IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 6, 2019, accepted May 20, 2019, date of publication May 23, 2019, date of current version June 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918506

Reinforcement Learning-Based Adaptive Modulation and Coding for Efficient Underwater Communications

WEI SU, JIAMIN LIN, KEYU CHEN, LIANG XIAO, (Senior Member, IEEE), AND CHENG EN, (Member, IEEE)

Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Xiamen University, Xiamen 361000, China
Department of Communication Engineering, Xiamen University, Xiamen 361000, China

Corresponding author: Wei Su (suweixiamen@xmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671398, Grant 61871336, and Grant 61671396.

ABSTRACT In this paper, we propose a reinforcement learning-based adaptive modulation and coding scheme for underwater communications; more specifically, based on the network states such as the quality of service requirement of the sensing message, the previous transmission quality, and the energy consumption. This scheme applies reinforcement learning to choose the modulation and coding policy in a dynamic underwater communication system. We provide the performance bound of this scheme and perform experiments in both pool and sea environments. The experimental data were collected and post-processed. Compared with the benchmark schemes, this scheme can improve the throughputs and reduce the BER with less energy consumption.

INDEX TERMS Reinforcement learning, adaptive modulation and coding, underwater communication.

I. INTRODUCTION

Underwater acoustic communications suffer from the fast time-variant channel states, the limited bandwidth, Doppler effect and sometimes low signal-to-noise ratio (SNR) [1], [2]. Adaptive modulation and coding (AMC) scheme that can improve the communication efficiency and reliability, depends on the predicted channel state to determine the modulation and coding policy from the feasible candidates at the transmitter. Being critical for the data throughput, the bit error rates (BERs) and the transmit energy consumption in the underwater acoustic communications, the modulation and coding policy must be optimized according to the communication performance of different modulation and coding methods under the current channel state.

For instance, an AMC scheme as presented in [3] named SB estimates the channel impulse response (CIR) and SNR of each received data block based on a deterministic prediction model, evaluates the current BER of each modulation and coding policy to choose the modulation and coding policy in terms of the throughput and the BER requirement. However, an underwater acoustic transmitter can rarely determine such information in time, especially under variant quality of services (QoS) requirements.

Reinforcement learning algorithms are useful to these problems [4]–[7]. For instance, a Q-learning-based AMC scheme as presented in [4] named QLM uses Q-learning to maximize the utility as a function of the SNR, the BER and the transmission time. However, this scheme does not improve the QoS of the messages in the underwater acoustic communications.

In this paper, we propose a reinforcement learning-based adaptive modulation and coding scheme. More specifically, this scheme applies reinforcement learning to choose the modulation and coding policy to optimize the long-term expected utility of the underwater transmitter, including the BERs, the energy consumption, the transmission time and the QoS requirements without knowing the underwater channel model.

Performance experiments are conducted in both pool and sea areas to evaluate its performance and compared with the SB scheme and the QLM scheme. Experimental results show that this scheme converges to the theoretical performance

The associate editor coordinating the review of this manuscript and approving it for publication was Guangjie Han.

bound, decreases the BERs, the transmission time, the energy consumption and increases the utility of the transmitter.

The rest of this paper is organized as follows. In Section II, the related work is introduced. In Section III, the communication model is presented. In Section IV, we propose a reinforcement learning-based modulation and coding scheme, and in Section V, present experimental results in both the pool and sea tests. In Section VI, the conclusions are drawn.

II. RELATED WORKS

Some underwater communication nodes, which install both underwater acoustic and optical modems, are reported in [8], [9]. They provide a flexible way to switch from a long-distance but relatively low-data-rate underwater acoustic link to a short-range but high-data-rate underwater optical link. In [10]–[13], the AMC schemes are designed under similar policy as presented in [3]. In these schemes, the BER of the future data package is predicted on the basis of the current channel state such as SNR and CIR with deterministic models. In [10], the future BER is predicted from the SNR of the current data package and the SNR is estimated before demodulation at the receiver. In [11], the future SNR is estimated after equalization and decoding at the receiver. In [12], both the SNR and the CIR are used to predict the future BER. The policy of this scheme is to maximize the throughput while maintaining a target average BER. In [13], pool experiments are conducted to verify the performance of AMC schemes. In [14] and [15], decision tree models are used to predict the future BER on the basis of the CIR and the SNR. The performance of this scheme is testified by post-processing sea experiments (REP15-Atlantic).

In [16], a reinforcement learning-based routing strategy called MARLIN is presented. The transmitters can use MARLIN to select optimal forward relay and the most suitable communication device for a reliable and low latency underwater networking. In simulation experiments, two communication devices, both operating on binary phase shift keying (BPSK) but at different central frequency are considered. In [17], an energy efficient and QoS-aware routing algorithm called EEQA is presented. This algorithm can ensure that different types of data are forwarded to the optimal relay node under the data requirements. In [18], a data collection scheme is presented for underwater wireless sensor networks. In this scheme, a trajectory adjustment mechanism and a reliable time mechanism are proposed respectively to address the high energy consumption problem in "hot region" and guarantee reliable data transmission. In [19], a high-availability data collection scheme based on multiple AUVs (HAMA) is proposed.

Reinforcement learning can be used to solve optimal problems in dynamic and complicated environments [5]. Q-learning has been successfully used to select the optimal channels for spectrum sensing and data transmission in cognitive radio networks [20], improve the anti-jamming transmission performance of wireless communications [5]–[7], minimize the total electricity cost [21], deal with the optimal battery management problem in smart residential environments [22], design the routing protocol in acoustic-optical underwater sensor networks [23] and so on. These systems can achieve goals with optimal or near-optimal policies through online learning. Moreover, they do not need to know the exact model and full prior knowledge of the environment.

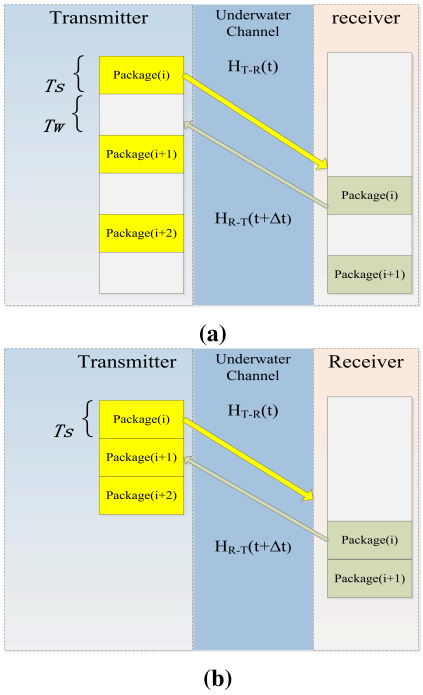


FIGURE 1. Communication modes. (a) Light-weight mode. (b) Fast mode.

III. SYSTEM MODEL

In this section, two communication modes with different bandwidth efficiency are introduced. The uncertainty of channel states prediction is analyzed and modeled. An overall performance optimization model is established.

A. COMMUNICATION MODEL

Two communication modes, the light-weight mode and the fast mode are shown in Fig. 1a and Fig. 1b. In both modes, the transmitter and the receiver move randomly. The distance between them is defined as $d(t)$. The time needed to transmit a task is defined as T_ζ, which is divided into fixed time slots. The time of each time slot is defined as T_s. In each time slot, a flexible-size data package is transmitted. The CIR estimation signals are inserted at the beginning, middle and the end of each time slot. Because of the channel reciprocity, we have $h(t) = h_{T-R}(t) = h_{R-T}(t + \Delta t)$ in a very small time Δt, where $h_{T-R}(t)$ is the CIR from the transmitter to the receiver at time t [24].

