applied sciences

MDPI

Article

A Self-Adaptive Vibration Reduction Method Based on Deep Deterministic Policy Gradient (DDPG) Reinforcement Learning Algorithm

Xin Jin, Hongbao Ma, Jian Tang and Yihua Kang *

School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
* Correspondence: yihuakang@hust.edu.cn

Abstract: Although many adaptive techniques for active vibration reduction have been designed to achieve optimal performance in practical applications, few are related to reinforcement learning (RL). To explore the best performance of the active vibration reduction system (AVRS) without prior knowledge, a self-adaptive parameter regulation method based on the DDPG algorithm was examined in this study. The DDPG algorithm is unsuitable for a random environment and prone to reward-hacking. To solve this problem, a reward function optimization method based on the integral area of the decibel (dB) value between transfer functions was investigated. Simulation and graphical experimental results show that the optimized DDPG algorithm can automatically track and maintain optimal control performance of the AVRS.

Keywords: self-adaptive; deep deterministic policy gradient (DDPG) algorithm; active vibration reduction system (AVRS)

check for updates

Citation: Jin, X.; Ma, H.; Tang, J.; Kang, Y. A Self-Adaptive Vibration Reduction Method Based on Deep Deterministic Policy Gradient (DDPG) Reinforcement Learning Algorithm. *Appl. Sci.* **2022**, *12*, 9703. https://doi.org/10.3390/app12199703

Academic Editors: Woon-Seng Gan, Dongyuan Shi and Jihui Aimee Zhang

Received: 6 September 2022
Accepted: 22 September 2022
Published: 27 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Due to the increasing requirement for accuracy and stability in research and manufacturing, the issue of micro-vibration, which widely exists in engineering applications, has become a critical factor that reduces the performance of high-precision machines. For example, as the state-of-art transmission electron microscope (TEM) has met the VC-G [1,2] or higher vibration standard (vibration velocity lower than 0.78 μm/s), the performance requirements of the AVRS are more stringent than before.

Active vibration control technology (AVCT) originated from the semi-active control methods represented by skyhook damping [3], for which it is difficult to obtain the absolute inertial reference frame (AIRF) that is required in practice. In a previous paper [4], a virtual skyhook damping isolator was proposed without the AIRF, which used a double cascade damping structure to suppress the resonance peak. In the engineering field, various high-precision instruments are equipped with passive vibration isolation systems, such as pneumatic or coil springs, which make them more sensitive to low-frequency vibrations. Because of the poor performance of passive and semi-active damping technology at low frequencies, the AVCT technique is urgently needed. The combination of AVCT and passive vibration isolators is one of the mainstream methods used today, in which passive isolators attenuate high-frequency vibrations while AVCT reduces the low-frequency vibrations. The AVRS overcomes the disadvantages of passive vibration reduction systems by integrating the active execution components.

Control algorithms [5–8] based on closed-loop PID have been proposed. Most of these methods are aimed at achieving the ideal dynamic capability of skyhook damping while balancing the robustness and stability of the system. Although appropriately increased feedback control (FBC) gain can enhance the vibration reduction effect, the adjustment range of parameters is greatly restricted by factors such as noise, decoupling accuracy, and ground and table disturbance [9]. Furthermore, the FBC response speed is greatly constrained by inherent algorithm delay and the hysteresis of the actuators.

The feed-forward control (FFC) strategy was first proposed to suppress the direct disturbances of AVRS, while the compensation control signal can be calculated from its regularity [10]. Under the premise of obtaining an accurate FF model, FFC can theoretically eliminate the influence of ground vibration on the platform. A previous paper [11] presented several prediction methods for FF signals and analyzed their tracking performance using filters. Irrespective of the FF signal prediction error changes with the increase in the frequency, proper parameters in FFC do not affect the closed-loop stability, which is merely determined by the loop transfer function [12].

To further improve the performance of AVRS, the FBC and FFC strategies were incorporated into a six-degree-of-freedom (DOF) isolation system that is based on absolute accelerator measurement [13], where FBC uses a genetic algorithm to suppress payload vibration, and FFC prevents ground vibration from transmitting to the platform.

Note that because of the hysteresis and the low response speed of the voice-coil motor, together with the high-order bending mode of the system, not only can the AVRS operating at fixed parameters be unstable, but it is also limited to suppressing the vibration below 100 Hz rather than the entire frequency range [14]. For the frequency above 100 Hz, vibration attenuation of AVRS mainly relies on passive isolation, while the AVCT focuses on the improvement of the low-frequency performance.

Furthermore, although the ideal model of AVRS is difficult to obtain, together with the rapid development of AVCT, various advanced control methods have been proposed for tracking the optimal parameters according to its specific applications [15–17]. However, few of these involve reinforcement learning (RL), which is one of the most popular branches in machine learning (ML) since Lee Sedol was defeated by AlphaGo in the board game Go. ML is excellent because it generates behavior directly from data of intelligent learning, rather than through complex programming [18]. RL is a trial-and-error learning algorithm with scarce guidance, whose goal is to train an agent to obtain the control objective or to enhance its performance through constantly interacting with the environment until achieving the best strategy, which is evaluated by the value of a reward function [19,20]. Although the Q-learning algorithm has been applied to realize automatic tuning of online parameters in the active structural control system [21], it is not suitable for continuous actions and its working frequency at 10 Hz is much lower than the ultra-precision vibration reduction range.

In this paper, based on the accurate mathematical derivation of the critical components in AVRS, a joint analysis model of ADAMS and MATLAB for AVRS is proposed. The mechanical structure of AVRS was presented elsewhere [13,14]. As the six DOF AVRS can be transformed into six independent single DOF sub-systems by the modal decomposition technique, a detailed single-channel block diagram of a vertical FB and FF control scheme can be produced by the transfer function of each component module, respectively. Then, the advanced DDPG algorithm is integrated into the FB and FF control scheme to determine the optimal control without considering the accuracy of the system model and the nonlinearity of parameters. DDPG, which is based on Actor-Critic [22] and Deep-Q-Network (DQN) algorithms [23] was proposed in [24]. It combines deep neural networks with a deterministic policy gradient (DPG) algorithm [25] and is one of the most popular algorithms for continuous actions. Simulation results show that the DDPG algorithm can achieve effective convergence and high stability in the optimal vibration reduction region. In practical applications, to overcome the invalidation of self-learning and the tendency toward reward-hacking of DDPG for random signals, in contrast to the tuning methods proposed in [26–28], this study introduced an optimal calculation method of the reward function. Experiment results validated that the agent trained by this method has superior decision-making performance and excellent vibration damping effects.

2. Principle of Active Vibration Isolation

2.1. The Principle of Active Vibration Control

The single DOF mass-spring damping system with active control used in this research is shown in Figure 1.

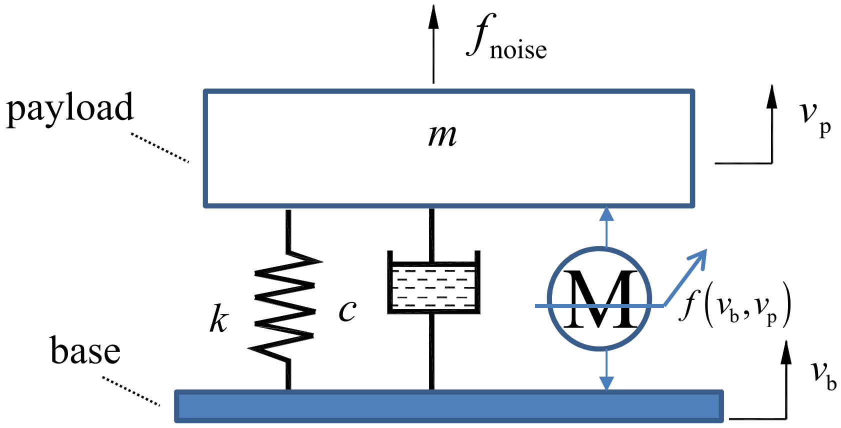


Figure 1. The single DOF mass-spring damping system with active control.

The vibration signals are measured by absolute velocity sensors and the control forces are generated by voice-coil motors. When ignoring the gravity and the static displacement caused by it, the linear ordinary differential equations in the equilibrium position of the system can be rewritten as below:

$$m\frac{\mathrm{d}v_{\mathrm{p}}}{\mathrm{d}t} + c(v_{\mathrm{p}} - v_{\mathrm{b}}) + k\left(\int v_{\mathrm{p}}\mathrm{d}t - \int v_{\mathrm{b}}\mathrm{d}t\right) = f(v_{\mathrm{p}}, v_{\mathrm{b}}) + f_{\mathrm{noise}} \tag{1}$$

where v_{p} and v_{b} are the measured absolute velocity of the payload and ground, respectively; m, c, and k are the system payload, equivalent damping, and stiffness, respectively; f_{noise} is the plate-top disturbance signal; $f(v_{\mathrm{p}}, v_{\mathrm{b}})$ is the motor driving force function with v_{p} and v_{b} as the dependent variables. Since the vibration amplitude of AVRS applied to ultra-precision equipment is small, the inverse electromotive force generated by the velocity difference between the coils and the magnet in the voice-coil motor can be ignored, so that the thrust coefficient K_{T} of the voice-coil motor is approximately constant. Then, the voice-coil motor output force $f(v_{\mathrm{p}}, v_{\mathrm{b}})$ can be considered to have an approximately linear relationship with v_{p} and v_{b}.

