A Dynamic Reinforcement Learning Scheme for UAV-Based Joint Communication and Radar Systems

Soo Yeon Woo
Department of Electronics Engineering
Tech University of Korea (TU Korea)
Siheung, Republic of Korea
jysy2125@tukorea.ac.kr

Su Min Kim
Department of Electronics Engineering
Tech University of Korea (TU Korea)
Siheung, Republic of Korea
suminkim@tukorea.ac.kr

Junsu Kim*
Department of Electronics Engineering
Tech University of Korea (TU Korea)
Siheung, Republic of Korea
junsukim@tukorea.ac.kr
(corresponding author)

***Abstract*—In a joint communication and radar (JCR) system that performs radar sensing and communication simultaneously using a single signal, unmanned aerial vehicle (UAV) can efficiently utilize limited frequency resources and provide high-quality services when they can be detected and controlled. In this paper, we discuss the JCR system in a dynamic environment that considers user mobility and air-to-ground channels, and propose a dynamic UAV control solution to maximize system performance. The position of the UAV is a major factor that affects system performance, then, the proposed dynamic reinforcement learning scheme adjusts the trajectory of UAV and subframe length to derive the optimal solution. The simulation results show that the proposed scheme can adapt to the dynamic environment and converge to optimal values, demonstrating the ability to find optimal solutions for UAV trajectory and subframe length control.**

Keywords—Joint communication and radar (JCR) system, unmanned aerial vehicle (UAV), reinforcement learning.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) have become increasingly popular for various applications, including communication and surveillance. In these applications, it is essential to enable high-speed data transmission between UAVs and ground-base stations (BSs), as well as the ability to control the trajectory of the UAVs. Therefore, recent research considered positioning of UAVs using reinforcement learning algorithm. Reinforcement learning algorithm learns through the environment and improves performance over time.

To estimate the position of UAVs, radar waveforms can be used. Active research is being conducted on joint communication and radar (JCR) systems, which perform both communication and radar sensing simultaneously, to provide high-quality services while efficiently utilizing limited frequency resources [1][2]. In [1], a frame structure of a JCR system was designed to perform radar sensing and data transmission. An optimal radar sensing time, i.e., a pilot part duration that enables maximum achievable Rate, was designed. In [2], the role of the pilot part of the signal, such as channel estimation, was considered in communication system using UAV, which was not taken into account in [1]. Additionally, the authors defined a utility function that considered both communication and radar performance and investigated the trade-off between these two-performance metrics. They analytically derived the optimal transmit power for both the pilot and data parts by simultaneously considering both performance metrics.

In [3], the optimal position of the UAV and pilot duration were investigated to maximize the performance of the JCR system. We derived a solution in a scenario where UAV was used to provide optimal service to specific user who did not have line-of-sight channels with the BS. By exploiting the reinforcement learning algorithm, we optimized the position of the UAV and pilot duration, resulting in the maximization of system performance. However, the work in [3] only considered user with fixed position and did not consider user mobility.

In this paper, we propose a dynamic reinforcement learning scheme to maximize the performance of the JCR system. We design a reinforcement learning framework for a practical scenario that considers both user mobility and the channel model in the air-to-ground (A2G) environment, where the environment changes rapidly. An agent learns to maximize both the communication system performance metric of throughput and the radar system performance metric of Cramer-Rao lower bound (CRLB) in the practical scenario. As a result, we obtained the optimal trajectory of the UAV and pilot duration that maximizes the performance of the JCR system.

II. SYSTEM MODEL

As shown in Fig. 1, the signal frame structure in JCR system includes a pilot part for radar signal transmission and a data part for communication signal transmission. The duration of each part is denoted as T_p and T_d, and the total transmission time can be expressed as $T = T_p + T_d$. The energy of the entire frame of the signal is defined as PT and it is assumed that the power of the pilot and data signals is the same, denoted as P. The pilot part is used by the radar function to estimate the position of the UAV, while the data part is used by the base station to provide communication service to user through the UAV.

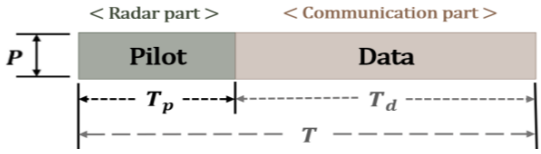

Fig. 1. Signal frame structure.

This work was supported in part by the MSIT, Korea, under the ICAN program (IITP-2023-RS-2022-00156326) supervised by the IITP and in part by the NRF funded by the Korea government (MSIT) (No. 2021R1A2C1013150).

Fig. 2 shows the system model considered in this paper. The system is composed of a ground BS, a UAV, and a ground user. The ground BS communicates with the ground user through the UAV acting as relay. On the JCR system, BS transmits a communication signal to send information to UAV while the radar receiver simultaneously receives the reflected signal to estimate the position change of the UAV.