In the light-weight mode as presented in Fig. 1a, after a package is transmitted, the transmitter waits for the feedback message from the receiver. This time is defined as T_w. Fixing $d(t)$ to D, we get $T_w \approx 2D_c$, where $D_c = D/C_v$ and $C_v \approx 1500$ m/s. In underwater acoustic channels, such

feedback operations will seriously affect the bandwidth efficiency.

Compared with the light-weight mode, the feedback channel in Fig. 1b is independent of the data channel. For example, the feedback channel is set to a lower central frequency with a slow bps. Using the fast mode, the data packages can be transmitted continuously, which makes it more bandwidth efficient compared with the light-weight mode.

In order to choose a most suitable modulation and coding method for the current time slot k, the current channel state must be predicted at the transmitter from the previous channel states. For example, the estimated CIR at the receiver of time slot $k-1$ is $\hat{\mathbf{h}}_{k-1} = [h(T_{k-1}), h(T_{k-1}+T_s/2), h(T_{k-1}+T_s)]^{\mathsf{T}}$. In light-weight mode, the transmitter will wait for the knowledge of $\hat{\mathbf{h}}_{k-1}$ from the receiver. In [12], the $\hat{h}(T_k)$ is predicted by a deterministic linear model, which is $\hat{h}(T_k) = \mathbf{w}\mathbf{h}_{k-1}^{\mathsf{T}}$, where the weights $\mathbf{w}$ are calculated and updated by recursive least squares (RLS) algorithm.

But in the fast mode, the data packages are transmitted continuously. When predicting the $h(T_k)$, the available knowledge is $[\mathbf{h}_{k-a}, \mathbf{h}_{k-a-1}, \cdots]$, where a changed with distance $d(t)$ and $a > 1$, which makes the prediction much harder compared with that in the light-weight mode. If the time delay is larger than the channel coherent time, the performance of deterministic CIR prediction models will degrade. Considering the uncertainty of prediction, the CIR of time slot k is modeled as

$$h(T_k) = f(\mathbf{h}_{k-a}, \mathbf{h}_{k-a-1}, \cdots) + \Upsilon(T), \tag{1}$$

where $f()$ represents a deterministic CIR prediction model and $\Upsilon(T)$ is the uncertainty caused by time T. For light-weight mode, we get $a = 1$ and $T = 2D_c$. And for fast mode, we get $a > 1$ and $T > 2D_c$. In this work, in order to improve the bandwidth efficiency, the fast mode is selected.

B. OPTIMIZATION MODEL

The QoSs of the AMC schemes are determined by the optimization policy.

The set of modulation and coding methods is defined as $\Xi = [\xi_1, \cdots, \xi_J]$. The modulated signal is defined as s_j^k, $1 \le j \le J$, where j is the serial number of the modulation and coding methods and k is the serial number of the time slot. The target of the overall performance optimization policy is s

$$\begin{aligned} &\min [\sum_{k=1} \min \left|\Theta_{l,j}^k \Omega^T\right|], \\ &\quad 1 \le k \le \infty, \quad 1 \le l \le L, \ 1 \le j \le J \\ &s.t\ \Theta_{l,j}^k(i) \le \Psi_l(i) \end{aligned} \tag{2}$$

where $\Theta_{l,j}^k = [\rho_j^k, T_{j,l}, p_j^k, \cdots]^T$ is the performance set of the modulation and coding method ξ_j at time slot k. l is the serial number of tasks. ρ_j^k is the BER performance. $T_{j,l}$ is the transmission time of lth task using a specific modulation and coding method ξ_j. p_j^k is the energy consumption. $\Omega^T = [\omega_\rho, \omega_T, \omega_p, \cdots]^T$ represents the costs of $\Theta_{l,j}$. $\Psi_l = [\psi_1, \cdots, \psi_i]$ is the lth QoS of message.

TABLE 1. Symbols and notations.

Notation	Description
ξ_j	The modulation and coding method
Ξ	The set of modulation and coding methods
J	The number of modulation and coding methods
s_j^k	The signal modulated by ξ_j and transmitted by transmitter at time slot k
$h(t)/\hat{h}(t)$	Real CIR at time t/ Estimated CIR at time t
h^k	Quantified CIR at time slot k
r^k	Quantified SNR at time slot k
ρ^k	Quantified BER at time slot k
ζ_l	lth task
L	Number of different type of tasks
Ψ_l	The lth QoS of message
ρ_j^k	The BER of jth action at time slot k
$T_{j,l}$	The time needed to finish this task
p_j^k	Energy consumption at time slot k
$\Theta_{j,l}^k$	Performance set of s_j^k.
ω_ρ	BER cost
ω_T	Transmission time cost
ω_p	Energy consumption cost
Ω_j^k	Cost set corresponding to $\Theta_{j,l}^k$
Λ_j^k	Transmitter state at time slot k
P_ζ^k	Extra punishment if the QoS of message are not met
$(\omega_{\zeta_l})_i$	Corresponding punishment
u^k	Utility of transmitter
α	Learning rate of Q-learning
δ	Discount factor of Q-learning

In this work, the QoS of message is set to three levels, which are $[0, a, \infty]$. When $\Psi_l(i) = 0$, this QoS of message must be satisfied. When $\Psi_l(i) = a$, for example, a specific BER can be tolerated. When $\Psi_l(i) = \infty$, we can ignore this requirement. In addition, for ease of reference, important notations are summarized in Table 1.

IV. ADAPTIVE MODULATION AND CODING SCHEME

In this section, a reinforcement learning-based AMC underwater acoustic communication scheme named RLMC is presented. Specifically, the hot-booting Q-learning algorithm is used to solve the optimization problem proposed by equation (2). The utility function and the cost function corresponding to equation (2) are designed for underwater acoustic communication environments. The performance bound of this optimization problem is calculated and analyzed.

A. REINFORCEMENT LEARNING-BASED MODULATION AND CODING

The CIR and SNR are estimated at the receiver and quantified to several levels, which are defined as h and r. These messages are fed back to the transmitter, which can be used to estimate the quantified BER ρ.

The quantified channel states are discrete and complete. Thus, the channel state can only transfer from one existing state to another existing state. Moreover, the current quantified channel state is transferred from and only related to previous quantified channel states. For example, the state of the quantified CIR h^k is only related to $[h^{k-n}, h^{k-n-1}, \cdots]$, where $n > 1$ in the fast communication mode, which means the state of h^k is random and occurs with a probability $Pr(h^k \big| h^{k-n}, h^{k-n-1}, \cdots)$. Therefore, the transitions of the quantified channel states can be modeled as a Markov chain and the problem expressed in equation (2) can be solved by reinforcement learning. More specifically, in this work, the hot-booting Q-learning algorithm is used.

The presented RLMC scheme can be divided into two stages: virtual learning stage and online learning stage.

In the virtual learning stage, we conducted numerous sea experiments in similar sea areas. These results are used to build a virtual Q-table named Q^*-table in order to accelerate the convergence speed of the online learning stage. The virtual learning stage includes two steps, the preparation step and the Q^*-table building step.

In the preparation step, firstly, a number of sea experiments are conducted using different modulation and coding methods. Then, each received signal is divided into fixed time slots. The SNR and the CIR of each time slot are estimated and the BER performance is calculated. After that, according to the requirements of QoS of message, they are quantified to several levels. From this step, we can connect the quantified BER performance with the quantified channel states.

In the Q^*-table building step, at the beginning, we initiate the Q^*-table $Q^*(\Lambda_j^k, s_j^k)$ to zero, where $k = 0$, j is the serial number of modulation and coding methods and $\Lambda_j^k = [\Psi_l^k, \rho_j^k, T_{j,l}{}^k, p_j^k]$ is the transmitter state, including the current QoS of message, the BER, the transmission time and the energy consumption of the current action. Then, an iterative process is performed to update the Q^*-table.