2.2. FB and FF Control Theory

Enlightened by the idea of independent channel control design that was based on the different disturbance transmission paths [29], and referring to the principle of force superposition of the FB and FF hybrid control system, the control force, generated by the voice-coil motor, can be decomposed into two parts: one provided by the FFC and the other by the FBC. It is worth noting that the driving forces of the FBC and FFC are not independent of each other in practical applications. It was found that the FFC will effectively prevent the ground disturbance that will transmit to the platform [11], and the corresponding FBC output force becomes smaller because the vibration amplitude of the payload decreases.

Here, s represents the Laplace operator. Assuming that the initial speed of the platform is zero, and taking f_{noise} and $v_{\mathrm{b}}(s)$ as external disturbance sources, the value of FF force $f_{\mathrm{ffc}}(s)$ is determined by $v_{\mathrm{b}}(s)$ and has nothing to do with $v_{\mathrm{p}}(s)$. Without considering the nonlinear factors of sensors, signal amplifier circuits, and actuators, the relationship between $f_{\mathrm{ffc}}(s)$ and the ground detection signal $v_{\mathrm{b}}(s)$ can be expressed as follows:

$$f_{\mathrm{ffc}}(s) = H_{\mathrm{ffc}}(s)v_{\mathrm{b}}(s) \tag{2}$$

where $H_{\text{ffc}}(s)$ denotes the velocity–force transfer function.

Now, the dynamic characteristics of the AVRS in terms of two aspects during the feedback control are investigated in the following paragraphs.

(1) When the AVRS is open-looped, according to the principle of speed superposition, the platform speed signal $v_{\text{p}}(s)$ can be decomposed into two parts, $v_1(s)$ and $v_2(s)$. Here, $v_1(s)$ is the component of platform velocity transmitted from ground disturbance $v_{\text{b}}(s)$ under the control of $f_{\text{ffc}}(s)$, while $v_2(s)$ is caused by f_{noise}. Then, the expression of the system's kinematic transfer functions described in Figure 1 is given by the Equation (3):

$$\begin{cases} msv_1(s) + c[v_1(s) - v(s)] + \frac{k}{s}[v_1(s) - v(s)] = -f_{\text{ffc}}(s) \\ msv_2(s) + cv_2(s) + \frac{k}{s}v_2(s) = f_{\text{noise}}(s) \end{cases} \tag{3}$$

The equations to calculate $v_1(s)$ and $v_2(s)$ can be presented as follows:

$$\begin{cases} v_1(s) = \frac{(cs+k)-sH_{\text{ffc}}(s)}{ms^2+cs+k}v_{\text{b}}(s) \\ v_2(s) = \frac{f_{\text{noise}}(s)s}{ms^2+cs+k} \end{cases} \tag{4}$$

Therefore, when the system is open-looped, the platform speed can be obtained by:

$$v_{\text{p}}(s) = v_1(s) + v_2(s) \tag{5}$$

By substituting Equation (4) into Equation (5), finally, $v_{\text{p}}(s)$ can be expressed by Equation (6) as below:

$$v_{\text{p}}(s) = \frac{s}{ms^2 + cs + k}\left[\left(\frac{cs+k}{s} - H_{\text{ffc}}(s)\right)v_{\text{b}}(s) + f_{\text{noise}}(s)\right] \tag{6}$$

Considering the combined effects of $v_{\text{b}}(s)$ and $f_{\text{noise}}(s)$, the FB force $f_{\text{fbc}}(s)$ can be calculated by $H_{\text{fbc}}(s)$ multiplied by $v_{\text{p}}(s)$, which is expressed in Equation (7) below:

$$f_{\text{fbc}}(s) = H_{\text{fbc}}(s)v_{\text{p}}(s) \tag{7}$$

where $H_{\text{fbc}}(s)$ stands for the forward channel transfer function between the platform velocity $v_{\text{p}}(s)$ and feedback force $f_{\text{fbc}}(s)$. Then, substituting Equation (6) into Equation (7), yields:

$$f_{\text{fbc}}(s) = H_{\text{fbc}}(s)\frac{s}{ms^2 + cs + k}\left\{\left[\frac{cs+k}{s} - H_{\text{ffc}}(s)\right]v_{\text{b}}(s) + f_{\text{noise}}(s)\right\} \tag{8}$$

From Equation (8), it can be seen that $f_{\text{fbc}}(s)$ contains the term $H_{\text{ffc}}(s)$, which indicates that the influence of ground disturbance has not been eliminated by $f_{\text{ffc}}(s)$. Therefore, the platform residual disturbance from ground disturbance, together with $f_{\text{noise}}(s)$, will be suppressed by the force component $f_{\text{fbc}}(s)$. Thus, before the closed-loop control is applied, the motor output force $f_{\text{M}}(s)$ is calculated as follows:

$$f_{\text{M}}(s) = f_{\text{ffc}}(s) + f_{\text{fbc}}(s) \tag{9}$$

According to the above formula, the relationship between the motor driving force $f_{\text{M}}(s)$ and ground vibration $v_{\text{b}}(s)$, together with surface disturbance $f_{\text{noise}}(s)$, can be obtained as below:

$$\begin{aligned} f_{\text{M}}(s) \quad &= H_{\text{fbc}}(s)\frac{s}{ms^2+cs+k}\left\{\frac{cs+k}{s} + \frac{ms^2+[c-H_{\text{fbc}}(s)]s+k}{H_{\text{fbc}}(s)s}H_{\text{ffc}}(s)\right\}v_{\text{b}}(s) \\ &+H_{\text{fbc}}(s)\frac{s}{ms^2+cs+k}f_{\text{noise}}(s) \end{aligned} \tag{10}$$

(2) In the closed-loop system, the motor force acts on the system at the value $f_{\text{M}}(s)$ shown in Equation (10), and the table vibration is rapidly attenuated due to the effect of $f_{\text{M}}(s)$. As the output of the controller remains the same before being updated, the dynamic process of the system from the perspective of computer control can be analyzed. When

ignoring the response delay of the system, assuming that at some point after the control loop is closed, the platform velocity under the combined effect of the motor driving force $f_{\mathrm{M}}{}'(s)$, $v_{\mathrm{b}}(s)$, and $f_{\mathrm{noise}}(s)$ is $v_p{}'(s)$. This means that dynamic Equation (1) still holds. Then, the following equations can be derived:

$$\begin{cases} msv_{\mathrm{p}}{}'(s) + c\left[v_{\mathrm{p}}{}'(s) - v_{\mathrm{b}}(s)\right] + \frac{k}{s}\left[v_{\mathrm{p}}{}'(s) - v_{\mathrm{b}}(s)\right] = f_{\mathrm{noise}}(s) - f_{\mathrm{M}}{}'(s) \\ f_{\mathrm{M}}{}'(s) = H_{\mathrm{ffc}}(s)v_{\mathrm{b}}(s) + H_{\mathrm{fbc}}(s)v_{\mathrm{p}}{}'(s) \end{cases} \tag{11}$$

Simplifying the above equations, $v_p{}'(s)$ can be obtained in Equation (12):

$$v_p{}'(s) = \frac{f_{\mathrm{noise}}(s) + [(c + k/s) - H_{\mathrm{ffc}}(s)]v_{\mathrm{b}}(s)}{ms + c + k/s + H_{\mathrm{fbc}}(s)} \tag{12}$$

Theoretically, if the system is ideal, the model of FF can be identified as:

$$H_{\mathrm{ffc}}(s) = \frac{cs + k}{s} \tag{13}$$

From the derived formula, it is observed that FFC can eliminate the impact of ground vibration on the system. Therefore, rather than deteriorating the system's stability, proper FFC will improve the robust performance of the system. To better reflect the influence of FF and FB parameters on the damping performance of AVRS, the vertical vibration transmissibility curve, which is used to analyze the relationship between the ground disturbance $v_{\mathrm{b}}(s)$ and the platform vibration $v_{\mathrm{p}}{}'(s)$ while ignoring $f_{\mathrm{noise}}(s)$, can be rearranged from the Equation (12) as:

$$\frac{v_{\mathrm{p}}{}'(s)}{v_{\mathrm{b}}(s)} = \frac{(c + k/s) - H_{\mathrm{ffc}}(s)}{ms + c + k/s + H_{\mathrm{fbc}}(s)} \tag{14}$$

Compared with the passive curve shown by the red line in Figure 2, the pure FF proportional (P) control effect in the dotted red line is not ideal because it amplifies the amplitude at the resonant frequency point while suppressing the vibration at high frequencies. The optimal coefficient value of P is a trade-off between the high- and low-frequency damping performance. Therefore, merely P in FFC is not sufficient, especially when considering the high-performance requirements of ultra-precision equipment.

