Journal of the KNST

2025; Vol.8, No.3; pp.314-322

ISSN: 2635-4926

https://doi.org/10.31818/JKNST.2025.9.8.3.314

Check for updates

Received: 2025/07/30
Revised: 2025/08/12
Accepted: 2025/09/15
Published: 2025/09/30

*Corresponding Author:
Youngmin Choo
Tel: +82-2-880-8380
E-mail: sonacer@snu.ac.kr

다중 무인 수중운동체 위협 환경에서 PPO 기반 강화학습을 이용한 잠수함의 최적 회피 및 의사결정 전략 연구

Optimal Evasion Decision and Strategies of Submarine Using PPO-based Reinforcement Learning under Multiple UUVs Threat Environment

강언약[1], 홍우영[2], 이귀영[3], 이종무[4], 백혁재[5], 배준호[6], 추영민[7*]

[1]세종대학교 해양시스템융합공학과 석사과정
[2]세종대학교 국방시스템공학과 교수
[3]해군 대위/세종대학교 해양시스템융합공학과 석사과정
[4]LIG넥스원 해양연구소 선임연구원
[5]LIG넥스원 해양연구소 수석연구원
[6]해군 소령/서울대학교 조선해양공학과 박사과정
[7]서울대학교 조선해양공학과 부교수

Eonyak Kang[1], Wooyoung Hong[2], Gwiyoung Lee[3], Jongmoo Lee[4], Hyukjae Baek[5], Junho Bae[6], Youngmin Choo[7*]

[1]M.S. student, Dept. of Ocean Systems Engineering, Sejong University
[2]Professor, Dept. of Defense Systems Engineering, Sejong University
[3]LT, ROK Navy/M.S. student, Dept. of Ocean Systems Engineering, Sejong University
[4]Research engineer, Maritime R&D Center, LIG Nex1
[5]Chief research engineer, Maritime R&D Center, LIG Nex1
[6]LCDR, ROK Navy/Ph.D. student, Dept. of Naval Architecture and Ocean Engineering, Seoul National University
[7]Associate Professor, Dept. of Naval Architecture and Ocean Engineering, Seoul National University

Abstract

본 연구는 다중 무인 수중운동체(UUV)의 위협 환경에서 잠수함의 생존율을 높이기 위해, 음향 탐지 모델을 보상 함수 설계에 통합한 강화학습 기반의 회피 전략을 제안한다. 구체적으로, 시뮬레이션 환경은 다중 UUV, 잠수함의 운동 모델링과 음향 탐지 모델로 구성하였다. 보상 함수는 잠수함이 적대적 UUV의 탐지 및 공격을 능동적으로 회피할 수 있도록 설계되었다. 시뮬레이션 결과, 제안한 강화학습 기반 회피 전략은 기존의 고정된 패턴 전략과 비교하여 잠수함의 생존율을 크게 높였다. 또한, 음향 탐지 모델에 따라 탐지 신호를 최소화하는 최적의 회피 기동을 학습함으로써, 효과적으로 회피하고 높은 생존율을 달성하였다.

This study proposes a reinforcement learning-based evasion strategy that integrates an acoustic detection model into the reward function design to enhance submarine survivability in a multi-unmanned underwater vehicle (UUV) threat environment. Specifically, the simulation environment was constructed with multi-UUV and submarine motion modeling, along with an acoustic detection model. The reward function was designed to enable the submarine to actively evade detection and attacks from hostile UUVs. Simulation results show that the proposed reinforcement learning-based evasion strategy significantly increased submarine survivability compared to conventional fixed-pattern strategies. Furthermore, by learning optimal evasion maneuvers that minimize detection signals based on the acoustic detection model, the strategy achieved effective evasion and high survivability.

Keywords

무인 수중운동체(Unmanned Underwater Vehicle),
잠수함(Submarine),
강화학습(Reinforcement Learning),
회피 전략(Evasion Strategy),
음향 탐지 모델(Acoustic Detection Model)

Acknowledgement

본 연구는 LIG Nex1의 지원을 받아 수행된 연구 결과임.

1. 서론

현대 해양전에서 능동 음향을 사용하는 무인 수중운동체(UUV, unmanned underwater vehicle)는 잠수함에 대한 가장 큰 위협 중 하나로 부상하고 있다[1]. UUV는 능동 소나를 사용하여 높은 정확도로 잠수함을 찾아낼 수 있다. 잠수함의 생존성 향상을 위한 전통적 대응법으로는 수동 운동체의 소나 신호를 교란하기 위한 기

만기 투하, 급선회 및 심도 변경 등이 활용되었다. 그러나 다중 UUV와 교전하는 시나리오에서는 전통적인 대응 전략만으로는 잠수함의 생존을 담보하기 어려운 실정이다.

이처럼, 다중 UUV에 대한 대응의 어려움으로 인해 최근 연구에서는 복잡한 해양전 문제를 강화학습으로 해결하려는 시도가 증가하고 있다. 2차원 격자 환경에서 인공신경망을 활용하는 연구[5]와 시뮬레이터 상에서 계층적 강화학습을 활용하여 회피 및 대응 전술을 고도화하는 연구[1] 등이 대표적인 사례들이다. 이처럼 강화학습 기반 접근법은 고정된 규칙에 의존하지 않고도 최적에 가까운 회피 행동을 실시간으로 도출할 수 있어 주목받고 있다. 하지만, 실제적인 운동체의 움직임과 복잡한 해양 환경에서의 음향 특성 등을 충분히 반영하지 못한다는 한계를 지닌다.

