Reinforcement learning based time-domain mutual interference avoidance for automotive radar

Xiao, He; Wang, Jianping; Li, Runlong; He, Yuan

DOI
10.1049/icp.2024.1476

Publication date
2023

Document Version
Final published version

Published in
IET Conference Proceedings

REINFORCEMENT LEARNING BASED TIME-DOMAIN MUTUAL INTERFERENCE AVOIDANCE FOR AUTOMOTIVE RADAR

He Xiao[1], Jianping Wang[2], Runlong Li[1], Yuan He[1]*

[1] *Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing, China*
[2] *Faculty of EEMCS, Delft University of Technology, Delft, Netherlands*
**yuanhe@bupt.edu.cn*

Keywords: AUTOMOTIVE RADAR, INTERFERENCE AVOIDANCE, TIME DOMAIN STRATEGY, DEEP REINFORCEMENT LEARNING.

Abstract

Due to the extensive usage of automotive radars on vehicles, mutual interference among radars on the road is becoming considerable. To address this , we propose a time domain strategy based on deep reinforcement learning (DRL). This approach helps avoid mutual interference for automotive radars in the time domain without extra communications. The numerical simulation results demonstrate that the proposed approach can avoid interference as effectively as frequency hopping. Moreover, the time domain strategy has more advantages than frequency hopping when encountering dynamic interference.

1 Introduction

Automotive radars are essential in the advanced driver assistance system. Among them, frequency modulated continuous wave (FMCW) radar has become one of the most popular choices due to its broad operational capability and low cost. However, as the number of FMCW radars equipped on vehicles increases rapidly, mutual interference among different radar devices arises inevitably in busy areas. Strong interference could mask weak targets and raise ghost targets, leading to a higher traffic accident risk. Therefore, it is crucial to mitigate interference for the safety purpose.

Various approaches have been investigated to counter mutual interference. Some studies have developed signal processing methods operated on the received signal to cancel interference [1], [2]. These methods exploit the differences between interference and target echoes in time, frequency, or time-frequency domain to suppress interference with slopes different from the victim radar, i.e., incoherent interference. However, when facing coherent interference which has an identical slope to the transmitting signal of the radar, these signal processing methods are no longer suitable. Other researchers have presented new radar systems or waveform designs [3], [4], [5], which spread interference in the frequency spectrum to avoid ghost targets. These methods are able to suppress coherent interference and improve detection performance. Nevertheless, they require new system and hardware designs as well as more complicated processing.

Over recent years, many resource allocation methods have been proposed to avoid interference. Some achieve cognitive radar approaches based on reinforcement learning (RL) [6], [7]. They exploit the information of the electromagnetic environment to implement spectrum allocation to prevent collisions in the frequency domain. However, their capabilities are limited by the spectrum resource. When facing more interference, spectrum allocation operated on finite bandwidth becomes inadequate to maintain both detection and anti-interference performance. In [8] and [9], time offset is introduced to avoid mutual interference in the time domain. They utilize radar and communication networks to realize centralized or localized resource allocation. However, these cooperative schemes heavily rely on communication.

In this paper, we propose a non-cooperative time domain method for automotive radar to avoid mutual interference. The proposed method only uses the information extracted from the received signal of the radar itself to make decisions, which does not demand any communications. The execution of the method is modelled as an MDP and implemented by deep Q-learning.

2. Methodology

2.1 Signal Model

In general, the transmitted signal of the FMCW automotive radar in one single chirp can be expressed as

$$x(t) = e^{j2\pi\left[f_c t+\frac{1}{2}K\left(t-\frac{T_{sw}}{2}\right)^2\right]}, \quad 0 \le t \le T_{sw} \tag{1}$$

where f_c, K, and T_{sw} denote the centre frequency, sweep slope, and sweep duration of one single chirp, respectively. Once the maximum detection range d_{max} is determined, the maximum time delay τ_{max} can be calculated as

$$\tau_{max} = \frac{2d_{max}}{c}, \tag{2}$$

where c denotes the speed of light. After dechirping, the maximum beat frequency $f_{b\,max}$ is determined as

$$f_{b\,max} = K\tau_{max}. \tag{3}$$

Ideally, we assume that the cut-off frequency of low-pass filter (LPF) used after dechirping is equal to $f_{b\,max}$. For simplicity, we assume that the transmitted signal is reflected by point targets, neglecting the multipath effect and clutters. Besides the echo of targets, the received signal could also contain

https://doi.org/10.1049/icp.2024.1476

transmitted signals from other automotive radars which are regarded as interference.

