Fu *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:116
https://doi.org/10.1186/s13634-023-01054-7

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Model-based optimal action selection for Dyna-Q reverberation suppression cognitive sonar

Yubin Fu[1,2*], Xiaochuan Ma[1,2*], Chao Feng[1], Xingxuan Pei[1,2] and Pengzhuo Li[1,2]

*Correspondence: fuyubin@mail.ioa.ac.cn; maxc@mail.ioa.ac.cn

[1] Laboratory of Autonomous Underwater Vehicles, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China

Abstract

The Doppler shift of low-speed targets is frequently disturbed by the reverberation Doppler spread clutter under the shallow sea. The clutter is generated by underwater scatterers, which increases the difficulty of Doppler estimation. To solve this problem, a reverberation target resolution function based on the Doppler spread clutter statistical model is proposed in this paper. Through the width of reverberation Doppler clutter, this function adjusts the waveform parameters by determining whether the target is discriminable. In addition, the reverberation Doppler spread clutter is time-spatial varying and affected by grazing angle, waves, wind speed, fish and other effects. Thus, the sonar waveform parameters need to be adjusted constantly. Therefore, this paper combines the cognitive sonar based on reinforcement learning with the reverberation target resolution function to evaluate different waveforms in different environments. Consequently, the sonar can adjust the waveform parameters in real-time and obtain the optimal waveform in different environments. Meanwhile, in this paper, the action selection strategy of Dyna-Q reinforcement learning is optimized, and the model-based maximum action selection Dyna-Q algorithm (Dyna-Q-Max-Action) is proposed. Compared with the traditional Dyna-Q and Q-learning algorithms, the proposed algorithm needs fewer episodes. Finally, numerical simulation verified the effectiveness of the proposed algorithm.

Keywords: Dyna-Q-Max-Action, Reinforcement learning, Reverberation Doppler spread clutter, Cognitive sonar

1 Introduction

For active sonar, low-speed small target detection is a challenging problem in engineering applications. In coastal and harbor areas, a large number of reefs, artificial facilities, fish and other strong scatterers, coupled with the movement of the sea surface, platforms and multipath effects, make the reverberation complex and variable. Consequently, it may cause severe Doppler spread, and the low-speed target is blended with high-level energy clutter. Due to the time-varying clutter, suppressing the Doppler spread clutter in real-time is the key to reverberation suppression.

Traditional methods of reverberation suppression include array design, waveform design and post-processing algorithm design. Typically, the essence of array design is

to design narrow beamwidth [1], reduce the spatial size of the resolving unit to increase the signal-to-reverberation ratio (SRR). Moreover, the essence of the waveform design method is to increase the product of pulse width bandwidth and improve the matched filter gain [2]. In addition, the post-processing algorithm design contains pre-whitening [3], spatial processing method [4], principal component inverse (PCI) subspace method [5], graphic method to suppress reverberation [6]. Pre-whitening converts reverberation from colored noise into white noise, improves output signal-to-noise ratio (SNR) with matched filters, improves receiver and increases active sonar detection capability. The PCI subspace method decomposes the echo into reverberation subspace and target echo plus white noise subspace. Since the energy of the reverberation subspace is stronger than the target plus white noise subspace components. Hence, setting a threshold to subtract the stronger components to suppress reverberation. In addition, the graphic approach is to improve the contrast between the reverberation and the target in the active sonar image, thus improving the detection capability. However, the above method unable to adjust the detection waveform according to the time-varying reverberation environments. Furthermore, the above method only adaptively adjusts the receiver side without joint the transmitter side for reverberation suppression, the active sonar computational resource is wasted and the parameter estimation efficiency is reduced.

In 2006 Haykin [7] proposed the concept of cognitive radar based on the echo-location system of bats. The cognitive radar consists of three elements: (1) the radar learns from environments; (2) the radar transmitter is interacted with receiver; (3) the radar preserves echo information. The above-mentioned adaptive radar methods make adaptive adjustments to the receiver. In contrast to the adaptive radar, the cognitive radar can adjust the transmit waveform parameters jointly with the "transmitter-receiver" over a long period of time.

The underwater acoustic channel is a severe time-doppler dual spread channel that enlarge the demand for active sonar waveform design freedom. In 2011, L. Xiaohua combined environmental information and target prior knowledge proposed the cognitive sonar, with reference to the cognitive radar. The cognitive sonar adjusts the transmit waveform parameters according to the echo [8]. In 2015, Tim Claussen [9] used Doppler processing and real-time interpolation to adjust the cognitive sonar transmit beamformer. 2016, X. Qing combined bionics and dolphin research [10] increased the freedom of cognitive sonar waveform design and proposed the cognitive sonar waveform (CSW). CSW combines the ambiguity function (AF) and Q function to constrain the waveform parameters, such as pulse width, frequency and the number of pulse trains, in order to suppress reverberation. Conversely, cognitive sonar cannot get optimal waveform rapidly, takes a lot of time does not adapt to a rapid time-varying environment.

Recently, with increased computer arithmetic and reduced difficulty in acquiring training data, machine learning has shown amazing performance in many areas, such as target recognition, acoustic confrontation, interference suppression, etc. Moreover, the optimal solution to the above problems is the non-deterministic polynomial (NP) hard problem. The sub-optimal method like reinforcement learning (RL) can provide solutions to the NP problem. RL learns by interacting with the environment through rewards and punishments, then continuously adjusting action towards a higher reward. RL is widely used. In the field of Go [11], to acquire decision-making

capabilities by learning previous game actions and using them to subsequent actions. As in the field of the game [12], RL can optimize the Super Mario's actions based on the environment to find the optimal path to the goal quickly. In 2018, Jason E. [13] proposed RL combined with cognitive sonar, which can reduce the time to acquire optimal waveform. In 2022, Jeff Tucker [14] combined RL with cognitive sonar for multi-target detection and tracking, by adjusting the sonar waveform to learn from the environment. The efficiency of target detection and tracking is greatly improved. Cognitive sonar based on RL has important implications and extensive research space.

RL is a Markov decision process (MDP) consisting of action, reward signal, transfer probability, model of environment, etc. The analogy of cognitive sonar based on RL is waveform, reward signal, transfer probability, echo information, etc.

The reward signal is defined by the goal of RL which represent the value of waveform selection. The underwater environment is complex, the fixed waveform cannot adapt to the time-varying environment, hence the waveforms and environment information are trained in RL to obtain the optimal waveform. The choice of the reward function is important, and irrational choice may lead to algorithm failure. In this paper, a target resolution function is proposed, which can directly detect whether the reverberation Doppler spread clutter contains a target. Since a Doppler-sensitive waveform can suppress the reverberation [15]. If the target cannot be detected, the cognitive sonar will change the waveform parameters to suppress the reverberation and distinguish the target, as shown in Fig. 1.

The convergence episodes of RL directly influence the real-time update efficiency of the active sonar. Traditional RL algorithms include the Q-Learning algorithm and the state action reward state action (SRASA) algorithm. These are all temporal-difference (TD) single-step update algorithms, which have a slow convergence speed. In order to make full use of existing knowledge, Sutton introduced the planning process in RL and proposed the model-based Dyna RL algorithm [16]. This type of algorithm stores experience knowledge by building a model of the environment and generates simulated experience to train the learning machine offline. Since the action selection of the model samples directly determines the efficiency of the algorithm. The introduction of the planning process gives RL the 'cognitive' ability, getting rid of simple trial-and-error learning and greatly improving the convergence efficiency of the algorithm.