본 논문에서는 3차원 운동 모델링과 음향 탐지 모델링을 통합한 환경에서 PPO(proximal policy optimization) 기반의 강화학습 에이전트를 훈련함으로써, 다중 UUV 위협에 대한 잠수함의 자율 회피 판단과 전략을 개발한다. 특히, 음향 탐지 모델을 보상함수로 활용하여 강화학습에 수중 음향 특성을 효과적으로 접목하였다. 이하에서는 운동 모델링과 음향 탐지 모델링을 포함한 시뮬레이션 환경을 설명한 뒤(2장), 제안한 강화학습 기반 회피 전략의 세부 설계(3장) 및 실험 결과와 분석(4장)을 차례로 제시한다.

2. 시뮬레이션 환경 및 구현

본 연구의 시뮬레이션 환경은 다중 UUV와 잠수함의 교전 상황을 가정하여 설계되었으며, 회피 기동과 음향 대항 체계를 포함하고 있다. 환경은 3차원 공간에서 이산적 시간 단위로 동작하며, 초기 거리 및 방위 조건에 따라 UUV와 잠수함의 움직임 및 상호 작용이 전개된다. 다음 절에서는 시뮬레이션 환경의 핵심 구성요소인 잠수함, UUV의 운동 모델, 소나 방정식 기반 음향 탐지 모델에 대해 설명한다.

2.1 잠수함의 운동 모델

잠수함의 제원은 박정민[2] 등의 연구를 참고하여 설계하였으며, 3차원 운동 모델을 기반으로 상하 기동과 선회 중심의 운동을 모사하였다. 잠수함은 탐지 경보 범위를 가지고 있으며, UUV가 이 반경 내로 진입하면 즉시 회피 기동을 시작할 수 있다. 이때, 속력은 증속 과정을 거쳐 변침과 심도 변경이 완료된 후에 최대 속력에 도달하게 된다. 위 연구에서는 잠수함의 심도 변경은 첫 회피에서만 이루어지며 이후 회피에서는 심도 변경을 하지 않는 것으로 가정하였다. 또한, 최대 4번의 변침이 가능하며 변침 시 부유식 기만기를 자동 발사한다. 부유식 기만기는 강력한 음향 신호를 방출하여 UUV가 신호를 식별하는 것을 어렵게 만드는 역할을 한다. 이러한 운동 모델과 기만기의 운용 방식은 Fig. 1에 제시되어 있다.

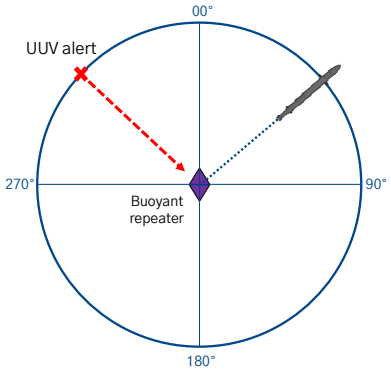


Fig. 1. Submarine evasive maneuver using buoyant repeater

2.2 UUV의 운동 모델과 음향 탐지 모델

UUV 운동 모델 역시 잠수함 운동 모델과 동일하게 3차원 운동 모델을 기반으로 하며, 상하 기동과 선회 동작을 포함하여 실제 운용 시나리오를 반영한 기동을 모사한다. 각 UUV는 교전 수행 과정에서 직주 단계, 탐색 단계, 공격 단계, 재탐색 단계의 네 단계로 구분되며, 각 단계는 능동 소나 기반의 음향 탐지 시스템과 연동되어 작동한다. 탐지 단계에서 UUV는 능동 소나를 통해 표적을 탐색하며, 이는 음향 펄스를 송신하고, 표적으로부터 반사되어 돌아오는 반향음을 수신하여 표적인 잠수함의 존재 여부를 판단한다. 본 연구에서는 수신된 신호의 세기를 기준으로 표적의 탐지 여부를 판단하며, 파형 분류나 추가적인 신호 식별 과정은 생략하였다. 탐지 성능 향상과 외부 신호 간섭 최소화를 위해 신호처리 과정에서 정합 필터(MF, matched filter)를 사용한다. 정합 필터가 적용된 소나 방정식을 이용하여 신호 초과(SE[signal excess])를 계산하며, 이는 식 (1)과 같이 나타낸다.

$$SE = SL - 2PL - NL + DI + TS + SG - DT \quad (1)$$

여기서, SL은 UUV에서 송신한 핑(ping)의 음원 준위(source level)이다. PL은 음파가 UUV에서 잠수함까지 전파되는 동안 발생하는 손실(propagation loss)을 의미하며, 기하학적 손실과 주파수에 의존적인 흡수 손실로 구성된다[3]. NL은 주변 환경 잡음과 UUV 자체 소음을 포함한 소음 준위(noise level)이다. DI는 빔 패턴의 지향성을 나타내는 지향 지수를 의미하며, DI^{off}는 표적이 빔의 중심축에서 벗어난 경우를 고려하여 $DI = DI^{on} - DI^{off}$ 식을 통해 계산된다[3]. TS는 표적 강도(target strength), SG는 정합 필터를 통한 신호 증폭 효과(signal gain)를 나타낸다. DT는 탐지 임계치(detection threshold)로, UUV가 표적을 탐지했다고 판단하기 위한 최소 신호대 잡음비(SNR, signal-to-noise ratio)를 나타낸다. 본 연구에서는 최적의 탐지 조건을 기준으로 DT 값을 설정하였는데, 입사하는 음향 빔이 잠수함의 측면에 도달하여 TS 값이 최대가 되는 조건이다.

