to τ_i. The larger the value of τ_i, the better the service quality. This algorithm, which selects suitable spectra for access based on the time factor of reserved spectrum holes after shared spectrum matching, can make resource-efficient decisions with good service quality.

3.2. Reinforcement Learning-Based Approach for Spectrum Decision Making

The shared spectrum matching algorithm described in the previous section lacks the consideration of system time variation in the face of complex and changing underwater environments, which can have an impact on the final communication quality. For this reason, combining the advantages of reinforcement learning in the field of resource allocation with the ability to obtain long-term gains, this section designs a spectrum decision-making method based on the reinforcement learning actor–critic algorithm, which reduces the number of spectrum switches while obtaining a higher quality of service.

The actor–critic algorithm, which is used in this paper, is based on the traditional reinforcement learning framework, which divides the intelligent agent into two parts, actor and critic, where the actor is responsible for selecting the action and the critic is responsible for evaluating the quality of the decision and providing the feedback signal. In this way, the actor–critic algorithm is able to achieve more refined decision making while ensuring learning efficiency. The principle of the spectral decision-making method based on the actor–critic algorithm is shown in Figure 6.

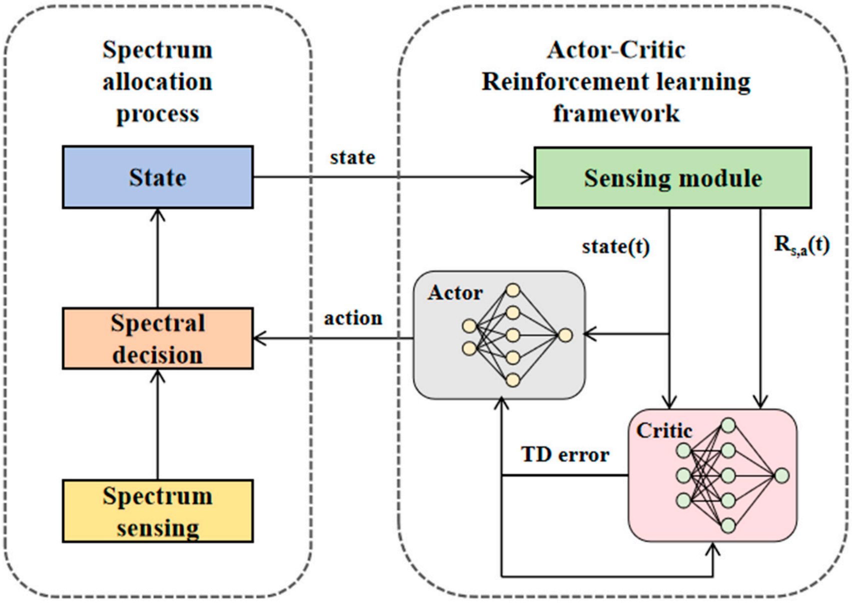


Figure 6. Schematic diagram of spectrum allocation based on reinforcement learning.

In the actor–critic learning framework, the node obtains the action A at the current moment by using the perception module to obtain the current environmental state information *state* (quality of service) input into the actor network, which is closely related to the acoustic modem. At this point, the state changes with the action and feeds back the reward return $R_{s,a}$ to the critic network. The critic network receives the reward return $\mathrm{R_{s,a}}$ and calculates the time-division error based on $R_{s,a}$, constantly updating the actor and the critic network.

In conjunction with the actor–critic framework, a Markov decision process (MDP) is constructed for the spectrum switching process, which is a process by which an agent interacts with the environment by making a decision to change the state of the environment so as to obtain a reward, and in this way, the process is shown in Equation (11).

$$M =< state, a, P_{s,a}, R_{s,a} > \tag{11}$$

In order to apply the reinforcement learning actor–critic algorithm to the spectral decision-making process, problem mapping is first performed:

1. Environment state *state*: In the actor–critic algorithm, the environment state is the key connecting point between the value function learning and the policy optimization, as shown in Equation (12), which defines the time factor of the reserved spectrum to be accessed in the system as the environment variable Q.

$$\mathrm{Q} = 1 + ln(P_i)/\lambda \mathrm{T_{r,i}} \tag{12}$$

The change in the value of Q corresponds to the change in the environmental state STATE and also represents the quality of service obtained by the node on the reserved spectrum.