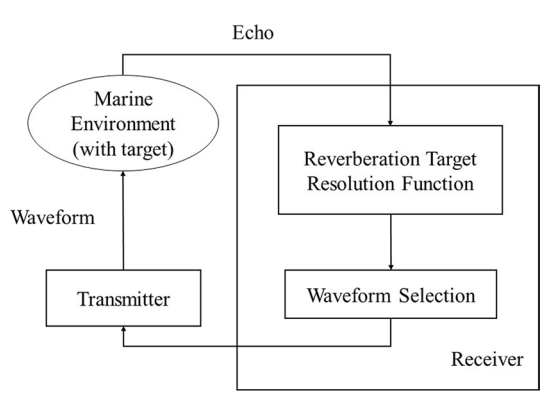


Fig. 1 Reverberation suppression cognitive sonar flow chart

In the Dyna framework, the action selection of the planning directly affects the convergence episodes of the algorithm. Sutton initially adopts the random-sample method [16], many iterations of the value function in planning are ineffective for the algorithm. Therefore, the maximum reward action selection strategy for Dyna-Q (Dyna-Q-Max-Action) is first proposed in this paper. The convergence episodes of the algorithm are shortened by reducing the probability of invalid random-sample action selection. Thus, this algorithm has fewer convergence episodes than the Dyna-Q algorithm convergence episodes. Cognitive sonar based on RL combines the advantage of RL to solve complex problems with the advantage of cognitive sonar to interact with the environment in real-time. It can solve the problem of difficulty in obtaining optimal waveform for complex environments by active sonar theoretically.

This paper is structured as follows. Section 2 describes the statistical model of reverberation Doppler spread clutter, and proves the statistical model with the real data from the Qiandao Lake. Furthermore, analysis of the factors of waveform parameters affecting the reverberation Doppler spread clutter. The target resolution function in reverberation Doppler spread clutter is proposed. The Dyna-Q-Max-Action algorithm is proposed by optimizing the action selection strategy of the Dyna-Q algorithm. Moreover, the principle and algorithm flow of the Dyna-Q-Max-Action algorithm is described in detail. Section 3 discusses numerical simulation results of the Dyna-Q-Max-Action cognitive sonar combing with the reverberation target resolution function. Meanwhile, analysis of the influence of the model training episodes and action selection probability on the Dyna-Q-Max-Action algorithm. Then, conclude a reasonable action selection probability. Conclusions are given in Sect. 4.

2 Methods and experiments

2.1 Reverberation modeling

The establishment of reverberation statistical model is significant for reverberation analysis and simulation. Etter and other researchers mentioned two difficulties in reverberation statistical modeling: the lack of analytical tools for solving complex boundary problems, and the difficulty of measuring and distinguishing too many influencing factors of reverberation [17–19]. According to the relationship between the size of the scatterer and the wavelength of the acoustic wave, the reverberation modeling can be divided into the point scattering model and the unitary scattering model. In this paper, a more realistic point scattering model is used for modeling.

The point scattering model assumes that scatterers are randomly distributed in the ocean and the reverberation is the superposition of all scatterer backscattered echoes. The point scattering model has a clear physical meaning and can directly assume the statistical properties of the scattered echo amplitude, phase, and Doppler shift. Considering the small amplitude of the multiple scattering, the multiple scattering effect is ignored. Under the assumption of the narrow-band waveform, the reverberation can be described as

$$R(t) = \sum_{i=1}^{N(t)} A_i s(t - \tau_i) \mathrm{e}^{\mathrm{j}2\pi\phi_\mathrm{i}\mathrm{t}} \tag{1}$$

where $N(t)$ is the number of scatterers contributing to the reverberation at time t. A_i,τ_i and ϕ_i, denote the amplitude, time delay, and Doppler shift of the i-th scatterer echo, respectively. $s(t)$ denotes the waveform. In addition, the statistical distribution of the echo amplitude A_i determines the statistical distribution model of the reverberation envelope. Moreover, the Doppler shift ϕ_i of the echo determines the Doppler shift of the reverberation, representing the severity of the Doppler spread. Each scatterer A_i obeys Gaussian distribution, τ_i and ϕ_i are independent of each other. Therefore, satisfying the above conditions is the wide-sense stationary uncorrelated scattering (WSSUS) channel reverberation.

According to reference [20], a reverberation Doppler spread clutter model is established. It is known that the mathematical model probability density function (PDF) of reverberation Doppler spread distribution conforms to the two-side exponential distribution. The two-side exponential distribution is the distribution model with the highest matching according to the real sea test data. The two-side exponential distribution is symmetrically distributed about $v = 0$, the parameters of the distribution can be used to characterize the severity of the Doppler spread.

$$\rho(v) = \frac{\mu}{2}\mathrm{e}^{-\mu|\mathrm{v}|} \tag{2}$$

where μ represents the slope ss of the Doppler spread. A large number of sea test data show that the ss is ranging between 6 and 20 dB/kn. The above equation is also known as the Doppler spreading function. μ can be solved by the definition of ss

$$ss = 20\lg\left[\frac{\rho(0)}{\rho(1)}\right] \tag{3}$$

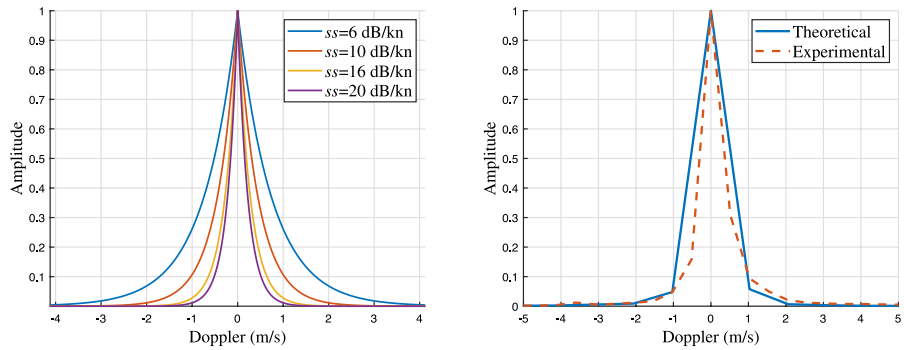
then

$$\mu = \ln\left(10^{\frac{ss}{20}}\right) \tag{4}$$

C. Zhang proposed a pseudo-random number generation method for reverberation numerical simulation in combination with the two-side exponential distribution model [21], and verified the model with the Qiandao Lake experiment data. Set y to be a uniformly distributed random number in the interval (0, 1), this leads to

$$v = \begin{cases} -\frac{1}{\mu}\cdot\ln y, v > 0 \\ \frac{1}{\mu}\cdot\ln y, v < 0 \end{cases} \tag{5}$$

According to Eq. (5), a random number v obeying the statistical model with the specified Doppler spread can be generated. Using the point scattering model, a scattering model with 10,000 points is established. v is converted into a Doppler shift ϕ_i using the formula $\phi_i = f_0\frac{2v}{\mathrm{c}}$ calculated by substituting the reverberation formula Eq. (1), where the amplitude obeys a Gaussian distribution with mean 0 and variance 1. The time delay obeys the uniform distribution, ss=20 dB/kn, 1 kn = 0.514 m/s. The Doppler shift obeys the two-side exponential distribution $\mu = 2.30$. The reverberation length is 3 s. Continuous wave (CW) pulse width is 0.58 s, and center frequency f_0= 4000 Hz. The reverberation is calculated 100 times using the Monte Carlo, and the time domain results of the



(a) The Doppler distribution curve (b) The Doppler spread clutter distribution curve

Fig. 2 Reverberation modeling and Doppler clutter distribution

reverberation correlation function are superimposed. Thus, the reverberation Doppler distribution curve is obtained, and the modeling results are as follows:

Figure 2a is modeled with Eq. (2), it is seen that the Doppler spread distribution is more concentrated with the increase of *ss*, and the spreading is not apparent. Figure 2b theoretical curve is modeled with the two-side exponential distribution by Eq. (5), it is obvious to find that the theoretical curve is basically consistent with the experimental Qiandao Lake data, indicating that the two-side exponential distribution model can be used for Doppler spread clutter modeling.

2.2 Waveform and reverberation Doppler distribution

Different waveform parameters influence the environment feedback reverberation. Correlation is widely used in reverberation analysis, the Doppler distribution of the reverberation is obtained by superimposing the correlation on the time delay. The following paper mainly focuses on the relationship of waveform and reverberation Doppler distribution.