2.3 시뮬레이션 흐름도

Fig. 2는 움직이는 잠수함과 UUV의 교전 상황을 평가하기 위한 시뮬레이션의 논리적 흐름을 나타낸 것이다. 시뮬레이션은 초기 변수 설정 및 시나리오별 파라미터 설정으로 시작되며 초기조건이나 회피 동작을 포함한 다양한 이동 관련 파라미터를 설정한다. 초기 설정 후, 시뮬레이션은 반복적인 시간 단계 루프(time-step loop)에 진입하여 다음과 같은 여러 단계로 구성된 프로세스를 순차적으로 수행한다.

첫 번째 단계는 운동 제어 단계로, 각 UUV의 현재 운용 단계에 따라 움직임을 업데이트한다. 동시에 표적도 위협 탐지 여부에 따라 자신의 위치, 수심 및 진행 방향을 동적으로 변경하여 움직임을 갱신한다.

두 번째 단계는 위치 계산 단계로, 앞서 갱신된 이동 파라미터를 기반으로 UUV와 표적의 위치를 재계산한다. 이 단계에서는 위협 탐지 또는 추적 전략으로 인해 수행되는 추가적인 기동도 반영하여 정확한 궤적을 산출한다.

세 번째 단계는 탐지 판단 단계로, UUV는 소나 방정식을 통해 신호 초과(SE)를 계산하여 유효한 탐지 여부를 평가한다. 앞서 설명한 것처럼, $SE > 0$ 조건을 만족하면 탐지 성공으로 간주하고 해당 UUV는 탐지된 표적에 대한 추적 단계(tracking phase)에 돌입하여 운용 단계를 탐색에서 공격 등으로 전환한다. 탐지가 실패한 경우에는 탐색 상태를 유지하거나, 표적을 다시 확보하기 위해 재탐색 단계로 되돌아간다.

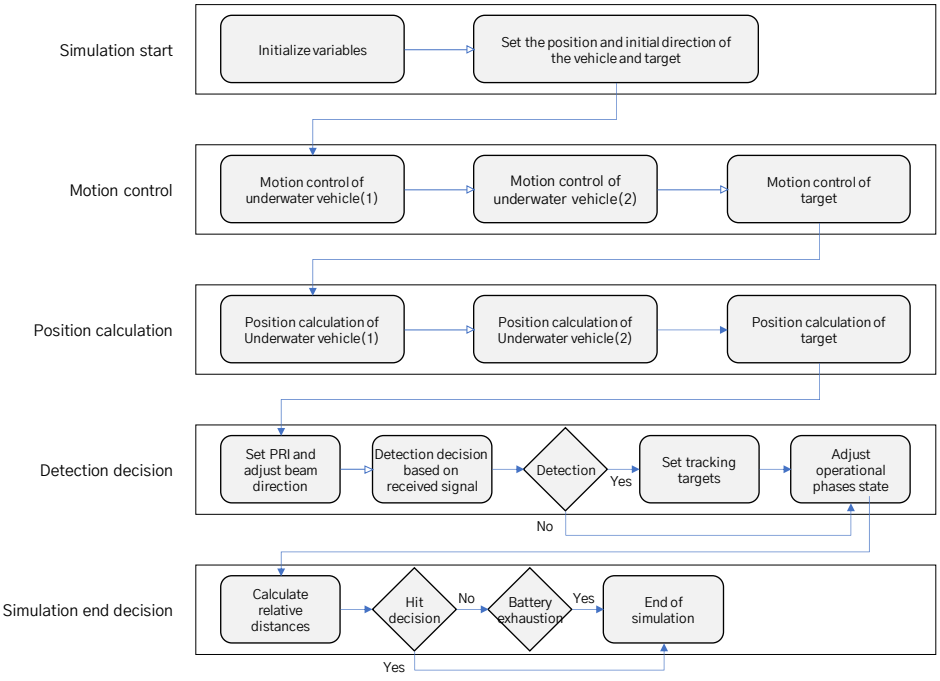


Fig. 2. Flowchart depicting the simulation logic for UUV

마지막 단계는 시뮬레이션 종료 판단 단계로, UUV가 표적을 성공적으로 공격하였는지, 배터리가 소모되었는지 또는 최대 시뮬레이션 시간이 초과되었는지 등의 조건을 확인하여 시뮬레이션 종료 여부를 결정한다. 이러한 조건 중 하나라도 만족하면 시뮬레이션이 종료되며, 그렇지 않으면 다음 시간 단계의 운동 제어 단계로 되돌아가 반복 과정을 계속 진행한다.

3. 강화학습 기반 최적의 회피 판단과 전략

본 연구의 제안 방법은 강화학습 알고리즘인 proximal policy optimization(PPO)를 사용하여 잠수함의 최적 회피 정책을 학습시키는 것이다. 앞서 서술한 시뮬레이션 환경을 환경 모델로 활용하고, 에이전트(잠수함)는 매 시뮬레이션 단계에서 관측값을 받아 회피 판단과 회피 기동 명령을 출력한다. 에이전트는 누적 보상을 최대화하도록 학습되며, 보상 신호는 음향 탐지 모델과 연계되어 설계된다. 본 장에서는 PPO 기반 에이전트의 구조와 학습 설정, 관측 공간 및 행동 공간의 구성, 그리고 핵심인 보상 함수 설계에 대해 설명한다.