At $k-1$th time slot, the transmitter chooses its action $s_j^{k-1} \in \Xi$, a task and a virtual quantified channel randomly. The signal is transmitted through this virtual channel. Then, the quantified CIR and the quantified SNR are estimated and calculated at the receiver and fed back to the transmitter.

At kth time slot, the transmitter uses the quantified h^{k-1} and r^{k-1} to obtain the BER (i.e.,ρ_j^k) of each available action.

In the RLMC scheme, the BER service quality is considered according to the current QoS of message. If the chosen action could not meet the corresponding BER service quality, it would get an extra punishment, which is

$$P_\zeta^k = (\omega_{\zeta_l})_i g(\rho_j^k, \Psi_i^k), \tag{3}$$

where $\Psi_i^k \in \Psi_l$ is the QoS of message presented in equation (2). $(\omega_{\zeta_l})_i$ is the corresponding punishment. $g(x, y)$ is an indicator function. If $x \geq y$, returns 1 and if $x < y$, returns 0.

Then, the transmitter calculates its utility and chooses its virtual action $s_j^k \in \Xi$. The long-term expected utility is calculated by

$$u^k = -(\omega_\rho \rho_j^k + \omega_p p_j^k + \omega_T T_{l,j} + P_\zeta^k), \tag{4}$$

where ρ is BER, p is the energy consumption, and $T_{l,j}$ is the transmission time. ω_ρ, ω_p and ω_T are their costs.

Let $V^*(\Lambda_j^k)$ denote the maximum value of the Q^*-table, the $Q(\Lambda_j^k, s_j^k)$ and $V(\Lambda_j^k)$ are updated by [4]:

$$\begin{aligned} Q(\Lambda_j^k, s_j^k) = (1-\alpha) Q(\Lambda_j^k, s_j^k) \\ + \alpha(u^k + \delta V(\Lambda_j^{k+1})) \end{aligned} \tag{5}$$

$$V(\Lambda_j^k) = \max_{s_j^k \in \Xi} Q(\Lambda_j^k, s_j^k), \tag{6}$$

where the $\alpha \in (0, 1]$ is learning rate, which shows the weight of the current experience. $\delta \in (0, 1]$ is the discount factor, which corresponds to the uncertainty about rewards to be received in the future.

At $k+1$th iteration, the transmitter chooses its action s_j^{k+1} randomly again. After enough iteration steps, the Q^*-table is stabilized and be saved.

In the online learning stage, the Q-table is initiated by Q^*-table at the beginning. At each time slot, the transmitter chooses its action based on the current state Λ_j^k and the updated Q-table. Specifically, the transmitter tries all the possible actions under each channel state repeatedly. Then, the transmitter chooses its action based on the ε-greedy policy [20]. The action which can maximize the Q value occurs with a large probability $1-\varepsilon$ and the other actions occur with a small probability $\varepsilon/(|\Xi|-1)$. The probability of action s_j^* is given by

$$p_r(s_j^k = \tau) = \begin{cases} 1-\varepsilon & \tau = s_j^* \\ \dfrac{\varepsilon}{|\Xi|-1} & \text{o.w.} \end{cases} \tag{7}$$

The optimal action s_j^* is given by

$$s_j^* = \arg\max_{s_j' \in \Xi} Q(\Lambda_j^k, s_j') \tag{8}$$

The proposed RLMC scheme is summarized in Algorithm 1.

B. PERFORMANCE BOUND

If the future channel states can be perfectly predicted, we can get the performance bound of the optimization problem established in equation (2). At the kth time slot, the state of the transmitter is $\Lambda_j^k = [\Psi_l^k, \rho_j^k, T_{j,l}, p_j]$. If the transmitter knows the state Λ_j^k precisely, we can find an optimal action

Algorithm 1 RLMC Algorithm

Stage 1: Virtual learning stage
Initialized $\alpha, \delta, \varepsilon, Q^*(\Lambda_j, s_j) = 0, V^*(\Lambda_j) = 0, \forall s_j \in \Xi, \Lambda_j$
for $i = 1, 2, 3..., I$ **do**
  Emulate a similar environment
  for $k = 1, 2, 3..., K$ **do**
    Chooses $s_j^{k-1} \in \Xi$ randomly
    if Modulation or coding method changes **then**
      Inform the receiver
    end if
    transmitter send modulated signals s_j^{k-1} to receiver
    Receive message $\{h^{k-1}, r^{k-1}\}$
    Calculate ρ_j^k, p_j^k and $T_{l,j}{}^k$
    $\Lambda_j^k = [\Psi_l^k, \rho_j^k, T_{j,l}{}^k, p_j^k]$
    Evaluate u^k via(4)
    $Q^*(\Lambda_j^k, s_j^k) = (1-\alpha)Q^*(\Lambda_j^k, s_j^k) + \alpha(u^k + \delta V^*(\Lambda_j^{k+1}))$
    $V^*(\Lambda_j^k) = \max_{s_j^k \in \Xi} Q^*(\Lambda_j^k, s_j^k)$
  end for
end for
Save $Q^*(\lambda_j, s_j)$
Stage 2: Online learning stage
$Q(\Lambda_j, s_j) = Q^*(\Lambda_j, s_j), V(\Lambda_j) = 0$
for $k = 1, 2, 3...$ **do**
  Choose $s_j^{k-1} \in \Xi$ via (8) and send it to receiver
  Update Λ_j^k
  Evaluate u^k via(4)
  Update $Q(\lambda_j^k, s_j^k)$ via(5)
  Update $V(\lambda_j^k)$ via(6)
  Choose $s_j^k \in \Xi$ via (8)
  $\Lambda_j^{k+1} = [\Psi_l^{k+1}, \rho_j^{k+1}, T_{j,l}{}^{k+1}, p_j^{k+1}]$
end for

which could satisfy the QoS of message and maximize the utility u.

$$\begin{aligned} u &= E\left|u_j^k\right| \\ &= E\left|\max[-(\omega_\rho \rho_j^k + \omega_p p_j^k + \omega_T T_{l,j} + P_\zeta^k)]\right| \end{aligned} \tag{9}$$

The probability of ζ_l is defined as $Pr(\zeta_l)$ and the probability of ρ_j^k is defined as $Pr(\rho_j^k)$. The performance bound is

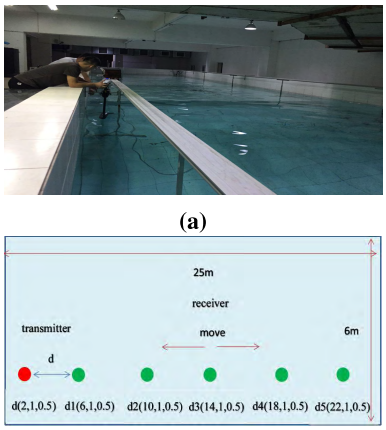
$$u = \sum_{k=1}^{\infty} u_j^k Pr(\zeta_l) Pr(\rho_j^k) \tag{10}$$

The optimal theoretical values u can be calculated by the Monte Carlo method using equation (10). We assume that each channel and each type of task occur with equal probability. First, all the received signals are divided into fixed time slots. Second, the channel states and the communication performance are quantified to some levels. Then, a Monte Carlo simulation is started. At each time slot, every possible modulation and coding method is tried, and the optimal values are saved. Then, their mean values are calculated. After a certain number of time slots, the mean values converged to the optimal values.