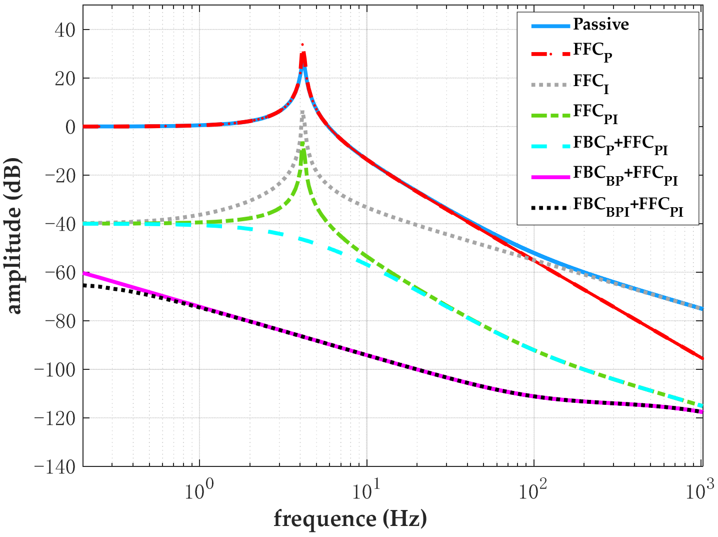


Figure 2. The velocity transmissibility of one DOF AVRS in the vertical direction in different situations: Passive, FF P control $\mathrm{FFC_P}$, FF I control $\mathrm{FFC_I}$, FF PI control $\mathrm{FFC_{PI}}$, FB P control based on FF PI control $\mathrm{FBC_P}$ + $\mathrm{FFC_{PI}}$, FB big P gain control based on FF PI control $\mathrm{FBC_{BP}}$ + $\mathrm{FFC_{PI}}$, FBC big P gain plus I control based on FFC PI control $\mathrm{FBC_{BPI}}$ + $\mathrm{FFC_{PI}}$.

It can be seen from the dotted gray line that the integral (I) control mode in FFC has a better attenuation effect in the low-frequency domain. Therefore, as shown in the dotted green line, it is necessary to adopt the proportional plus integral (PI) control mode in FFC, rather than P or I mode. It is worth noting that the simulation results show that the FFC has a very limited ability to suppress vibration at the structural resonance frequency. As indicated in the dotted blue curve in the figure, the closed-loop P control was added based on feed-forward PI control, which can effectively suppress the natural frequency of the platform. It can be seen from the pink line and dotted black line in Figure 2 that, with the increase in P, the vibration attenuation effect is better. Not only can the adoption of the closed-loop PI mode regulate the equivalent damping, but it changes the equivalent stiffness of the system. Although the performance of closed-loop PI control is superior to P in the frequency range below 1 Hz, due to the constraints of the ultra-low frequency detection accuracy of sensors, the closed-loop system usually adopts the P control mode.

In the absolute velocity feedback control system, although the performance of FBC will be improved as the feedback control gain increases, the system will be unstable at high gains if the influence of the model, the bandwidth of sensors and actuators, the sampling rate, etc., are considered [30,31]. In addition, under the control of fixed parameters, AVRS may be unstable due to load changes, environmental interference, human operation, and other factors. To ensure stability and improve the performance of AVRS, this study aimed to introduce the DDPG reinforcement learning algorithm to realize the automatic optimization of control parameters through continuous interaction with the working environment.

3. MATLAB and ADAMS Co-Simulation

3.1. The Single DOF Model of AVRS in the Vertical Direction

Due to the complex coupling characteristics of the parameters in the system having six degrees of freedom, the AVRS should be transformed into six independent subsystems. To conduct the co-simulation, ADAMS and MATLAB were employed to establish a single DOF model of AVRS in the vertical direction to study the characteristics of FFC and FBC. In this study, ADAMS software was used to simulate the main structure of the system, which is supported by four parallel steel springs, and MATLAB was used to build the control system. The vertical driving force was established at the center of each spring to simulate the vertical linear motor in the actual system. By applying white noise excitation to the base, the passive velocity transmissibility curve, which is critical for model validation, was acquired and is shown in Figure 3.

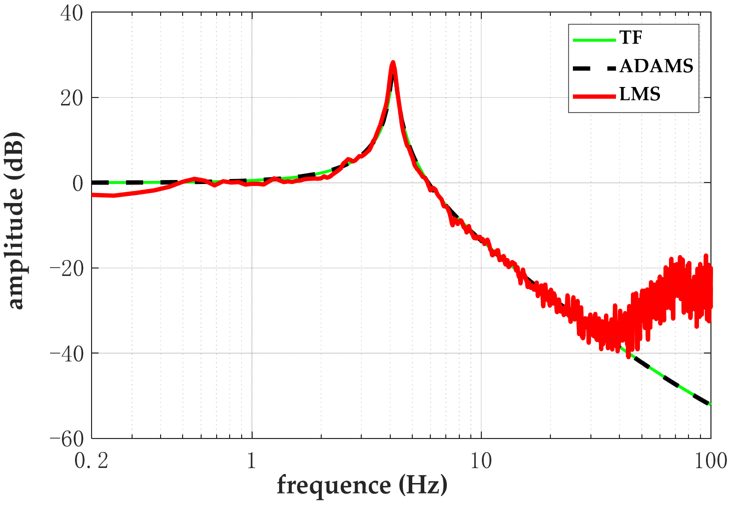


Figure 3. Validation of consistency between calculated transfer function (TF) and ADAMS model, and contrast with the LMS measured result.

The black dotted line represents the passive transfer function curve that was obtained from the sweep frequency test of the ADAMS model. The curve is consistent with the mathematical transfer function model shown by the green line in the same figure. The red line is the experimental transmissibility curve that was vertically acquired by external VSE-15D sensors and the LMS-SCADAS test system. The results demonstrate that the simulation model can still well fit the realistic system below 40 Hz even considering the influence of the system's high-frequency noise and high-order modes.

3.2. DDPG Reinforcement Learning Algorithm

The DDPG algorithm is an upgraded version of the Actor-Critic (AC) algorithm shown in Figure 4. It integrates the advantages of DQN and overcomes the shortcomings of the Actor-Critic algorithm. It uses a replay memory buffer and soft-update target networks to realize stable and robust control in large-scale states or continuous action space. In addition, it adopts the mini-batch method to eliminate the correlation between samples. Through batch normalization to gain hyper-parameters, the agent is enabled to generalize across the environments with different scales of state values. The following paragraphs briefly describe the DDPG algorithm process according to the control flow presented in Figure 4. Details were previously presented elsewhere [24,32]. It is worth noting that the DDPG algorithm adopts the AC structure which includes a policy-based neural network (Policy-net) system and a value-based neural network (Q-net) system. Each consists of an online network and a target network. The two target networks are updated using the same software approach, while the two online networks are different: the online Q-net updates the parameters by minimizing the mean squared error (MSE) as the loss function (LF) to obtain the maximum value of Q, whereas the online Policy-net updates the network parameters by training the mini-batch data to achieve an unbiased estimate of the DPG according to the Monte Carlo method.

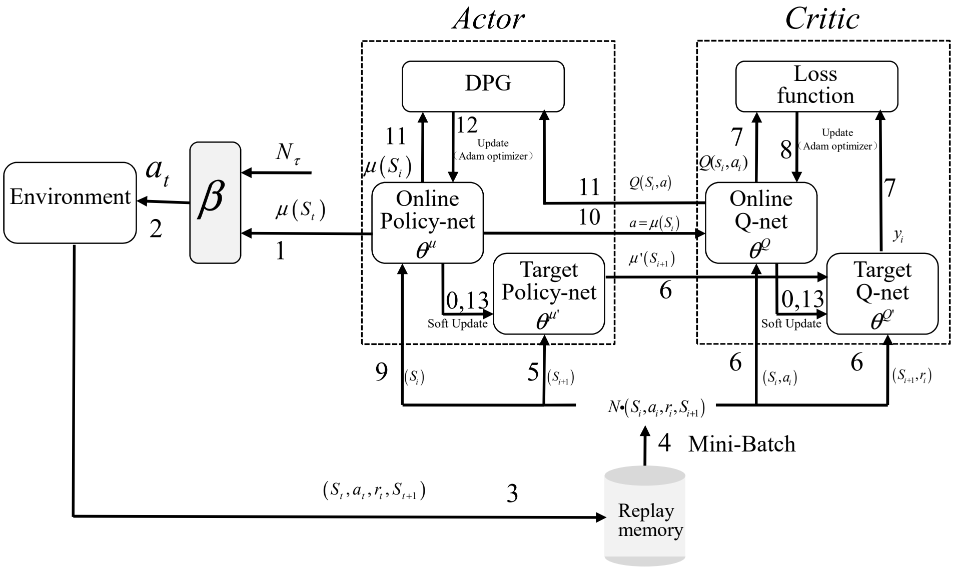


Figure 4. The block diagram of the DDPG algorithm.