2. Action decision a: There are two types of actions that a representing node can make to affect the state of the environment, the $a_t \in \{0, 1\}$. When $a_t = 0$, this represents the decision of accessing the matched shared spectrum, and $a_t = 1$ indicates the decision of needing to switch to the reserved spectrum, totaling two dimensions of output data.
3. $p[state(t), a(t), state(t+1)] \sim p[state(t+1)|\ state(t), a]$ represents the probability of choosing action $a(t)$ in state $state(t)$ and getting state $state(t+1)$. That is, the action is

made based on the quality of service at the current moment and the quality of service at the next moment is obtained.

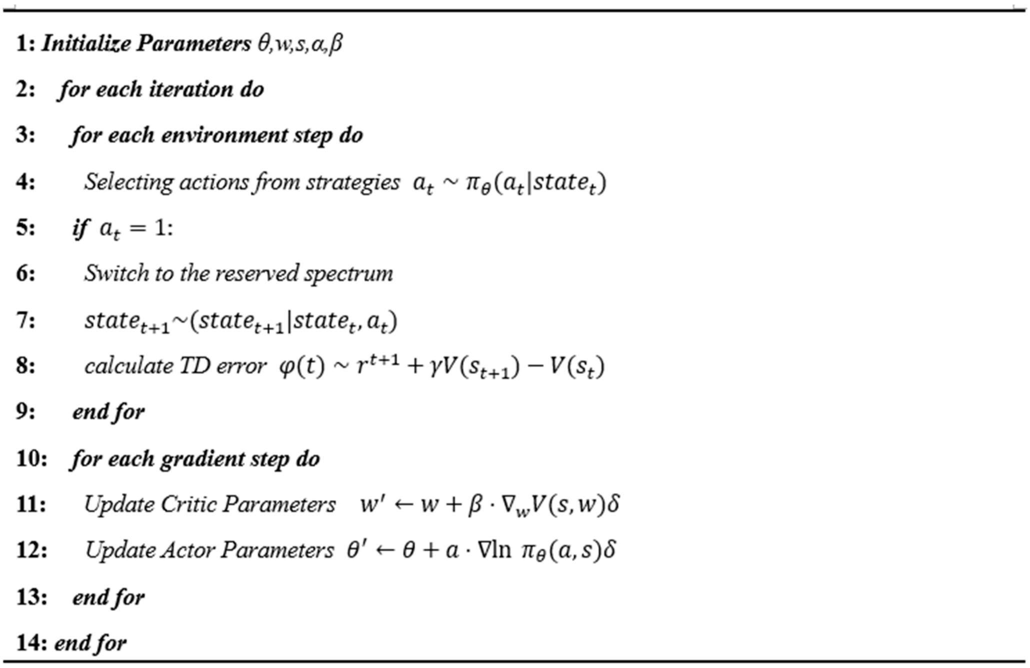
4. Reward value $R_{s,a}(t)$: Obtaining a reward value is the goal of reinforcement learning. The $R_{s,a} = E[R_{t+1}|\ s,a]$ denotes the immediate reward obtained by the agent after taking a certain action. As shown in Equation (13), when accessing the reserved spectrum or accessing the shared spectrum, the Q value increases, giving it a reward value to encourage it to continue making decisions that will improve the quality of service, and when the Q value decreases, the reward value is 0.

$$R_{s,a}(t) = \begin{cases} 1,\ Q(t+1) > Q(t) \\ 0,\ Q(t+1) < Q(t) \end{cases} \tag{13}$$

The core of the actor–critic algorithm is that critic evaluates the actor's actions based on the state and calculates the reward value to perform network updates to establish a more accurate relationship between the state and the action, so that each time a state is experienced, a better decision can be made by using previous experience to maximize the return of obtaining the cumulative rewards, as shown in Equation (14).

$$max\sum_{t=0}^{n} R_{t+1} = max\sum_{t=0}^{n} E[R_{t+1}|S_t = state, A_t = a] \tag{14}$$

The pseudo-code of the running process of the spectrum decision-making method proposed in this paper is shown in Figure 7.



```
Initialize Parameters $\theta, w, s, \alpha, \beta$
 for each iteration do
   for each environment step do
     Selecting actions from strategies $a_t \sim \pi_\theta(a_t|state_t)$
   if $a_t = 1$:
     Switch to the reserved spectrum
     $state_{t+1} \sim (state_{t+1}|state_t, a_t)$
     calculate TD error $\varphi(t) \sim r^{t+1} + \gamma V(s_{t+1}) - V(s_t)$
   end for
   for each gradient step do
     Update Critic Parameters $w' \leftarrow w + \beta \cdot \nabla_w V(s,w)\delta$
     Update Actor Parameters $\theta' \leftarrow \theta + a \cdot \nabla \ln \pi_\theta(a,s)\delta$
   end for
end for
```

Figure 7. Pseudo-code diagram of spectrum decision method based on reinforcement learning.