3.1 관측 공간 및 행동 공간

강화학습 관측 공간(observation space)은 에이전트가 인식할 수 있는 상태 벡터로 정의된다. 본 환경에서 관측 벡터는 Fig. 3 및 식 (2)와 같이 잠수함과 UUV들의 상대적인 3차원 위치 정보와 시뮬레이션 시간으로 구성된다. 이러한 관측 공간 설계는 잠수함과 UUV 사이의 거리와 상대적인 위치 관계를 에이전트가 실시간으로 파악할 수 있도록 한다. 특히 잠수함과 각 UUV 간의 거리, 방위각, 심도 차이를 간접적으로 포함하고 있어, 에이전트는 이를 기반으로 회피 기동을 결정할 수 있다.

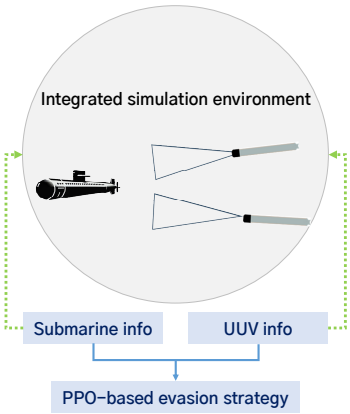


Fig. 3. PPO for evasion decisions and strategy in the environment

$$s_t = [P_s, P_{uuv1}, P_{uuv2}t, n_{eva}] \in R^{14} \tag{2}$$

여기서, P_s는 잠수함의 3차원 위치와 침로를 포함하며, P_{uuv1}, P_{uuv2}는 UUV들의 3차원 위치와 침로 정보, n_{eva}는 침로 변경 횟수로 14차원의 관측 벡터로 구성된다. 이때, UUV의 개수를 늘리면 관측 상태의 정보 역시 늘어나게 된다.

식 (3)은 행동 공간에 대한 설명으로, 행동 공간은 잠수함이 취할 수 있는 회피 판단과 회피 기동을 나타낸다.

$$a_s = [a_{ang}, a_{decision}] \tag{3a}$$

$$a_{ang} \in [0°, 360°] \tag{3b}$$

$$a_{decision} \in [-1, 1] \tag{3c}$$

여기서, 출력으로는 연속적인 회피 침로각(evasion angle, a_{ang})과 회피 판단의 여부(evasion decision, $a_{decision}$)로 설정하였다. 회피 판단의 여부를 위한 출력값은 –1과 1 사이의 범위를 가지며, 0보다 클 경우 회피 기동을 시작한다. 회피 침로각의 출력은 0도부터 360도까지의 범위를 가지는 연속형 액션으로, 잠수함이 회피를 위하여 선회할 방향을 의미한다.

3.2 보상 함수 설계

에이전트의 학습을 위해서는 주어진 목적에 부합하는 보상 함수를 설계해야 한다. 본 연구의 궁극적인 목표는 잠수함이 UUV에 피격되지 않고 일정 시간 동안 생존하는 것이므로, 보상 함수는 이를 달성하도록 유도하는 방향으로 구성하여 Table 1에 정리하였다.

잠수함과 어느 한 UUV 간 거리가 일정 거리 이하로 좁혀지면 충돌로 간주한다. 이 경우 에피소드는 즉시 종료되며, 에이전트에게 부정적 보상을 준다. 이는 UUV에 피격되는 실패 상황에 벌점을 줌으로써, 에이전트가 이러한 상황을 최대한 피하도록 학습시키기 위함이다. 에피소드가 최대 시뮬레이션 시간까지 종료되지 않고 정해진 생존 시간을 모두 버티면 잠수함이 성공적으로 회피한 것으로 간주된다. 이때 에피소

드를 종료하고 에이전트에게 큰 양의 보상을 부여한다. 이를 통해 에이전트는 UUV를 끝까지 따돌리는 생존 전략을 최우선 목표로 학습하게 된다.

Table 1. Reward function designed for agent

Situation	Reward	Timing
Hit	−10	Episode termination
Survival complete	+100	Episode termination
Sonar detected	−0.1	Every time step (if detected)
Decreased distance after changing angle	−0.01	Every time step (if true)
Increased distance after changing angle	+0.01	Every time step (if true)

에피소드 도중 실패나 성공으로 종료되지 않은 중간 단계에서는, 에이전트의 행동이 얼마나 효과적으로 회피를 수행하고 있는지에 따라 작은 보상을 부여한다. 구체적으로는, 잠수함과 UUV 사이의 거리가 멀어질수록 약간의 양의 보상을 주고, 탐지되어 추적당하고 있는 상황에서는 페널티를 준다. 예를 들어, 잠수함이 회피 판단 후 회피 각도로 변침할 경우, UUV와 잠수함 간의 거리가 증가하면 소정의 보상을 추가하고, 반대로 거리가 줄어들면 벌점을 부여한다. 또한 UUV의 능동 소나에 탐지되어 추적당하는 상태(detected)에는 추가적인 부정 보상을 부여하여, 에이전트가 탐지 회피의 이점을 인식하도록 설계하였다. 이러한 보상들은 학습 초기 단계에서 탐색을 독려하고 안정적인 학습을 유도하는 역할을 한다. 단, 보상 스케일은 주요 목표(생존/격침 보상)에 비해 충분히 작게 설정하여, 주된 학습 목표를 흐리지 않는 수준에서 동작하도록 조절하였다.