V. EXPERIMENTAL RESULTS

In this work, the fast variant channel states, the computational complexity and the robustness of channel prediction algorithms under severe channel states are considered when selecting the set of modulation and coding methods. Frequency modulation methods have low computational complexity. In addition, these methods can work well without CIR estimation results. Thus, even if the bandwidth efficiency of the frequency modulation methods are less than 0.5, these modulation methods are still a research interest and is used in underwater acoustic communication modems [25], [26]. In [25]–[27], multiple frequency shift keying (MFSK) modulation methods with different orthogonal or semi-orthogonal coding methods are presented. In [25], the Hadamard-MFSK is used in underwater acoustic modem ATM-850. In [26], a joint time-frequency coding method is presented. In [27], the complementary-code-keying is used. In [28], [29], a pattern-time-delay-shift coding method is presented. The performance of these methods is testified by pool and sea experiments.

In [30], the single-carrier-phase-coherent underwater acoustic communication method using decision-feedback-equalizer with a second-order digital phase-locked-loop (DFE+PLL) is presented, the bandwidth efficiency of which can be large than 1. This technique has been successfully used in a recently developed underwater acoustic communication modem [31]. The performance of CIR estimation and CIR updating algorithms are essential to phase coherent underwater acoustic communication. Compared with block processing methods, such as single-carrier-frequency-domain-equalization (SCFDE), orthogonal frequency division multiplexing (OFDM), affected by the peak-to-average ratio problem and the orthogonal signal-division multiplexing (OSDM) [32], [33], the CIR can be updated much faster when using [31], which makes it more suitable for fast variant underwater acoustic communication environments. In recent years, semi-blind and blind CIR estimation and equalization methods are researched in order to improve the bandwidth efficiency when block processing is used. But their performance is extremely dependent on accurate channel updating results. Thus, MFSK and single carrier coherent modulation methods are selected in this paper. In both pool and sea experiments. The MFSK signals were modulated by 1-of-4 method presented in [26] with different coding methods [27]. The symbol duration was 100 ms. The number of subcarriers was 160, and the bandwidth was 3.2 kHz (18.4- 21.6 kHz). The central frequency of MPSK was set to 20 kHz, and the symbol duration was set to 0.5 ms. The sampling rate was fixed to 100 kHz. Each modulation method had uncoded and different coded versions in the pool and sea experiments.



(a)

(b)

FIGURE 2. Pool experiments environments. (a) Pool experiment. (b) Network topology.

A number of sparse channel estimation algorithms have been presented in recent years [34]–[36]. But in this study, considering that the sparsity of the channel changed with time and only the quantified channel characters were fed back to the transmitter, the modified least square (LS) channel estimation algorithm proposed in [32] was used.

In the pool and sea experiments, the performance of the RLMC scheme was compared with the SB scheme, the QLM scheme and the theoretical solutions.

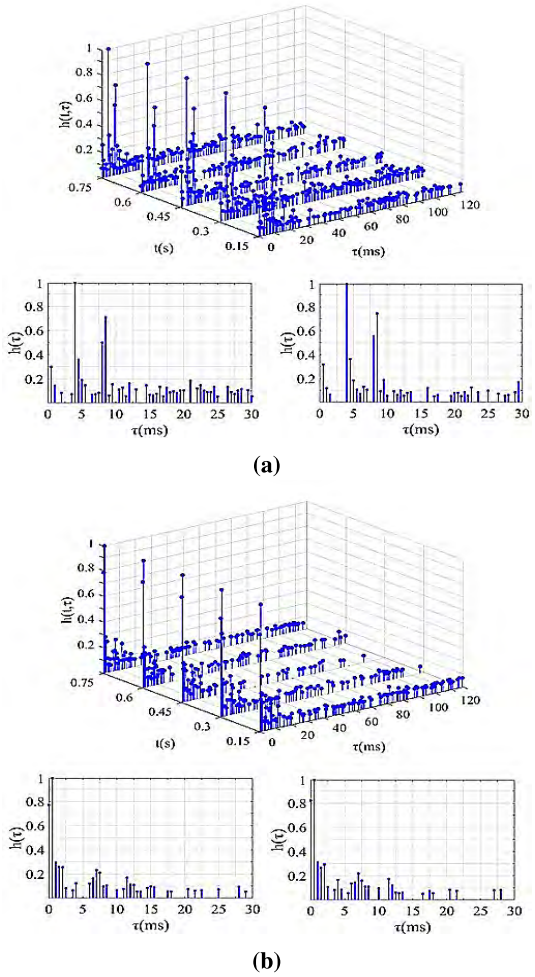
A. POOL EXPERIMENTS

The pool experiments were conducted in a 25-m-long, 6-m-wide, and 1.6-m-deep static non-anechoic water pool (Fig. 2).

In the experiments, the water depth was approximately 1 m and the depth of the transmitter and receiver was approximately 0.45 m. The locations of the transmitter and the receiver were alongside the pool wall as shown in Fig. 2b. During the pool experiments, the location of transmitter was fixed at d. The receiver moved and the experimental data were collected between d1 and d5 for the virtual learning stage.

In the online learning stage, the receiver randomly moved between the five locations d1, d2, d3, d4, and d5. The voltages applied to the transducer were changed from 0.5 V to 10 V. At each location, all types of modulation and coding methods with different transmitting powers were transmitted. The performance of the proposed RLMC scheme was testified by post-processing.

Fig. 3 showed the CIR $h(t, \tau)$ estimated at locations d1 and d3 every 0.15 s. The amplified figures in Figs. 3a and Fig. 3b respectively corresponded to $t = 0.15$ s and $t = 0.45$ s. In the pool environments, the channel didn't change at a fixed location. However, at different locations, the channel changed considerably. This would have a strong impact on the communication performance. The BER performance of different modulation and coding methods at d1 and d3 was presented in Fig. 4. The different SNRs were obtained by changing the voltage applied to the transducer.



(a)

(b)

FIGURE 3. CIR estimation at pool. (a) CIR estimation at d1. (b) CIR estimation at d3.

The performances of the presented RLMC scheme were analyzed and compared in Figs.5 and Fig. 6. In the pool experiments, we set $\omega_p = 0.1$, $\omega_T = 0.1$, $\omega_\rho = 3$, the BER service quality level vector $\Psi_i^k \in [0.02, 0.05, 0.08]$, and $(\omega_{\zeta_l})_i \in [1, 2, 3]$.

In Fig. 5, the BER performance, the energy consumption, the transmission time and the utility were given at each time slot. In Fig. 5, the red dotted lines represent the optimal theoretical values obtained by equation (10). The green lines represent the performance of the SB scheme presented in [3]. The blue lines represent the performance of the QLM scheme presented in [4]. The red lines represent the performance of our presented RLMC scheme. We can get the following results from Fig. 5:

(1): The BER performances were shown in Fig. 5a. Because the policy of the SB scheme was to maximize the throughout while maintaining the BER under a certain value (0.02) without considering the QoS of message and the energy consumption, the BERs obtained by the SB scheme at each time slot were below this value and fluctuating (shown by the green line).

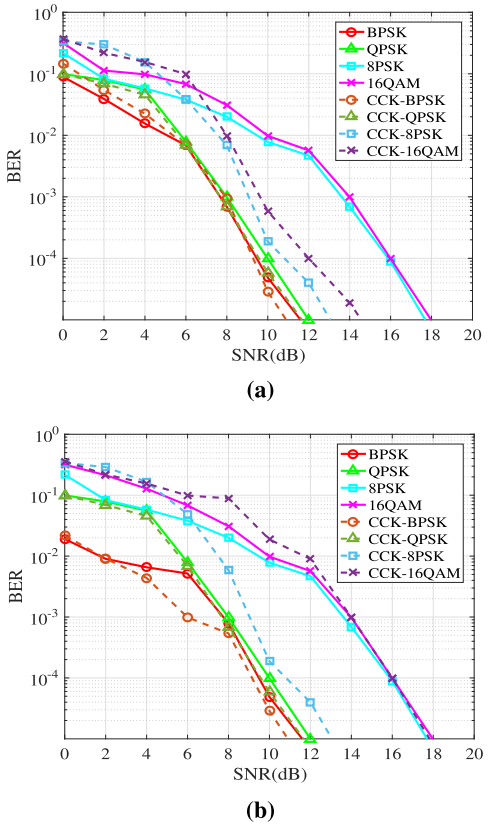


FIGURE 4. BER performance at d1 and d3. (a) Receiver at d1. (b) Receiver at d3.