2.2.1 AF of waveform

The most important method to measure the detection capability of the waveform is the AF. The expression of the AF is obtained by doing time-Doppler domain matched filtering of the transmit waveform [22] as

$$|\chi_s(\tau,\phi)|^2 = |\int s(t)s^*(t+\tau)e^{j2\pi t\phi}dt|^2 \tag{6}$$

The above equation is the AF of the transmit waveform $s(t)$. τ is the time delay and ϕ is the Doppler shift. Through the AF of the waveform, the measurement accuracy, error ellipse, and intrinsic resolution constant of the waveform can be found. Within the 3 dB range, there is a time-Doppler resolvability range $\chi(0,0)$ for the waveform. The Doppler resolution of the waveform can be calculated

$$\chi_s(\tau,\phi) = \frac{\sqrt{2}\chi_s(0,0)}{2} \tag{7}$$

In this paper, without any special explanation, the square wave envelope CW is $s(t) = T^{-1/2}\mathrm{rect}(\mathrm{t/T})\mathrm{e}^{\mathrm{j}2\pi \mathrm{f}_0\mathrm{t}}$. According to the equation $\int_{-T/2}^{T/2} \mathrm{e}^{\mathrm{j}2\pi \mathrm{ft}}\mathrm{dt} = \mathrm{Tsinc}(\pi \mathrm{fT})$. The AF of the CW is

$$|\chi_s(\tau,\phi)|^2 = \begin{cases} \left(1-\frac{|\tau|}{T}\right)^2 \mathrm{sinc}^2\left[\pi\phi T\left(1-\frac{|\tau|}{T}\right)\right], |\tau| \le T \\ 0, |\tau| > T \end{cases} \tag{8}$$

The Doppler resolution of the CW can be calculated as $\phi_{\mathrm{Hz}} = \pm 0.44/T(\mathrm{Hz})$. According to the formula $\phi_{\mathrm{wave}} = \frac{\phi_{\mathrm{Hz}}\mathrm{c}}{2f_0}$, the Doppler resolution $\phi_{\mathrm{wave}} = \frac{0.44\mathrm{c}}{2Tf_0}(\mathrm{m/s})$ can be calculated. c is the speed of the acoustic. Based on ϕ_{wave}, it can be seen that the CW AF is mainly affected by the joint influence of pulse width T and frequency f_0. In this paper, the CW center frequency f_0 =4000 Hz is fixed, only the effect of pulse width on Doppler resolution is considered.

2.2.2 Reverberation Doppler distribution

In general, the echo $X_0(t)$ can be composed of target $Tar(t)$, reverberation $Rev(t)$, and noise $Noi(t)$

$$X_0(t) = Tar(t) + Rev(t) + Noi(t) \tag{9}$$

$Noi(t)$ is the environment noise, the target echo is $Tar(t) = s(t-\tau_t)\mathrm{e}^{\mathrm{j}2\pi\phi_\mathrm{t}\mathrm{t}}$, τ_t and ϕ_t, are the target time delay and target Doppler shift, respectively. Each scatterer echo is a copy of the waveform with time delay and Doppler shift of different intensities, so the above echo $X_0(t)$ can be reduced to

$$X_0(t) = A_t s(t-\tau_t)\mathrm{e}^{\mathrm{j}2\pi\phi_\mathrm{t}\mathrm{t}} + \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{Rev_j s}\left(\mathrm{t}-\tau_\mathrm{j}\right)\mathrm{e}^{\mathrm{j}2\pi\phi_\mathrm{j}\mathrm{t}} + \mathrm{Noi(t)} \tag{10}$$

A_t, Rev_j represent the target, reverberation amplitude. τ_t, τ_j represent the target, reverberation time delay. ϕ_t, ϕ_j represent the target, reverberation Doppler shift. N represents the number of reverberation scatterers. Correlating the echo with the time delay and Doppler shift copy, obtain

$$\mathrm{R}_{sX}(\tau,\phi) = A_t\chi_s(\tau_t,\phi-\phi_t) + \sum_{j=1}^{N} Rev_j\chi_s\left(\tau_j,\phi-\phi_j\right) \tag{11}$$

Since reverberation is the environment interference of active sonar. In the field of acoustic engineering, the normalized reverberation channel is a typical WSS channel [23]. Under the assumption that the scatterers are independent of each other, reverberation channel satisfy both US channel and WSS channel. Thus, the reverberation characteristics can be expressed by the reverberation channel scattering function.

The time-varying impulse response of the reverberation channel is $\mathrm{g}\left(\tau',\mathrm{t}\right)$, τ' is the time delay at different times t, and the received reverberation is $Rev(t)$, then the reverberation can be expressed as the convolution of the transmitted waveform and the impulse response

$$Rev(t) = \int \mathrm{g}(\tau', \mathrm{t})\mathrm{s}(\mathrm{t} - \tau')\mathrm{d}\tau' \tag{12}$$

The Fourier transform $\mathrm{g}(\tau', \mathrm{t})$ of the time-varying impulse response $P_s(\tau', \phi')$ is called the spreading function.

$$\mathrm{g}(\tau', \mathrm{t}) = \frac{1}{2\pi} \int \mathrm{P_s}(\tau', \phi') \exp\left(\mathrm{j}2\pi\phi'\mathrm{t}\right)\mathrm{d}\phi' \tag{13}$$

Ignoring the constant term, substituting Eq. (13) into Eq. (12) yields

$$Rev(t) = \iint P_s(\tau', \phi')s(t - \tau') \exp\left(\mathrm{j}2\pi\phi'\mathrm{t}\right)\mathrm{d}\tau'\mathrm{d}\phi' \tag{14}$$

The above equation shows that the reverberation is related to the channel spreading function and the waveform in the time-Doppler domain. Thus, the reverberation can be composed by the channel spreading function $P_s(\tau', \phi')$ multiplied by the time delay, Doppler shift weighted waveform.

By correlating $Rev(t)$ with the waveform $s(t)$, obtain

$$\begin{aligned}
R_{Rs}(\tau, \phi) &= \int Rev^*(t + \tau)s(t) \exp(\mathrm{j}2\pi\phi\mathrm{t})\mathrm{d}\mathrm{t} \\
&= \iiint P_s^*(\tau', \phi')s^*(t - \tau' + \tau) \exp\left(-\mathrm{j}2\pi\phi'(\mathrm{t} + \tau)\right)s(t) \exp\left(\mathrm{j}2\pi\phi\mathrm{t}\right)\mathrm{d}t\mathrm{d}\phi'\mathrm{d}\tau' \\
&= \iiint P_s^*(\tau', \phi') \exp\left(-\mathrm{j}2\pi\phi'(\mathrm{t} + \tau) + \mathrm{j}2\pi\phi\mathrm{t}\right)s^*(t + \tau - \tau')s(t)\mathrm{d}t\mathrm{d}\phi'\mathrm{d}\tau' \\
&= \iiint P_s^*(\tau', \phi') \exp\left(\mathrm{j}2\pi\mathrm{t}(\phi - \phi') - \mathrm{j}2\pi\phi'\tau\right)s^*\left(t + (\tau - \tau')\right)s(t)\mathrm{d}t\mathrm{d}\phi'\mathrm{d}\tau' \\
&= \iint P_s^*(\tau', \phi')\chi_s(\tau - \tau', \phi - \phi') \exp\left(-\mathrm{j}2\pi\phi'\tau\right)\mathrm{d}\phi' d\tau'
\end{aligned} \tag{15}$$

Do autocorrelation on Eq. (15).

