Q-Learning Based Adaptive Channel Selection for Underwater Sensor Networks

Antony Pottier[†], Paul D. Mitchell[‡], Francois-Xavier Socheleau[†], Christophe Laot[†]
[†]IMT-Atlantique, Lab-STICC, UBL, [‡]University of York, Department of Electronic Engineering
{antony.pottier,fx.socheleau,christophe.laot}@imt-atlantique.fr, paul.mitchell@york.ac.uk

Abstract—**In this paper, we provide self-configuration and adaptation capabilities to UWSN thanks to Q-learning. UWSN deployed for the long term over large areas for environmental monitoring are possible applications of our work. Sensor nodes deployed on the sea bottom are devoted to measure a physical quantity of interest transmitted to surface buoys considered as access points. Packet transmission are asynchronous and low overheads are desirable so as to save throughput and battery life. Prior to a transmission, the nodes choose, depending on the channel conditions, which access point maximizes the probability of successful decoding a the receiver side. Results show that Q-learning is able to perform close to an ideal "genie-aided" scheme, without the need of a detailed knowledge on the environment.**

I. Introduction

Underwater sensor networks (UWSN) have drawn the attention of the underwater acoustic (UWA) communications research community for more than two decades. As the number of surveys on the topic can testify [1]–[8], research has been conducted in many fields from the physical to the routing layer and a wide variety of applications have been investigated. The specificity of the UWA environment calls for the development of dedicated protocols, such as those proposed in [9]–[12]. Furthermore, the UW channel varies on different time scales [13], [14] and the transmission ranges and depths may change due to mobility. Therefore, some adaptation capabilities of the modems and networks are desirable.

Adaptive medium access (MAC) protocols have been proposed in [15], [16]. Synchronization between transmitters is assumed in [15], which is difficult to maintain underwater. In [16] a delay-tolerant handshaking mechanism is implemented, which necessitates some overheads. More recently, [17]–[21] have investigated decentralized spectrum sharing between non-cooperative UW communication links. Communications take place outside any network, without protocol, synchronization or information exchanges between different links. Spectrum sharing is achieved by implementing an equilibrium strategy of a properly defined game. A limitation is the possible unfairness and inefficiency of noncooperative equilibria. Recently, reinforcement learning (RL) [22], [23] has successfully provided adaptive MAC layers for radio wireless sensor networks [24]–[26]. In these works, Q-Learning is used to enable dynamic spectrum access in LTE networks [24] and to schedule TDMA-based MAC schemes [25], [26]. In the UWA community, [27], [28] use Q-Learning to adapt transmission parameters to the temporal variations of the channel in a single user context.

In this paper, we exploit similar ideas for application in UWSN where asynchronism, low overheads and adaptability are desirable. Many things can motivate asynchronous operations in UWA communications: lack of GPS, low capacity links and long propagation delays which inhibits signaling for distributed synchronizations, drift/motion of devices, etc. Low overheads improve data throughput and battery life. Self-configuration and adaptation capabilities are also needed so as to enable operations in dynamic environments. UWSN deployed for the long term over large areas for environmental monitoring are possible applications of our work. Sensor nodes deployed on the sea bottom are devoted to measure a physical quantity of interest transmitted to surface buoys or sink nodes. To allow for adaptability, we consider nodes with choices of transmission strategies. This choice could be an access point (like a surface buoy or a sink node), a frequency bandwidth, a time slot, etc .., or a combination of different parameters. As an example, in this paper we focus on choices of access points. This defines the actions sets of the nodes considered as learning agents. Transmissions are asynchronous and data packets might collide at the reception. However, the signal-to-interference-plus-noise-ratio (SINR) may still be sufficient to decode the packet of interest, depending on the path losses suffered. It is considered that decoding all the packet is not sensible, but we rather want to maximize the number of successfully transmitted packets. This assumption is made in the light of works on random access compressed sensing [29], [30] where the phenomenon under monitoring admits a sparse representation in the spatial domain, allowing for reconstruction of the measures without all the packets to be received correctly. When a node transmits a data packet, it must choose which access point is best to establish a reliable communication depending on the channel conditions.