In the above process of spectrum decision making, the input of the critic network is the current state and the output is the state value function under the current state. At the same time, utilizing the characteristic of a neural network fitting a nonlinear function, a multilayer neural network is established for state value function fitting, and the current state value function is calculated by the critic network, as shown in Equation (15).

$$V(state_t) = E[R_{t+1} + \gamma V(state_{t+1})\ |\ S_t = state] \tag{15}$$

where γ is the discount rate. Meanwhile, for the next moment, the state value function V state *t + 1* is performed, and the time-division error (TD error) of both is calculated to obtain the loss function. Finally, the method shown in Equations (16) and (17) is used to update the parameters of the critic network with a function optimizer.

$$F_{\text{Loss}} = V(state_t)\phi(t) \tag{16}$$

$$w' \leftarrow w + \beta \cdot F_{\text{Loss}} \tag{17}$$

where β is the critic network learning rate. $\phi(t)$ is the TD error, which is calculated as shown in Equation (18).

$$\phi(t) = r^{t+1} + \gamma V(state_{t+1}) - V(state_t) \tag{18}$$

where r is the reward return. In the network design, a fully connected neural network, as shown in Figure 8, is used to fit the state value function. Among them, the hidden layer has six neurons, and the activation function adopts the hyperbolic tangent function (tanh), as shown in Equation (19).

$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{19}$$

Meanwhile, the critic network optimizes the loss function by the gradient descent method and updates the network parameters. Its structure is shown in Figure 8.

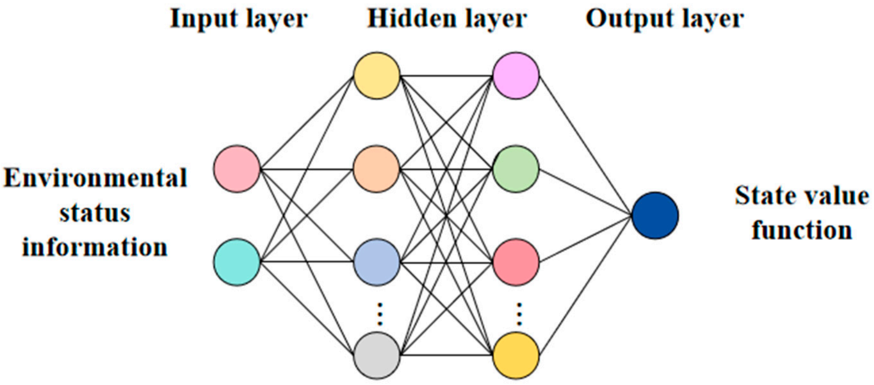


Figure 8. Critic network architecture diagram.

The input to the actor network is also the current state, but its output is the result of a spectral decision. With *a* parameterized behavioral strategy, P_θ denotes the action selected by the sensor node. θ is a function of the strategy $P_\theta(state|a)$, which denotes the strategy in which the θ parameter shows the probability that the sensor node is in the state *state* to perform action *a*. Also, it can be combined with the TD error obtained in the critic network $\phi(t)$ to compute the loss function value as a way to update the parameters of the actor network, as shown in Equation (20).

$$\theta' \leftarrow \theta + \alpha \cdot \left(r^{t+1} + \gamma V_\pi(state_{t+1}) - V_\pi(state_t)\right) \nabla \ln P_\theta(a_t \mid state_t) \tag{20}$$

where α is the learning rate of the actor network. The actor network uses a fully connected neural network, as shown in Figure 9, which also has six neurons in the hidden layer, and the activation function uses the tanh function and adopts a gradient descent strategy to optimize the loss function to update the network parameters.

When the loss values converge, the training process is ended and the trained network is used to make switching and access decisions. Finally, the flowchart of the R2S algorithm proposed in this paper is shown in Figure 10.