$$\begin{aligned}
\langle R_{Rs}(\tau, \phi)R_{Rs}^*(\tau, \phi)\rangle = &\iiiint R_{ps}(\tau', \phi')\delta(\tau' - \tau_1)\delta(\phi' - \phi_1) \\
&\chi_s(\tau - \tau', \phi - \phi') \exp\left(-\mathrm{j}2\pi\phi'\tau\right) \\
&\chi_s^*(\tau - \tau_1, \phi - \phi_1) \exp\left(\mathrm{j}2\pi\phi_1\tau\right)\mathrm{d}\phi' d\tau'\mathrm{d}\phi_1\mathrm{d}\tau_1 \\
= &\iint R_{ps}(\tau', \phi')|\chi_s(\tau - \tau', \phi - \phi')|^2 d\phi'\mathrm{d}\tau'
\end{aligned} \tag{16}$$

where $E\left[P_s(\tau_1, \phi_1)P_s^*(\tau', \phi')\right] = R_{ps}(\tau', \phi')\delta(\tau' - \tau_1)\delta(\phi' - \phi_1)$. δ represents the Dirac delta function. $R_{ps}(\tau, \phi)$ is the scattering function of the channel which is used to describe the time-Doppler distribution. When the autocorrelation is at τ, ϕ, Eq. (16) can reduce to

$$\langle R_{Rs}(\tau, \phi)R_{Rs}^*(\tau, \phi)\rangle_{\tau,\phi} = R_{ps}(\tau, \phi) * |\chi_s(\tau, \phi)|^2 \tag{17}$$

The above equation shows that the autocorrelation of the $Rev(t)$ matched filter is the two-dimensional convolution of waveform AF and reverberation channel scattering function. Under the assumption of WSSUS, the autocorrelation function of the reverberation echo $\sum_{j=1}^{N} Rev_j\chi_s(\tau_j, \phi - \phi_j)$ can be reduced to the two-dimensional

convolution of the channel scattering function and the waveform AF. Doing autocorrelation on Eq. (11) obtain

$$\mathrm{Rr}_{sX}(\tau,\phi) = A_t^2|\chi_s(\tau_t,\phi-\phi_t)|^2 + A_c^2\left(R_{ps}(\tau,\phi) * |\chi_s(\tau,\phi)|^2\right) \tag{18}$$

A_c represents the amplitude of the reverberation. $\mathrm{Rr}_{sX}(\tau,\phi)$ is superimposed on the time delay to obtain the Doppler distribution curve $\varphi(\phi)$ of the reverberation

$$\varphi(\phi) = \int \mathrm{Rr}_{sX}(\tau,\phi)\mathrm{d}\tau = \int A_t^2|\chi_s(\tau_t,\phi-\phi_t)|^2 + A_c^2\left(R_{ps}(\tau,\phi) * |\chi_s(\tau,\phi)|^2\right)\mathrm{d}\tau \tag{19}$$

The scattering effect of the channel is a fuzzy effect. According to the approximate derivation of the reverberation point scattering model [23], the reverberation channel scattering function can be reduced to

$$R_{ps}(\tau,\phi) = K\rho(\tau,\phi) \tag{20}$$

where K is a constant. $\rho(\tau,\phi)$ is the joint distribution of the scatterer with the time and Doppler shift. The equation indicates that the reverberation scattering function is determined by the joint PDF of the τ and ϕ of the scatterer. Similarly, if the spatial distribution of the scatterer and the Doppler distribution are independent of each other, then we have

$$R_{ps}(\tau,\phi) = K\rho(\tau)\rho(\phi) \tag{21}$$

If the scatterers are uniformly distributed by distance, this equation can be further reduced to

$$R_{ps}(\tau,\phi) = \frac{K'}{T}\rho(\phi) \tag{22}$$

where T is the pulse width of the waveform. K' is a constant. The result of the reverberation correlation function can be reduced to

$$\mathrm{Rr}_{sX}(\tau,\phi) = A_t^2|\chi_s(\tau_t,\phi-\phi_t)|^2 + A_c^2\frac{K'}{T}\left(\rho(\phi) * |\chi_s(\tau,\phi)|^2\right) \tag{23}$$

The reverberation Doppler distribution curve can be reduced to

$$\varphi(\phi) = \int A_t^2|\chi_s(\tau_t,\phi-\phi_t)|^2 + A_c^2\frac{K'}{T}\left(\rho(\phi) * |\chi_s(\tau,\phi)|^2\right)\mathrm{d}\tau \tag{24}$$

In summary, it can be seen that the reverberation Doppler distribution is influenced by the waveform pulse width T, the PDF $\rho(\phi)$ of the scatterer and the waveform Doppler resolution. Moreover, the target Doppler resolution interval and reverberation correlation function Doppler 3 dB width determine the Doppler distribution; when the target Doppler resolution interval has more overlapping areas with 3 dB reverberation correlation function, it means the target is not easy to distinguish; when the waveform Doppler resolution interval has less overlapping areas with 3 dB reverberation correlation

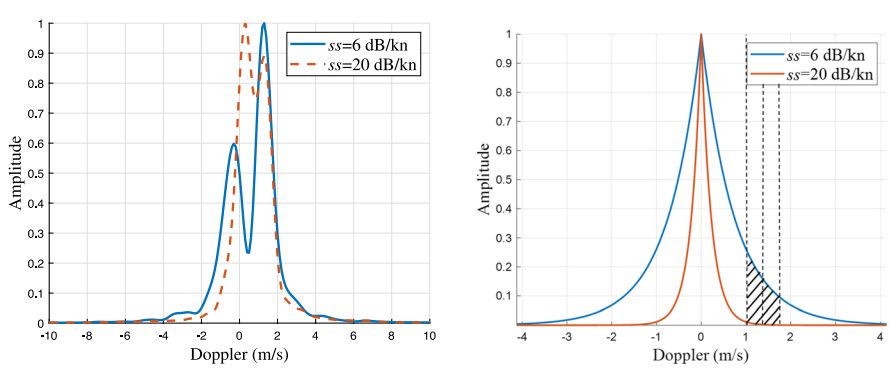
function, it means the reverberation Doppler clutter suppression is better and the target is easy to distinguish.

2.2.3 Effect of scatterer Doppler PDF $\rho(\phi)$ on reverberation Doppler distribution

According to Eq. (24), the reverberation Doppler distribution is jointly influenced by the waveform AF and the Doppler PDF of the scatterer. In this section, assuming that the waveform parameters are consistent, and the influence of the waveform AF on the reverberation Doppler distribution is eliminated, the relationship between different Doppler distribution $\rho(\phi)$ of the scatterer and the reverberation Doppler distribution is discussed. The scatterer Doppler distribution $\rho(\phi)$ is modeled according to the two-side exponential distribution, and two extreme cases of scatterer Doppler distribution are considered, which are discussed as ss=6 dB/kn, μ=0.69 and ss=20 dB/kn, μ=2.30. The waveform CW pulse width is T=0.23 s, the target is 1.29 m/s, and the signal-to-reverberation ratio (SRR) is SRR=4 dB.

Moreover, it is assumed that the target time information is known, and the target time region is extracted with twice the pulse width before subsequent signal processing.

Figure 3a is modeled with Eq. (5) and shows the relationship between different scatterer Doppler distributions and the reverberation Doppler distribution. It is easy to find that the 6 dB/kn curve is easier to distinguish the target than the 20 dB/kn curve, which indicates that the wider the scatterer Doppler distribution function is, the easier it is to distinguish the target. Figure 3b is modeled with Eq. (2) and shows the relationship between the target Doppler resolution and the scatterer Doppler distribution. The shaded area is the overlapping area between the waveform Doppler resolution and the scatterer Doppler distribution, which can be considered a good matching area for high-energy clutter. Outside the shaded area, the reverberation Doppler clutter is suppressed due to the waveform Doppler filtering effect. For the same waveform, if the scatterer Doppler distribution is wider, the better the clutter suppression effect is [15].



(a) Reverberation Doppler distribution curve containing the target

(b) Relationship between $\rho(\phi)$ and waveform Doppler resolution

Fig. 3 Reverberation suppression of different scatterer Doppler distribution

2.2.4 The effect of AF on the reverberation Doppler distribution

In this section, the Doppler scatterer distribution $\rho(\phi)$ is assumed to be constant, and the two-side exponential distribution model is set ss=10 dB/kn, $\mu = 1.15$. The effect of different waveform parameters on the reverberation Doppler distribution is discussed.