The proposed algorithm enables nodes to autonomously make choices of access points in order to maximize their probability of successful transmission. Q-learning [23] offers a way to allow nodes to learn about the environment and react to changes in the network topology, channel fading statistics or transmission geometry. Overheads are minimal as it does not require synchronization or information exchange between nodes or between nodes and their receivers, apart from a 1-bit feedback signal from the receivers. We evaluate the performance in terms of the average probability of successful transmission through simulations and compare with the ideal adaptive channel selection scheme which would have perfect knowledge of the channels before every transmissions.

978-1-5386-6442-1/18/$31.00 ©2018 IEEE

Average path losses are simulated with Bellhop [31] This give an average channel gain around which random log-normal fluctuations are added to simulate time-varying UWA channels.

II. Transmission models

A set $\mathcal{I}$ of I non-cooperative transmitters (TXs) is supposed to transmit packets to a set $\mathcal{A}$ of A receivers (RXs). The RXs are access points such as surface buoys, sinks or relay nodes, and they collect data from sensor nodes considered as TXs. Each TX $i \in \mathcal{I}$ accesses only one RX $a \in \mathcal{A}$ at a time, and may eventually cause interference to the others. It is assumed that $I > A$, so that RXs are accessed by several concurrent TXs, which may also interfere with each other when transmitting at the same time. The receivers decode each transmitter separately using a single-user decoding (SUD) scheme, the others being considered as an unknown interference.

Each TX generates, asynchronously and independently of the others, λ packets per second whose inter-arrivals are exponentially distributed. On average, there are $I \times \lambda$ packets per second transmitted within the network. For each packet generated, the corresponding TX must choose an RX so as to maximize the probability of successful transmission. Direct sequence spread spectrum (DSSS) communications are considered, so as to enhance transmissions robustness by taking advantage of the processing gain [32], [33]. A spreading code is associated with each RX. All the TXs know the spreading codes so that they can choose an RX by choosing the sequence with which they spread their signals. We suppose all the sequences to have the same length and the same spreading gain G. The RXs are able to decode only packets spread with their own sequences. Transmissions are asynchronous, so that the codes of the different RXs cannot be considered as perfectly orthogonal. Thus, the TXs may interfere with all the RXs.

Let T_p denote the packet duration and consider a TX $i \in \mathcal{I}$ transmitting a packet to the RX a. The packet is transmitted with power P_i uniformly spread on a bandwidth B centered on f_c, and arrives at RX a at a time t_i. It has incurred an attenuation depending on the time-varying transfer function of the channel between i and a. The UWA channel time variations can be decomposed into two parts [34], [35]: a large-scale fading component, which makes the received power vary slowly compared to the packet length, and a small-scale fading whose coherence time is much shorter than T_p. Let $H_{i,a}(t, f)$ be the randomly time-varying channel transfer function between TX i and RX a. The average attenuation on the packet arrived at time t_i can be expressed as

$$\bar{\rho}_{i,a}(t_i) = \frac{1}{T_p} \int_{t_i}^{t_i+T_p} \rho_{i,a}(t) \mathrm{d}t \tag{1}$$

where

$$\rho_{i,a}(t) = \frac{1}{B} \int_{f_c-B/2}^{f_c+B/2} |H_{i,a}(f,t)|^2 \, \mathrm{d}f \tag{2}$$

is a realization of a random process which depends on the fading processes underlying $H_{i,a}(f, t)$.

During the reception of a packet from TX i, collisions from others TXs might occur as long as it exists some $j \neq i \in \mathcal{I}$ such that the arrival time of its packet at the RX chosen by i, denoted by t_j, is within $[t_i - T_p, t_i + T_p]$. Let the interference power perceived on the TX i's packet denoted by the random variable

$$I_{i,a}(t_i) = \frac{1}{T_p} \int_{t_i}^{t_i+T_p} \sum_{j \neq i} \rho_{j,a}(t) P_j \mathbb{1}_{[t_j, t_j+T_p]}(t) \, \mathrm{d}t \tag{3}$$

which also depends on the channels fading processes, as well as on the arrival times of all the packets generated.

