Application of reinforcement learning in phase control of mid-air ultrasonic haptics

Mifuka Nakamura‡, Nobuya Sato, Daisuke Mizushima* (Aichi Inst. Tech.)

1. Introduction

Haptics is a technology that simulates the tactile sensations. It is being utilized not only in entertainment fields such as VR but also in telemedicine, robot operation, and many other areas. Among these, mid-air ultrasonic haptics can create tactile sensations without contact by applying the acoustic radiation pressure of focused ultrasonic to the skin. Due to this feature, it is also being used to clarification of tactile principles and sensitivity thresholds[(1)].

In order to focus mid-air ultrasonic waves, it is essential to use beamforming, which changing a phase of ultrasonic waves emitted from the ultrasonic speakers and moving the focus as desired. In existing products, the focus is calculated by obtaining the target position information from external sensors, such as cameras. In contrast, a method of focus control that does not rely on external sensors is beamforming by using reinforcement learning. Previous research has adopted PPO (Proximal Policy Optimization), which is a policy-based reinforcement learning algorithm. It is assumed the phase to a continuous value. However, the result is that no policy is found to maximize the reward and a unique focus is not obtained [(2)].

In this study, phases are considered as discrete values ranging from 0 to 360 degree, and the goal is to obtain a unique focus for discrete state spaces by using value-based reinforcement learning algorithms, SARSA and DQN (Deep Q-Network).

2. Reinforce Learning

Reinforcement learning mimics the human decision-making process. The decision maker is called the agent, the possible options of the agent are actions *a*, the state *s* is the result of the actions acting on the environment, and the desirability index of the actions is called the reward *R*. The value of taking action *a* in a given state *s* is referred to as the state-action value *Q*, and the table listing these values is called the Q-table. In reinforcement learning, high value *Q* is learned to be obtained continuously. In this study, the action is the phase difference between the sound sources and the state is the sound pressure of the actuator.

There are two types of reinforcement learning algorithms: value-based and policy-based. In value-based methods, Q-learning and SARSA are representative algorithms, while in policy-based methods, PPO is a typical algorithm. Value-based algorithms include Q-learning and SARSA, while PPO is an example of a policy-based algorithm.

On the other hand, in policy-based methods, the algorithm learns an optimal policy and improves it to enhance the value. In reinforcement learning, it is important to balance the search for actions and the application of the learning results. We used the ε-greedy method as the search method. This method takes random actions with probability ε and selects the best action based on the learned information with probability (1-ε). As learning progresses, ε is gradually reduced to shift from exploration to optimization.

2.1 SARSA

SARSA is a representative method of value-based reinforcement learning. The agent takes an action a_t from the current state s_t, receives a reward R_t, transitions to the next state s_{t+1}, and selects the next action a_{t+1}, updating the Q-table in the process. This sequence of symbols is referred to as SARSA. The Q-value update formula for SARSA is given in Equation (1).

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[R_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (1)$$

In the above equation, α represents the learning rate and indicates the width of the Q-value updates. γ is the discount rate, which is a measure for discounting the rewards obtained from a sequence of actions following the chosen action. In other words, SARSA updates the Q-table values by estimating the future rewards received during actions. In actual action selection, whether to choose the action with the maximum Q-value or to continue exploring is determined according to the ε-greedy method.

2.2 DQN

DQN is an algorithm that combines Q-learning with deep learning. The Q-value update formula in Q-learning is given in Equation (2).