First, the nodes are initialized with the quality of service requirements Pi and reinforcement learning parameters, etc., and then the shared spectrum matching is performed, and after matching the spectrum *j* to the node *i*, the judgement is performed by using the action value a, which is output by the actor–critic network. When *a* > 0.5, the reserved spectrum is accessed according to the size of τi, and when $a < 0.5$, it is accessed on the matched spectrum, by which a higher quality of service at the next moment is obtained and constraints are imposed on the switching behavior. When the execution of the algorithm is complete, the spectrum or relay link that needs to be accessed has been determined, and the spectrum switching and access operation can be performed next.

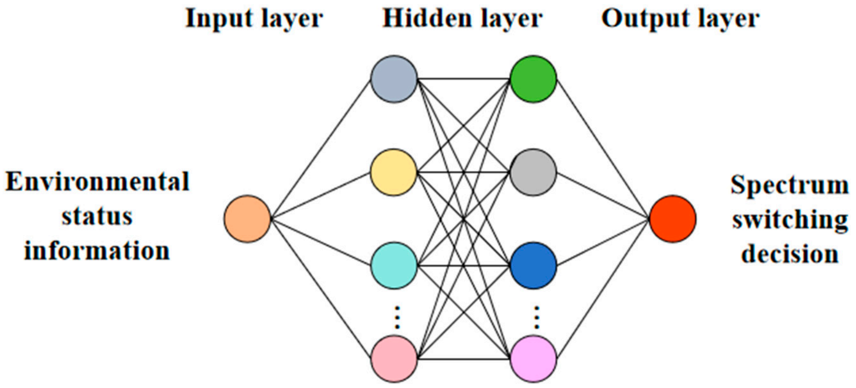


Figure 9. Actor network architecture diagram.