정리하면, 보상 함수는 생존 시간 최대화, 격침 회피, 거리 유지 및 탐지 회피라는 목표를 반영한다. 큰 보상(또는 페널티)은 에피소드의 성공/실패 결과에 대해 주어지며, 작은 보상들은 행동의 질에 따라 부가적으로 주어진다. 이를 통해 에이전트는 UUV와의 거리를 벌리며 탐지되지 않도록 기동하는 전략을 학습하게 되고, 최종적으로 최대 시간 생존을 목표로 하는 최적 회피 정책을 찾아나가게 된다.

3.3 PPO 기반 에이전트 구조

본 연구에서 채택된 PPO 알고리즘은 강화학습 기반 에이전트의 안정적이고 효율적인 학습을 위해 설계된 정책 경사(policy gradient) 방법의 하나이다. PPO는 trust region policy optimization(TRPO)의 아이디어를 기반으로 하면서도 구현이 더 간단하고, 학습 과정에서 정책 업데이트의 안정성을 보장한다. 주요 특징은 정책 업데이트 시 '클리핑(clipping)' 기법을 적용하여 새로운 정책이 기존 정책에서 과도하게 벗어나지 않도록 제한함으로써, 정책 붕괴(policy collapse)나 불안정한 학습을 방지하는 점이다. 이로 인해 PPO는 높은 샘플 효율성을 가지며, 특히 연속 행동 공간(continuous action space)에서의 복잡한 의사결정 문제에 적합하다.

본 연구는 MATLAB의 Reinforcement Learning Toolbox에서 제공하는 rlPPOAgent 구현체를 활용하여 에이전트를 구축하였다. 이는 표준 PPO 알고리즘을 기반으로 하며, 사용자 정의 환경(시뮬레이션 모델)과의 연동이 용이하다. 에이전트의 전체 구조는 액터-크리틱(Actor-Critic) 아키텍처로 설계되었으며, 이는 정책(actor)과 가치 함수(critic)를 별도의 신경망으로 분리하여 학습하는 방식이다. 이러한 구조는 정책 경사를 계산할 때 가치 함수를 활용하여 분산(variance)을 줄이고, 안정적인 학습을 촉진한다. Fig. 4에 시각화된 바와 같이, 학습 루프는 시뮬레이션 환경으로부터 관측 상태를 입력받아 행동을 출력하고, 보상을 통해 신경망을 업데이트하는 반복 과정으로 구성된다.

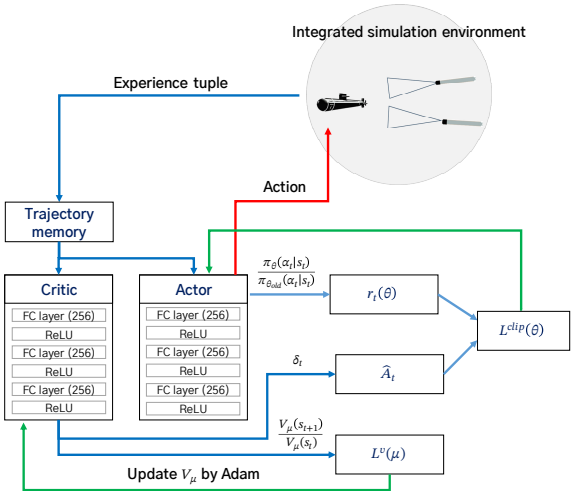


Fig. 4. Training processes based on Actor-Critc PPO structure

3.3.1 신경망 아키텍처

에이전트의 actor와 critic은 모두 다층 퍼셉트론(multi-layer perceptron, MLP) 기반의 신경망으로 구현되었다. 입력층은 상태 벡터 s_t를 관측 공간(observation space)으로부터 받으며, 이는 3.1절에서 정의된 바와 같이 잠수함과 UUV의 위치와 침로 그리고 침로 변경 횟수 등을 포함한다(식 (2) 참조).

구체적인 신경망 구조는 입력층(input layer), 은닉층(hidden layers), 출력층(output layer)으로 구분된다. 입력층은 관측 상태 벡터 s_t의 차원에 따라 결정되며, 은닉층은 3개의 완전 연결층(fully connected layer)으로 구성되어, 층마다 256개의 뉴런(노드)을 가진다. 활성화 함수로는 ReLU(rectified linear unit)가 사용되어 비선형성을 부여한다. ReLU는 학습 속도를 높이고, 그라디언트 소실 문제를 완화하는 장점이 있다. 출력층은 actor(정책망)와 critic(가치망)으로 구분된다. Actor는 행동 공간(식 (3) 참조)에 따라 2차원 가우시안 분포(평균 벡터 μ와 표준편차 벡터 σ)를 출력한다. Critic은 단일 스칼라 값인 상태 가치 $V(s)$를 출력한다. 이는 현재 상태에서 미래 누적 보상의 기대값을 나타내며, 선형 활성화(linear activation)를 사용한다.

위 신경망 구조는 실험적으로 검증된 것으로, 1-2층과 같이 너무 얕은 층은 복잡한 상태-행동 매핑을 학습하기에는 부족하고, 너무 깊은 층은 오버피팅 위험이 있으므로 3층으로 균형을 맞췄다.

3.3.2 정책 표현 및 행동 샘플링

Actor 신경망은 식 (4)와 같이 상태 s_t에 대한 정책 $\pi(a\,|\,s)$를 가우시안 분포로 근사한다.

$$\pi_\theta(a_t\,|\,s_t) = N\left(a_t\,\middle|\,\mu_\theta(s_t), diag(\sigma_\theta(s_t)^2)\right) \tag{4}$$

여기서, $\mu_\theta(s_t)$는 행동의 평균, $\sigma_\theta(s_t)$는 표준편차를 나타낸다.