(2): The BERs of the presented RLMC scheme can converge to sub-optimal values, which were just a little bit higher (0.0005) than the theoretical value. The presented RLMC scheme had the fastest learning speed (converged after 200 time slots in Fig. 5a).

(3): The performance curves of the transmission time, the energy consumption and the utility also showed the same results. For example, the utility learning speed of RLMC scheme was 45% higher than the QLM scheme, and the utility converged to sub-optimal values after 1400 time slots (below the theoretical value of approximately 0.05, in Fig. 5d). At the same time, the energy consumption and the transmission time decreased by approximately 23.5% and 53.6% respectively compared with the SB scheme.

The learning speed of BER, the transmission time, and the energy consumption were different because their costs were set manually.

Furthermore, the performance of these schemes in different underwater acoustic communication environments were analyzed and the results were shown in Fig. 6. In this experiment, the underwater acoustic communication environments were classified based on the BER performance. The BER of different modulation and coding methods under a specific channel state with a certain SNR was defined as $r_m = [r_{1,m}, \cdots, r_{j,m}, \cdots, r_{J,m}]$, where j was the serial number of

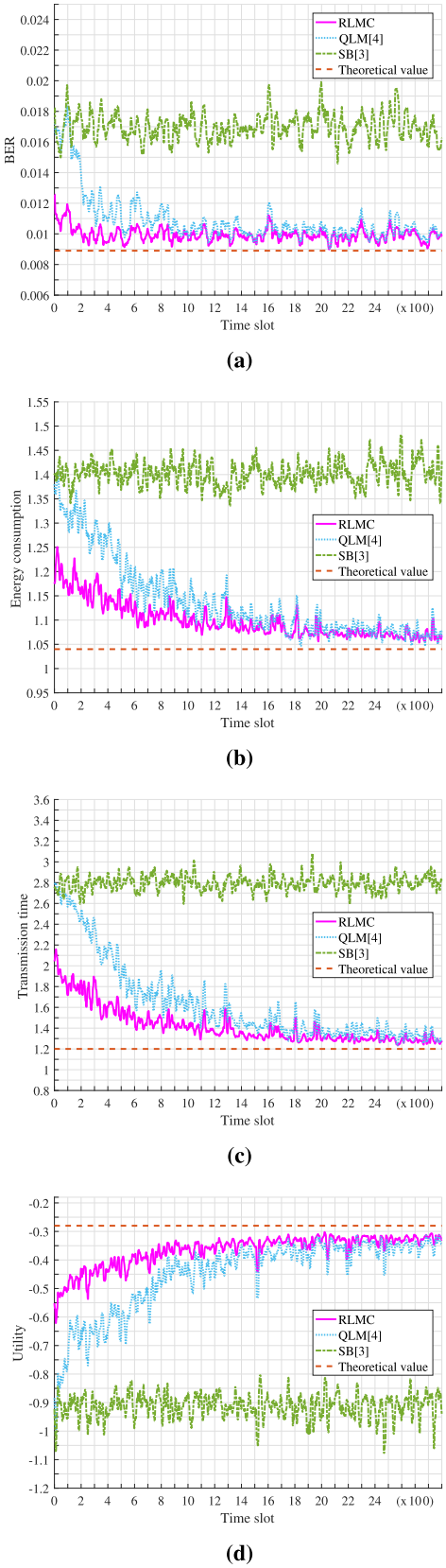
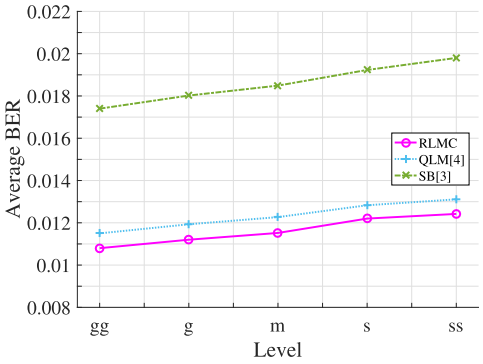
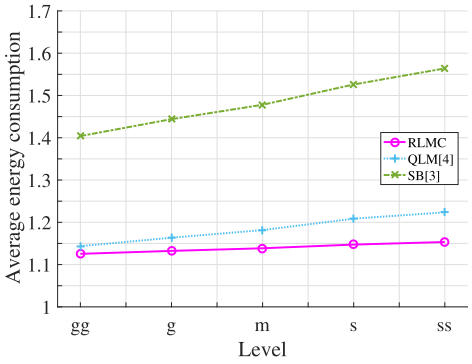


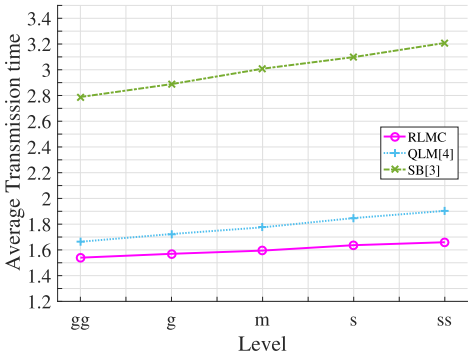
FIGURE 5. The performance of AMC schemes in pool experiments. (a) BER. (b) Energy consumption. (c) Transmission time. (d) Utility.



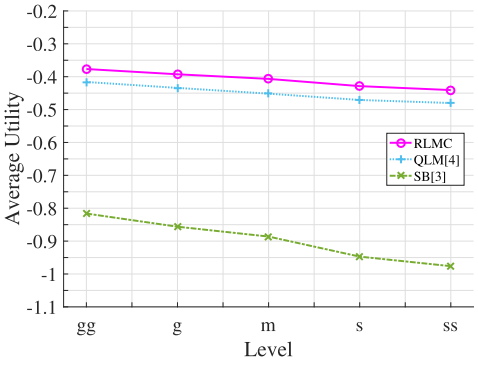
(a)



(b)

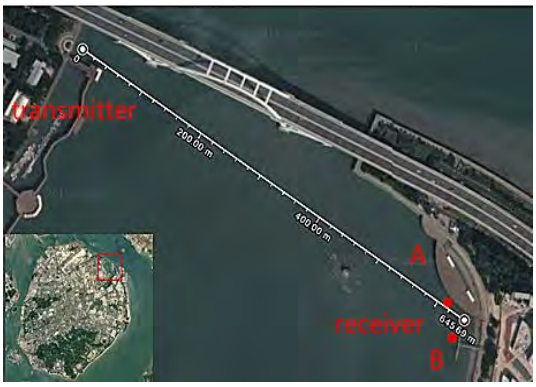


(c)



(d)

FIGURE 6. The performance of AMC schemes in different underwater acoustic communication environments. (a) Average BER. (b) Average energy consumption. (c) Average transmission time. (d) Average utility.



(a)

(b)

(c)

FIGURE 7. Sea experiments environment. (a) Sea environment. (b) Part of equipments. (c) Received signal.

modulation and coding method and m was the serial number of channel state. The median value of r_m was calculated and defined as $\dot{r}_m$. In this experiment, the environments were classified into five levels from severe to good ([ss, s, m, g, gg]) by sorting the $\dot{r}_m$ in descending order.

In each communication environment, the average BER, transmission time, energy consumption and utility were calculated after convergence and shown in Fig. 6. From Fig. 6, we can see the performance decreased in severe communication environments. However, in each communication environment, the presented RLMC scheme always had the best performance. For example, the average utility of the RLMC scheme shown in Fig. 6d was 0.05 higher than the QLM scheme in a very severe communication environment.