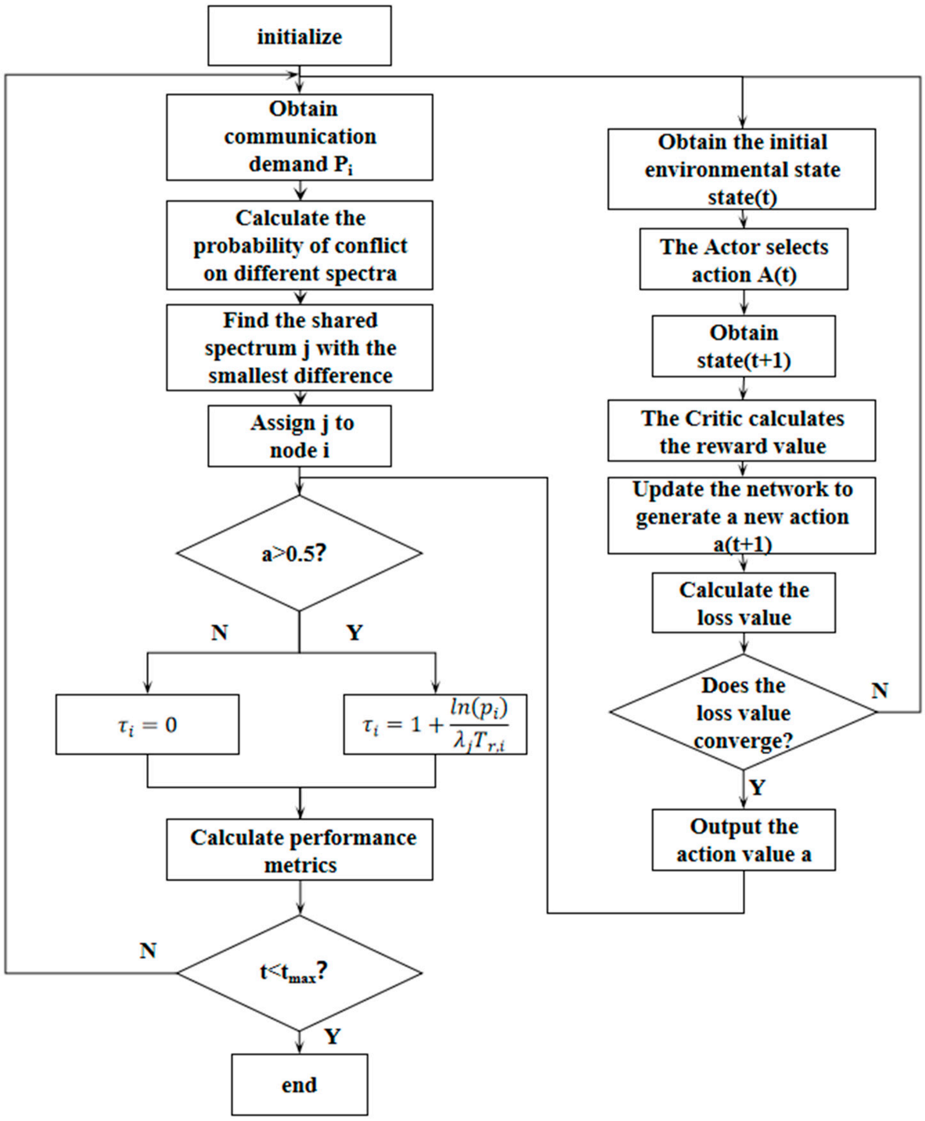

initialize
Obtain communication demand P_i
Calculate the probability of conflict on different spectra
Find the shared spectrum j with the smallest difference
Assign j to node i
a>0.5?
N
Y
$\tau_i = 0$
$\tau_i = 1 + \frac{ln(p_i)}{\lambda_j T_{r,i}}$
Calculate performance metrics
N
$t<t_{max}$?
Y
end
Obtain the initial environmental state state(t)
The Actor selects action A(t)
Obtain state(t+1)
The Critic calculates the reward value
Update the network to generate a new action a(t+1)
Calculate the loss value
Does the loss value converge?
N
Y
Output the action value a

Figure 10. R2S algorithm flowchart.

4. Experimental Results and Analysis

In this section, the proposed algorithm is simulated and evaluated, and the experimental procedure is carried out in the Windows 10 system through MatlabR2022a. We make the following assumptions: the upper limit of the number of sensors is M, there is a total of N available channels, the spectrum and relay link times obey an exponential distribution with parameters λ_j and λ_{SD}, and the time required for each service is $T_{r,i}$. The simulation parameters used are shown in Table 1.

Table 1. Experimental parameters.

Maximum number of sensor nodes M	50
Number of channels N	10
Time parameters for shared spectrum λ_j	Uniform distribution between [1/2000, 1/200]
Time parameters for relay links λ_{SD}	Uniform distribution between [1/1000, 1/100]
$T_{r,i}$ of the sensor nodes	Uniform distribution between [10, 300]

In the above parameter settings, the $T_{r,i}$ of the sensor node is included in the value domain of the shared spectrum and the relay link, and the difference between the shared spectrum and the relay link is reflected by the use of exponential distributions of different parameters. The t_{max} was set to 50, the discount rate of the actor–critic algorithm used was set to 0.9, and the actor and critic network learning rates were both set to 0.005. Experiments were conducted by traversing each sensor node in the sensor network under the condition that a relay-assisted mechanism has been employed, with the aim of matching the most appropriate spectrum voids for each service time demand.

In controlled experiments, the algorithm in this paper is compared with the shared spectrum matching algorithm and the spectrum allocation algorithm based on the VCG bidding mechanism, which are mentioned as independent algorithms in the previous section. In the VCG algorithm, the sensor nodes need to calculate the valuation and bidding price for the spectrum based on the environmental state information such as noise power, channel coefficient, etc., respectively. When the bidding price is higher than the estimated price, it is proven that the spectrum being auctioned is applicable to the sensor node. Next, the node with the largest bidding price is found in the set of nodes applicable to the spectrum to be accessed into the auction spectrum. The process is looped until all users waiting to be switched are given a spectrum or the number of spectra used for the auction is zero. This model has a wide range of applications in the field of economics and also has a relatively good performance in the study of spectrum resource allocation, which has some reference value.

Next, Figure 11 shows the performance comparison of the algorithms on the switch count metric:

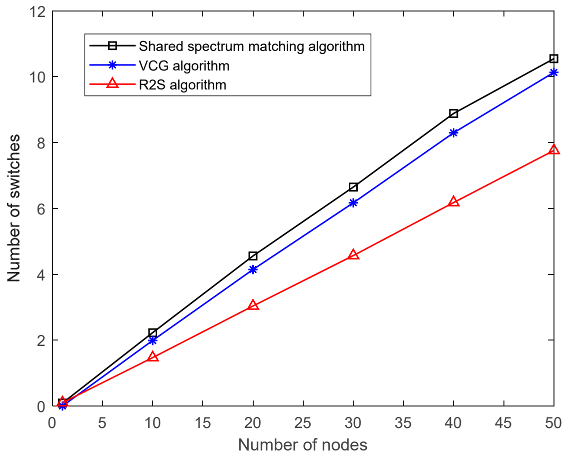


Figure 11. Comparison of switching times.