학습 초기에는 σ_θ값을 초기화하여 다양한 행동을 탐험(exploration)하도록 유도한다. 학습이 진행됨에 따라 PPO 업데이트가 누적 보상 정보를 기반으로 σ_θ를 자연스럽게 줄여, 결정론적(deterministic) 정책으로 수렴한다. 행동 샘플링 시 가우시안 분포로부터 무작위 샘플링을 수행하나, 클리핑 기법으로 인해 과도한 탐험이 제한된다.

3.3.3 가치 함수 및 어드밴티지 추정

Critic 신경망은 상태 가치 함수 $V(s)$를 예측하며, 이는 정책 π 하에서 상태 s_t로부터 얻을 수 있는 장기 기대 보상이다. 학습 시 critic은 에피소드에서 수집된 실제 보상 데이터를 바탕으로 업데이트된다. 어드밴티지(advantage) 추정에는 식 (5)와 같이 generalized advantage estimation(GAE) 방법을 사용한다.

$$\widehat{A_t} = \sum_{k=0}^{\infty} (\gamma\lambda)^k \left[r_{t+k} + \gamma V_\mu(s_{t+k+1}) - V_\mu(s_{t+k})\right] \tag{5}$$

여기서, $V_\mu(s)$는 critic 신경망에 의해 파라미터 μ로 근사된 가치 함수를 나타내며, 무한 합은 미래의 모든 타임 스텝에 대한 가중된 TD 오류(temporal difference error)를 반영한다. γ는 할인율(discount factor: 0.99), λ는 GAE factor(0.95)이다(Table 2 참조).

GAE는 bias-variance trade-off를 조절하여 안정적인 어드밴티지 추정을 제공한다.

3.3.4 손실 함수 및 최적화

PPO의 핵심은 클리핑된 대리 목적 함수(surrogate objective)를 통해 정책을 업데이트하는 것이다. 정책 비율 $\gamma_t(\theta)$는 $r_t(\theta) = \dfrac{\pi_\theta(a_t\,|\,s_t)}{\pi_{\theta_{old}}(a_t\,|\,s_t)}$ 로 표현되며 이에 기반한 클리핑 손실 $L^{clip}(\theta)$은 식 (6)과 같다.

$$-E_t\left[\min\left(r_t(\theta)\widehat{A_t}, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right] \tag{6}$$

여기서 ϵ은 클립 팩터(0.2, Table 2 참조)로, 정책 변화의 상한을 설정한다.

Critic의 손실은 평균 제곱 오차(MSE)를 최소화하는 방식으로 진행되며, 식 (7)과 같다.

$$L^v(\mu) = E_t\left[\left(R_t - V_\mu(s_t)\right)^2\right] \tag{7}$$

여기서, $(V_\mu(s_t))^2$는 타겟 가치(GAE 기반) 이다.

탐험성을 유지하기 위한 엔트로피를 포함하여 최종적으로 학습에 사용되는 총 손실 함수는 식 (8)과 같이 표현된다.

$$L(\theta,\mu) = L^{clip}(\theta) + c_v L^v(\mu) + L^{ent}(\theta) \quad (8)$$

여기서, $L^{ent}(\theta)$는 정책 엔트로피 손실을 나타내며, C_v는 가치 손실의 가중치이다.

위 손실 함수를 기반으로, Adam 옵티마이저를 적용하여 학습을 수행하였다. 정리하면, actor 망은 회피 판단과 회피 기동을 결정하는 정책을 표현하고, critic 망은 그 정책의 성과를 평가하는 가치 함수를 근사한다. 두 신경망 모두 앞서 정의한 상태 벡터 입력에 기반하며, 은닉층 규모, 깊이 등 실험적으로 검증된 안정적인 학습 설정을 채택하였다. 실험에 사용된 하이퍼파라미터들은 Table 2에 정리하였다.

Table 2. Hyperparameters used in the experiment

Parameters	Value
Clip factor	0.2
Entropy loss weight	0.03
GAE factor	0.95
Discount factor	0.99
Experience horizon	2048
Optimizer	Adam
Activation function	ReLU
Num epoch	5
Mini batch size	128
Learn rate (actor)	3e-4
Learn rate (critic)	3e-4
Sample time	0.1 sec

4. 시뮬레이션 결과 및 분석

제안된 PPO 기반 잠수함 회피 에이전트의 성능을 평가하기 위한 시뮬레이션을 수행하였다.

주요 실험은 두 개의 UUV의 공격 시나리오로, 이는 다중 위협 상황에 대한 기본적인 구성이다. 비교 대상은 상대 거리가 기준인 규칙 기반 회피 전략으로, 기존 연구에서 사용되는 일정한 지그재그 회피 기동과 고속 항주를 조합한 방식이다. 이 전략은 고정된 탐지 거리 임계값에 도달할 때마다 회피를 개시하며, 구체적으로 UUV와의 거리가 5,000 m, 3,500 m, 2,000 m, 500 m에 도달할 때 순차적으로 강제 기동을 수행하도록 설계되었다. 이 방식은 기계적으로 동일한 회피 패턴을 반복하며, 위협 상황 변화에 유연하게 대응하지 못하는 한계가 있다. 반면, PPO 기반 강화학습 에이전트는 회피 시점과 기동 형태를 상태(탐지 신호, 상대 거리 등)에 따라 결정한다.