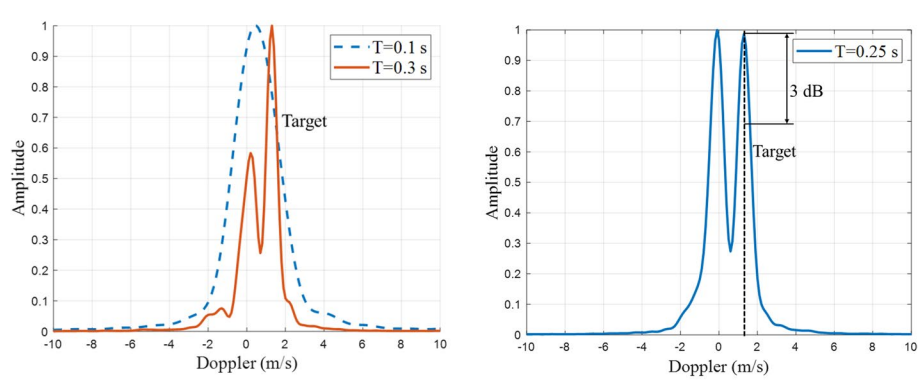
Figure 4a set the target Doppler 1.29 m/s, SRR = 4 dB, Monte Carlo experiment 100 times to compare the reverberation Doppler distribution curve in the case of $T = 0.1$ s and $T = 0.3$ s. When the CW pulse width T =0.1 s, the target is not easy to distinguish. When the CW pulse width $T = 0.3$ s, the reverberation Doppler distribution becomes narrower and the waveform Doppler resolution increases, it is obvious that the 0.707 (3 dB) position target is easy to distinguish. Therefore, without considering noise interference, the 0.707 (3 dB) Doppler distribution width of the reverberation can be used to determine whether the target is distinguishable. Thus, the reverberation target resolution function is proposed:

$$\phi_{\text{data}} = 2\phi_{\text{wave}}(T) \tag{25}$$

$\phi_{\text{data}}$is the reverberation Doppler 0.707 (3 dB) width, and $\phi_{\text{wave}}(T)$ is waveform Doppler resolution half of the 0.707 (3 dB) width. When the 0.707 (3 dB) reverberation width is equal to two times the waveform Doppler resolution, it represents that the target is distinguishable.

With the change of waveform pulse width, the reverberation Doppler distribution is also changing. For the single-peak target, 3 dB position can be directly judged as Fig. 4a, if the 3 dB position consists of a double-peak as Fig. 4b, one of the peaks near 0 m/s, another peak with 3 dB width can be judged, if the secondary peak width meets Eq. 25, on behalf of the target can be distinguished.

From the above analysis, it can be seen that increasing the waveform pulse width can reduce the reverberation Doppler spread clutter width to distinguish the target. For high-speed moving targets, the Doppler clutter will basically not exist. As the ocean environment changes, the reverberation scatterer Doppler distribution changes in real time, and different waveforms need to be used according to different environments to help target discrimination. The more waveforms retained in the active sonar, the higher the flexibility of active sonar waveform selection.



(a) Weak reverberation target resolution

(b) Strong reverberation target resolution

Fig. 4 Relationship between different pulse width waveforms and reverberation Doppler distribution

Although the target resolution function is proposed, the optimal waveform cannot be quickly achieved. Then combining the RL with cognitive sonar, the active sonar can adjust the waveform parameters according to the reverberation Doppler width to distinguish the target from the Doppler spread clutter quickly.

2.3 The Dyna-Q-Max-Action reverberation suppression cognitive sonar

The underwater environment is complex, Doppler spread is serious and time-varying. Thus, active sonar requires freedom for waveform design, so combining active sonar with the Dyna-Q algorithm can use different parameters waveforms for different reverberation Doppler spread clutter. However, the random action selection strategy of the Dyna-Q algorithm influences the convergence episodes. Therefore, in this section, the action selection strategy of the Dyna-Q algorithm is improved and the Dyna-Q-Max-Action algorithm is proposed.

2.3.1 The Dyna-Q-Max-Action five tuples

Five tuples are the basic component of RL. For better integration with cognitive sonar, the Dyna-Q-Max-Action can be equated to five tuples of MDP, which consists of a model of environment $\mathbf{S}_{\text{revb}}$, action $\mathbf{A}_{\text{wave}}$, transfer probability P, reward signal R_{reward}, discount factor γ.

1. Model of environment $\mathbf{S}_{\text{revb}}$: Model of environment $\mathbf{S}_{\text{revb}}$ is the reverberation Doppler clutter width set.

$$\mathbf{S}_{\text{revb}} = [S_{r1}, S_{r2}, \ldots, S_{rm}, \ldots]^{\text{T}} \tag{26}$$

   S_{rm} represents the Doppler clutter width of the m-state reverberation.
2. Action $\mathbf{A}_{\text{wave}}$: $\mathbf{A}_{\text{wave}}$ is a collection of waveforms, which consists of waveforms with different parameters.

$$\mathbf{A}_{\text{wave}} = [A_{w1}, A_{w2}, \ldots, A_{wm}, \ldots]^{\text{T}} \tag{27}$$

   When active sonar selects waveform A_{wm}, then automatically jumps to the corresponding state S_{rm}. In practice, in order to improve the efficiency of target detection, without considering the blind area, all waveforms can be sent out at once, and the optimal waveform is achieved by RL calculating.
3. Transfer probability P: The P transfer probability is the probability of choosing waveform A_{wm} causing the state transfer from S_{rm} to S_{rn}. In the absence of prior knowledge and the transfer probability is unknown, the initial transfer probability P is equal probability.
4. Reward signal R_{reward}:

The reward signal is the reward value of different waveforms A_{wm}. The reward signal setting is significant, unreasonable setting cause the waveform transmit strategy to fall into the local optimal solution. According to the Doppler spread clutter width and the previous target resolution function Eq. 25 to define the reward signal

$$\mathrm{R}_{\mathrm{reward}}(\mathrm{T}) = \begin{cases} 10, 0 \leq |2\phi_{\mathrm{wave}}(T) - \phi_{\mathrm{data}}|^{q} \leq 0.01 * 2\phi_{\mathrm{wave}}(T) \\ \mathrm{C}|2\phi_{\mathrm{wave}}(\mathrm{T}) - \phi_{\mathrm{data}}|^{\mathrm{q}}, \text{ else} \end{cases} \tag{28}$$

When $\mathrm{R}_{\mathrm{reward}}(\mathrm{T}) \in [0, 0.01 * 2\phi_{\mathrm{wave}}(\mathrm{T})]$, the reverberation Doppler clutter target is distinguishable and the reward signal is set to 10. When $\mathrm{R}_{\mathrm{reward}}(\mathrm{T}) \notin [0, 0.01 * 2\phi_{\mathrm{wave}}(\mathrm{T})]$, the reward factor C=-1, the waveform reward signal is obtained through Eq. 28. q is the magnification factor, which is used to enlarge the difference in reward signals between different waveforms. $\mathrm{R}_{\mathrm{reward}}$ can be used to evaluate the difference between different waveforms, and the closer to the optimal waveform is, the higher $\mathrm{R}_{\mathrm{reward}}$ tends to guide the RL to converge to the optimal waveform.

5) Discount factor γ:

The discount factor $\gamma \in [0, 1]$, which determines the decay of future reward signal, when γ tends to 0, the active sonar tends to obtain immediate rewards; when γ tends to 1, the active sonar tends to obtain long-term gains, indicating that almost all reward signals are influencing the Q-value.