In hydroacoustic sensor networks, as the number of nodes increases, the number of switches in the whole network also increases. The more nodes there are in the network, the more complex and difficult it is to manage the network, which leads to performance degradation. The R2S algorithm proposed in this paper shows significant advantages over the traditional spectrum switching algorithm when the number of nodes is 50. The R2S algorithm reduces the number of switches by 26.4% compared to the shared spectrum matching algorithm and also reduces the number of switches by 23.4% compared to the spectrum switching algorithm based on the VCG bidding mechanism.

The decrease in the number of spectrum switches is attributed to the reinforcement learning-based spectrum decision-making method in the R2S algorithm, which not only performs well in selecting the spectrum with a switching probability that meets the requirements but also constrains the switching behavior of the spectrum based on the long-term gain. In this way, the R2S algorithm effectively reduces the number of unnecessary switches, reduces the extra overhead in the communication process, and improves the overall performance of the system.

Next, the system quality of service of each algorithm is comparatively analyzed under different numbers of nodes to further verify the effectiveness and practicality of the R2S algorithm in solving the spectrum switching problem of multi-node sensor networks.

From the analysis of Figure 12, it can be concluded that with the gradual increase in the number of nodes in the hydroacoustic sensor network, the overall quality of service of the system shows a decreasing trend. This is due to the fact that as the number of nodes increases, the complexity of communication and the amount of data transmitted in the network increases, leading to an increased likelihood of network congestion and spectrum resource conflicts, which in turn affects the overall quality of service.

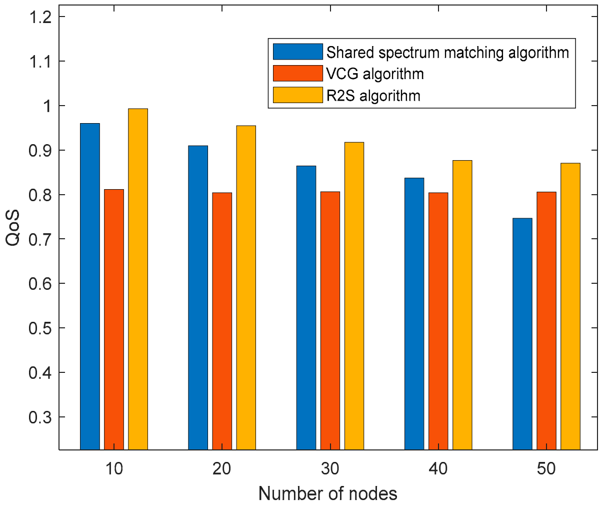


Figure 12. Comparison QoS.

When the number of nodes reaches 50, the proposed algorithm in this paper improves the quality of service of the system by 16.6% compared to the shared spectrum matching algorithm and also improves by 8.07% compared to the spectrum switching algorithm based on the VCG mechanism. In this paper, the algorithm is able to improve the performance of the system by adjusting the allocation strategy of spectrum resources according to the variation in the quality of service of the whole system in time through the spectrum decision-making method based on reinforcement learning.

Next, a controlled experiment is designed to focus on the change in the outage probability of the system as the probability of a marine mammal to engage in communication behaviors continues to increment in continuous time. This experiment is used to verify the stability and reliability of the relay assistance mechanism and the proposed algorithm in response to different communication demands and network states, and the results are shown in Figure 13.