The successful transmission of a packet by the TX i to the RX a will be evaluated through the average SINR on the packet

$$\gamma_{i,a}(t_i) = \frac{\bar{\rho}_{i,a}(t_i) P_i}{\sigma_a^2 + I_{i,a}(t_i)} \times G, \tag{4}$$

where σ_a^2 is the ambient noise power at the RX a. It is considered that the packet of TX i is successfully decoded at the chosen RX a if $\gamma_{i,a} \geq \Gamma_i$, where Γ_i is some SINR constraint.

In a practical set-up, the SINR should be estimated at the RX side using pilot symbols included within the packets. The choice of a satisfaction criterion based on the SINR is not mandatory since the type of metric chosen does not have an influence on the general behavior of the proposed algorithm. One could choose, for example, a bit error rate constraint on the packet or whether the packet is correctly decoded using a cyclic redundancy check. Therefore, the assumption of perfect knowledge of the SINR is not sensible in what will be presented next. However, the choices of the values for rewards/punishments of successful/failed transmissions (see (5)) may have an influence but these discussions go beyond the scope of this paper. These values will be chosen in a very pragmatic manner in the next section.

III. Channel selection with Q-Learning

We propose to use the RL algorithm called Q-Learning to enable the nodes to autonomously choose the receivers which maximize their probability of successful transmission. The TXs are considered as independent agents having several actions to their disposal. In the scenario previously described, the possible actions correspond to a choice of a RX in $\mathcal{A} = \{1, \cdots, A\}$ to transmit the current packet at a given time. As each RX is associated to a spreading sequence, to choose an RX is equivalent to choose a spreading sequence, which can be written in the sensors memory prior to deployment. For all $i \in \mathcal{I}$, the action space is thus defined as $\mathcal{A}_i = \mathcal{A}$. In order to ease the presentation, we take the point of view of a particular agent i in the following.

Let $\mathcal{T}_i = 0, 1, 2, \cdots$ be the set of time indexes (possibly mapped to the real line) at which TX i 's packets arrive at any RX (neglecting the propagation delays). The algorithm proceeds as follows. The TX first chooses the RX $a_{i,t} \in \mathcal{A}$ to sent the packet which arrives at time $t \in \mathcal{T}_i$. Note that the packet was sent at time $t - \Delta_{i,a_i}$, where Δ_{i,a_i} is the propagation delay between TX i and RX a_i. The RX is supposed able to compute the SINR of the agent i on the basis of Eq (4). The agent i receives a reward $R_{i,t+1}$ which depends on whether the SINR contraint is met or not. The reward is thus defined as

$$R_{i,t+1} \triangleq \begin{cases} +1 & \text{if } \gamma_i(a_{i,t}) \geq \Gamma_i \\ -1 & \text{if } \gamma_i(a_{i,t}) < \Gamma_i \end{cases} \tag{5}$$

and can take the form of a 1 bit feedback from the RX $a_{i,t}$ to the TX i to acknowledge the successful reception of the packet. On the basis of the cumulated received rewards, the agent will update the action chosen for the next packets by favoring those that generate positive rewards the most often.

The agents do not know what can be the channel gains and the arrival times of other packets, and they are not supposed to observe them. So, from the point of view of agent i and for a given action a_i, $R_{i,t+1}$ is a random variable with unknown distribution. The agent i seeks to maximize its expected reward through a judicious choice of action, *i.e.* it solves

$$\max_{a_{i,t} \in \mathcal{A}_i} \mathbb{E}\left[R_{i,t+1}\right]. \tag{6}$$

The expected rewards must be estimated. Let $Q_{i,t}$ be an estimator of $\mathbb{E}\left[R_{i,t+1}\right]$ when considered as a function of a_i, the action taken at time t and for which a reward $R_{i,t}$ is received. For all $a_i \in \mathcal{A}_i$