B. SEA EXPERIMENTS

The sea experiments were conducted in a shallow water area (Wuyuan Bay, Xiamen, China). The distance between the transmitter and the receiver was about 640 m. Both the transmitter and the receiver drifted with sea current. Under the influence of tide, the sea depth changed from 6 m to 9 m during the experiments.

Because the channels were variant in the sea experiments, the data of different modulation and coding methods could not be collected at continuous location and depth. We designed a special sea experiment to ensure that the experimental results can prove the effectiveness of the proposed RLMC scheme. This sea experiment consists of two steps. In the first step, the data were collected from a number of communication experiments. Using these data, we can

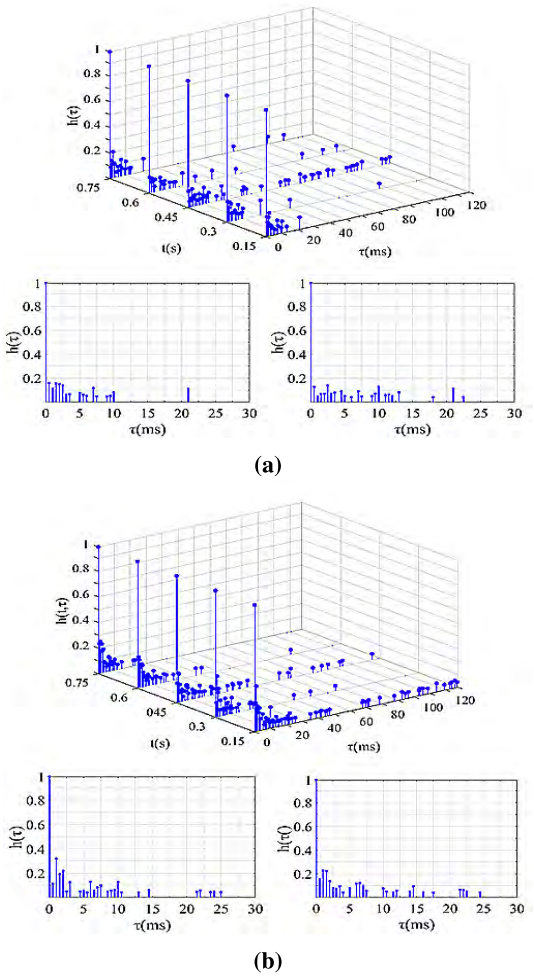


FIGURE 8. CIR estimation in sea environments. (a) CIR estimation at location A. (b) CIR estimation at location B.

calculate the performance of each modulation and coding method under a specific quantified channel state. In the second step, the transmitter transmitted the channel estimation signals continuously. At the same time, the receiver moved randomly. Using these received data, we can determine the estimated channel states at each time slot. In post-processing, the changing channel states were obtained from continuous mobile communication experiments. Also, the performance of different modulation and coding methods on a specific quantified channel was known. Thus, all the conditions required to prove the performance of the RLMC algorithm were guaranteed.

Fig. 8 presented the CIRs estimated every 0.15 s. The amplified figures of Fig. 8a and Fig. 8b respectively corresponded to $t = 0.15$ s and $t = 0.45$ s. From the amplified figures, we can see that the CIRs changed even in a small-time window (0.3 s). We can also see that only parts of the CIRs have changed and the future CIRs were not totally uncorrelated with the current CIRs. Thus, the future CIRs can be modeled by equation (1). Fig. 9 presented the BER performance of the selected modulation and coding methods

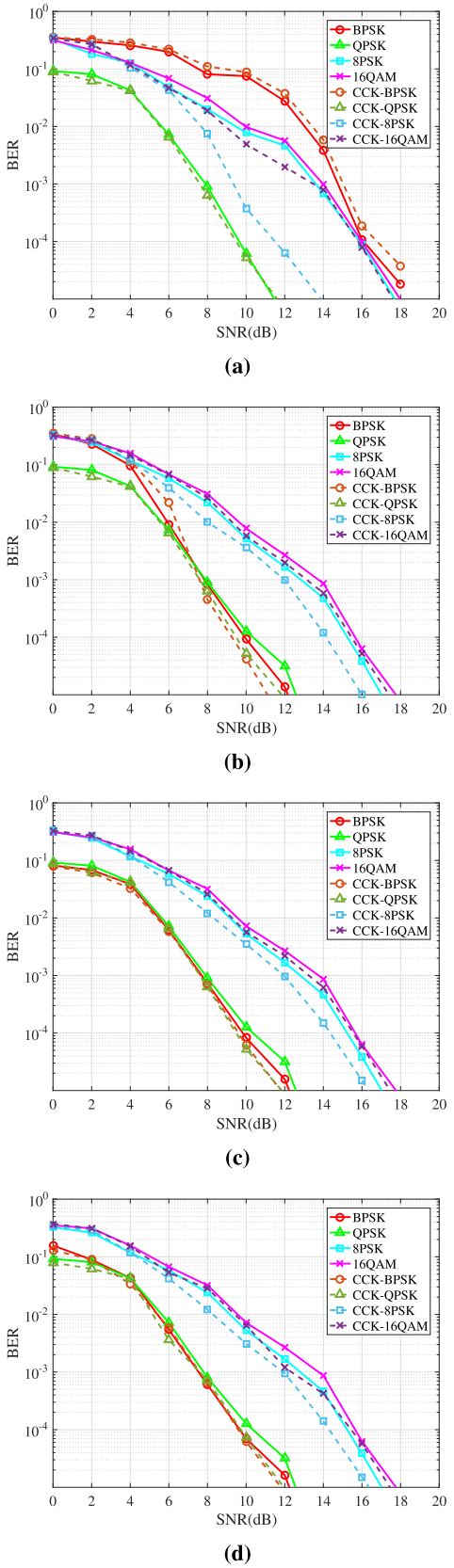


FIGURE 9. BER performance at A and B. (a) At location A, $t = 0.15$ s. (b) At location A, $t = 0.45$ s. (c) At location B, $t = 0.15$ s. (d) At location B, $t = 0.45$ s.

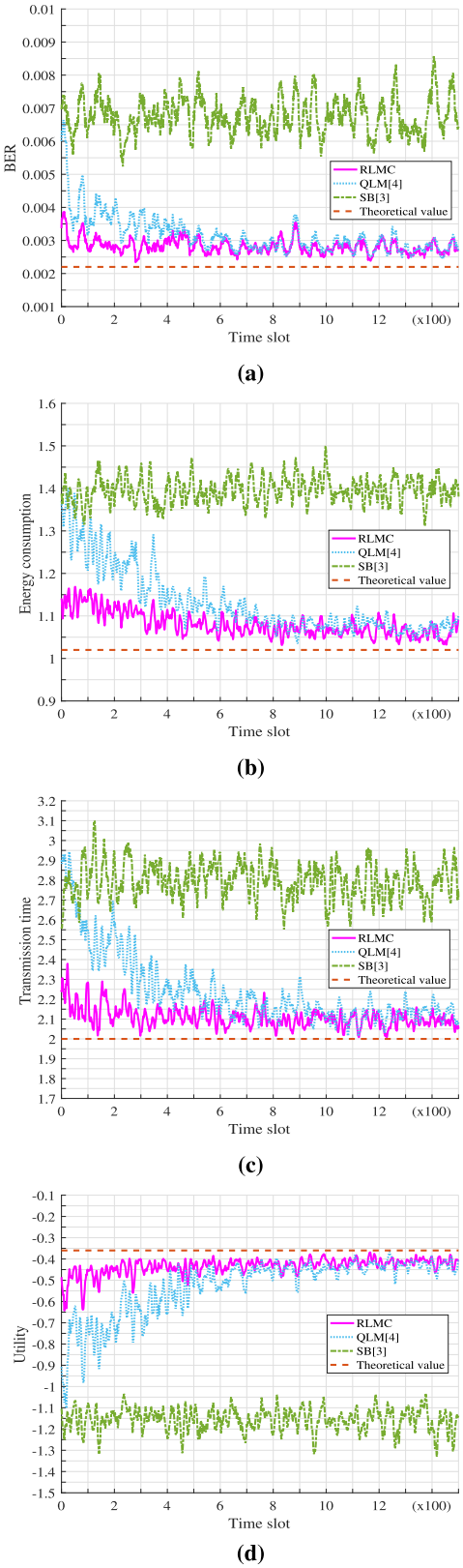


FIGURE 10. The performance of AMC schemes in sea experiments. (a) BER. (b) Energy consumption. (c) Transmission time. (d) Utility.