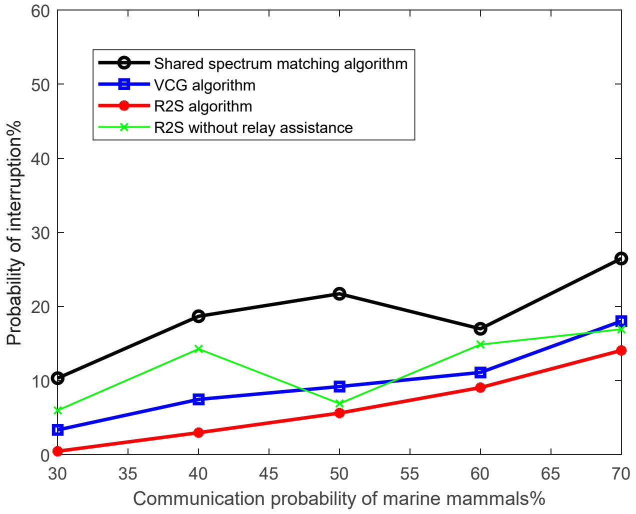


Figure 13. Comparison of interruption probability.

The experimental results show that as the probability of marine mammal communication increases, the probability of network disruption, although fluctuating due to the randomness of the experimental process, shows an overall upward trend. This phenomenon is mainly due to the fact that an increase in the probability of marine mammal communication implies an increase in the communication activity of the primary users in the network, which leads to an increase in competition and conflict for spectrum resources, which in turn raises the likelihood of communication disruption.

The results show that compared to the shared spectrum matching algorithm, the method proposed in this paper decreases the outage probability by 12.4% when the communication probability of marine mammals is 70%. Meanwhile, compared with the VCG-based spectrum switching algorithm, the outage probability of this paper's algorithm under the same marine mammal communication probability also decreases by 3.95%, which further validates the advantage of this paper's algorithm in reducing the outage probability.

Due to the introduction of the relay assistance mechanism, the outage probability of this paper's method decreases by 2.85% at a marine mammal communication probability of 70% and brings about a decrease in outage probability at any marine mammal communication probability. This shows that the relay assistance mechanism can effectively enhance the connectivity and stability of the network.

Finally, the effectiveness of the dynamic maximum occupancy duration proposed in the design of the avoidance algorithm is verified by utilizing the control variable method and setting b to a fixed value. According to the definition of b, b = 0.3, 0.5, and 0.7 are used to represent the change in the magnitude of the difference between Th and Tr from large to small, respectively, and the curve of the maximum occupancy duration M is plotted as shown in Figure 14.

It can be seen that for any value of b, the proposed recession method reduces the maximum occupancy duration as the probability of marine mammal vocalization increases. With the same probability of vocalization in an oceanic mammal, the smaller the value of b, the longer the maximum occupancy duration, i.e., the larger the T_h with respect to T_r, the larger the maximum occupancy duration. Such results validate the effectiveness of the proposed method, i.e., the ability to dynamically adjust the maximum occupancy duration according to the probability of oceanic mammal vocalization at each moment and the relative relationship between T_h and T_r to achieve the goal of bio-friendliness.

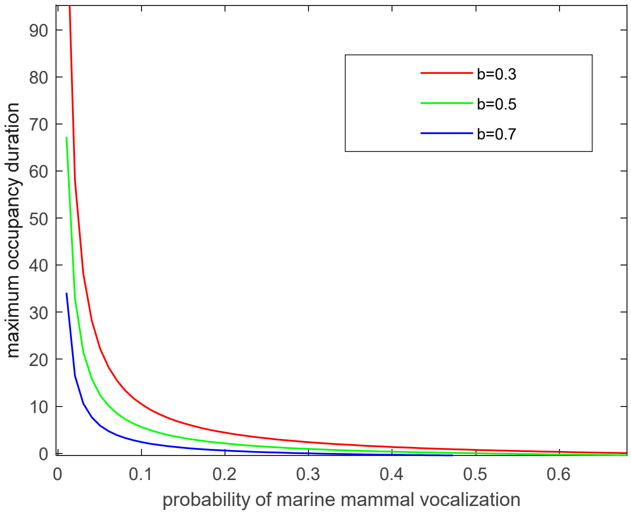


Figure 14. Dynamic occupancy duration.

5. Conclusions

In order to reduce the impact of communication activities of underwater sensor networks on marine mammals, this paper proposes a marine mammal conflict avoidance strategy and designs a stop-occupancy rule for the active spectrum of marine mammals to achieve the goal of bio-friendliness. In this paper, we also design a relay-assisted spectrum switching method (R2S) based on reinforcement learning to dynamically adjust the spectrum switching and access strategy to reduce the number of switching times and the interruption probability of the system while ensuring the quality of service of the sensor nodes, so as to realize the reasonable and efficient use of the limited spectrum resources. However, the research process is mainly centered on the probabilistic model, and the performance and overhead of the proposed algorithm should be further examined in the next step in conjunction with the actual marine environment. Meanwhile, the latest research results for reinforcement learning algorithms can be further combined with the algorithms in this paper to improve the performance of the communication system.