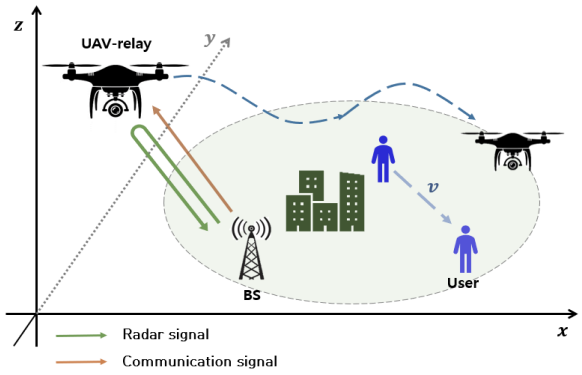

Fig. 2. System model.

A. User's Mobility

Even if the position of the UAV is optimized for a specific configuration, user movements can lead to a change in the optimal UAV position, and if the UAV is not properly repositioned to match user movements, it can have a negative impact on the performance of the communication system. Therefore, it is important to consider the effect of user mobility on the optimization of the UAV position and to design the framework that can adapt to changing user position in real-time.

The scenario of user movements is as follows: the ground BS is fixed at the position [20, 0], and the ground user initially remains stationary at the position [300, 300]. At a certain time, the user starts moving at a constant speed towards the final destination at the position [50, 50]. The positions of the user and BS are shown in Fig. 3. In other words, as the user moves closer to the BS, it is expected to result in an improvement in system performance.

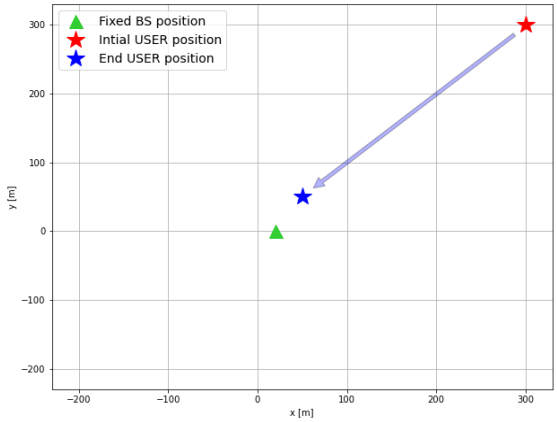

Fig. 3. A scenario of user movements.

B. Air-to-Ground (A2G) Path Loss Model

The A2G channel can be affected by the surrounding environment, which may result in the presence or absence of line-of-sight (LoS) path. Therefore, in scenarios where accurate information about the environment, such as the number of obstacles, is not available, both LoS and non-line-of-sight (NLoS) path probabilities must be considered in the channel modeling. In the paper, we adopt the A2G channel model proposed in [4]. The probability of LoS is given by

$$P_{Los}(\theta) = \frac{1}{1+\alpha \exp(-\beta(\theta-\alpha))}, \quad (1)$$

where α and β are constants depending on the environment, and θ is the elevation angle given by $\frac{180}{\pi}\arcsin(\frac{h_{UAV}}{d})$. The parameters h_{UAV} and d denote the altitude of UAV and the distance between the transmitter and receiver, respectively.

The path loss model for the UAV channel in A2G environments, which takes into account the probabilities of LoS and NLoS, can be expressed as

$$PL(d) = 20\log_{10}\left(\frac{4\pi d f_0}{c}\right) + P_{Los}(\theta)\eta_{LoS} + P_{NLos}(\theta)\eta_{NLoS} \ [dB], \quad (2)$$

where f_0 is the carrier frequency, c is the speed of light, and $P_{NLos}(\theta) = 1 - P_{Los}(\theta)$, respectively. Furthermore, η_{LoS} and η_{NLoS} represent the additional losses based on the environment, respectively.

C. Performance Metric

In this paper, we adopt a utility function proposed in [2] to evaluate the performance of the JCR system. We use throughput as the communication performance metric, while the CRLB is used as the radar performance metric. Throughput and CRLB are given by

$$R = P_{Los}(\theta) * \frac{\sum \tau B \log_2(1+SNR_C)}{T}, \quad (3)$$

and

$$\sigma^2_{CRLB} \equiv \frac{c^2}{32\pi^2 B^2 T_p SNR_r} \leq \sigma_r^2, \quad (4)$$

where B is the bandwidth of the transmit signal, τ is time slot length, SNR_c is signal-to-noise ratio for communication, and SNR_r is signal-to-noise ratio for radar, respectively.

The utility function is defined using the sum of two weighted performance metrics, as given by

$$U = w_c \cdot R - w_r \cdot \log_{10}\sigma^2_{CRLB}, \quad (5)$$

where w_c and w_r denote the weights for communication and radar performance, respectively.

III. Dynamic Reinforcement Learning Scheme

Reinforcement learning is a method of training an agent to take actions that result in maximum rewards at the end of an episode, by repeatedly taking actions in a state and receiving rewards for the resulting state transitions. In this paper, we aimed to maximize the performance of the JCR system in the dynamic environment. We use the reinforcement learning algorithm to find the optimal 3D position of the UAV and pilot

duration to maximize the JCR system performance. The proposed reinforcement learning environment consists of an agent, state, action, and reward, which are detailed as follows.