2.2 The Time Domain Strategy for Interference Avoidance

Since several methods [1], [2], [7] have been proposed to mitigate incoherent interference, we concentrate on coherent interference here, i.e., interference shares the same sweep slope and pulse repetition interval (PRI) with the victim radar.

According to the corresponding relationship of τ_{max} and $f_{b_{max}}$, it can be inferred that if the time delay between the transmitted signal and received signal is not within the range of $[0,\ \tau_{max}]$, the received signal will be removed by LPF. Inspired by this, we propose the time domain strategy for interference avoidance. As depicted in Fig. 1, the radar adjusts its transmitting time to avoid interference. Practically, the transmitted signal of the victim radar is delayed by t_d, which changes the beat frequency between the transmitted signal and interference so that the interference would be filtered out by LPF after dechirping. For convenience, we set a time reference here. The transmitting time delay t_d of the radar and the arrival time delay τ_{intf} of the interference are all defined relative to the time reference.

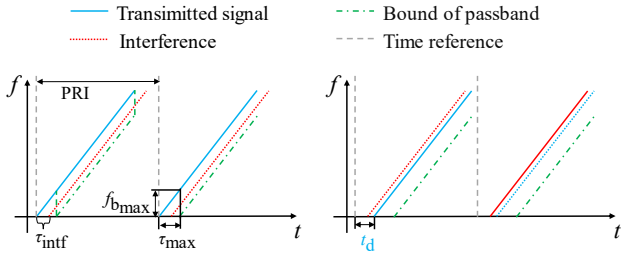


Fig. 1. Time domain strategy.

2.3 Markov Decision Process Modelling of Interference Avoiding

The execution process of the joint strategy is formulated as an MDP model which can be described by the tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$. The state space $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_{N_S}\}$ is the set of all possible states that the radar can reach. The state $\boldsymbol{s} = [\hat{f}_c, t_d, SINR] \in \mathcal{S}$ consists of the radar's frequency domain state $\hat{f}_c$ (i.e., centre frequency), time domain state t_d (i.e., transmitting time delay), and the SINR of the received signal. The action space $\mathcal{A}$ can be defined as $\{\boldsymbol{a} | \boldsymbol{s} \xrightarrow{\boldsymbol{a}} \boldsymbol{s}', \boldsymbol{s} \in \mathcal{S}, \boldsymbol{s}' \in \mathcal{S}\}$, in which the action $\boldsymbol{a}$ consists of the radar's actions in the frequency domain and time domain. Given n available frequency domain states and m available time domain states, the size of $\mathcal{A}$ is $(mn)^2$. However, the size of $\mathcal{S}$ is not countable due to the SINR with continuous values. The transition probability function $p(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ presents the probability distribution of reaching states $\boldsymbol{s}'$ from state $\boldsymbol{s}$ by taking action $\boldsymbol{a}$. The reward function $r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ presents the immediate reward obtained after transitioning to state $\boldsymbol{s}'$ from state $\boldsymbol{s}$ by taking action $\boldsymbol{a}$.

At each time step, SINR is used to evaluate the radar's action. The metric is defined as follows:

$$SINR = 10\log_{10}\frac{|\boldsymbol{x}_t|^2}{|\boldsymbol{x}_i + \boldsymbol{n}|^2}, \tag{4}$$

where $\boldsymbol{x}_t$, $\boldsymbol{x}_i$, and $\boldsymbol{n}$ are respectively the beat signal of targets, interference, and noise reserved after dechirping and low-pass filtering. The immediate reward r is calculated by the SINR:

$$r = \begin{cases} \dfrac{2(SINR - SINR_1)}{(SINR - SINR_1) + (SINR_2 - SINR_1)}, & SINR \geq SINR_1, \\ 0.1(SINR - SINR_1), & SINR < SINR_1, \end{cases} \tag{5}$$

where $SINR_1$ is the threshold to give positive or negative reward, which can be set based on the requirement for target detection practically. $SINR_2$ presents an upper bound of available SINR for reward normalization, which is usually relative to the noise level. In this paper, we assume that the SINR of the received signal can be estimated accurately, and the estimation is not going to be discussed.