$$\begin{aligned} Q_{i,t}(a_i) &= Q_{i,t-1}(a_i) + \alpha\left[R_{i,t} - Q_{i,t-1}(a_i)\right] \\ &= (1-\alpha)^{t-1} Q_{i,0} + \sum_{k=1}^{t-1} \alpha(1-\alpha)^{t-k} R_{i,k} \mathbb{1}_{[a_{i,k}=a_i]}. \end{aligned} \tag{7}$$

with $\alpha \in [0,1]$ the step size (or learning rate) parameter and $Q_{i,0}$ an initial value. By maintaining a table with Q-Values $Q(a_i)$ associated to each action $a_i \in \mathcal{A}_i$ and updating it over time, an agent can devise how worth it is to play a given action. The step-size parameter is used to weight differently the new estimates compared to the old ones, allowing to track non-stationary problems [22]. By selecting actions and getting rewards accordingly, the agent learns about the environment. This is expressed through the successive improvements in the estimation of the expected rewards by the Q-Values. When the agent exploits its current knowledge, its selects the "greedy" action at the time considered :

$$a_{i,t} = \underset{a_i \in \mathcal{A}_i}{\operatorname{argmax}}\, Q_{i,t-1}(a_i). \tag{8}$$

Nevertheless, it is necessary to try each action sufficiently often to guarantee a good estimation of the Q-Values. Exploration is thus needed. Here, we consider the exploration strategy consisting in exploring with probability ϵ and choosing the greedy action with probability $1-\epsilon$. Algorithm 1 sums up the procedure previously described (the agent's index is omitted). If several Q-Values are maxima, we randomize between the corresponding actions.

IV. Numerical Results

The proposed scheme is evaluated through simulations. A set $I = 80$ TXs and $N = 8$ RXs are considered. The transmitters can be considered, for example, as sensor nodes immersed at the sea bottom and the receivers are surface buoys. Communications take place on a bandwidth $B = 8$ kHz centered at $f_c = 12$ kHz. Each node sends packet at a rate $\lambda = 0.1$. There are thus 8 packets per seconds on average transmitted by the whole set of nodes to the 8 buoys. Packets are spread with a DSSS sequence of $L = 128$ chips, offering a processing gain of $G = 21$ dB. We approximate the chip duration by $T_{chip} \approx 1/B = 0.125$ μs so that the symbol duration is $L \times T_{chip} = 16$ μs. The packet duration is set to $T_p = 1$ s, so that 62 symbols are transmitted per packets. In a practical set-up where the sensors are devoted to measure a physical quantity, such a packet length can be sufficient to transmit one or several measurements, depending on the constellation size. The terminals locations are randomly drawn in an area of 64 km^2, with a minimum distance separating TXs and RXs of 100 m and 500 m respectively. The water depth is 1000 m. TXs are immersed randomly between 1000 and 800 meters, while RXs are immersed between 5 and 20 meters.

Algorithm 1 Q-Learning

```
parameters: $\epsilon, \alpha$
$t = 0$
$\forall\ a_i \in \mathcal{A}_i \quad Q_{i,t}(a_i) = 0$
for $t = 1, 2, \cdots$ do
    $a_{i,t} = \begin{cases} \underset{a_i \in \mathcal{A}_i}{\operatorname{argmax}}\, Q_{i,t-1}(a_i) & \text{w. p. } 1-\epsilon \\ \text{a random action} & \text{w. p. } \epsilon \end{cases}$
    $R_{i,t} \leftarrow$ reward for $a_{i,t}$ based on Equation (5)
    $Q_{i,t}(a_{i,t}) = Q_{i,t-1}(a_{i,t}) + \alpha\left[R_{i,t} - Q_{i,t-1}(a_{i,t})\right]$
end for
```