Here, the LF can be expressed as follows:

$$LF = \frac{1}{\mathrm{N}}\sum_i \left[\underbrace{r_i + \gamma Q'\left(S_{i+1}, \underbrace{\mu'\left(S_{i+1}|\theta^{\mu'}\right)}_{Targetpolicy-net:\ 5\to 6} \middle| \theta^{Q'} \right)}_{TargetQ-net:\ 6\to 7} - \underbrace{Q\left(S_i, a_i|\theta^{Q}\right)}_{OnlineQ-net:\ 6\to 7} \right]^2 \tag{15}$$

and the DPG can be written as shown in Equation (16):

$$\nabla_{\theta^{\mu}} \mu|S_i \approx \frac{1}{N}\sum_i \underbrace{\underbrace{\nabla_a Q\left(S,a|\theta^{Q}\right)\Big|_{S=S_i,a=\mu(S_i)}}_{OnlineQ-net:\ 9\to10\to11} \nabla_{\theta^{\mu}} \underbrace{\mu(S|\theta^{\mu})|_{S_i}}_{OnlinePolicy-net:\ 9\to11}}_{DPG:\ 11\to12} \tag{16}$$

Thus, according to the subscript in Equations (15) and (16) and the structure in Figure 4, the calculation process of the DDPG algorithm can be clearly understood. In this paper, the adaptive moment estimation (Adam) optimizer used in the online neural networks of DDPG enables faster convergence to the global optimum.

3.3. Simulation of DDPG Algorithm in Active Vibration Control

To build a simulation model that is consistent with the realistic AVRS, the motor drive model $M(s)$ and the low-frequency extensive model of the sensor $S(s)$ should be considered. Suppose that the control blocks of FF and FB are $FF(s)$ and $FB(s)$; then, the classic structure of AVRS is shown in the black-dotted box in Figure 5.

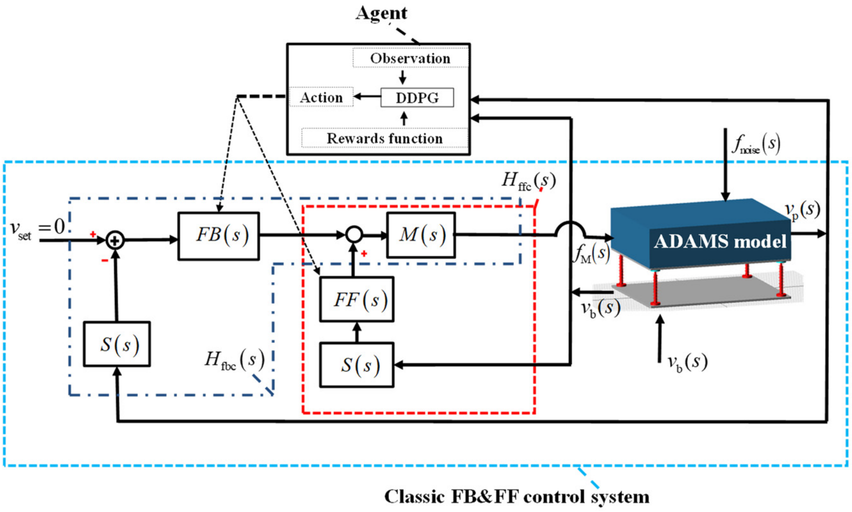


Figure 5. The principle of DDPG algorithm application in co-simulation.

Based on the aforementioned FB and FF co-simulation control model, the DDPG algorithm module was constructed in MATLAB. Through the plant export function, the M file exported from the model of ADAMS was imported into MATLAB to obtain the environment required for the simulation and to realize the information exchange between ADAMS and MATLAB.

Through continuous iterative analysis of the vibration reduction performance of AVRS, the critical control parameters of $FB(s)$ and $FF(s)$ were updated to make the neural network learning parameters converge to the global optimum.

3.3.1. The Structure of the Neural Network

The neural network structure of online Policy-net and online Q-net in DDPG is shown in Figure 6.

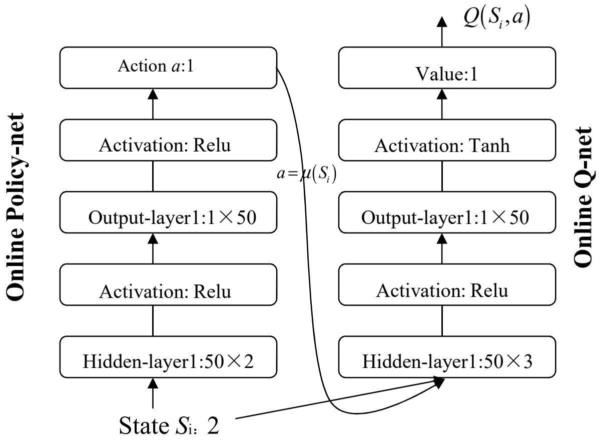


Figure 6. The neural network structure of online Q-net and online Policy-net.

Online Policy-net has one hidden layer with fifty nodes and receives the current state signal S and outputs the action signal a after the Gaussian distribution and amplitude restriction. The activation function of hidden layers in the online Policy-net is Relu, and that of the output layer is also Relu. The value of the online Q-net has the same numbers of hidden layers and nodes as in the online Policy-net and represents the action reward, which is used to evaluate the effectiveness of the selected action. The activation function of the hidden layer in online Q-net is Tanh, and that of the output layer is linear-activated.

3.3.2. Reward Function

The setting of the reward function is crucial; a good design can not only speed up the training process of the reinforcement learning algorithm but also avoid the situation of reward-hacking and convergence to a local optimum. The control objective for active vibration reduction is that the absolute speed of the payload is infinitely close to zero. Then, the reward function can be set as a linear function of v_{p} in simulation, and its expression is shown in Equation (17):

$$R = -|v_{\mathrm{p}}| - 0.1{v_{\mathrm{p}}}^2 \tag{17}$$

3.4. The Results of Simulation

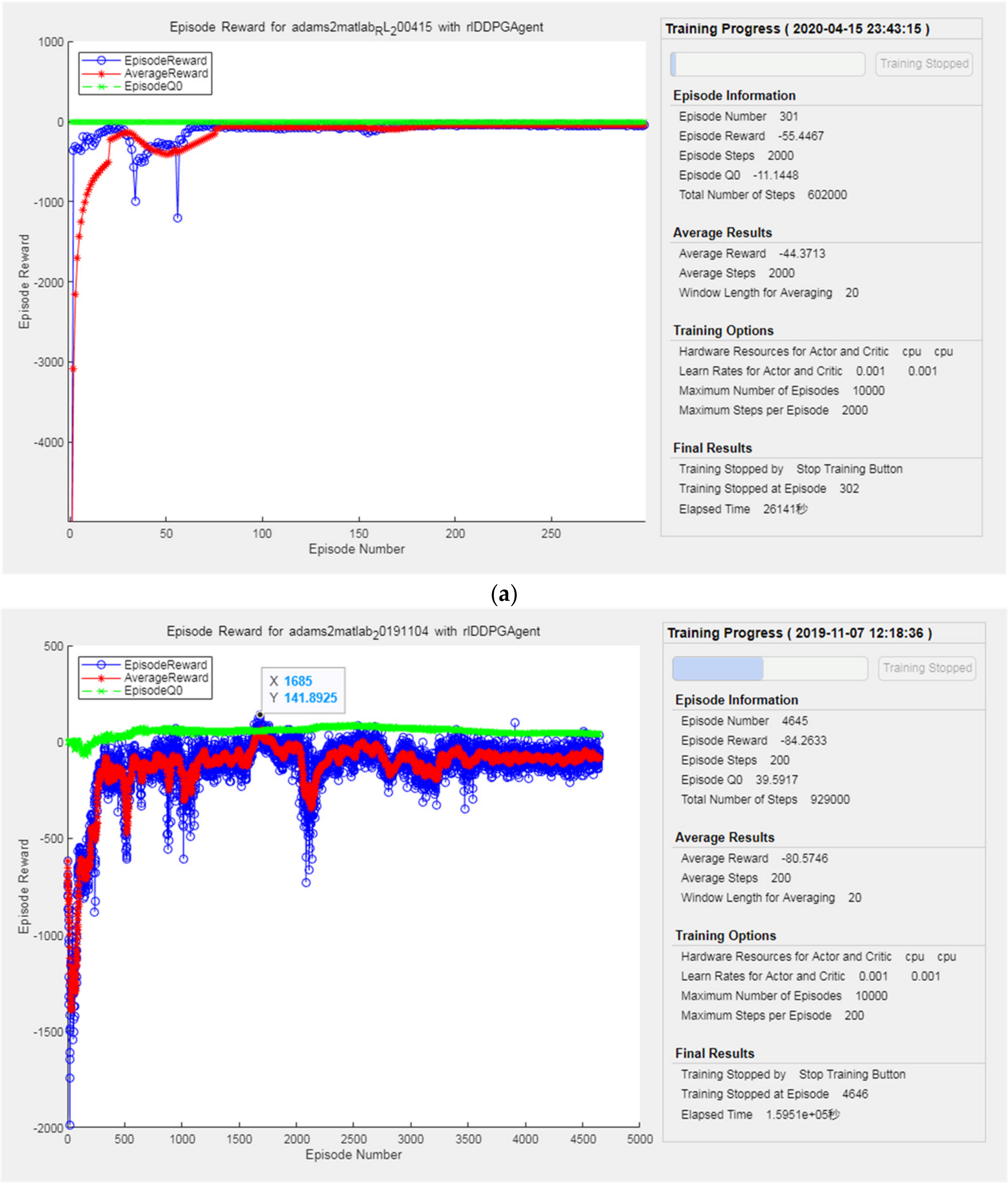
In the simulation, white noise signals with amplitudes of 1 and 200 were used to simulate ground excitation and table disturbance, respectively. More detailed parameters are shown in Table 1.