A policy π is utilized to choose the action based on the current state: $\boldsymbol{a}_t = \pi(\boldsymbol{s}_t)$. The state-action sequence based on policy π in one episode is defined as a trajectory $\{\boldsymbol{s}_0, \boldsymbol{a}_0, \boldsymbol{s}_1, \boldsymbol{a}_1, \dots, \boldsymbol{s}_{t_\infty - 1}, \boldsymbol{a}_{t_\infty - 1}, \boldsymbol{s}_{t_\infty}\}$, $\boldsymbol{s}_0$ and $\boldsymbol{s}_{t_\infty}$ denote the initial state and the terminal state of the episode, respectively. The cumulative reward starting from time step t on this trajectory is

$$G_t(\pi) = \sum_{k=t}^{t_\infty - 1} \gamma^{k-t} r_{k+1} = \sum_{k=t}^{t_\infty - 1} \gamma^{k-t} r(\boldsymbol{s}_k, \boldsymbol{a}_k, \boldsymbol{s}_{k+1}), \tag{6}$$

where $\gamma \in [0,1]$ is the discount factor weighting the future reward. To chase a high SINR, the radar is expected to perform a trajectory getting as much cumulative reward as possible with an optimal policy which is

$$\pi^* = \arg\max_{\pi} G_t(\pi). \tag{7}$$

There may be more than one optimal policy, but they share the same state-action value function [10]. Here, the Q function $Q_\pi(\boldsymbol{s}, \boldsymbol{a})$ denotes the state-action value function for a policy π which is defined as the expectation of G_t starting from $\boldsymbol{s}$, taking the action $\boldsymbol{a}$, and thereafter following the policy π:

$$\begin{aligned} Q_\pi(\boldsymbol{s}, \boldsymbol{a}) &= \mathbb{E}_\pi[G_t | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}] \\ &= \sum_{\boldsymbol{s}' \in \mathcal{S}} p(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') \{r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') \\ &\quad + \gamma \mathbb{E}_\pi[G_{t+1} | \boldsymbol{s}_{t+1} = \boldsymbol{s}']\}, \end{aligned} \tag{8}$$

Then, the goal for the optimal policy π^* is to find the optimal Q function Q^* enabling the radar to choose the optimal action $\boldsymbol{a}^*$ at each time step:

$$\boldsymbol{a}^* = \arg\max_{\boldsymbol{a} \in \mathcal{A}} Q_{\pi^*}(\boldsymbol{s}, \boldsymbol{a}) = \arg\max_{\boldsymbol{a} \in \mathcal{A}} Q^*(\boldsymbol{s}, \boldsymbol{a}). \tag{9}$$

2.4 Deep Reinforcement Learning Based Implementation of the Proposed Method

We choose Q-learning to optimize the Q function for its faster convergence and sample reusability.At every time step, Q-learning updates the value function as follows:

$$Q(\boldsymbol{s}_t, \boldsymbol{a}_t) \leftarrow Q(\boldsymbol{s}_t, \boldsymbol{a}_t) + \beta \delta_t, \tag{10}$$

where β denotes the learning rate, and δ_t is the TD error for updating:

$$\begin{aligned} \delta_t = r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) \\ + \gamma \max_{\boldsymbol{a}_{t+1} \in \mathcal{A}} Q(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - Q(\boldsymbol{s}_t, \boldsymbol{a}_t). \end{aligned} \tag{11}$$

The current state $\boldsymbol{s}$ could only present the interference situation in the current passband, but it can be inadequate for radar to make an effective decision. To provide more information about interference, we extend the state $\boldsymbol{s}_t$ to $\boldsymbol{S}_t =$

$[\boldsymbol{s}_{t-k+1}, \boldsymbol{s}_{t-k+2}, \dots, \boldsymbol{s}_t]$, which contains the states of the latest k time steps within one episode. While adding more observations, the extended state $\boldsymbol{S}_t$ greatly enlarges the state space. To tackle such a complex state space, Q-network is used as the approximation of the Q function. The architecture of the network is shown in Fig. 2. At each time step, the extended state $\boldsymbol{S}_t$, i.e., a sequence of some recent states, is input into a gated recurrent unit (GRU) layer to extract information about interference. Then, the output of the GRU layer is input into dense layers to calculate the Q values of all actions.