2.3.2 The Dyna-Q-Max-Action algorithm flow

The traditional Dyna-Q algorithm treats planning as an improvement of the action, but the action selection within the model of the Dyna-Q algorithm is random [24]. In this paper, the action selection strategy of the Dyna-Q algorithm is improved, then the Dyna-Q-Max-Action algorithm is proposed. The algorithm flow is as follows:

Algorithm 1 The Dyna-Q-Max-Action

1: Initialize Q(S_{rm},A_{wm}) and Model(S_{rm},A_{wm}) for all $S_{rm} \in \mathbf{S}_{\mathrm{revb}}$ and $A_{wm} \in \mathbf{A}_{\mathrm{wave}}(S_{rm})$;
2: **while** True **do**
3: $S_{rm} \leftarrow$ current (nonterminal) state
4: $A_{wm} \leftarrow \varepsilon(S_{rm}, \mathrm{Q})$
5: Execute action A_{wm}; observe resultant reward $\mathrm{R}_{\mathrm{reward}}$ and state S_{rn}
6: $Q(S_{rm}, A_{wm}) \leftarrow Q(S_{rm}, A_{wm}) + \alpha[\mathrm{R}_{\mathrm{reward}} + \gamma \max_{A_{wn}} Q(S_{rn}, A_{wn}) - Q(S_{rm}, A_{wm})]$
7: Model(S_{rm},A_{wm}) $\leftarrow \mathrm{R}_{\mathrm{reward}}$, S_{rn} (assuming deterministic environment)
8: **for** $ii = 1$ to n **do**
9: $S_{rn} \leftarrow$ random previously observed state
10: $A_{wm} \leftarrow \varepsilon$ The reward signals of each action in the model of different states are summed and compared, and the maximum value is selected as the next action
11: $\mathrm{R}_{\mathrm{reward}}$, $S_{rn} \leftarrow$ Model(S_{rm},A_{wm})
$Q(S_{rm}, A_{wm}) \leftarrow Q(S_{rm}, A_{wm}) + \alpha[\mathrm{R}_{\mathrm{reward}} + \gamma \max_{A_{wn}} Q(S_{rn}, A_{wn}) - Q(S_{rm}, A_{wm})]$
12: **end for**
13: **end while**

α is the step size, $\alpha \in (0, 1)$, which determines the effect of estimation error on Q(S_{rm} ,A_{wm}). S_{rm} and A_{wm} are the state and the action at this step. S_{rn} and A_{wn} are the state and the action at the next step. $\mathrm{R}_{\mathrm{reward}}$ is the reward signal. ε is the greediness of action selection, $\varepsilon \in [0, 1]$, representing the probability of selecting the maximum reward signal action. Greedy action selection uses current knowledge to maximize immediate reward and does not sample worse actions. The state-action value function is abbreviated as Q-value. The Q table consists of (S_{rm}, A_{wm}) corresponding position to the Q-value.

Steps 1–7 of Dyna-Q-Max-Action algorithm are identical to the traditional Dyna-Q algorithm, differing only in steps 8–12. Steps 8–12 can be summarized as n updates of the Q-value using the model already learned. Inspired by the action selection strategy of the Q-learning algorithm, introducing the ε action selection strategy to

the Dyna-Q-Max-Action algorithm. The Dyna-Q-Max-Action algorithm superimposes the reward signals of each action in all environments, and compares the sum of reward signals of different actions. ε probability selects the next action with the maximum sum of the reward signals, 1-ε probability selects the next action randomly

$$A_{wm} = \varepsilon \mathrm{Max}\left(\sum (\mathrm{R}_{\mathrm{reward}})\right) \tag{29}$$

The Dyna-Q-Max-Action algorithm can select an optimal waveform by integrating the waveform into all environments through Eq. 29, which shortens the convergence episodes of the Q-value. The Dyna-Q-Max-Action algorithm architecture is shown in Fig. 5 below.

Figure 5 'Real Experience' to 'Value Functions (Q-value)' represents direct learning based on real experience to improve Q-values and actions. 'Real Experience' to 'Model' to 'Simulated Experience' to 'Value Functions (Q-value)' is a model-based learning process. The model learns from real experience and generates simulated experience. Finally, the simulated experience is used to update the Q-value.

The core idea of the Dyna-Q-Max-Action algorithm is the Q-value learning from real experience and planning from simulated experience. Learning and planning are deeply integrated in the sense that they share almost all the same machinery, differing only in the source of their experience. The ε model action selection strategy can shorten the convergence episodes of the active sonar.

3 Results and discussion

In this section, the relationship between the reverberation Doppler clutter width and CW waveform pulse width is simulated by the Dyna-Q-Max-Action algorithm. According to the Dyna-Q-Max-Action algorithm iteration, the optimal waveform is achieved. Set SRR = 4 dB. The following table shows the reverberation Doppler clutter widths obtained by numerical simulation in different environments, and calculates the reward signals for the corresponding waveform pulse widths according to Eq. 28.

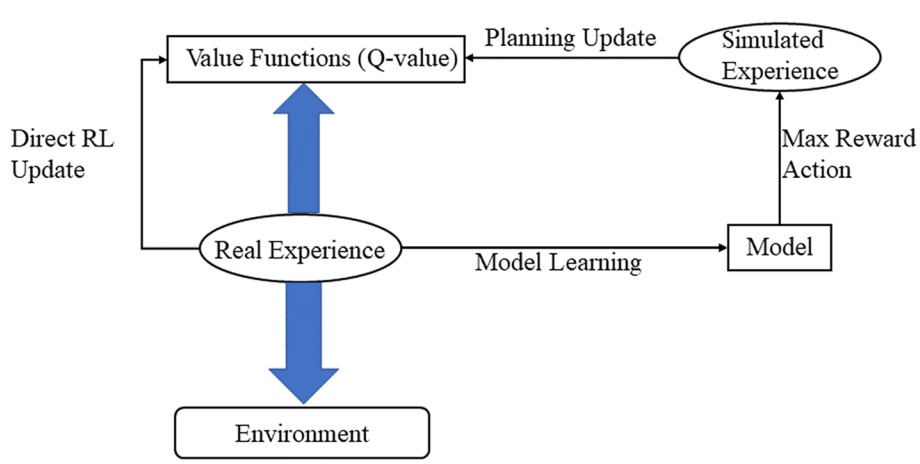


Fig. 5 The Dyna-Q-Max-Action algorithm architecture

The reward signals show that only the 300ms CW signal is the optimal waveform in this environment. The rest of the waveforms are given different R_{reward}, guiding the active sonar to converge to optimal waveform 4.

In this paper, there are two kinds of convergence: step convergence and Q-value convergence. When the sum of the reward per episode, and the converged steps per episode are converged, means the Dyna-Q-Max-Action algorithm step is convergence. While, when the loss function is converged, means the Dyna-Q-Max-Action algorithm Q-value is convergence.

The loss function is the error between predicted Q-value and real Q-value.

$$\mathrm{Loss} = \mathrm{Sum}(\mathrm{Q}_{\mathrm{predict}} - \mathrm{Q})/\mathrm{Steps} \tag{30}$$

$\mathrm{Q}_{\mathrm{predict}}$ is the predicted value of Q-value, Q is the real value of Q-value, and Steps is the number of steps to reach the optimal in this episode.

Define the average reward signal per episode R_{average}

$$R_{\mathrm{average}} = \mathrm{Sum}(\mathrm{R}_{\mathrm{reward}})/\mathrm{Steps} \tag{31}$$

The next subsection will discuss the different training times and different ε greediness of the Dyna-Q-Max-Action algorithm.

3.1 Different training times of the Dyna-Q-Max-Action algorithm with ε=1

Set ε=1 for the Dyna-Q-Max-Action algorithm, means each action is selected from the model with the maximum reward signal. Then compare the number of the Dyna-Q-Max-Action algorithm convergence episodes for different model training times.

In Fig. 6a–d Env=0 refers to the Q-learning algorithm, Env=5 represents the Dyna-Q-Max-Action algorithm with 5 times of model training, and the x-axis represents the number of training episodes. According to the results of the number of convergence steps, the sum of rewards and the average reward, the step convergence efficiency increases as the number of model training times increases.

In Fig. 6a, b, the step of the Q-learning algorithm needs about 30 episodes to converge, while after 5 times of model training the step of the Dyna-Q-Max-Action algorithm only needs about 20 episodes to converge. Even after 90 times of model training, the step of the Dyna-Q-Max-Action algorithm just needs several episodes to converge.