- *Agent:* We assume that UAV is agent of the proposed reinforcement learning framework.
- *State:* The states are represented as the position of the agent and pilot duration, and can be expressed as

$$s(t) = \{x(t),\ y(t),\ z(t),\ T_p(t)\}. \quad (6)$$

- *Action:* The actions consist of eight possible options, including positive or negative movement in the x, y, and z directions, as well as an increase or decrease in the pilot duration, and can be expressed as

$$a(t) = \{\pm\Delta x,\ \pm\Delta y,\ \pm\Delta z,\ \pm\Delta T_p\}. \quad (7)$$

  We applied the decaying ε-greedy policy to determine the action that the agent should take. Specifically, we set the value of the exploration probability variable, ε, relatively high at the beginning of the learning process and gradually decreased it as the learning progressed, allowing the agent to explore more in the early episode. As the episode progresses, the agent can choose the action with the optimal reward by decreasing the probability of exploration and increasing the probability of exploitation.

- *Reward:* The reward is defined as JCR system performance metric using (5).

During our learning process, we have events which are the movements of the user. At a certain time, the user starts moving gradually and changes their position. Then, the agent takes action that leads to increased reward, following the moving user. While the user is moving, the agent continues to take actions to find the optimal position and pilot duration. An episode ends when the user arrives at the destination and stops moving for a certain amount of time. The proposed dynamic reinforcement learning scheme was trained by repeating these episodes.

IV. Simulation Results

In simulation, the environmental parameters for A2G path loss model are suburban environment, with $\alpha = 4.8800$, $\beta = 0.4290$, $\eta_{LoS} = 0.1\ dB$, and $\eta_{NLoS} = 21\ dB$, respectively. The agent, UAV, was initialized at a random location, and we performed 2000 episodes of reinforcement learning. Each episode consists of a total of 120 steps of action and corresponding reward, and ends after the last action in the episode. Fig. 4 shows the convergence and optimality of the proposed scheme. The notations 'A' and 'B' represent the time when the user starts moving and the time when the user reaches the destination and stops, respectively. Just before the time A, since the reward has converged to its optimal value, the UAV takes the action to adjust the pilot duration at the optimal location. When time A is reached, user movement occurs, which causes the UAV's optimal location to become a suboptimal location. The UAV then repeats the process of finding the optimal location where the reward is maximized by following the movement of the user.

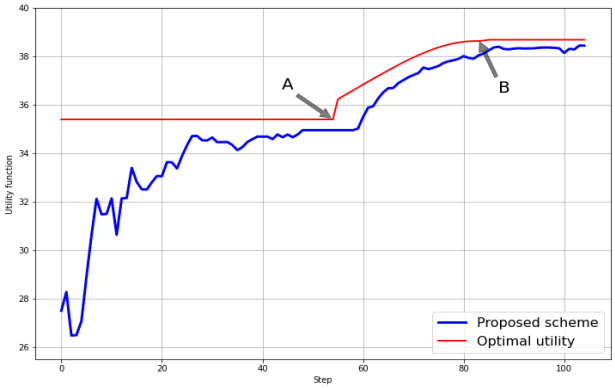

Fig. 4. Performance of the proposed scheme.

Table 1. Optimal parameters.

Optimal parameters	A	B
Position $[x, y, z]$ [m]	[190, 290, 185]	[50, 35, 180]
Pilot duration T_p [ms]	76	48

The simulation results demonstrate that the reward converges to the optimal value by optimizing the UAV position and pilot duration. The optimal parameters found by the proposed reinforcement learning are presented in Table 1.

V. Conclusions

In this paper, a dynamic reinforcement learning scheme was proposed to maximize the performance of the JCR system. The performance of the JCR system in dynamic environments that considered user mobility and A2G channels could be improved by controlling the UAV. Reinforcement learning was conducted through a practical simulation scenario that considered the dynamic environment. The simulation results demonstrated that the proposed scheme converged to the optimal reward value obtained at the optimal UAV location. Furthermore, by following the user's movement and maximizing the reward, the proposed scheme was able to find the optimal UAV location and subframe length based on the user's location in a dynamic environment. In future work, this paper can be further developed by applying it to a large-scale IoT network environment and multi-UAV scenarios.

References

[1] P. Kumari, D. H. N. Nguyen, and R. W. Heath, Jr., ''Performance trade-off in an adaptive IEEE 802.11AD waveform design for a joint automotive radar and communication system,'' in Proc. *IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2017.

[2] J. M. Park, J. Cho, S. Noh and H. Yu, "Optimal Pilot and Data Power Allocation for Joint Communication-Radar Air-to-Ground Networks," *IEEE Access*, vol. 10, pp. 52336-52342, May. 2022.

[3] S. Y. Woo, S. M. Kim, and J. Kim, "Optimization of UAV based Joint Communication and Radar System using Reinforcement Learning," in Proc. *KICS Winter Conf.*, Feb. 2023.

[4] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wirel. Commun. Lett.,* vol. 3, no. 6, pp. 569-572, Dec. 2014.