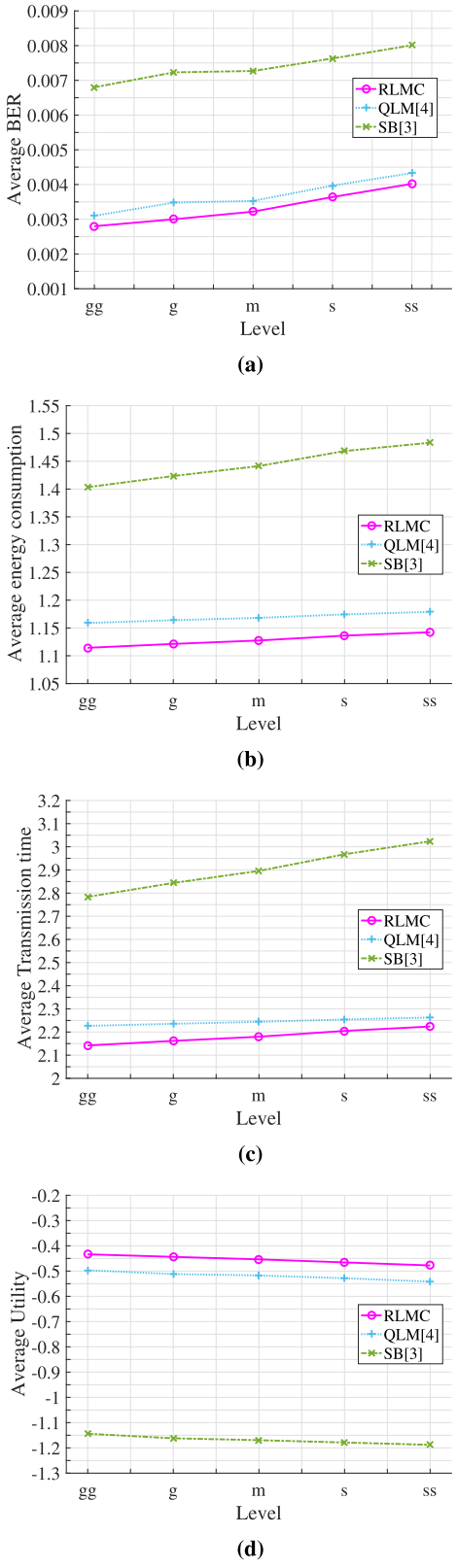


FIGURE 11. The performance of AMC schemes in typical sea environments. (a) Average BER. (b) Average energy consumption. (c) Average transmission time. (d) Average utility.

under different channel states. The performance of a specific modulation and coding method varied under these channels.

In sea experiments, we set $\omega_p = 0.1$, $\omega_T = 0.1$, $\omega_\rho = 2$, $\Psi_i^k \in [0.005, 0.01, 0.05, 0.08]$ and $(\omega_{\zeta_i})_i \in [1, 2, 3, 4]$. In Fig. 10, the BER performance, the energy consumption, the transmission time and the utility were given at each time slot.

From Fig. 10 we can see that the BERs of the presented RLMC scheme converged to sub-optimal values and just a little bit higher (0.0015) than the theoretical value; The energy consumption and the transmission time were 5% and 0.25% higher than the theoretical values. Moreover, the presented RLMC scheme had a higher learning speed than the QLM scheme. Compared with pool experiments, the performance curves fluctuated stronger in sea environments. The reason was that the channel states changed much faster in sea environments.

Furthermore, the performances of these schemes in typical sea environments were analyzed, and the results were shown in Fig. 11. The average BER, transmission time, energy consumption and utility were calculated after convergence. In Fig. 11a, the average BER, transmission time and the energy consumption increased under severe channel conditions, and the utility decreased. However, in each sea environment, the presented RLMC scheme had the best performance. For example, the average utility of the RLMC scheme shown in Fig. 11d was 0.05 higher than the QLM scheme.

In general, all the performance curves of sea experimental results converged to sub-optimal values and the results of sea experiments were similar to pool experiments. All of these experimental results can prove the performance of the proposed RLMC scheme.

VI. CONCLUSIONS

In this paper, we have proposed a reinforcement learning-based adaptive modulation and coding scheme to improve the efficiency of underwater communications in terms of BER, transmission time and the energy consumption of the transmitter according to the QoS requirements. We proved that this scheme enables an underwater transmitter to jointly optimize its modulation and coding policy without being aware of the underwater channel model and provided its convergence performance bound. Experiments performed both in a static non-anechoic water pool and a bay area to evaluate its performance and compared with two benchmark schemes. Experimental results verify the analysis results and show that our proposed scheme increases the throughput, decreases the BER and the transmission time, and saves the energy consumption compared with the SB scheme and the QLM scheme. For instance, the proposed scheme reduces the BER by approximately 44%, saves the transmission time by approximately 53%, saves the energy consumption by approximately 25%, thus increases the utility by approximately 63% compared with SB scheme in pool experiments

REFERENCES

[1] M. Chitre, S. Shahabudeen, L. Freitag, and M. Stojanovic, "Recent advances in underwater acoustic communications networking," in *Proc. OCEANS*, Sep. 2008, pp. 1–10.

[2] M. N. Rajesh, B. K. Shrisha, N. Rao, and H. V. Kumaraswamy, "An analysis of BER comparison of various digital modulation schemes used for adaptive modulation," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 241–245.

[3] M. Huda, N. B. Putri, and T. B. Santoso, "OFDM system with adaptive modulation for shallow water acoustic channel environment," in *Proc. IEEE Int. Conf. Commun., Netw. Satell. (Comnetsat)*, Oct. 2017, pp. 55–58.

[4] J. Lin, W. Su, L. Xiao, and X. Jiang, "Adaptive modulation switching strategy based on Q-learning for underwater acoustic communication channel," in *Proc. 13th ACM Int. Conf. Underwater Netw. Syst.*, Dec. 2018, pp. 38-1–38-5. [Online]. Available: http://doi.acm.org/10.1145/3291940.3291976

[5] L. Xiao, D. Jiang, X. Wan, W. Su, and Y. Tang, "Anti-jamming underwater transmission with mobility and learning," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 542–545, Mar. 2018.

[6] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.

[7] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[8] F. Campagnaro, F. Guerra, F. Favaro, V. S. Calzado, P. Forero, M. Zorzi, and P. Casari, "Simulation of a multimodal wireless remote control system for underwater vehicles," in *Proc. 10th Int. Conf. Underwater Netw. Syst.*, Oct. 2015, pp. 32-1–32-8. [Online]. Available: http://doi.acm.org/10.1145/2831296.2831298

[9] R. Diamant, P. Casari, F. Campagnaro, O. Kebkal, V. Kebkal, and M. Zorzi, "Fair and throughput-optimal routing in multimodal underwater networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1738–1754, Mar. 2018.