4.1 회피 기동 방식에 따른 피격률 비교 분석

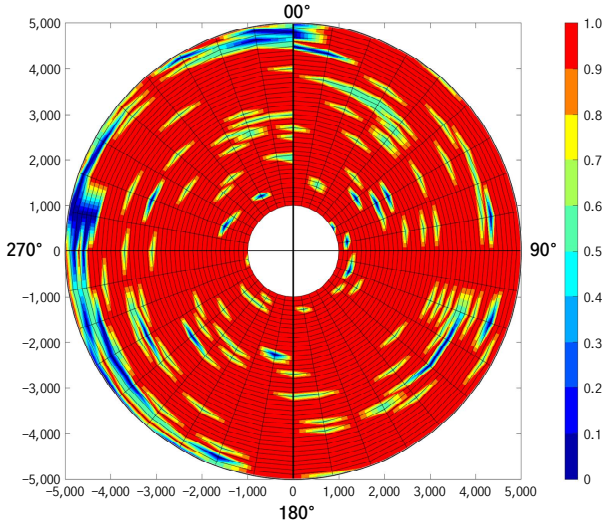
실험 환경은 Table 3와 같으며, 심해 환경을 고려하여 잠수함은 초기에 진북 방향(00°)으로 향하며 최대 수심은 250 m, UUV는 최대 300 m로 심도 하한을 설정하였다. 거리는 최소 1,000 m에서 100 m 간격으로 최대 5,000 m로 설정하였고, 방위는 10° 간격으로 전방위에 대한 실험을 진행하였다.

Table 3. Parameters for the experimental environment

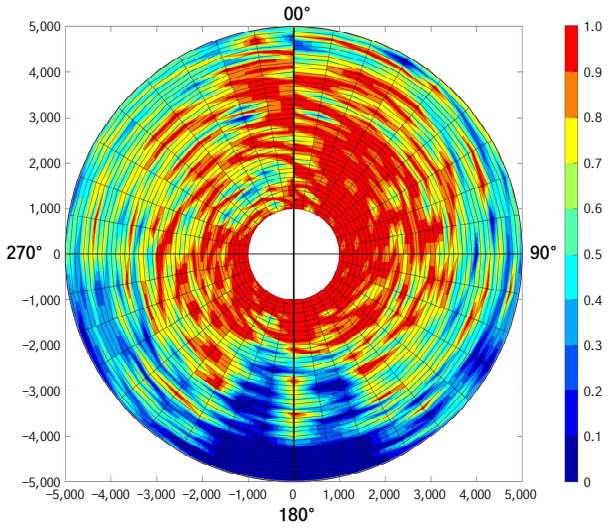
Parameters	Value
Range	1,000 m – 5,000 m
Angle	0° – 360°
Depth limit	300 m

Fig. 5(a)는 규칙 기반 회피 시, Fig. 5(b)는 강화학습 기반 회피 시 각 전략의 생존율을 거리와 방위에 따른 피격률을 통해 보여준다. 생존율은 100 %에서 명중률을 감하여 산출하였으며, 전체 명중률은 모든 셀의 명중률 값을 산술 평균하여 근사치로 사용하였다. Fig. 5(a)에서 볼 수 있듯이, 상대 거리가 기준인 규칙 기반 회피 기동의 생존율은 7.48 %에 불과한 반면, 강화학습 기반 회피 기동의 생존율은 29.4 %를 기록하며 약 4배 가까운 성능 격차를 보였다. 잠수함의 초기 진행 방향(00°)인 앞부분에서 UUV가 나타날 경우에서는 두 회피 기동 방법 모두 높은 피격률을 보였으나, UUV가 잠수함의 뒷부분인 90° – 270°에 위치한 경우 강화학습 기반의 생존율 결과가 급격히 좋아진 것을 보여준다.

Fig. 6는 거리 5,000 m, 방위 120°에서의 회피 궤적(빨강색)과 UUV의 궤적(파랑색, 초록색)을 보여주며, Table 4는 Fig. 6의 상황에서의 회피 로그를 보여준다. Fig. 6와 Table 4에서 볼 수 있듯이, 회피 시점과 기동 방식이 상황에 따라 유연하게 변화하였다.



(a) Hit rate of rule-based evasion strategies



(b) Hit rate of reinforcement learning-based evasion strategies

Fig. 5. Comparative analysis of hit rate according to evasive maneuver method

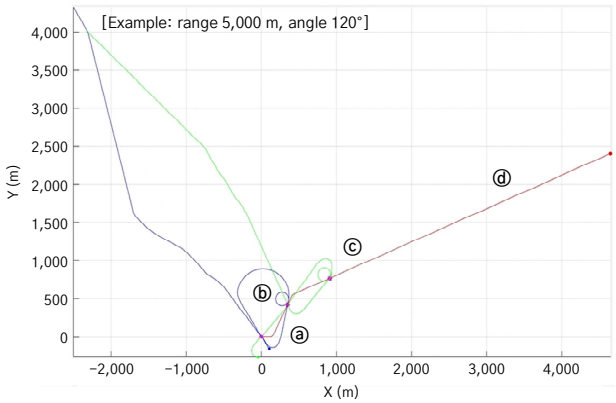


Fig. 6. Evasion trajectory during PPO-based evasion maneuver

Table 4. Evasion log in combat situations

Evasion log	Angle	Evade decision
ⓐ	53°	Yes
ⓑ	41°	Yes
ⓒ	0°	Yes
ⓓ	None	No

ⓐ 시점에서는 잠수함이 첫 UUV를 인식하고 53°로 비교적 큰 각도의 회피 기동을 수행하였다. 이는 탐지 신호의 강도가 상당히 높은 상황에서 급격히 방향을 전환하여 메인 로브 탐지각에서 벗어나려는 의도가 반영된 것으로 보인다.

