A Deep RL Approach for Duplex Mode Selection in Hybrid FD/HD ISAC Networks

Julius Ssimbwa[†], Byungju Lim[‡] and Young-Chai Ko[†]
[†]School of Electrical and Computer Engineering, Korea University, Seoul, Korea
[‡]Department of Electronic Engineering, Pukyong National University, Busan, Korea
(e-mail: kayjulio@korea.ac.kr, limbj@pknu.ac.kr, koyc@korea.ac.kr).

Abstract

This paper investigates an adaptive duplex mode selection strategy for hybrid full-duplex/half-duplex (FD/HD) integrated sensing and communication (ISAC) networks. The problem is formulated as an optimization task aimed at maximizing network throughput while minimizing interference. However, the resulting formulation is NP-hard and nonconvex, making it intractable to solve in polynomial time. To address this challenge, we employ deep reinforcement learning (RL) to develop an adaptive duplex mode control policy. Simulation results demonstrate that the proposed approach significantly enhances network spectral efficiency, highlighting the promising potential of RL-based strategies in future ISAC-enabled communication networks.

I. INTRODUCTION

Over the years, radar technology operating within its dedicated spectrum has played a crucial role in numerous applications such as remote sensing, surveillance, and traffic control, among others [1]. Meanwhile, communication technology has reached its operational limits due to rapid advancements that demand ever-increasing data rates within limited spectral resources. To address the challenge of spectrum scarcity, attention has shifted toward sharing the spectral bands traditionally occupied by incumbent radar systems, an approach that has given rise to *Integrated Sensing and Communication (ISAC)* systems. ISAC technology extends conventional radar capabilities by enabling simultaneous communication, localization, tracking, and other functions within a unified framework. However, effective coexistence between sensing and communication components remains a key challenge, particularly in the areas of resource allocation and interference management. Numerous studies on ISAC have primarily focused on aspects such as beamforming design, time and bandwidth allocation, and waveform optimization, with most investigations confined to the half-duplex (HD) domain. In recent works, full-duplex (FD) operation has been introduced into ISAC systems to leverage its potential advantages, including enhanced spectral efficiency and reduced latency, made possible by enabling simultaneous transmission and reception over the same frequency band [2]. However, the performance of FD-based ISAC remains constrained by various forms of interference, such as self-interference (SI) and multi-user interference, which can significantly degrade system reliability.

Several SI cancellation techniques have been proposed in literature, including analog, digital, and antenna-based methods. However, many of these approaches achieve SI mitigation at the expense of increased power consumption, while others fail to completely eliminate residual SI. This limitation not only affects system performance but also leads to higher operational costs for network operators. In this work, we propose the adoption of a hybrid FD/HD ISAC network to mitigate the inherent limitations of FD operation while preserving its potential advantages. Specifically, we formulate an optimization framework aimed at maximizing the overall network throughput while minimizing interference effects. To achieve this, we employ a reinforcement learning–based approach to design an adaptive duplex mode selection strategy, as elaborated in the subsequent sections of this paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a single cell ISAC system constituting a FD-BS equipped with a two uniform linear array (ULA) antenna structure divided into N_t ULA transmit and N_r ULA receive antennas respectively, as shown in Fig. 1. The BS operates as a dual-function radar-communication (DFRC) handling sensing and communication via detection of a target while simultaneously receiving and transmitting signals from K_u uplink (UL) signals and to K_d downlink (DL) single antenna users, respectively. To implement mode selection, we introduce a binary mode selection variable α such that when $\alpha = 1$, our system operates in FD mode while $\alpha = 0$ indicates HD operation.

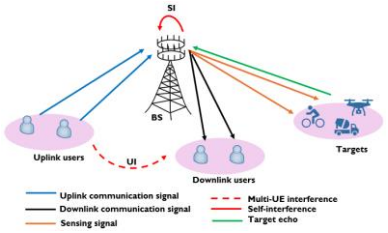

Fig. 1. An illustration of an FD ISAC network

B. Problem Formulation

According to Fig.1, during FD operation, the DL communication signal is affected by multi-user interference, UL-to-DL interference, and interference due to sensing, while the UL communication signal is affected by SI and interference due to target reflection. On the contrary, DL and UL transmission is separated via equal time division to mitigate interference during HD operation. With that said, we examine the performance of our proposed system by formulating an optimization problem (P1) to analyze the sum-rate of the network subject to sensing mutual information (MI) constraint C1 and duplex mode selection constraint C2 as follows.

