2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Deep Reinforcement Learning based Personalized Locomotion Planning for Lower-Limb Exoskeletons

Javad K. Mehr, *Student Member, IEEE*, Eddie Guo, *Student Member, IEEE*, Mojtaba Akbari, Vivian K. Mushahwar, *Member, IEEE*, and Mahdi Tavakoli, *Senior Member, IEEE*

***Abstract*— This paper introduces intelligent central pattern generators (iCPGs) that can plan personalized walking trajectories for lower-limb exoskeletons. This can make walking more comfortable for the users by resolving one of the significant shortcomings of most commercially available exoskeletons, which is the use of pre-defined fixed trajectories for all users. The proposed method combines reinforcement learning (RL) with previously introduced adaptable central pattern generators (ACPGs) to learn a user's physical interaction behaviour and refine the exoskeleton's walking trajectories. The ACPG method embeds physical human-robot interaction (pHRI) in CPGs to make changing gait trajectories in real-time, possible. However, to effectively refine gait trajectories based on pHRIs, the parameters must be precisely identified and updated as a user interacts with the exoskeleton. Our proposed method uses RL to modify (amplify/attenuate) the pHRI energy based on a user's interaction behaviour, and form an effective energy value which can facilitate reaching desired gait pattern for users via iCPG dynamics. The proposed method can resolve the aforementioned challenges with ACPGs and personalized trajectory generation. The simulation and experimental results provide evidence that the proposed method can effectively adapt to the user's behaviour in different walking scenarios with the Indego lower-limb exoskeleton.**

I. Introduction

Neurological impairments, such as spinal cord injury, stroke, and multiple sclerosis, result in mobility impairments that reduce the quality of life of millions worldwide. The use of assistive and rehabilitative exoskeletons can help individuals maintain their independence and improve their physical fitness. Several powered exoskeletons such as Indego [1], Exo H3 [2], ReWalk [3], HAL [4], and Ekso GT [5] have been developed in recent years for user assistance and rehabilitation in clinics. Despite the great capability of these devices, there still exists a need for software improvement to increase the demand for their use.

The ideal exoskeleton controller must understand a user's intention and adapt to their gait pattern. Que *et al.* [6] used electroencephalogram (EEG) and electrocardiogram (ECG) signals to determine a user's intention and appropriately adjust the exoskeleton's assistance level. The method developed by Gue *et al.* [7] used EEG and ECG signals to select between three predefined trajectories (static, normal walking, high leg lifting) with a neural network classifier. Although using these types of sensors is a promising way of understanding intention, their usage is limited due to difficulties in attaching the sensors to the user's body in addition to signal processing. In a different approach, some studies [8, 9] considered the body features, e.g., age and weight, to reshape the exoskeleton's walking pattern. However, to be sufficiently accurate to capture all features of user's locomotion, a large number of parameters need to be considered, which makes implementing the method challenging. An alternative solution for these challenges is to use advanced motion planning methods in combination with machine learning (ML) based intention estimation.

Central pattern generators (CPGs) are among the major motion planning algorithms for exoskeleton control due to their ability to generate synchronized rhythmic motions between different joints [10]–[12]. Inspired by the motion of the salamander, the dynamics of CPGs in robots were first proposed by Ijspeert *et al.* [13]. Due to the inherent rhythmic motion generation feature, CPGs have been widely used in high-level control of exoskeletons to plan a fixed walking pattern [14]–[17]. Different control approaches, e.g., impedance control [15, 16] and admittance control [17], are being used in low-level control strategies to track the desired trajectories. Although these methods provide a control strategy for exoskeletons, the trajectories need to be changeable to address the needs of different people. Some studies have used adaptable CPGs (ACPGs) to let users refine gait trajectory via physical interaction with the robot [10, 12, 18]. These studies showed that ACPGs could solve the adaptability issue in the motion planning of exoskeletons if the initialization is precise and the user's interaction behaviour does not change considerably over time. These requirements limit ACPGs in providing personalized locomotion trajectories in long periods of walking during which changes in the users' interaction behavior occurs.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC), Canadian Institutes of Health Research (CIHR), Canada Foundation for Innovation (CFI), and the Alberta Jobs, Economy and Innovation Ministry's Major Initiatives Fund to the Center for Autonomous Systems in Strengthening Future Communities. *(Corresponding author: Javad K. Mehr)*

Javad K. Mehr is with the Department of Electrical and Computer Engineering, and the Department of Medicine, University of Alberta, Edmonton, Alberta, Canada (e-mail: J.Khodaeimehr@ualberta.ca)