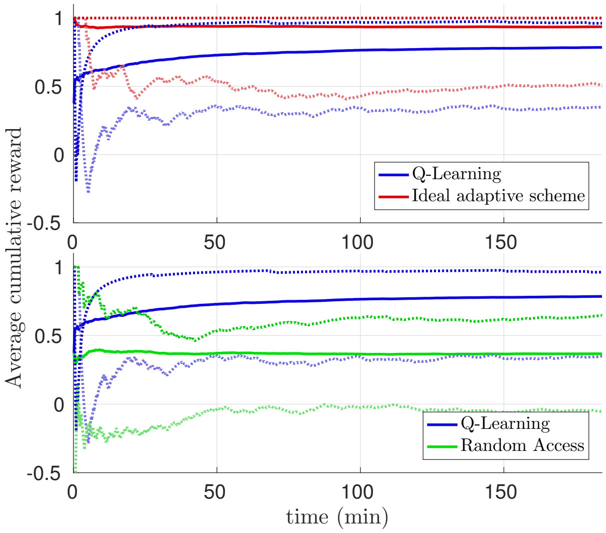

Fig. 1. Average cumulative rewards over time for channel selections based on Q-Learning and two other schemes. Solid lines depicts the performance averaged on the number TXs. Dashed lines are minimum and maximum performance among TXs.

The channels gains coefficients $\rho_{i,n}(t)$ are simulated as order 1, log-normal auto-regressive processes such that $\forall\ i \in \mathcal{I}$, $\forall\ n \in \mathcal{N}$ (with t understood as discrete time indexes in the following):

$$\rho_{i,n}(t) = 10^{\frac{1}{10}(g_{i,n}(t)+\bar{g}_{i,n})} \tag{9}$$

where

$$g_{i,n}(t) = \phi\, g_{i,n}(t-1) + \epsilon_{i,n}(t) \tag{10}$$

with $\epsilon_{i,n}(t) \sim \mathcal{N}(0, \sigma_{dB}^2)$. This models the large-scale fading of the channel. The constant ϕ is computed so as to have a coherence time of $\tau_c = 60$ s and the power spread is $\sigma_{dB}^2 = 10$ dB. The large-scale fading process is centered around a mean $\bar{g}_{i,n}$ corresponding to the transmission loss returned by the Bellhop ray-tracing simulator [31]. This simulator traces acoustic rays from a TX to an RX given the transmission geometry parameters (depths, ranges) and a sound speed profile. This SSP was acquired in North Atlantic at longitudes $[-70°, -60°]$ and latitudes $[22°, 30°]$[1] and was truncated at a 1 km depth for the needs of the simulation. An impulse response is then computed on the basis of the ray tracing. Transmission losses can then be computed by integrating over frequencies to give the coefficient $\bar{g}_{i,n}$. The transmission power is constrained to 170 dB ref μPa and the noise power is computed according to the path loss to have a reference signal to noise ratio (SNR) of 10 dB at 1 km. The SINR constraint is set to $\Gamma_i = 10$ dB.

Each node runs Algorithm 1. The step-size is set to $\alpha = 0.1$ and the exploration is $\epsilon = 0.05$. Usually, ϵ is small so as to benefit of exploitation when the algorithm as converged to consistent choices. The step-size is set empirically here, as several simulations have shown no sensibility regarding the average long-term cumulated rewards. This is explained by the channels stationarity when considered on sufficiently long duration, as the log-normal fading always fluctuates around the same average path loss. When a packet is received by the chosen RX, the SINR is computed with Equation (4) and the corresponding reward is sent back to the TX for updating its action. All propagation delays are taken into account.

Performance is evaluated in terms of the average cumulated rewards. Comparisons are made with the naive random (uniform) selection of RXs and with a "genie-aided" adaptive scheme where, before a packet is sent, the TXs know perfectly what will be the channel gains at the RXs side and choose the one with the best gain. This ideal scheme would necessitate information exchanges between TXs and RXs prior to every packet transmission, which would produce large overheads. Nevertheless this a reasonable upper-bound for comparison. Figure 1 shows the results. It can be seen that the Q-Learning selection is beneficial compared to the random access procedure, as expected. Most importantly, it performs quite close to the ideal scheme but with much less knowledge required about the environment (only the rewards encoded into a 1-bit feedback). The cumulated rewards of Q-learning translate asymptotically into a probability of successful transmission of 90,3%. This probability is 97% for the ideal scheme and 68,4% for random access. The ideal scheme shows less deviation of individual nodes performance from the average. The reward value standard deviations tend to 8.5% in the ideal scheme, 14.6% in Q-learning and 13.2% in random access.