$$\begin{aligned}(\text{P1})\ \max_{\alpha}\ & \alpha f\left(R_{\text{DL}}^{\text{FD}}, R_{\text{UL}}^{\text{FD}}\right) + (1-\alpha) f\left(R_{\text{DL}}^{\text{HD}}, R_{\text{UL}}^{\text{HD}}\right) \\ \text{s.t. C1:}\ & \alpha R_{\text{rad}}^{\text{FD}} + (1-\alpha) R_{\text{rad}}^{\text{HD}} \le R_{\text{rad}}^{\text{th}} \\ \text{C2:}\ & \alpha \in \{0,1\},\end{aligned}$$

where the function $f\left(R_{\text{DL}}^{(\cdot)}, R_{\text{UL}}^{(\cdot)}\right) = R_{\text{DL}}^{(\cdot)} + R_{\text{UL}}^{(\cdot)}$, with $R_{\text{d}}^{(\cdot)} = \sum_{d=1}^{K_d} \log_2\left(1+\rho_d^{(\cdot)}\right)$, $R_{\text{UL}}^{(\cdot)} = \sum_{u=1}^{K_u} \log_2\left(1+\rho_u^{(\cdot)}\right)$, $R_{\text{rad}}^{(\cdot)} = \log_2\left(1+\rho_{\text{rad}}^{(\cdot)}\right)$ representing the DL rate, UL rate and the sensing MI, respectively. Meanwhile, $\rho_d^{(\cdot)}$, $\rho_u^{(\cdot)}$, and $\rho_{\text{rad}}^{(\cdot)}$, denote the DL, UL, and sensing signal-to-interference-plus-noise ratio (SINR), and $R_{\text{rad}}^{\text{th}}$ is the minimum radar sensing MI.

III. PROPOSED RL ALGORITHM DESIGN

A. Markov Decision Process Model Formulation

It is evident that problem (P1) is nonconvex and NP-hard, which makes it difficult to solve. Although off-shelf CVX solvers can be employed to solve such a problem, they are limited by complexity. We therefore propose the use of deep RL to address the problem. In this setting, we model the problem as a Markov decision process (MDP) defined by $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma >$, where $\mathcal{S}$

is the environment state space, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the transition probability from one state to another, $\mathcal{R}$ is a reward function while γ is a discount factor [3]. Generally, at time step t an agent interacts with the environment in state $s_t \in \mathcal{S}$ and takes an action $a_t \in \mathcal{A}$ to obtain a reward $r_t \in \mathcal{R}$. Consequently, the goal of the agent is to learn a policy π that maximizes the cumulative reward. In the sequel, we deploy a model-free MDP based RL scheme to attain the optimal policy π^* that yields the maximum cumulative reward, and the details of the proposed MDP model are as follows.

a) **State**: It consists of the channel state information and the duplex mode selection variables such that at time step t

$s_t = [\mathrm{H}_{\mathrm{DL}}(t), \mathrm{H}_{\mathrm{UL}}(t), \mathrm{H}_{\mathrm{rad}}(t), \mathrm{H}_{\mathrm{SI}}(t), \alpha(t-1)]$,

where $\{\mathrm{H}_{\mathrm{DL}}(t), \mathrm{H}_{\mathrm{UL}}(t), \mathrm{H}_{\mathrm{rad}}(t), \mathrm{H}_{\mathrm{SI}}(t)\}$ represent DL, UL, radar, and SI channel matrices respectively.

b) **Action**: The action space contains the duplex mode selection variable such that at time step t

$$a_t = \alpha(t) \in \{0,1\}$$

c) **Reward**: The reward is motivated by the objective of this work associated with maximizing the network throughput via desirable duplex mode selection. The reward at time step t is therefore given by

$r_t =$

$\alpha(t) f\left(R_{\mathrm{DL}}^{\mathrm{FD}}, R_{\mathrm{UL}}^{\mathrm{FD}}\right) + \left(1-\alpha(t)\right) f\left(R_{\mathrm{DL}}^{\mathrm{HD}}, R_{\mathrm{UL}}^{\mathrm{HD}}\right) - \lambda_{\mathrm{r}} E_{\mathrm{r}} - \lambda_{\mathrm{s}} E_{\mathrm{s}}$,

where

$E_{\mathrm{r}} = \max\left(0, R_{\mathrm{rad}}^{\mathrm{th}} - \left(\alpha(t) R_{\mathrm{rad}}^{\mathrm{FD}} + \left(1-\alpha(t-1)\right) R_{\mathrm{rad}}^{\mathrm{HD}}\right)\right)$