[10] A. Svensson, "An introduction to adaptive QAM modulation schemes for known and predicted channels," *Proc. IEEE*, vol. 95, no. 12, pp. 2322–2336, Dec. 2007.

[11] L. Wan, H. Zhou, X. Xu, Y. Huang, S. Zhou, Z. Shi, and J.-H. Cui, "Adaptive modulation and coding for underwater acoustic OFDM," *IEEE J. Ocean. Eng.*, vol. 40, no. 2, pp. 327–336, Apr. 2015.

[12] A. Radosevic, R. Ahmed, T. M. Duman, J. G. Proakis, and M. Stojanovic, "Adaptive OFDM modulation for underwater acoustic communications: Design considerations and experimental results," *IEEE J. Ocean. Eng.*, vol. 39, no. 2, pp. 357–370, Apr. 2014.

[13] M. Sadeghi, M. Elamassie, and M. Uysal, "Adaptive OFDM-based acoustic underwater transmission: System design and experimental verification," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw.*, Jun. 2017, pp. 1–5.

[14] K. Pelekanakis, L. Cazzanti, G. Zappa, and J. Alves, "Decision tree-based adaptive modulation for underwater acoustic communications," in *Proc. IEEE 3rd Underwater Commun. Netw. Conf.*, Aug. 2016, pp. 1–5.

[15] K. Pelekanakis and L. Cazzanti, "On adaptive modulation for low SNR underwater acoustic communications," in *Proc. OCEANS MTS/IEEE Charleston*, Oct. 2018, pp. 1–6.

[16] S. Basagni, V. D. Valerio, P. Gjanci, and C. Petrioli, "Finding MARLIN: Exploiting multi-modal communications for reliable and low-latency underwater networking," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[17] W. Zhang, Y. Liu, G. Han, Y. Feng, and Y. Zhao, "An energy efficient and QoS aware routing algorithm based on data classification for industrial wireless sensor networks," *IEEE Access*, vol. 6, pp. 46495–46504, 2018.

[18] G. Han, X. Long, C. Zhu, M. Guizani, Y. Bi, and W. Zhang, "An AUV location prediction-based data collection scheme for underwater wireless sensor networks," *IEEE Trans. Veh. Technol.*, to be published.

[19] G. Han, X. Long, C. Zhu, M. Guizani, and W. Zhang, "A high-availability data collection scheme based on multi-AUVs for underwater sensor networks," *IEEE Trans. Mobile Comput.*, to be published.

[20] A. Li, F. H. Panahi, T. Ohtsuki, and G. Han, "Learning-based optimal channel selection in the presence of jammer for cognitive radio networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[21] Y. Liu, C. Yuen, N. U. Hassan, S. Huang, R. Yu, and S. Xie, "Electricity cost minimization for a microgrid with distributed energy resource under different information availability," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2571–2583, Apr. 2015.
[22] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.
[23] T. Hu and Y. Fei, "MURAO: A multi-level routing protocol for acoustic-optical hybrid underwater wireless sensor networks," in *Proc. 9th Annu. IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw. (SECON)*, Jun.2012, pp. 218–226.
[24] T. C. Yang, "Temporal resolutions of time-reversal and passive-phase conjugation for underwater acoustic communications," *IEEE J. Ocean. Eng.*, vol. 28, no. 2, pp. 229–245, Apr. 2003.
[25] K. F. Scussel, J. A. Rice, and S. Merriam, "A new MFSK acoustic modem for operation in adverse underwater channels," in *Proc. MTS/IEEE Conf.*, Oct. 1997, pp. 247–254.
[26] D. Wang, X. Hu, W. Su, X. Jiang, and Y. Xie, "Research on multi-channel time frequency shift keying for underwater acoustic communication," in *Proc. MTS/IEEE Washington OCEANS*, Oct. 2015, pp. 1–5.
[27] J. Cai, X. Jiang, H. Huang, Q. Ding, and W. Su, "A non-real-time underwater acoustic communication system based on WeChat," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Oct. 2017, pp. 1–4.
[28] J.-W. Yin, S. Yang, D. Yao, and X. Zhang, "Study of underwater acoustic communication based on vector Pattern time delay shift coding," in *Proc. MTS/IEEE SEATTLE OCEANS*, Sep. 2010, pp. 1–6.
[29] Y. Wu, H. Song, and F. Xu, "Time delay estimation in underwater positioning for pattern time delay shift coding," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, Dec. 2013, pp. 1618–1622.
[30] M. Stojanovic, J. A. Catipovic, and J. G. Proakis, "Phase-coherent digital communications for underwater acoustic channels," *IEEE J. Ocean. Eng.*, vol. 19, no. 1, pp. 100–111, Jan. 1994.
[31] F. Liu, W. Cui, and X. Li, "China's first deep manned submersible, JIAOLONG," *Sci. China Earth Sci.*, vol. 53, no. 10, pp. 1407–1410, Oct. 2010.
[32] J. Han, S. P. Chepuri, Q. Zhang, and G. Leus, "Iterative per-vector equalization for orthogonal signal-division multiplexing over time-varying underwater acoustic channels," *IEEE J. Ocean. Eng.*, vol. 44, no. 1, pp. 240–255, Jan. 2019.
[33] J. Han, L. Zhang, Q. Zhang, and G. Leus, "Low-complexity equalization of orthogonal signal-division multiplexing in doubly-selective channels," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 915–929, Feb. 2019.
[34] M. Stojanovic, "Adaptive channel estimation for underwater acoustic MIMO OFDM systems," in *Proc. IEEE 13th Digit. Signal Process. Workshop*, Marco Island, FL, USA, Jan. 2009, pp. 132–137.
[35] P. C. Carrascosa and M. Stojanovic, "Adaptive channel estimation and data detection for underwater acoustic MIMO–OFDM systems," *IEEE J. Ocean. Eng.*, vol. 35, no. 3, pp. 635–646, Jul. 2010.
[36] M. Biagi, S. Rinauro, and R. Cusani, "Channel estimation or prediction for UWA," in *Proc. MTS/IEEE OCEANS*, Jun. 2013, pp. 1–7.

WEI SU received the Ph.D. degree from the Communication Engineering Department, Northwestern Polytechnic University, in 2009.

He is currently an Assistant Professor with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Xiamen University, Ministry of Education, Xiamen, China. His research interests include the general area of underwater acoustic communication and networking, spanning from the communication networks, and multimedia signal processing and communication.

JIAMIN LIN received the bachelor's degree in electronic information engineering from Fujian Normal University. She is currently pursuing the master's degree with the Key Lab of Underwater Acoustic Communication and Marine Information Technology, Xiamen University, Ministry of Education, Xiamen, China. Her research interest includes the general area of underwater acoustic communication and networking.

KEYU CHEN received the Ph.D. degree in communication engineering from Xiamen University, Ministry of Education, Xiamen, China, in 2013, where he is currently an Engineer with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology. His current research interests include underwater acoustic communication and networking, cross-layer design for localization, and mobile communication.

LIANG XIAO received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 2009.

She was a Visiting Professor with Princeton University, Virginia Tech, and the University of Maryland, College Park, MA, USA. She is currently a Professor with the Department of Communication Engineering, Xiamen University, Xiamen, China.

Dr. Xiao was a recipient of the Best Paper Award for 2016 INFOCOM Big Security WS and 2017 ICC. She has served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

CHENG EN received the Ph.D. degree from the Communication Engineering Department, Xiamen University, Ministry of Education, Xiamen, China, in 2006, where he is currently a Professor with the Communication Engineering Department and also the Director of the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Xiamen University. His research interests include the general area of underwater acoustic communication and networking, spanning from the communication networks, multi-media signal processing and communication, video/image quality measurement, and embedded systems design.