Author Contributions: Methodology, H.W.; software, H.W.; formal analysis, H.W.; investigation, J.L.; resources, Y.X.; data curation, B.L.; writing—original draft preparation, H.W.; writing—review and editing, Y.X.; visualization, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Science & Technology Development Project of Jilin Province (Grant No. 20220101141JC). Any correspondence related to this paper should be addressed to Yihu Xu.

Data Availability Statement: Data are contained within the article.

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

References

1. Zhang, H.; Chen, S. Overview of research on marine resources and economic development. *Mar. Econ. Manag.* **2022**, *5*, 69–83. [CrossRef]
2. Khan, M.R.; Das, B.; Pati, B.B. Channel estimation strategies for underwater acoustic (UWA) communication: An overview. *J. Frankl. Inst.* **2020**, *357*, 7229–7265. [CrossRef]
3. Liu, Y.; Wang, H.; Cai, L.; Shen, X.; Zhao, R. Fundamentals and advancements of topology discovery in underwater acoustic sensor networks: A review. *IEEE Sens. J.* **2021**, *21*, 21159–21174. [CrossRef]
4. El-Fikky, A.E.R.A.; Eldin, M.E.; Fayed, H.A.; Abd El Aziz, A.; Shalaby, H.M.; Aly, M.H. NLoS underwater VLC system performance: Static and dynamic channel modeling. *Appl. Opt.* **2019**, *58*, 8272–8281. [CrossRef] [PubMed]
5. Khalighi, M.A.; Uysal, M. Survey on free space optical communication: A communication theory perspective. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 2231–2258. [CrossRef]
6. Qu, L.; Xu, G.; Zeng, Z.; Zhang, N.; Zhang, Q. UAV-assisted RF/FSO relay system for space-air-ground integrated network: A performance analysis. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 6211–6225. [CrossRef]

7. Xu, G.; Zhang, N.; Xu, M.; Xu, Z.; Zhang, Q.; Song, Z. Outage Probability and Average BER of UAV-Assisted Dual-Hop FSO Communication With Amplify-and-Forward Relaying. *IEEE Trans. Veh. Technol.* **2023**, *72*, 8287–8302. [CrossRef]
8. Atalay, C.M. UCUNCUM Transmitter and Receiver Design for Underwater Communication. *J. Mach. Intell. Data Sci. (JMIDS)* **2022**, *3*, 25–35.
9. Etter, P.C. *Underwater Acoustic Modeling and Simulation*; CRC Press: Boca Raton, FL, USA, 2018; pp. 230–238.
10. Alfa, A.S.; Ghazaleh, H.A.; Maharaj, B.T. Performance Analysis of Multi-Modal Overlay/Underlay Switching Service Levels in Cognitive Radio Networks. *IEEE Access* **2019**, *7*, 78442–78453. [CrossRef]
11. Tan, Z.; Li, Y.; Sun, S.; Zhang, R.; Yao, X. Dynamic Spectrum Allocation Mechanism of Joint Power Control and Channel Allocation. *J. Internet Technol.* **2021**, *22*, 165–171.
12. Li, X.; Sun, Y.; Guo, Y.; Fu, X.; Pan, M. Dolphins first: Dolphin-aware communications in multi-hop underwater cognitive acoustic networks. *IEEE Trans. Wirel. Commun.* **2016**, *16*, 2043–2056. [CrossRef]
13. Xue, L.; Cao, C. Joint Channel and Power Assignment for Underwater Cognitive Acoustic Networks on Marine Mammal-Friendly. *Appl. Sci.* **2023**, *13*, 12950. [CrossRef]
14. Elhachmi, J. Distributed reinforcement learning for dynamic spectrum allocation in cognitive radio-based internet of things. *IET Netw.* **2022**, *11*, 207–220. [CrossRef]
15. Wang, H.; Li, Y.; Qian, J. Self-adaptive resource allocation in underwater acoustic interference channel: A reinforcement learning approach. *IEEE Internet Things J.* **2019**, *7*, 2816–2827. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.