Table 1. Specifications of simulation.

	MATLAB		
Solver	ode4	Type	Fixed-step
Episode time	2s	Fixed-step size	0.001s
	ADAMS Model		
ADAMS Solver type	Fortran	Simulation mode	Continuous
Animation mode	batch	Communication interval	0.001s
	DDPG		
TargetSmoothFactor	1×10^{-3}	Agent sample time	0.001s
MaxEpisodes	10,000	MaxStepsPerEpisode	20
MiniBatchSize	64	DiscountFactor	0.998

In this study, the DDPG reinforcement learning algorithm simulation was carried out based on the joint control model, which mainly includes the following two parts: (1) FBC parameter that is optimized through DDPG under the condition of proper FFC parameter; (2) FFC and FBC parameters optimize the system simultaneously.

From Figure 7, EpisodeQ0 stands for the prospective discount reward, whose stability reflects whether the critic network was designed properly. The reinforcement learning reward curve of the FBC proportional parameter under the specified feed-forward parameter is shown in Figure 7a, and the self-learning episode reward curve of multiple parameters that contains both the feed-forward gain and the feedback gain is shown in Figure 7b. The result shows that when multiple parameters are learned at the same time, the system can still converge to the optimal value range in a long step size.



Training Progress (2020-04-15 23:43:15)

Episode Information
Episode Number 301
Episode Reward -55.4467
Episode Steps 2000
Episode Q0 -11.1448
Total Number of Steps 602000

Average Results
Average Reward -44.3713
Average Steps 2000
Window Length for Averaging 20

Training Options
Hardware Resources for Actor and Critic cpu cpu
Learn Rates for Actor and Critic 0.001 0.001
Maximum Number of Episodes 10000
Maximum Steps per Episode 2000

Final Results
Training Stopped by Stop Training Button
Training Stopped at Episode 302
Elapsed Time 26141秒

(**a**)

Training Progress (2019-11-07 12:18:36)

Episode Information
Episode Number 4645
Episode Reward -84.2633
Episode Steps 200
Episode Q0 39.5917
Total Number of Steps 929000

Average Results
Average Reward -80.5746
Average Steps 200
Window Length for Averaging 20

Training Options
Hardware Resources for Actor and Critic cpu cpu
Learn Rates for Actor and Critic 0.001 0.001
Maximum Number of Episodes 10000
Maximum Steps per Episode 200

Final Results
Training Stopped by Stop Training Button
Training Stopped at Episode 4646
Elapsed Time 1.5951e+05秒

(**b**)

Figure 7. The learning process of episode reward in DDPG: (**a**) single parameter and (**b**) multi-parameter.

4. Experiments

The DDPG reinforcement learning algorithm, which is a type of model-free RL method, achieves optimal control through continuous interaction with the environment, receives action from the agent and returns the corresponding reward value of the current state. The agent only needs to collect and analyze the vibration signals from the ground and platform to realize self-learning control. The purpose of AVRS is to minimize the vibration of the platform by regulating the real-time dynamic parameters and changing the output control force value corresponding to the detection signals. However, in practical applications, considering the nonlinear characteristics of the inertial system and its multivariable coupling, it is difficult to accurately evaluate the system performance when the updated frequency of parameters is too high.

To improve the control robustness of the system and realize the engineering application of the DDPG control method in the field of vibration reduction, this study aimed to transform real-time dynamic parameter regulation into optimal parameter adjustment. Optimizing the reward function overcomes the shortcomings of the DDPG algorithm, which is highly dependent on the system operating speed and hardware performance, and makes the intelligent integration of the system more economical.

4.1. Reward Function Optimization

Not only the DDPG algorithm is unsuitable in a random environment, but it facilitates reward-hacking when the reward function is improperly set in the process of reinforcement learning. Since the transmissibility curve reflects the modal information about AVRS, and its corresponding technical indicators are not affected by the form of environmental excitation signal under the same parameters, it is appropriate for the amplitude curve of transmissibility to be set as the reference of the reward function. By integrating the difference decibel value of the amplitude between the passive transmissibility curve and the active control curve within the specified frequency range, the area size of the enveloped curve directly reflects the attenuation performance of the AVRS.

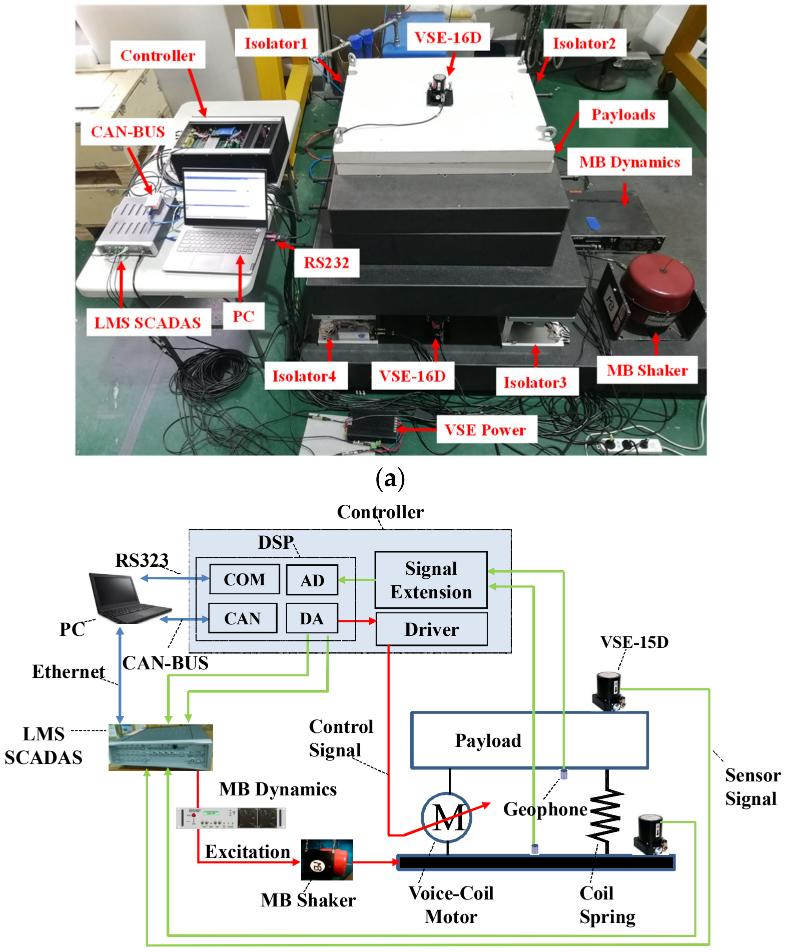
Then, the reward function is defined as:

$$R = \begin{cases} 0 & f_i < f_\mathrm{H}, f_i \geq f_\mathrm{H} \\ \sum\limits_{i=0}^{n} \{[F_\mathrm{o}(f_i) - F_\mathrm{c}(f_i)] * [lg(f_{i+1}) - lg(f_i)]\} & f_\mathrm{L} < f_i < f_\mathrm{H} \end{cases} \quad (18)$$

In Equation (18), $F_\mathrm{o}(f_i)$ and $F_\mathrm{c}(f_i)$ represent the amplitude of passive and active control transmissibility at f_i, respectively. According to the requirements of different technical indicators, the corresponding boundary frequency range $[f_\mathrm{L}, f_\mathrm{H}]$ and the calculation method can be modified, and the parameters optimized through the DDPG self-learning algorithm, to achieve the control performance that meets the application requirements.

4.2. Experimental Setup

The AVRS experimental setup consists of four independent vibration isolation units, controller, payloads, PC, etc. Each isolator is integrated with a steel spring, geophones, and voice-coil motors. The schematic diagram of the experiment is shown in Figure 8.



(**a**)

(**b**)

Figure 8. Experimental system: (**a**) experimental setup and (**b**) sketch of the experiment.

In this experiment, excitation signals generated by the LMS signal system were amplified and acted on the AVRS substrate by the MB shaker. All internal signals were acquired by the AD port in the DSP controller and, after frequency expansion, signal decoupling, and algorithm processing, they were output from the DA port as motor drive signals.

In addition, the PC was used to establish an intelligent agent, which communicates with the DSP in the controller through the serial port and the CAN port. The CAN port realizes vibration signal transmission, and the serial port is used to execute the parameter's updated commands. The online network parameters are updated at the end of each 100-step episode. Before starting self-learning control, 2000 random samples should be stored in replay memory. When the sample size reaches the set value, the agent randomly selects 32 mini-batch samples for DDPG reinforcement learning. Next, two external high-resolution VSE-15D servo velocity seismometers are used to verify the vibration reduction effect of the AVRS. Table 2 shows the parameters of the AVRS.