Figure 2 shows the Q-Values and actions choices of two randomly chosen TXs in case of a breakdown of RX #4, which was initially preferred. The breakdown appears after 15 minutes of operation. It can be seen that the node modifies its behavior after receiving sufficient punishments. The number of successive punishments needed to consider RX #4 as a bad choice depends on α. The choice of another RX depends then on the exploration previously performed. It is also interesting to see that convergence behaviors can be quite different from one node to another. Node #55 seems to converge towards the choice of a single RX after the breakdown, while node #2 seems to converge to a probability distribution over the RXs choice. This shows the adaptation capability of the proposed scheme.

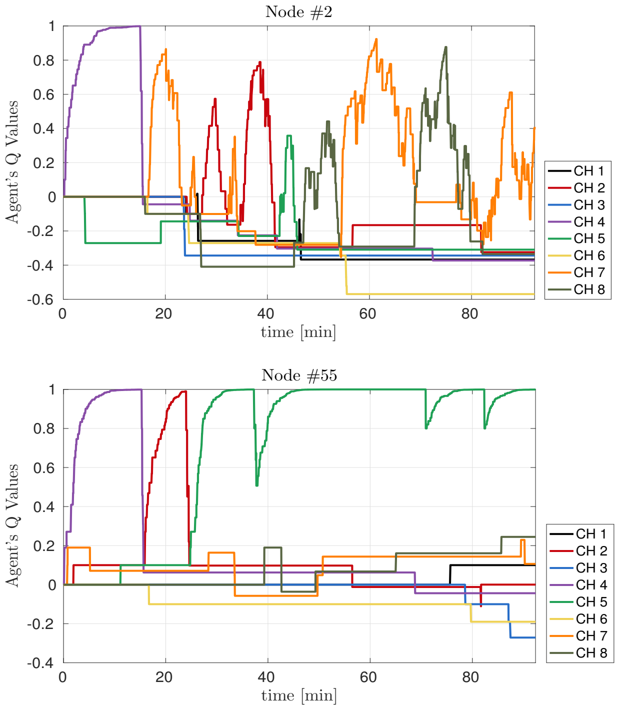

Fig. 2. Evolution of the Q-Values of two TXs in case of a breakdown of the access point (RX) #4 after 15 minutes.

V. Conclusion

In this paper, we have shown the benefits of RL to underwater acoustic networks through the study of a particular application scenario. These algorithms are able to offer some self-configuration and adaptation capabilities in a decentralized way, without the need of large overheads or messages exchanges between network nodes to retrieve information about the channel state. However, we believe that what we have proposed is not limited to the type of underwater sensor network described here. For example, one could consider nodes having different transmission strategies than choices of an access point in a discrete set. The same method could be used in any problem where UW transmitters have some degrees of freedom in their transmission strategy and can easily get some feedback about how good it was to take a particular action at a time. Another strength of some RL algorithms such as Q-learning is that they do not rely on a detailed model of the problem. This is desirable in UW where there is no consensus on the channel statistical properties and where it is usually difficult to predict precisely the conditions in which modems or networks will operate.

[1] see also the example provided by [36]

References

[1] D. Kilfoyle and A. Baggeroer, "The State of the Art in Underwater Acoustic Telemetry," *IEEE J. of Oceaninc Eng.*, vol. 25, no. 1, pp. 4–27, Jan 2000.

[2] E.M. Sozer, M. Stojanovic, and J.G. Proakis, "Underwater Acoustic Networks," *IEEE J. of Oceanic Eng.*, vol. 25, no. 1, pp. 72–83, January 2000.

[3] Ian F. Akyildiz, Dario Pompili, and Tommaso Melodia, "Challenges for Efficient Communication in Underwater Acoustic Sensor Networks," *SIGBED Rev.*, vol. 1, no. 2, pp. 3–8, July 2004.