ⓑ 시점의 41° 기동은 다중이 된 UUV를 인식하고 추가적인 회피가 이루어지는 모습을 보여준다. 다중 UUV가 접근 중인 것을 고려한 회피를 시도한 사례로 해석된다.

ⓒ 시점에서는 회피각이 0°로 기록되었으나, 회피 의사결정은 'Yes'로 나타났다. 즉, 탐지가 예상되었으나 기동 각도 변경 대신 부유식 기만기를 통해 탐지 억제를 선택한 사례에 해당한다. 실제 강화학습 에이전트는 회피 행동을 반드시 방향 변화로만 정의하지 않으며, 침로를 고정한 채 기만기만 발사하는 등의 전략도 회피 행동으로 간주한다.

ⓓ 시점에서는 회피각이 'None'으로 표시되고 회피 의사결정이 'No'로 나타났다. 이는 위험 신호가 충분히 낮거나 탐지 범위를 벗어났다고 판단해, 회피 기동을 보류한 것으로 해석된다. 이러한 선택은 잠수함의 에너지 보존과 같은 탐지 회피의 효율을 고려해 학습된 PPO 에이전트의 행동 패턴을 반영한다.

종합적으로 로그 분석을 통해 강화학습 기반 회피 에이전트가 단순히 일정 거리나 탐지 신호에 기계적으로 반응하는 것이 아니라, 상대적 위협 수준과 기동 비용을 종합적으로 고려해 상황별로 다양한 대응 전략을 구사하고 있는 것으로 나타났다. 특히 동일한 회피 의사결정('Yes') 하에서도 기동 각도가 다르거나 0°로 유지되는 사례는 전통적인 규칙 기반 기동과 질적으로 다른 유연하고 적응적인 회피 전략이 학습되었음을 보여준다. 또한, 위험이 낮다고 판단되는 시점에 회피를 보류('No')하는 사례도 나타나, 불필요한 기동으로 인한 탐지 노출을 억제하는 학습된 행동이 확인되었다. 이러한 특성은 PPO 에이전트가 보상 설

계에 따라 탐지 회피뿐 아니라 생존 시간 등 여러 요소를 통합적으로 고려해 전략을 최적화했음을 보여준다.

5. 결론

본 연구에서는 다중 UUV 위협 환경에서 잠수함의 생존성을 극대화하기 위한 강화학습 기반 회피 전략을 제안하고 검증하였다. 특히, 기존의 고정 패턴 기반 회피 기동과 달리 PPO 알고리즘을 적용하여 잠수함이 상황별로 최적의 기동을 유연하게 선택할 수 있도록 학습시켰다.

시뮬레이션 결과, 강화학습 에이전트는 기존 규칙 기반 전략에 비해 약 4배 가까운 우수한 생존율을 보였다. 질적 분석 측면에서도 강화학습 에이전트는 탐지 신호 강도, UUV의 접근 각도 등의 상태 정보를 종합적으로 고려해 회피 시점과 기동 형태를 결정하였다. 예를 들어, 일정 거리에서 기계적으로 회피를 개시하는 기존 전략과 달리, PPO 에이전트는 위험이 급증하는 순간에 즉시 큰 각도로 기동하거나, 상대적으로 낮은 위협에는 현 침로를 유지하고 기만기만 발사하는 등 정교한 회피 전략으로 이어졌다. 이를 통해, 음향 탐지 모델인 소나 방정식과 강화학습을 결합함으로써 고도화된 자동 회피 정책을 도출할 수 있음을 입증하였다.

향후 연구로는 음선을 고려한 해수면, 해저 지형 반사, 잔향이 소음보다 지배적인 환경 등 보다 복잡한 해양 환경 요소를 포함해 시뮬레이션의 현실성을 높인 연구가 필요하다.

참고문헌

[1] B. Kang and W. Yun, “Hierarchical Reinforcement Learning for Submarine Torpedo Countermeasures and Evasive Manoeuvres,” IEEE Access, 2024.
[2] J.-M. Pak, B.-H. Ku, Y.-H. Lee, D.-G. Ryu, W.-Y. Hong, H.-S. Ko, and M.-T. Lim, “Effectiveness Analysis for a Lightweight Torpedo Considering Evasive Maneuvering and Torpedo Acoustic Counter Measures of a Target,” Journal of the Korea Society for Simulation, Vol. 20, No. 4, pp. 1-11, 2011.
[3] A. Mjelde, “A Homing Torpedo: The Effect of the Tactical Situation and the Torpedo Parameters on the Torpedo Effectiveness,” Ph.D. dissertation, Naval Postgraduate School, Monterey, CA, USA, 1977.
[4] K. R. Armo, “The Relationship Between a Submarine’s Maximum Speed and Its Evasive Capability,” M.S. thesis, Naval Postgraduate School, Monterey, CA, USA, 2000.
[5] J.-H. Chung, G.-S. Kim, S.-H. Park, J.-H. Kim, and W. Yun, “Reinforcement Learning-Based Deception Tactics for Torpedo Threat Evasion,” Journal of the Korean Institute of Communications and Information Sciences, Vol. 49, No. 3, pp. 333-345, 2024.
[6] R. J. Urick, Principles of Underwater Sound, 3rd ed. New York, NY, USA: McGraw-Hill, 1983.
[7] H. Medwin and C. S. Clay, Fundamentals of Acoustical Oceanography. New York, NY, USA: Academic Press, 1998.