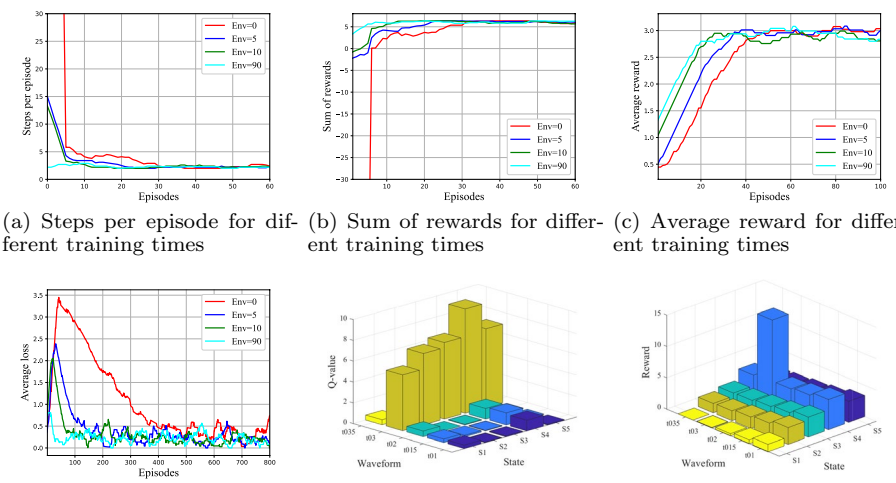
According to the average loss curve in Fig. 6d, the Q-learning algorithm Q-value needs about 400 episodes to converge. While after 5 times of model training, the Dyna-Q-Max-Action algorithm Q-value needs about 150 episodes to converge, and only about 30 episodes to converge after 90 times of training.

Therefore, as the number of training times increases, the Dyna-Q-Max-Action algorithm with ε=1 shortens the number of training episodes and increases the efficiency.

Figure 6e, f the histograms are plotted based on the Q table and the table of model reward signals after 90 times of model training. Figure 6e indicates the Q-value of different actions in different environments. Figure 6f indicates the model reward of different actions in different model environments. S1, S2, S3, S4, and S5 represent five different reverberation Doppler clutter widths, and t01, t015, t02, t03, and t035 represent five waveforms, which corresponding to pulse widths of 100 ms, 150 ms, 200 ms, 300 ms, and 350 ms, respectively in Table 1.

Table 1 The Dyna-Q-Max-Action algorithm numerical simulation parameters

Num.	*T*(ms)	Doppler resolution (m/s)	Doppler clutter width (m/s)	R_{reward}
1	100	0.825	4.2	$-(2.55)^q$
2	150	0.55	2.7	$-(1.6)^q$
3	200	0.4125	2.2	$-(1.375)^q$
4	300	0.275	0.553	10
5	350	0.236	2	$-(1.528)^q$



(a) Steps per episode for different training times (b) Sum of rewards for different training times (c) Average reward for different training times

(d) Average loss for different training times (e) 3D histogram of Q-value (f) 3D histogram of model reward

Fig. 6 The Dyna-Q-Max-Action algorithm with ε=1

According to the histogram, it can be seen that the optimal waveform t03 in different environments of the Q table of the Dyna-Q-Max-Action algorithm has the largest Q-value and can provide the absolute maximum action selection to the environment. The histogram of the model reward shows that only the optimal action and its corresponding optimal environment reward is the largest. Therefore, it makes each model action selection directly select the t03 optimal waveform and speed up the step convergence speed and shorten the Q-value convergence episodes.

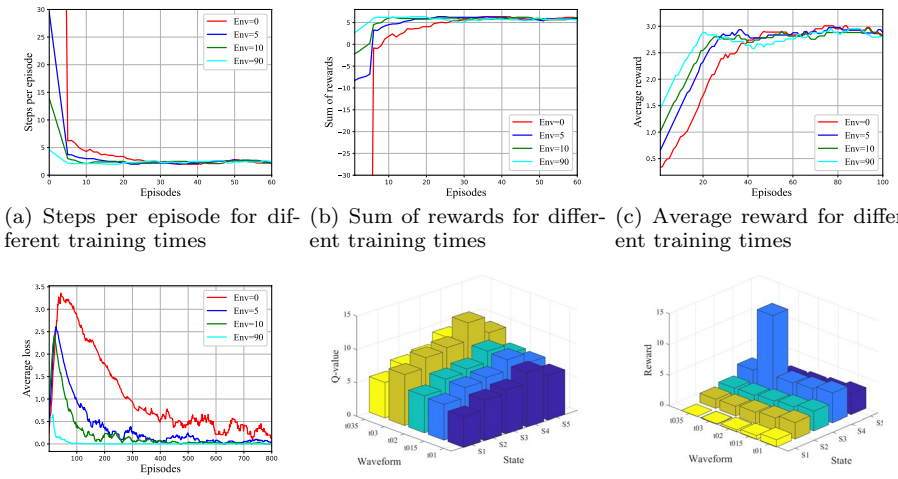
3.2 Different training times of the Dyna-Q-Max-Action algorithm with ε=0.6

Set ε= 0.6 and compare the results of different model training times.

According to Fig. 7a–d, it can be seen that the Dyna-Q-Max-Action algorithm converges with fewer episodes as the number of training times increases, and the model training times 90 is optimal.

In Fig. 7a, b, the step of the Q-learning algorithm needs about 30 episodes to converge, the step of the Env=5 Dyna-Q-Max-Action algorithm need about 20 episodes to converge. It is almost the same as Fig. 6a, b.

In addition, according to Figs. 6d and 7d, it can be seen that after 90 times model training, the ε=0.6 curve is flatter compared to the ε=1 curve after convergence.



(a) Steps per episode for different training times (b) Sum of rewards for different training times (c) Average reward for different training times

(d) Average loss for different training times (e) 3D histogram of Q-value (f) 3D histogram of model reward

Fig. 7 The Dyna-Q-Max-Action algorithm with ε=0.6

The histogram results in Fig. 7e show that the Q table of ε=0.6 the Dyna-Q-Max-Action algorithm shows a stepwise growth in different environments, and the Q-value tend to the optimal waveform t03. Comparing with Fig. 6e, it can be seen that ε=0.6 Q table is smoother than ε=1 Q table. In addition, the maximum model reward in Fig. 7f is the optimal environment corresponding to the optimal waveform, which is consistent with ε=1 in Fig. 6f.

3.3 The Dyna-Q-Max-Action algorithm with different greediness ε

The model was set to training 10 times to compare the results for different ε action selection probabilities [0, 0.3, 0.6, 1], where ε=0 represents the Dyna-Q algorithm.

In Fig. 8, comparing with ε=0.3, ε=0.6, ε=1 and the Dyna-Q algorithm curves show that the Q-value of the Dyna-Q-Max-Action algorithm converge with fewer episodes

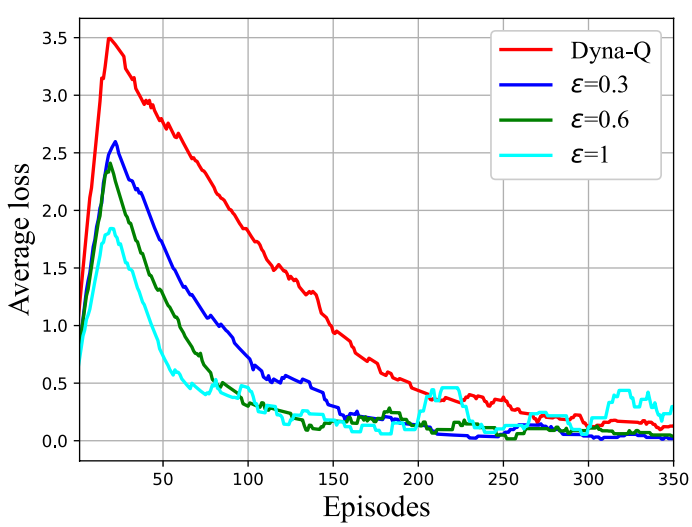


Fig. 8 Average loss curve for different action selection probabilities ε

than the Dyna-Q algorithm Q-value convergence episodes, and ε=1 converges with the fewest episodes. However, ε=1 is too large leading to overfitting, which makes the average loss curve fluctuate. Therefore, finding a suitable action selection probability ε, can reduce the fluctuation and shorten the convergence episodes. According to the results in Fig. 8, ε=0.6 is more appropriate.