[4] H. Riksfjord, O. T. Haug, and J. M. Hovem, "Underwater Acoustic Networks - Survey on Communication Challenges with Transmission Simulations," in *2009 Third International Conf. on Sensor Technologies and Applications*, June 2009, pp. 300–305.

[5] M. Stojanovic, S. Shahabudeen, and M. Chitre, "Underwater Acoustic Communications and Networking: Recent advances and future challenges," *Marine Technology Society J.*, vol. 42, no. 1, 2008.

[6] P. Casari and M. Zorzi., "Protocol Design Issues in Underwater Acoustic Networks," *Computer Communications*, vol. 34, pp. 2013–2025, 2011.

[7] R. Otnes, A. Asterjadhi, P. Casari, M. Goetz, T. Husoy, I. Nissen, K. Rimstad, P. van Walree, and M. Zorzi, *Underwater Acoustic Networking Techniques*, Springer Briefs in Electrical and Computer Eng., 2012.

[8] Emad Felemban, Faisal Karim Shaikh, Umair Mujtaba Qureshi, Adil A. Sheikh, and Saad Bin Qaisar, "Underwater sensor network applications: A comprehensive survey," *International J. of Distributed Sensor Networks*, vol. 11, no. 11, 2015.

[9] B. Peleato and M. Stojanovic, "Distance Aware Collision Avoidance Protocol for Ad-Hoc Underwater Acoustic Sensor Networks," *IEEE Comm. Letters*, vol. 11, no. 12, December 2007.

[10] Yishan Su, Yibo Zhu, Haining Mo, Jun-Hong Cui, and Zhigang Jin, "UPC-MAC: A Power Control MAC Protocol for Underwater Sensor Networks," in *Wireless Algorithms, Systems, and Applications*, Kui Ren, Xue Liu, Weifa Liang, Ming Xu, Xiaohua Jia, and Kai Xing, Eds., pp. 377–390. Springer Berlin Heidelberg, 2013.

[11] M. Molins and M. Stojanovic, "Slotted FAMA: a MAC protocol for underwater acoustic networks," in *OCEANS 2006 - Asia Pacific*, May 2006, pp. 1–7.

[12] A. A. Syed, W. Ye, and J. Heidemann, "T-lohi: A new class of mac protocols for underwater acoustic sensor networks," in *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, April 2008.

[13] P. Qarabaqi and M. Stojanovic, "Adaptive Power Control for Underwater Acoustic Communications," in *Proc. of IEEE OCEANS 2011*, June 2011.

[14] F-X. Socheleau, C. Laot, and J-M. Passerieux, "Parametric Replay-Based Simulation of Underwater Acoustic Communication Channels," *IEEE J. of Oceaninc Eng.*, vol. 40, no. 4, pp. 4838–4839, 2015.

[15] S. Jiang, F. Liu, and S. Jiang, "Distance-alignment based adaptive MAC protocol for underwater acoustic networks," in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[16] X. Guo, M. R. Frater, and M. J. Ryan, "Design of a Propagation-Delay-Tolerant MAC Protocol for Underwater Acoustic Sensor Networks," *IEEE Journal of Oceanic Engineering*, vol. 34, no. 2, pp. 170–180, April 2009.

[17] A. Pottier, F.-X. Socheleau, and C. Laot, "Robust Noncooperative Spectrum Sharing Games in Underwater Acoustic Interference Channels," *IEEE J. of Oceanic Eng.*, vol. 42, no. 4, pp. 1019 – 1034, October 2017.

[18] A. Pottier, F-X. Socheleau, and C. Laot, "Quality-of-Service Satisfaction Games for Noncooperative Underwater Acoustic Communications," *IEEE Access*, vol. 6, pp. 2541–2558, May 2018.

[19] A. Pottier, F.-X. Socheleau, and C. Laot, "Distributed Power Allocation Strategy in Shallow Water Acoustic Interference Channels," in *IEEE Int. workshop on Signal Processing Advances in Wireless Communications (SPAWC) 2016*, Jul. 2016.