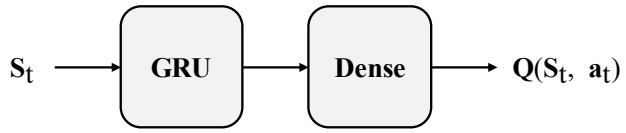


Fig. 2. The architecture of Q-network.

In training, several skills are employed to train the Q-network:

1) Double Q-networks: Two networks, evaluation network Q_{eval} and target network Q_{tar}, are used to relieve training instability. Specifically, Q_{eval} will be updated every time step while Q_{tar} will be updated to Q_{eval} by an interval of numbers of time steps. When updating Q_{eval}, Q_{tar} is used to calculate the value of $\boldsymbol{a}_{t+1}$ chosen by Q_{eval}. δ_t for updating is rewritten as

$$\begin{aligned}\delta_t = {} & r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) \\ & + \gamma Q_{tar}\left(\arg\max_{\boldsymbol{a}_{t+1}\in\mathcal{A}} Q_{eval}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})\right) \\ & - Q_{eval}(\boldsymbol{s}_t, \boldsymbol{a}_t),\end{aligned} \tag{12}$$

and the parameters of Qeval, $\boldsymbol{w}$, is updated by

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \beta \frac{\partial \delta_t^2}{\partial \boldsymbol{w}}. \tag{13}$$

2) ϵ-greedy exploration: To avoid the Q-network being trapped in suboptimal actions, the radar will randomly choose actions by the exploring probability $\epsilon \in [0,1]$:

$$\boldsymbol{a}_t = \begin{cases} \arg\max_{\boldsymbol{a}_t\in\mathcal{A}} Q_{eval}(\boldsymbol{s}_t, \boldsymbol{a}_t) & \text{with probability } (1-\epsilon), \\ random\ action & \text{with probability } \epsilon. \end{cases} \tag{14}$$

3) Experience replay: $\boldsymbol{e}_t = [\boldsymbol{S}_t, \boldsymbol{a}_t, \boldsymbol{S}_t, r_{t+1}]$ is defined as the transition of the time step t. The replay buffer stores a number of recent transitions. If the number of stored transitions reaches the upper limit, the oldest transition will be replaced by the latest one. At every time step, Q_{eval} is trained on a batch of transitions randomly picked from the replay buffer. Besides improving sample efficiency, experience replay can also break the correlation among consecutive samples that could harm the training.

3 Numerical Simulation

3.1 Simulation Settings

3.1.1 Radar Settings: Parameter settings of the victim radar are listed in Table 1. We assume 3 available states for radar in both frequency and time domains. One-hot vector is used to represent these parameter states. The interference shares identical sweep parameters with the victim radar, including centre frequency, sweep slope, sweep duration, bandwidth, and PRI.

Table 1 Parameter settings of victim radar

Parameter	Value
Centre frequency $\hat{f_c}$ [GHz]	76.5, 77.0, 77.5
Sweep slope [MHz/us]	10
Sweep duration [us]	50
Bandwidth [MHz]	500
Pulse repetition interval [us]	60
Chirp number per frame	128
Sampling frequency [MHz]	20
Maximum detection range [m]	120
Maximum beat frequency [MHz]	8
Maximum delay τ_{max} [us]	0.8
Time domain state t_d [$\times\tau_{max}$]	0, 1, 2

3.1.2 Scenario Settings: Simulation settings of training and test are shown in Table 2. Here, $\boldsymbol{\mathcal{U}}\{\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_l}\}$ denotes the discrete uniform distribution on the finite set $\{\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_l}\}$, and $\boldsymbol{\mathcal{U}}(\boldsymbol{b_1}, \boldsymbol{b_2})$ denotes the continuous uniform distribution on the interval $(\boldsymbol{b_1}, \boldsymbol{b_2})$. Since the arrival delay of the interference signal accounts for both the distance and transmitting time of the interference source, the delay time $\boldsymbol{\tau_{intf}}$ is given directly for simplicity. Here, static scenario and dynamic scenario are defined. In the simulation, the targets and interference will be randomly initialized at the beginning of each episode. An episode has a number of time steps, and only one frame of the signal will be transmitted, received, and processed in each time step. The frame's duration is less than 10 ms, which is too short for moving targets and interference sources to significantly influence the signal in the passband. Therefore, the duration of each time step is neglected in the static scenario, which means that both targets and interference remain constant in one episode. In the dynamic scenario, the interval between time steps is enlarged so that the relative motion of targets and interference sources could have a noticeable effect. For instance, one interference can move into or out of a passband, which could influence the radar's behaviour. Specifically, the distance of the target and the delay time of interference will be updated before each time step:

$$d(t_n) = d(t_n) + v_{target}\Delta t, \tag{15}$$

$$\tau_{intf}(t_n) = \tau_{intf}(t_{n-1}) + \frac{v_{intf}\Delta t}{c}, \tag{16}$$

where $d(t_n)$ and $\tau_{intf}(t_n)$ denote the distance of the target and the delay time of interference at time step t_n. v_{target} and v_{intf} are the relative velocities of the target and interference source. For simplicity, the amplitudes and velocities of targets and interference would not change in one episode. If the distance of a target or τ_{intf} of interference is out of the range shown in Table 2, the item will be replaced by a new one randomly initialized.

Table 2 Simulation settings

Parameter		Value	
		Training	Test
Point-like target	Number	$\mathcal{U}\{1,2,3\}$	
	Amplitude	$\mathcal{U}(0.5,1.5)$	
	Distance [m]	$\mathcal{U}(0.3,120)$	
	Velocity relative to the victim radar [m/s]	$\mathcal{U}(-16,16)$	
Interference	Number	$\mathcal{P}(12)$	$6\sim18$
	Amplitude	$\mathcal{U}(1,12)$	
	Velocity relative to the victim radar [m/s]	$\mathcal{U}(-32,32)$	
	Delay time [$\times\tau_{max}$]	$\mathcal{U}(0,3)$	
Signal-to-noise ratio [dB]		10	
Time steps per episode		40	200
Time step interval [ms]		125	

3.1.3 RL Settings: The settings of RL are listed in Table 3. The following radar agents are set in training and test:

1) Radar-Fixed: The agent's parameter state $(\hat{f}_c, t_d)$ are fixed to (77 GHz, τ_{max}).
2) Radar-F: The agent executes frequency hopping, but its t_d is fixed to τ_{max}.
3) 3) Radar-T: The agent executes the time domain strategy, but its $\hat{f}_c$ is fixed to 77 GHz.

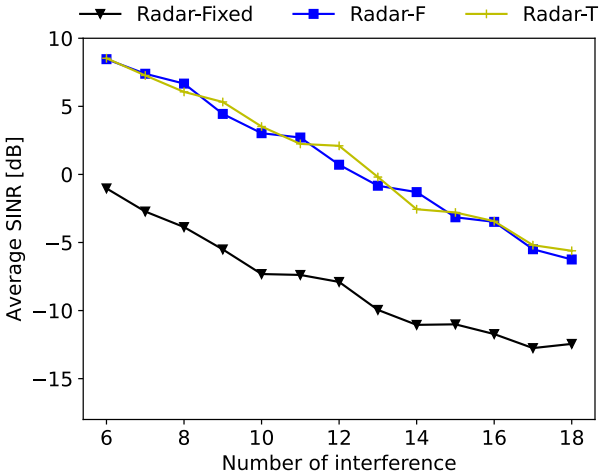
The Q-network consists of one GRU layer and two dense layers. A ReLU activation is used after the first dense layer. The input of the Q-network is the one-hot vector of the current parameter state concatenated with the SINR, and the output is the Q-values of each available parameter states. So the output shapes of Radar-F or Radar-T, i.e., the neuron numbers of the second dense layer, is 3. During training, the learning rate β and the exploration probability ϵ are exponentially decaying to the minimum episode by episode.

Table 3 RL settings

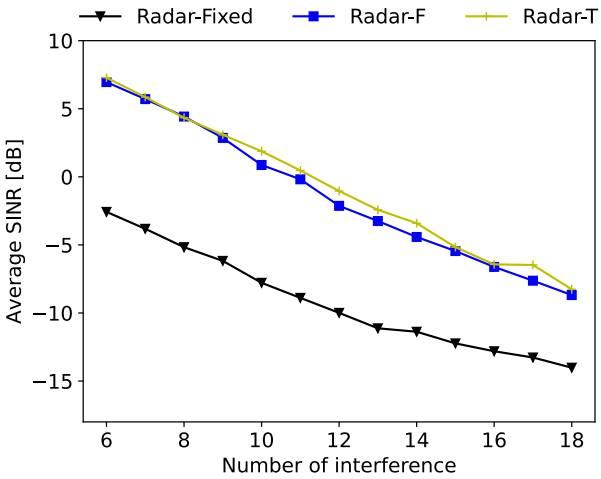
Item		Value
	Layer	Number of neurons
Q-network architecture	GRU	16
	Dense 1	32
	Dense 2	3
$(SINR_1, SINR_2)$ [dB]		(0, 11)
γ		0.99
β		0.01 ~ 0.00001
ϵ		1 ~ 0.05
Maximum capacity of $\boldsymbol{S}_t$		20 time steps
Replay buffer capacity		50000
Replay batch size		4 × 32
Q_{tar} updating interval		200 time steps