4 Conclusion

The discrimination of low-speed weak targets in reverberation Doppler spread clutter is a difficult problem in active sonar signal processing. This paper combines the point scattering model and two-side exponential distribution to model the reverberation Doppler spread clutter, and verifies the effectiveness of the model through the real data of the Qiandao Lake. In this paper, a target resolution function is proposed for the reverberation Doppler clutter target resolution problem, which can quickly identify the target in the Doppler spread clutter, and the target resolution function is combined with the Dyna-Q-Max-Action algorithm, which enables the active sonar to adjust the waveform parameters according to different reverberation. Meanwhile, the relationship between the Dyna-Q-Max-Action algorithm and the greediness of action selection and the number of model training times are discussed. Based on the numerical simulation results, it is found that the step convergence efficiency of the Dyna-Q-Max-Action algorithm combined with the action selection greediness converges more rapidly than the step of the Q-learning algorithm. According to the results, the Dyna-Q-Max-Action algorithm converges with fewer episodes than the Dyna-Q algorithm converges episodes. Providing a theoretical basis for future engineering applications of RL based reverberation suppression cognitive sonar.

Abbreviations

Dyna-Q	Integrating planning, acting, and learning, Q-learning
Dyna-Q-Max-Action	Dyna-Q maximum reward action selection strategy
AF	Ambiguity function
NP	Non-deterministic polynomial
RL	Reinforcement learning
MDP	Markov decision process
SRASA	State action reward state action
TD	Temporal difference
WSSUS	Wide-sense stationary uncorrelated scattering
PDF	Probability density function
CW	Continuous wave
SRR	Signal-to-reverberation ratio

Acknowledgements
The authors would like to express their gratitude to the Deep Sea Observation Project, 2019JCJQZD02400.

Author contributions
Yubin Fu put forward the original idea of the paper and complete the manuscript. Xiaochuan Ma contributed to the validation, review, editing and supervision. Xingyuan Pei and Pengzhuo Li finished the revisions.

Funding
The research was supported by the Deep Sea Observation Project, 2019JCJQZD02400.

Availability of data and materials
The data are not publicly available due to the private reasons.

Declarations

Ethics approval and consent to participate
All procedures performed in this paper were in accordance with the ethical standards of research. community

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 27 May 2023 Accepted: 8 September 2023
Published online: 14 November 2023

References

1. Z. Hao ke, M. Qu li, L. Hai lin, Study on robust space-time adaptive reverberation suppressing, in *IEEE 10th International Conference on Signal Processing Proceedings*, pp. 2407–2410 (2010)
2. X. Cui, C. Chi, S. Li, Y. Li, H. Huang, Coprime pulse trains of frequency-modulated for suppressing reverberation, in *OCEANS 2021: San Diego Porto*, pp. 1–4 (2021)
3. B.W. Choi, E.H. Bae, J.S. Kim, K.K. Lee, Improved prewhitening method for linear frequency modulation reverberation using dechirping transformation. J. Acoust. Soc. Am. **123**(3), 21–25 (2008)
4. J.N. Maksym, M. Sandys-Wunsch, Adaptive beamforming against reverberation for a three-sensor array. J. Acoust. Soc. Am. **102**(6), 3433–3438 (1997)
5. Y. Li, H. Huang, C. Zhang, S. Li, New schur-type-based pci algorithms for reverberation suppression in active sonar, in *Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, vol. 4, pp. 641–6444 (2005)
6. Z.-Q. Wang, L. An, J.-R. Lu, Signal detection based on mathematical morphology in oceanic reverberation, in *2007 14th International Conference on Mechatronics and Machine Vision in Practice*, pp. 8–12 (2007)
7. S. Haykin, Cognitive radar: a way of the future. IEEE Signal Process. Mag. **23**(1), 30–40 (2006)
8. L. Xiaohua, L. Yaan, L. Guancheng, Y. Jing, Research of the principle of cognitive sonar and beamforming simulation analysis, in *2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–5 (2011)
9. T. Claussen, V.D. Nguyen, Real-time cognitive sonar system with target-optimized adaptive signal processing through multi-layer data fusion, in *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 357–361 (2015)
10. X. Qing, D. Nie, G. Qiao, J. Tang, Dolphin bio-inspired transmitting waveform design for cognitive sonar and its performance analysis, in *2016 IEEE/OES China Ocean Acoustics (COA)*, pp. 1–7 (2016)
11. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science **362**(6419), 1140–1144 (2018)
12. M. Taylor, Teaching reinforcement learning with mario: An argument and case study, in *Proceedings of the National Conference on Artificial Intelligence* **2** (2011)
13. J.E. Summers, J.M. Trader C.F. Gaumond, J.L. Chen, Deep reinforcement learning for cognitive sonar. *J. Acoust. Soc. Am.* **143**(3-Supplement), 1716–1716 (2018)
14. J. Tucker, V. Chavali, K.E. Wage, J.K. Nelson, Multiple objective optimization for fully adaptive active sonar, in *OCEANS 2022, Hampton Roads*, pp. 1–9 (2022)
15. T.C. Yang, J. Schindall, C.-F. Huang, J.-Y. Liu, Clutter reduction using Doppler sonar in a harbor environment. J. Acoust. Soc. Am. **132**(5), 3053–3067 (2012)
16. R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2016)
17. X. He, Y. Xu, M. Liu, C. Hao, C. Hou, Adaptive estimation of k-distribution shape parameter based on fuzzy statistical normalization processing. IEEE Trans. Aerosp. Electron. Syst. **58**(5), 4566–4577 (2022)
18. P.C. Etter, C.H. Haas, D.V. Ramani, Evolving trends and challenges in applied underwater acoustic modeling, in *OCEANS 2015 - MTS/IEEE Washington*, pp. 1–10 (2015)
19. F. Cao, X. Zhang, J. Han, S. Lv, Experimental analysis of statistical property of low frequency reverberation envelope in shallow water, in *2021 OES China Ocean Acoustics (COA)*, pp. 534–538 (2021)
20. J.J. Murray, A theoretical model of linearly filtered reverberation for pulsed active sonar in shallow water. J. Acoust. Soc. Am. **136**(5), 2523–2531 (2014)
21. C. Zhang, X. Ma, X. Li, F. Zhan, S. Zhang, Modified asymmetric statistical model for the reverberation doppler spread spectrum. Shengxue Xuebao/Acta Acustica **43**, 943–950 (2018)
22. J. Zhang, X. Qiu, C. Shi, Y. Wu, Cognitive radar ambiguity function optimization for unimodular sequence. EURASIP J. Adv. Signal Process. **2016**, 1–13 (2016)
23. N.U.R. Junejo, M. Sattar, S. Adnan, H. Sun, A.B.M. Adam, A. Hassan, H. Esmaiel, A survey on physical layer techniques and challenges in underwater communication systems. *J. Marine Sci. Eng.* **11**(4) (2023)
24. X. Li, C. Yang, J. Song, S. Feng, W. Li, H. He, A motion control method for agent based on dyna-q algorithm, in *2023 4th International Conference on Computer Engineering and Application (ICCEA)*, pp. 274–278 (2023)

Yubin Fu male, PhD student, fuyubin@mail.ioa.ac.cn. His main research interests are signal processing, parametemr estimation, waveform design, cognitive sonar, etc.

Xiaochuan Ma male, researcher, PhD. maxc@mail.ioa.ac.cn. His main research interests include acoustic signal processing, autonomous navigation and control of underwater vehicles, array signal processing, pattern recognition, etc.

Chao Feng male, PhD. fengchao@mail.ioa.ac.cn. His main research interests are signal processing, noise and vibration control, etc.

Xingyuan Pei male, PhD student. peixingyuan@mail.ioa.ac.cn. His main research interests are signal processing, acoustic array signal processing, etc.

Pengzhuo Li male, PhD student. lipengzhuo@mail.ioa.ac.cn. His main research interests are signal processing, parameter estimation, underwater object perception, etc.