$E_{\mathrm{s}} = |\alpha(t) - \alpha(t-1)|$

B. Proposed Proximal Policy Optimization Algorithm

Due to the discrete nature of our action in the proposed MDP model, it would be beneficial to adopt a desirable RL method. Some of such potential schemes include deep Q-network (DQN) and proximal policy optimization (PPO) [3,4]. DQN is sample efficient due to its use of an experience replay buffer to store and reuse past experiences, however, due to the complexity of our problem, we use PPO owing to its demonstrated benefits in balancing exploration and policy improvement. PPO derives its performance from the application of clipping to a surrogate objective function, which ensures stability and full control when optimizing the policy. This objective function can be given by

$$L_{\mathrm{clip}}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right],$$

where ϵ is the clipping threshold while $\hat{A}_t = \delta_t + \gamma\lambda_q \hat{A}_{t+1}$, is the advantage function with δ_t and λ_q denoting the temporal-difference (TD) error and the balancing weight in generalized advantage estimation (GAE), respectively. Additionally, $r_t(\theta)$ is the probability ratio between new and old policies such that

$$r_t(\theta) = \pi_\theta(a_t|s_t)\left(\pi_{\theta_{old}}(a_t|s_t)\right)^{-1},$$

where $\pi_\theta(a_t|s_t)$ is the action probability distribution for action a_t given state s_t under policy network parameter θ. The learning procedure of our proposed MDP model follows the layout similar to the PPO algorithm discussed in [4]. Furthermore, Fig. 2 shows the execution of PPO-RL for our proposed MDP model.

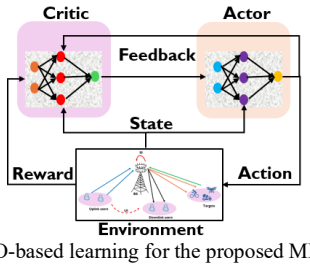

Fig. 2. PPO-based learning for the proposed MDP model.

IV. PERFORMANCE EVALUATION

In our simulations, we consider a single BS equipped with $N_t = N_r = 8$ antennas, a single target and $K_d = K_u = 2$ users, operating under 20GHz frequency. The UL and DL transmit power are set 23 dBm and 30 dBm respectively, the bandwidth set to 200MHz, the threshold set to 0.7 nats/s/Hz, and residual SI variance set to 110dBm. We evaluate the performance of the proposed scheme (PPO) via rewards and sum rate in comparison with DQN as follows.

Fig. 3 illustrates the variation of the rewards and sum rate with the number of updates (episodes). It shows that PPO is generally outperforming DQN by a 2% gap due to its ability to perform more exploration with stability and full control in policy improvement. This shows that DQN incurs more constraint violations compared to PPO. Briefly, PPO obtains an average reward of 350 and average rate of 1.75 compared to DQN's 343 and 1.71 average reward and average rate respectively. This performance renders PPO more efficient for our proposed duplex mode selection RL framework.

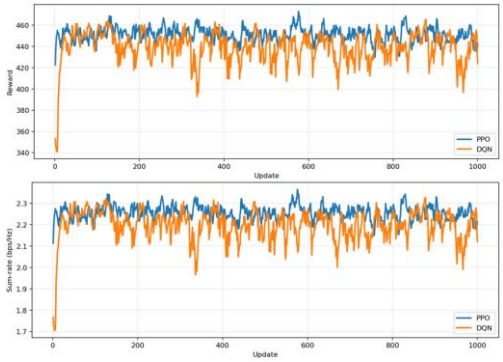

Fig. 3. Reward and sum rate versus the number of updates.

V. CONCLUSION

In this article, we have examined a duplex switching strategy for a hybrid FD/HD ISAC network. We have derived an optimization problem to maximize network throughput while minimizing the interference. We have further proposed a PPO-based RL to optimize the duplex selection variable and improve spectral efficiency. The simulation results show promising potential of including RL techniques in future network generations to improve network performance amidst the high complexity and network dynamics. In our future work, we hope to investigate our system model by including beamforming with highly mobile users and targets.

Acknowledgment

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (2022-0-00704, Development of 3D-NET Core Technology for High-Mobility Vehicular Service).

References

[1] L. Zheng, M. Lops, Y. C. Eldar, and X. Wang, "Radar and communication coexistence: An overview: A review of recent methods," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 85–99, Sep. 2019.

[2] C. B. Barneto et al., "Full Duplex Radio/Radar Technology: The Enabler for Advanced Joint Communication and Sensing," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 82-88, Feb. 2021.

[3] X. Wang et al., "Deep Reinforcement Learning: A Survey," *IEEE Trans. Neural Netw.*, vol. 35, no. 4, pp. 5064-5078, April 2024,

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv*: 1707.06347, 2017.