3.2 Simulation Results

The random initialization of interference in each episode leads to a large variance during sampling, subsequently making RL training difficult and unstable. To deal with the instability, the number of time steps per episode is reduced in training, and it takes tens of thousands of episodes for agents to reach convergence. In particular, the radar agents are first trained to converge in the static scenario and then retrained in the dynamic scenario. After training, the radar agents are tested with the interference number varying from 6 to 18. For each number of interference, the agents are tested for one thousand episodes in both dynamic and static scenarios. The test results are shown in Fig. 3, and the overall performance of interference avoidance is demonstrated in Table 4. The avoidance ratio is defined as the proportion of the time steps without interference in the passband. Besides the average SINR, the avoidance ratio presents the interference avoiding performance of the radar agent.



(a)



(b)

Fig. 3. The test result versus various numbers of interference. (a) Average SINR in the static scenario. (b) Average SINR in the dynamic scenario.

Table 4 The overall performance of interference avoidance

Radar agent	SINR [dB]		Avoidance ratio [%]	
	static	dynamic	static	dynamic
Fixed	−5.23	-5.96	18.64	16.30
F	0.65	-0.88	39.29	33.70
T	0.76	-0.51	39.86	35.22

Among the three agents, Radar-Fixed has the worst performance on interference avoidance. Compared to Radar-Fixed, Radar-F and Radar-T all perform much better. Notably, their performance is almost the same in the static scenario, which proves the ability of the proposed time domain strategy. The reason is that the numbers of available states in the time

and frequency domains are the same, and the interference is distributed uniformly in both domains when initialized. In the dynamic scenario, all these agents suffer performance degradation due to the continuously changing interference. However, Radar-T slightly outperforms Radar-F. The dynamic of interference in the time domain is more regular than that in the frequency domain, so it is easier to avoid interference in the time domain, especially when facing more interference.

Fig. 4 shows the state transitions and the performance on SINR of the three radar agents in one episode as the interference number is 12 in the dynamic scenario. The average SINR of Radar-Fixed, Radar-F, and Radar-T are -17.69 dB, 0.33 dB, and 1.33 dB, respectively. Both Radar-F and Radar-T will react immediately if a sudden drop in SINR caused by dynamic interference occurs. They almost always get a higher SINR than Radar-Fixed during the whole episode. However, Radar-T performs more stably than Radar-F. Since it is more difficult to predict the dynamic of interference, Radar-F switches its parameter state much more frequently, even if it does not suffer a low SINR.

4 Conclusion

In this paper, we propose a non-cooperative time domain strategy to avoid interference for FMCW automotive radar. The proposed strategy is implemented by utilizing Markov Decision Process model and deep reinforcement learning. The numerical simulation demonstrates the effectiveness of the method. The proposed time domain strategy shows a better and more stable performance than frequency hopping in the dynamic scenario.

5 Acknowledgements

Acknowledgements should be placed after the conclusion and before the references section. Details of grants, financial aid and other special assistance should be noted.

6 References

[1] Xu, Z., Xue, S., and Wang, Y.: ' Incoherent Interference Detection and Mitigation for Millimeter-Wave FMCW Radars ', Remote Sensing, vol. 14, no. 19, 2022.

[2] Wang, J., Li, R., He, Y., et al.: 'Prior-Guided Deep Interference Mitigation for FMCW Radars', IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022.

[3] Luo, T.-N., Wu, C.-H. E., and Chen,Y.-J. E.: 'A 77-ghz cmos automotive radar transceiver with anti-interference function', IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 12, pp. 3247–3255, 2013.

[4] Kitsukawa, Y., Mitsumoto, M., Mizutani, H., et al.: 'An Interference Suppression Method by Transmission Chirp Waveform with Random Repetition Interval in Fast-Chirp FMCW Radar', 2019 16th European Radar Conference (EuRAD), Paris, France, 2019, pp. 165-168.