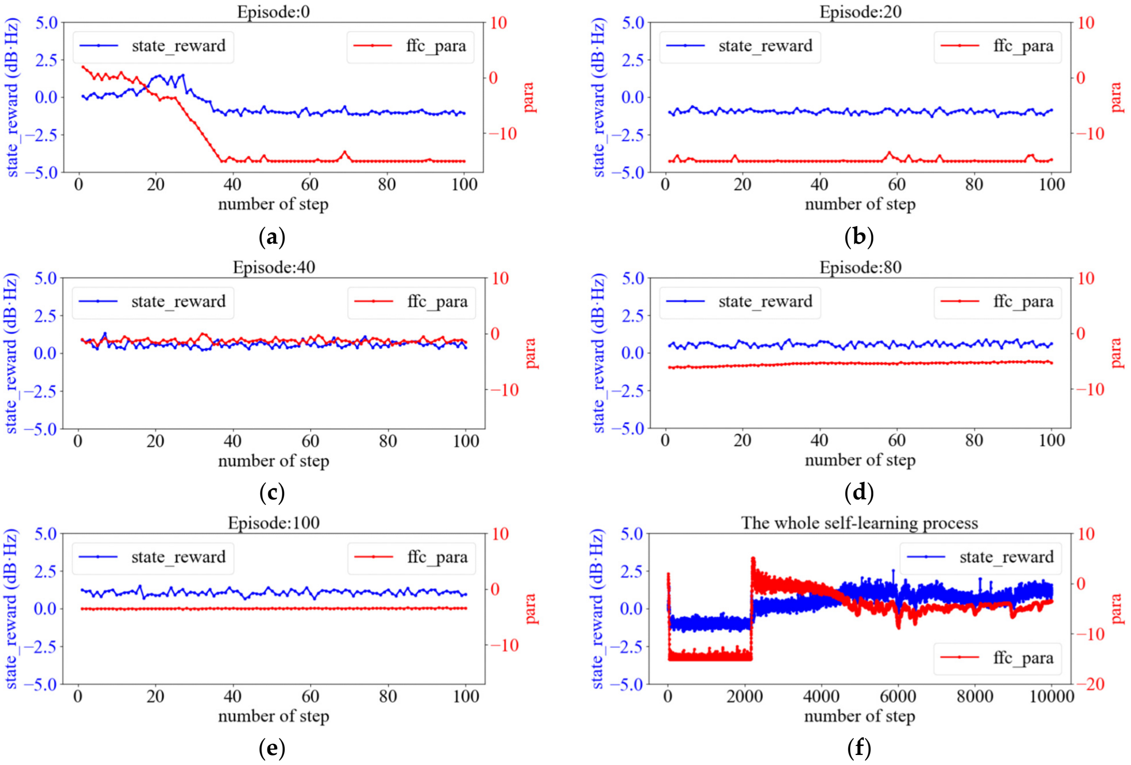
Table 2. Specifications of components in AVRS.

Components	Symbol	Parameter	Value
Payload	m	Mass	1750 kg
Helical spring	k c	Stiffness Damping	300,000 N/m 490 Ns/m
Voice-coil motor	K_T R_M L_M	Force constant Resistant Inductance Moving distance Frequency range	80 N/A 34.5 Ω 28 mh 28 mm ≤200 Hz
Geophone	f_g G_g R_d	Natural frequency sensitivity Internal resistant	4.5 Hz 100.4 V/m/s 2450 Ω

4.3. Experimental Results

The content of the experiment mainly comprises the following three parts: (1) the FFC parameter, which is optimized through DDPG under the condition of a proper FBC parameter; (2) the FBC parameter is optimized in a reasonable FFC parameter; (3) FFC and FBC parameters optimizing simultaneously.

Figure 9 shows the experimental results of FFC parameter self-learning by the DDPG algorithm when the FBC proportional gain is fixed at 1.8. Figure 9a–e are the learning curves of the 0/20/40/80/100th episode, where the horizontal axis is the number of steps, the red line is the value of the FFC parameter, and the blue line is the corresponding value of the action reward. Figure 9f reflects the convergence of the FFC parameter in the whole self-learning process.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 9. The experimental results of FFC single parameter self-learning in DDPG: (**a**–**e**) are the learning curves of the 0/20/40/80/100th episode, respectively; (**f**) are the curves of the whole self-learning process.

Figure 10 shows that there is an approximate normal envelope distribution relationship between the FFC parameter and the state reward value. When the parameter is located between [−15, −5], the active vibration reduction performance becomes better with the increase in the FFC parameter. When it is bigger than −4, the control performance gradually deteriorates. When the parameter is at the limit value of −15 or 5, the system state value is negative, indicating that the system is in a state of vibration amplification or instability. The parameter range of the maximum state reward value is [−5, −4], as marked in the red circle, which is consistent with the region corresponding to the parameter convergence value in Figure 9f, indicating that the DDPG algorithm can track the optimal value range of single-parameter FFC during self-learning.

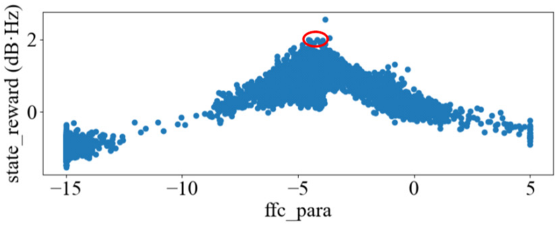
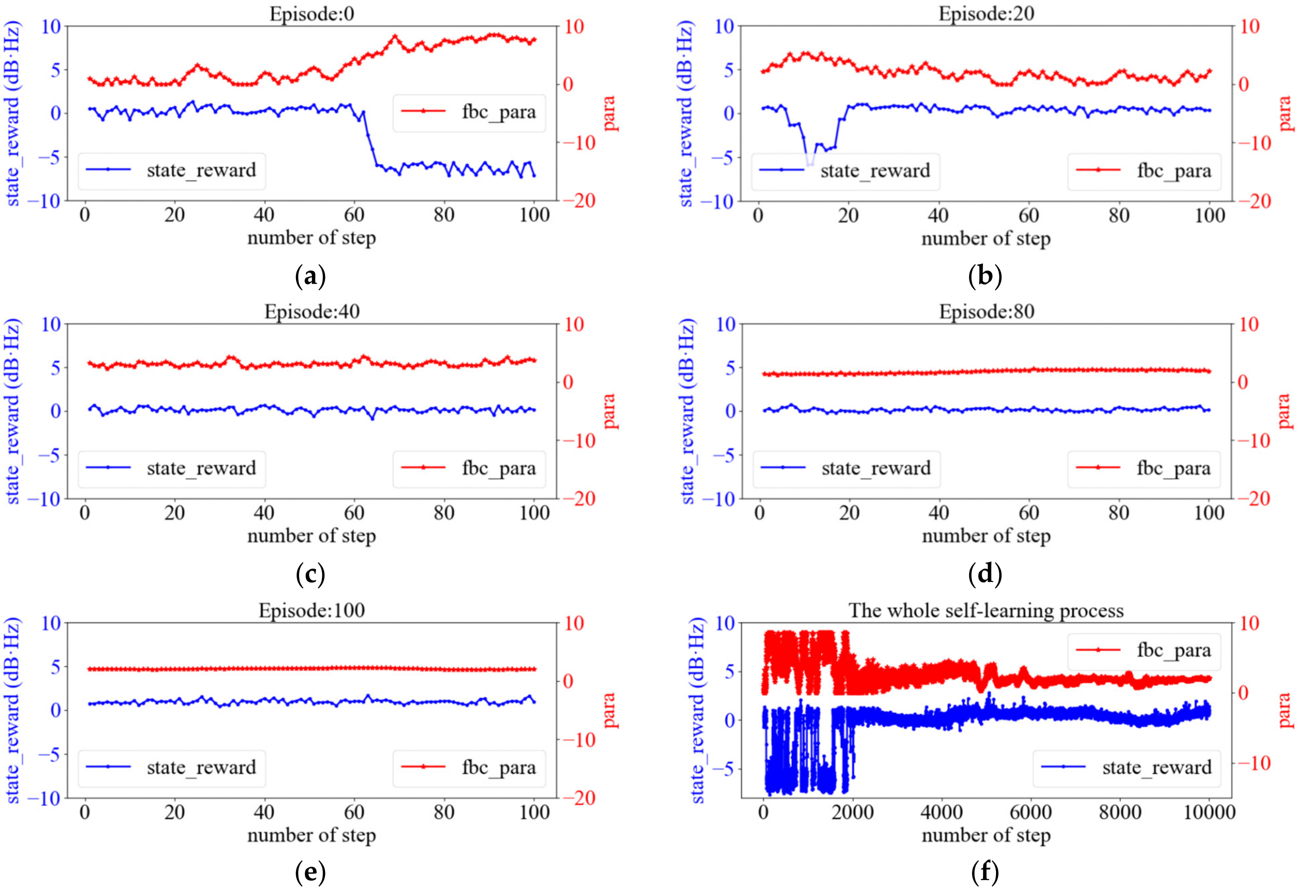


Figure 10. Scatter diagram of state reward and feed-forward proportional parameter.

As shown in Figure 11, the FFC proportional control parameter is fixed to −4.5, and the DDPG optimization learning is carried out when the initial value of the FBC proportional parameter is 0. Figure 11a–e are also the learning curves of the 0/20/40/80/100th episode. Figure 11f reflects the convergence range of the FBC parameter at [2, 3].