[20] A. Pottier, F.-X. Socheleau, and C. Laot, "Adaptive Power Allocation for Noncooperative OFDM Systems in UWA Interference Channels," in *Proc. 2nd Underwater Acoustic Communications and Networking Conf. (UComms)*, Sept. 2016.

[21] A. Pottier, F.-X. Socheleau, and C. Laot, "Power-Efficient Spectrum Sharing for Noncooperative Underwater Acoustic Communication Systems," in *MTS/IEEE OCEANS'16 Monterey*, Sept. 2016.

[22] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[23] Christopher J. C. H. Watkins and Peter Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.

[24] N. Morozs, T. Clarke, and D. Grace, "Cognitive spectrum management in dynamic cellular environments: A case-based Q-learning approach," *Eng. Applications of Artificial Intelligence*, vol. 55, pp. 239 – 249, 2016.

[25] Y. Chu, S. Konusalp, P.D. Mitchell, D. Grace, and T. Clarke, "Application of reinforcement learning to medium access control for wireless sensor networks," *Eng. Applications of Artificial Intelligence*, vol. 46, pp. 23 – 32, 2015.

[26] S. Konusalp, Y. Chu, , P.D. Mitchell, D. Grace, and T. Clarke, "Use of Q-learning approaches for practical medium access control in wireless sensor networks," *Eng. Applications of Artificial Intelligence*, vol. 55, pp. 146 – 154, 2016.

[27] V. DiValerio, C. Petrioli, L. Pescosolido, and M. VanDerShaar, "A Reinforcement Learning-based Data-Link Protocol for Underwater Acoustic Communications," in *In Proc. of the 10th International Conference on Underwater Networks and Systems (WUWNET '15), ACM, New York, NY, USA*, 2015.

[28] C. Wang, Z. Wang, W. Sun, and D. Fuhrmann, "Reinforcement Learning-Based Adaptive Transmission in Time-Varying Underwater Acoustic Channels," *IEEE Access*, vol. 6, pp. 2541–2558, May 2018.

[29] F. Fazel, M. Fazel, and M. Stojanovic, "Random Access Compressed Sensing for Energy-Efficient Underwater Sensor Networks," *IEEE J. on Selected Areas in Communications*, vol. 29, no. 8, pp. 1660–1670, September 2011.

[30] F. Fazel, M. Fazel, and M. Stojanovic, "Random Access Compressed Sensing over Fading and Noisy Communication Channel," *IEEE Trans. on Wireless Communications*, vol. 12, no. 5, pp. 2114–2125, May 2013.

[31] M. B. Porter, "The BELLHOP Manual and Users Guide: Preliminary draft," *Heat, Light and Sound Research, Inc., La Jolla, CA, USA, Tech Report*, 2011.

[32] M. Stojanovic, J.G. Proakis, J.A. Rice, and M.D. Green, "Spread spectrum underwater acoustic telemetry," in *OCEANS'98 Conf. Proceedings*, 1998, vol. 2, pp. 650–654.

[33] F. Frassati, C. Lafon, P-A Laurent, and J-M Passerieux, "Experimental assessment of ofdm and dsss modulations for use in littoral waters underwater acoustic communications," in *Oceans 2005-Europe*, 2005, vol. 2, pp. 826–831.

[34] P. Qarabaqi and M. Stojanovic, "Statistical Characterization and Computationally Efficient Modeling of a Class of Underwater Acoustic Communication Channels," *IEEE J. Ocean. Eng.*, vol. 38, no. 4, pp. 701–717, October 2013.

[35] F.-X. Socheleau, C. Laot, and J.-M. Passerieux, "Stochastic Replay of non-WSSUS Underwater Acoustic Communication Channels Recorded at Sea," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4838–4849, 2011.

[36] B. Dunshaw, "Worldwide Sound Speed, Temperature, Salinity, and Buoyancy from the NOAA World Ocean Atlas," http://staff.washington.edu/dushaw/WOA/.