[5] Uysal, F.: 'Phase-Coded FMCW Automotive Radar: System Design and Interference Mitigation', IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 270–281, 2020.

[6] Thornton, C. E., Kozy, M. A., Buehrer, R. M., A. F. et al.: 'Deep Reinforcement Learning Control for Radar Detection and Tracking in Congested Spectral Environments', IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 4, pp. 1335– 1349, 2020.

[7] Liu, P., Liu, Y., Huang, T., et al.: 'Decentralized Automotive Radar Spectrum Allocation to Avoid Mutual Interference Using Reinforcement Learning', IEEE Transactions on Aerospace and Electronic Systems, vol. 57, no. 1, pp. 190–205, 2021.

[8] Khoury, J., Ramanathan, R., McCloskey, D., et al.: 'RadarMAC: Mitigating Radar Interference in Self-Driving Cars', 2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), London, UK, 2016, pp. 1-9.

[9] Aydogdu, C., Keskin, M. F., Garcia, N., et al.: 'RadChat: Spectrum Sharing for Automotive Radar Interference Mitigation', IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 1, pp. 416–429, 2021.

[10] Sutton, R. S., and Barto, A. G., 'Reinforcement Learning: An Introduction', MIT press, 2018.

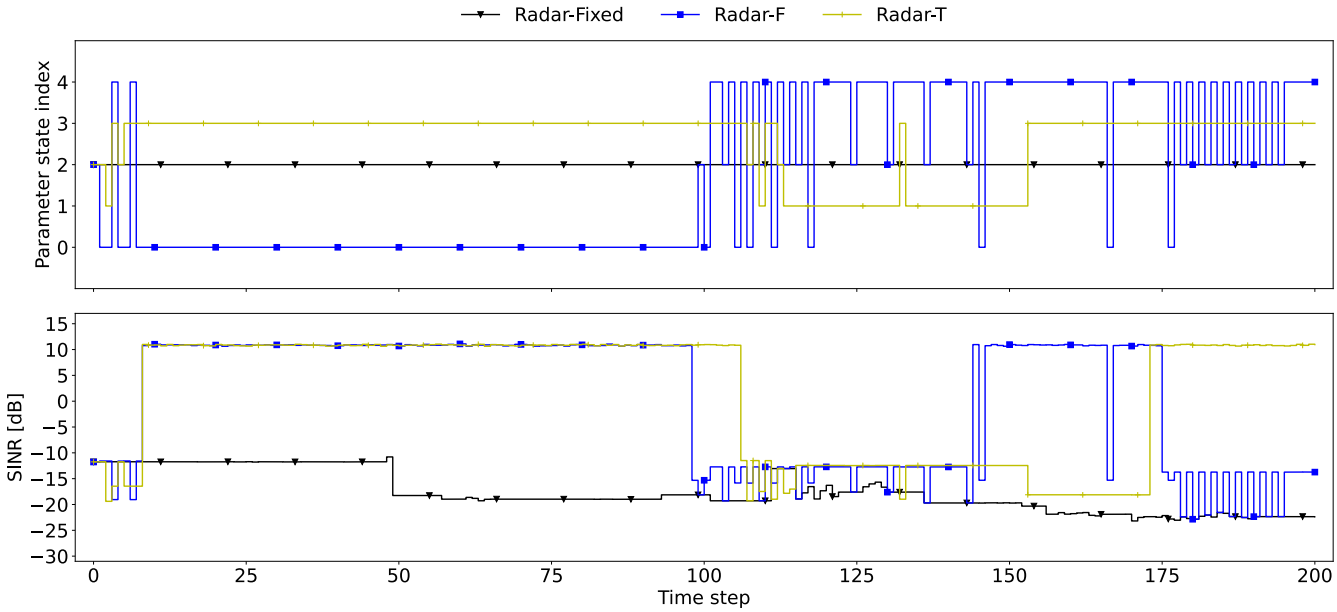


Fig. 4. Test result of one episode in the dynamic scenario. Parameter state index setting is {Index : ($\widehat{f_c}$ $[GHz], t_d$ $[\times \tau_{max}]$) | 0 : (76.25, 1), 1 : (76.75, 0), 2 : (76.75, 1), 3 : (76.75, 2), 4 : (77.25, 1)}.