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 11. The experimental results of FBC single parameter self-learning in DDPG: (**a**–**e**) are the learning curves of the 0/20/40/80/100th episode, respectively; (**f**) are the curves of the whole self-learning process.

The experimental learning curve in Figure 12 can be divided into three regions: the stable region, the transition region, and the unstable region. It is observed that when the FBC parameter is less than 2.5, the system is in the stable region, and with the increase in the FBC parameter, the control performance of the system improves gradually. In the transition region of [2.5, 6], it is found that there is a similar hysteresis characteristic between the system state reward and the FBC proportional parameter. Therefore, when the FBC parameter increases monotonously from 2.5 to 4.5, the system performance deteriorates. When it is greater than 4.5, the vibration reduction performance of the system starts to decline sharply, and the vibration amplitude of the platform is constant and amplified compared with the initial value. When the FBC parameter increases to above 6, the system is in an unsteady state. Conversely, when the parameter decreases monotonously from 6 to 4.5, the system is still unstable; when it is less than 3, the system gradually returns to the stable state.

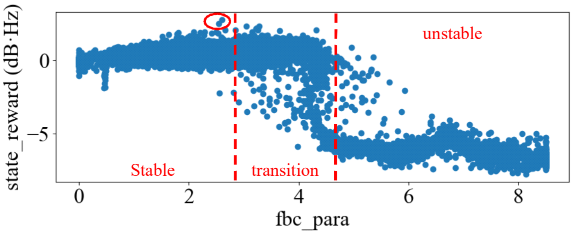


Figure 12. Scatter diagram of state reward and feedback proportional parameter.

In Figure 13, the results of the multi-parameter self-learning experiment are shown when the initial proportional control parameters P of the FBC and FFC are both 0. From Figure 13a–f, the optimal parameter of FBC is in the range of [2.0, 3.0], and the FFC parameter is located in [−5.0, −4.0]. The results show that the parameters learned simultaneously have good consistency with the values learned separately.

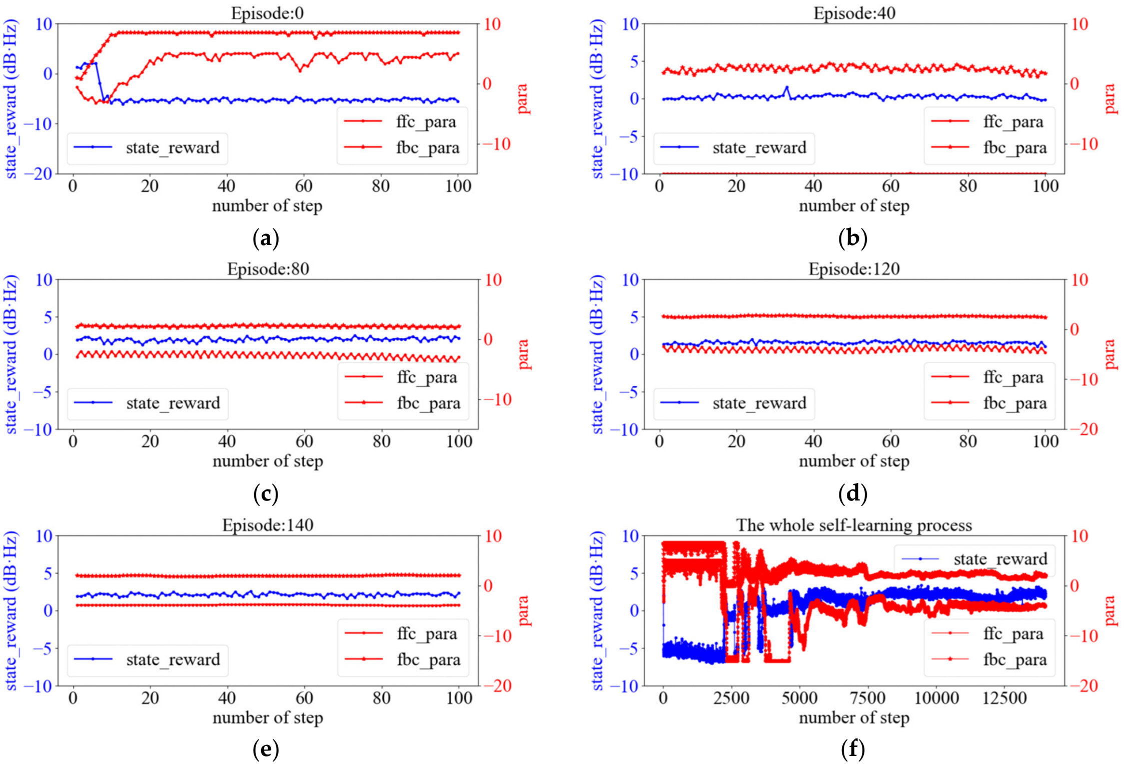
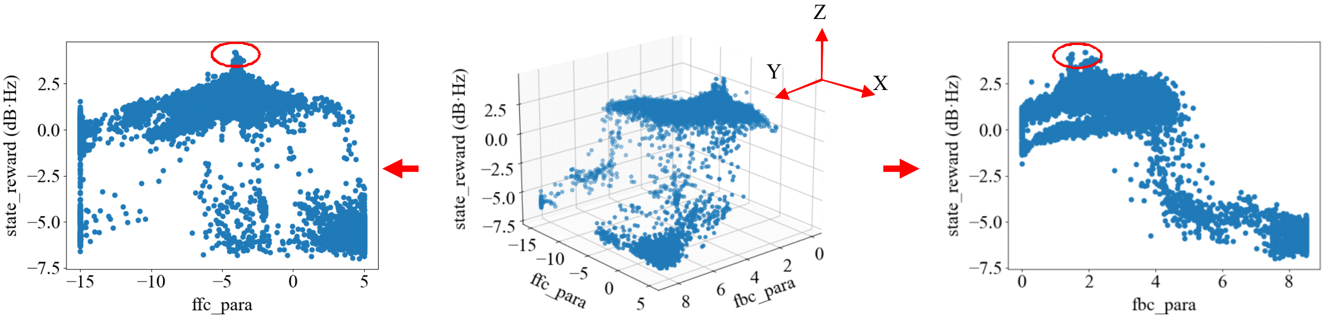


Figure 13. Experimental results of multi-parameter self-learning in DDPG: (**a**–**e**) are the learning curves of the 0/40/80/120/140th episode, respectively; (**f**) are the curves of the whole self-learning process.

Figure 14 is the 3D projected scatter diagram of state reward and the multiple parameters. The red circles in the figure indicate the area where the optimal values are located, and the identification results are consistent with the final convergence values of the learning curves. Due to the high-order, nonlinearity, and strong coupling characteristics of the actual system, its dynamic characteristics are very complex, and it is difficult to obtain accurate system control parameters. Nevertheless, the above experiments show that a small optimal value interval can be determined.



(b) (a) (c)

Figure 14. The learning results: (**a**) 3D scatter diagram of state reward and multiple parameters; (**b**) the projection of (a) on the opposite of the "Y" direction; (**c**) the projection of (**a**) in the "X" direction.

Based on the DDPG self-learning results above, which take the integral area of the difference between the passive and active vibration transmissibility curves as the reward function, the vertical FBC parameter of the system is selected as 2.5 and the FFC parameter as −4.0.

Figure 15 shows the comparison results of vibration transmissibility curves about internal geophones and external VSE-16D sensors under passive, FFC, FBC, and hybrid control. The curves in the figure show that the decoupling calculation value of the internal distributed sensors, which can correctly reflect the actual vibration of the system, has good consistency with the direct test value of the external sensors. The experimental results show that the application of the DDPG algorithm in active vibration control is feasible and effective.

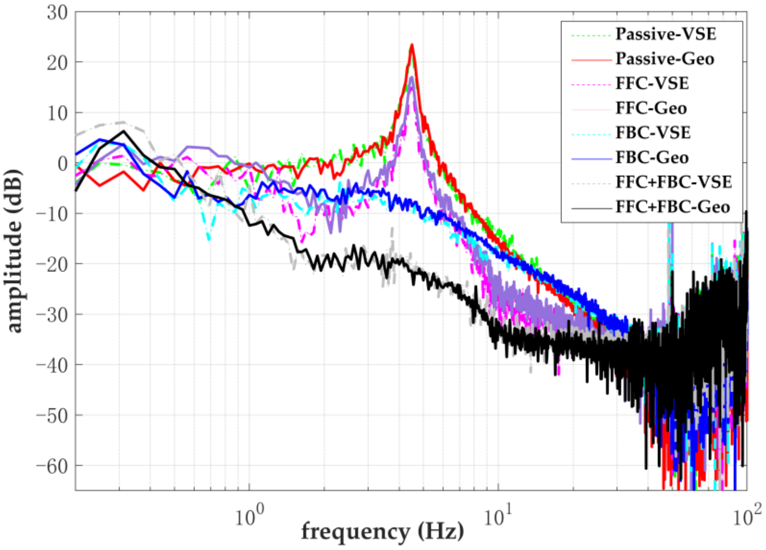


Figure 15. Contrasts between the transmissibility curves of internal geophones and external VSE-16D sensors under optimal learning parameters: Passivity, FFC, FBC, hybrid control (FBC + FFC).