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[R_t + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (2)$$

In Q-learning, updating the Q-table by selecting the action with the highest Q-value. It requires

E-mail: ‡v24723vv@aitech.ac.jp
*d-mizushima@aitech.ac.jp,

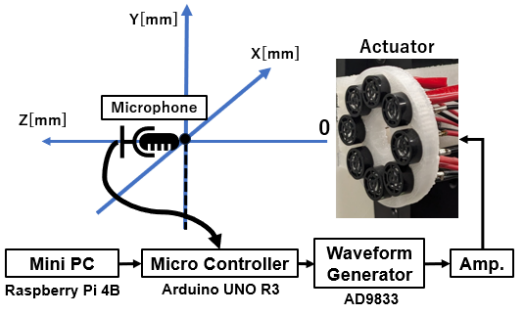

Fig. 1 Experimental Setup.

computing Q-values for all possible actions, which is computationally intensive. DQN approximates these Q-values using deep learning. As in the case of SARSA, for actual action decisions, the ε-greedy method is used to decide whether to choose the action with highest Q-value or to continue exploring.

3 Experimental Setup

The configuration of the experimental setup is shown in **Fig. 1.** Eight ultrasonic speakers are installed in a concentric circle with a radius of 12.7 mm as actuators. SPL SU1007 (Resonant frequency is 40 kHz) is used as ultrasonic speaker. The control unit that performs reinforcement learning is a Mini PC (Raspberry Pi 4B). Arduino Uno R3 is a controller that transmits the phase information to the waveform generator IC AD9833. The signal output from the waveform generator IC is amplified through the amplification circuit to a maximum voltage of 60 Vpp and applied to the speakers, which then emit ultrasonic.

The actuators and microphones are positioned directly opposite each other, and the center of the actuator's circle as the origin of three-dimensional space (x, y, z). Microphones are placed at any point in the three-dimensional space. The voltage obtained from the microphones is used as the reward in reinforcement learning. Initially, voltage is applied to two speakers to determine the optimal phase difference. This process is then repeated for each additional speakers, applying voltage to them and finding its optimal phase difference, continuing this for all speakers.

4. Results

SARSA and DQN were implemented on the Mini PC, and the sound pressure at each focus was measured. Prior to implementation, simulations were conducted to optimize the learning rate and the decay rate of ε. Since the reward setting has a significant impact on the learning results, the following three reward patterns were tested in the actual system: I. Estimate the target voltage based on the number of speakers and distance, and use the ratio of the obtained voltage to this target voltage as the reward. II. Use the obtained voltage directly as the reward. III. If the obtained voltage is higher than before, assign a reward of +200; if it is lower, assign a reward of -1.

Table 1 Obtained sound pressure with SARSA, DQN algorithms.

	Sound pressure [Pa]	
	SARSA	DQN
Reward I	3501	2409
Reward II	4239	3437
Reward III	2730	1028

Table 1 shows the focal sound pressures obtained after learning with SARSA and DQN. The focus is set at (x, y, z) = (0, 0, 30). In both cases, the highest sound pressure is achieved with reward pattern II. The sound pressure distributions at the focal plane for SARSA-II and DQN-II at the focal plane is measured, and the center of sound pressure at the focus can be observed. SARSA is a wider half-value width of the sound pressure compared to DQN.

5. Conclusion

In this study, the phase differences of the sound sources were assumed to be discrete states ranging from 0 to 360 degrees. A value-based reinforcement learning algorithm was used to achieve a unique focus. Policy-based algorithms, such as PPO, are designed for continuous states and are better suited for multivariate problems. Therefore, policy-based methods were considered redundant.

Reward pattern II resulted in the highest sound pressure. This is likely because it provided larger reward compared to the ratio with the target value. Additionally, SARSA achieved higher sound pressure compared to DQN, which is suspected to be due to DQN's experience replay. Randomly sampling experiences from the replay buffer may have led to repeated learning of suboptimal sound pressures, introducing bias into the exploration process.

References

1) W. Frier, A.Abdouni, D.Pittera, O.Georgiou, and R.Malkin, 2022, IEEE Access, vol. 10, pp.15443-15456.
2) R.Sato and K.Okita, Abstracts of the 56th Annual Conference of the College of Industrial Science and Engineering, Nihon University, pp.638-639, 2023 [in Japanese].
3) M.Deepanshu, IJERT Vol. 8 Issue 12, 2019, pp.717-722.