Through the DDPG control method, the influence of the control parameters on the system performance within the set range can be analyzed. When the system structure and payload change, the agent can repeat the learning process until the new optimal control parameters are obtained.

In addition, the final learned parameters can be easily solidified into the static flash to achieve high-performance offline operation. The theoretical results of the method proposed in this paper are in good agreement with the experimental measurements and have strong control robustness in the active vibration reduction field. This method has good application prospects in mature commercial products.

5. Conclusions

In this study, the dynamic control mechanism of FB and FF was investigated, and a detailed block diagram of the FB and FF control system was proposed according to the classification of vibration transmission paths. In addition, based on the united FF and FB control case, the DDPG algorithm was introduced to perform real-time self-learning on the system parameters. Since DDPG is not suitable for random signal learning, this paper proposed a reward function optimization method that is still applicable to the multi-parameter self-learning process. Simulations and experiments show that the DDPG algorithm can continuously optimize system parameters through interaction with the isolator units to achieve effective suppression of platform vibration without the need for artificial prior knowledge. Therefore, the method presented in this paper has great potential value in future practical applications because it avoids human interference. The reward function optimization is crucial in speeding up the DDPG learning process. The setting method of the reward function proposed in this paper still has the disadvantages of random errors at low frequencies and calculation errors at high frequencies. Therefore, reasonable value selection and optimal calculation of the frequency interval will be important directions in subsequent research.

Author Contributions: Conceptualization, X.J.; methodology, X.J. and Y.K.; software, H.M.; validation, X.J. and J.T.; writing—original draft preparation, X.J.; writing—review and editing, J.T.; visualization, X.J.; supervision, Y.K.; project administration, Y.K.; funding acquisition, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC), grant number 51875226.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank Shaojun Wang of the Wuhan University of Technology for his writing assistance in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gordon, C.G. Generic Vibration Criteria for Vibration-Sensitive Equipment. In Proceedings of the SPIE's International Symposium on Optical Science, Engineering, and Instrumentation International Society for Optics and Photonics, Denver, CO, USA, 28 September 1999; pp. 22–33.
2. Amick, H.; Gendreau, M.; Busch, T.; Gordon, C.; Gordon, C.; Lane, S.; Bruno, S. Evolving Criteriafor Research Facilities: I—Vibration. In Proceedings of the Reprinted from Proceedings of SPIE Conference 5933: Buildings for Nanoscale Research and Beyond, San Diego, CA, USA, 31 July–1 August 2005.
3. Karnopp, D.; Crosby, M.J.; Harwood, R.A. Vibration Control Using Semi-Active Force Generators. *J. Eng. Ind.* **1974**, *96*, 619–626. [CrossRef]
4. Griffin, S.; Gussy, J.; Lane, S.A.; Henderson, B.K.; Sciulli, D. Virtual Skyhook Vibration Isolation System. *J. Vib. Acoust.* **2002**, *124*, 63–67. [CrossRef]
5. Zuo, L.; Slotine, J.-J.E.; Nayfeh, S.A. Model reaching adaptive control for vibration isolation. *IEEE Trans. Contr. Syst. Technol.* **2005**, *13*, 611–617. [CrossRef]
6. Jin, Q.B.; Liu, Q. IMC-PID Design Based on Model Matching Approach and Closed-Loop Shaping. *ISA Trans.* **2014**, *53*, 462–473. [CrossRef] [PubMed]

7. Ding, R.; Wang, R.; Meng, X.; Chen, L. A Modified Energy-Saving Skyhook for Active Suspension Based on a Hybrid Electromagnetic Actuator. *J. Vib. Control* **2019**, *25*, 286–297. [CrossRef]
8. Yan, B.; Brennan, M.J.; Elliott, S.J.; Ferguson, N.S. Active Vibration Isolationofa Systemwitha Distributed Parameter Isolator Using Absolute Velocity Feedback Control. *J. Sound Vib.* **2010**, *329*, 1601–1614. [CrossRef]
9. Zuo, L.; Slotine, J.-J.E. Robust Vibration Isolation via Frequency-Shaped Sliding Control and Modal Decomposition. *J. Sound Vib.* **2005**, *285*, 1123–1149. [CrossRef]
10. Yasuda, M.; Osaka, T.; Ikeda, M. Feedforward Control of a Vibration Isolation System for Disturbance Suppression. In Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 13 December 1996; Volume 2, pp. 1229–1233.
11. Butler, H. Feedforward Signal Prediction for Accurate Motion Systems Using Digital Filters. *Mechatronics* **2012**, *22*, 827–835. [CrossRef]
12. Sun, L.; Li, D.; Gao, Z.; Yang, Z.; Zhao, S. Combined Feedforward and Model-Assisted Active Disturbance Rejection Control for Non-Minimum Phase System. *ISA Trans.* **2016**, *64*, 24–33. [CrossRef]
13. Yoshioka, H.; Takahashi, Y.; Katayama, K.; Imazawa, T.; Murai, N. An Active Microvibration Isolation System for Hi-Tech Manufacturing Facilities. *J. Vib. Acoust.* **2001**, *123*, 269–275. [CrossRef]
14. Ding, J.; Luo, X.; Chen, X.; Bai, O.; Han, B. Design of Active Controller for Low-Frequency Vibration Isolation Considering Noise Levels of Bandwidth-Extended Absolute Velocity Sensors. *IEEE/ASME Trans. Mechatron.* **2018**, *23*, 1832–1842. [CrossRef]
15. Li, P.; Lam, J.; Cheung, K.C. H_∞ Control of Periodic Piecewise Vibration Systems with Actuator Saturation. *J. Vib. Control* **2017**, *23*, 3377–3391. [CrossRef]
16. Tang, D.; Chen, L.; Tian, Z.F.; Hu, E. Adaptive Nonlinear Optimal Control for Active Suppression of Airfoil Flutter via a Novel Neural-Network-Based Controller. *J. Vib. Control* **2018**, *24*, 5261–5272. [CrossRef]
17. van der Poel, T.; van Dijk, J.; Jonker, B.; Soemers, H. Improving the Vibration Isolation Performance of Hard Mounts for Precision Equipment. In Proceedings of the 2007 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Zurich, Switzerland, 4–7 September 2007; pp. 1–5.
18. Schölkopf, B. Learning to See and Act. *Nature* **2015**, *518*, 486–487. [CrossRef]
19. Littman, M.L. Reinforcement Learning Improves Behaviour from Evaluative Feedback. *Nature* **2015**, *521*, 445–451. [CrossRef]
20. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the Game of Go without Human Knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
21. Khalatbarisoltani, A.; Soleymani, M.; Khodadadi, M. Online Control of an Active Seismic System via Reinforcement Learning. *Struct. Control Health Monit.* **2019**, *26*, e2298. [CrossRef]
22. Peters, J.; Schaal, S. Natural Actor-Critic. *Neurocomputing* **2008**, *71*, 1180–1190. [CrossRef]
23. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
24. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. *arXiv* **2015**, arXiv:1509.02971.
25. Silver, D.; Lever, G. Deterministic Policy Gradient Algorithms. In Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 21–26 June 2014.
26. Kofinas, P.; Vouros, G.; Dounis, A.I. Energy Management in Solar Microgrid via Reinforcement Learning Using Fuzzy Reward. *Adv. Build. Energy Res.* **2018**, *12*, 97–115. [CrossRef]
27. Marashi, M.; Khalilian, A.; Shiri, M.E. Automatic Reward Shaping in Reinforcement Learning Using Graph Analysis. In Proceedings of the 2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 18–19 October 2012; pp. 111–116.
28. Sumino, S.; Mutoh, A.; Kato, S. Evolutionary Approach of Reward Function for Reinforcement Learning Using Genetic Programming. In Proceedings of the 2011 International Symposium on Micro-NanoMechatronics and Human Science, Nagoya, Japan, 6–9 November 2011; pp. 385–390.
29. Smith, M.; Wang, F.-C. Controller Parameterization for Disturbance Response Decoupling: Application to Vehicle Active Suspension Control. *Control Syst. Technol. IEEE Trans.* **2002**, *10*, 393–407. [CrossRef]
30. Tjepkema, D.; van Dijk, J.; Soemers, H.M.J.R. Sensor Fusion for Active Vibration Isolation in Precision Equipment. *J. Sound Vib.* **2012**, *331*, 735–749. [CrossRef]
31. Yin, H.; Wang, Y.; Zhang, X.; Li, P. Feedback Delay Impaired Reinforcement Learning: Principal Components Analysis of Reward Positivity. *Neurosci. Lett.* **2018**, *685*, 179–184. [CrossRef] [PubMed]
32. Hao, G.; Fu, Z.; Feng, X.; Gong, Z.; Chen, P.; Wang, D.; Wang, W.; Si, Y. A Deep Deterministic Policy Gradient Approach for Vehicle Speed Tracking Control With a Robotic Driver. *IEEE Trans. Automat. Sci. Eng.* **2021**, *19*, 2514–2525. [CrossRef]