Eddie Guo is with Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada and the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada (email: edward.guo1@ucalgary.ca)

Mojtaba Akbari, and Mahdi Tavakoli are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada (e-mail: Akbari@ualberta.ca, Mahdi.Tavakoli@ualberta.ca)

Vivian K. Mushahwar is with the Department of Medicine, Division of Physical Medicine and Rehabilitation, University of Alberta, Edmonton, Alberta, Canada T6G 2E1 (e-mail: Vivian.Mushahwar@ualberta.ca)

All authors are with the Sensory Motor Adaptive Rehabilitation Technology (SMART) Network, University of Alberta, Edmonton, Alberta, Canada T6G 2E1

The inherent characteristic of reinforcement learning (RL) is learning while interacting in real-time making it a good fit for personalization applications. The method proposed by Shen *et al.* [19] modelled the human-exoskeleton system as a leader-follower system and used an RL-based control algorithm to adjust the walking assistant level. Huang *et al.* [20] used a mass-spring-damper model to estimate physical interaction between humans and exoskeletons, and they used RL to learn the spring and damper coefficients in the model. They employed the estimated interaction in high-level control of an exoskeleton [20] (the impedance model changes the trajectory generated by dynamical movement primitives; DMPs). RL has also been used in manipulating motion planning algorithm parameters by Zhang *et al.* [21]. Here, the RL algorithm adjusted the gain in trajectory generated by DMPs while taking the stability of thesystem into consideration [21].

Among RL algorithms used for robotic applications, deep deterministic policy gradient (DDPG) has been commonly used, including control of a biped robot [9] and motion control of a six-degree-of-freedom arm robot [22]. However, DDPG suffers from overestimating future rewards, and optimal policy convergence [23]. The twin delayed deep deterministic policy gradient (TD3) addresses these limitations. TD3 is a model-free, off-policy, actor-critic algorithm used for online learning in an environment with continuous action spaces. Thus, TD3 is an improvement over DDPG, and related algorithms by increasing its robustness through clipped double-Q learning and decreasing the likelihood of Q-function exploitation via policy smoothing [23, 24].

This paper introduces the intelligent CPG (iCPG), which combines reinforcement learning with ACPGs for personalized motion planning of exoskeletons. This method resolves the need for precise initialization in ACPGs, which is necessary for effective human-robot interactions (HRIs). Furthermore, our proposed method can adapt to changes in the interaction behaviour of users. The contributions of the paper are summarized as follows:

- We introduce a novel RL-based method to modify HRI energy based on the user's interaction behaviour.
- The ACPG structure is improved, and the iCPG method is introduced for the first time to resolve challenges with previous ACPGs [10, 12].

II. Intelligent CPG dynamics

A multi-degree-of-freedom lower-limb exoskeleton interacting with a human user can be modeled as follows:

$$M_q(q)\ddot{q} + C_q(q)\dot{q} + G_q(q) = \tau_{\text{mot}} + \tau_{\text{hum,p}} + \tau_{\text{hum,a}} \quad (1)$$

where $M_q(q)$, $C_q(q)$, and $G_q(q)$ are the inertia matrix, the matrix of Coriolis, centrifugal, and damping terms, and the vector of gravitational torques, respectively. Further, q is the vector of the exoskeleton joint positions, τ_{mot} is the exoskeleton's motor torque, and $\tau_{\text{hum,p}}$ and $\tau_{\text{hum,a}}$ are the passive and active parts of the human torque vector, respectively.

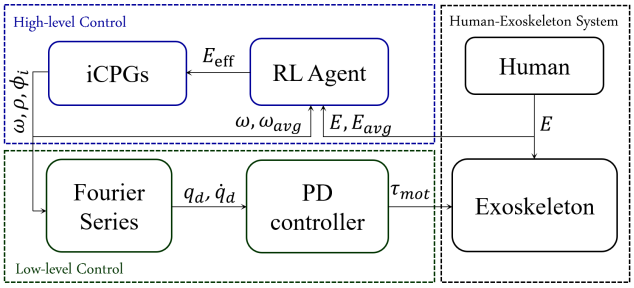


Fig. 1: Schematic of the proposed iCPG method for personalized motion planning.

The ACPG was used to plan the exoskeleton joints' motion in real-time during walking [10, 12, 18]. Although it could refine gait trajectories based on pHRIs, the parameter values play an important role in the method's effectiveness. Furthermore, precise parameter identification in conjunction with minimum changes in the user's interaction behaviour is critically important for the ACPG's performance. To address these issues and provide personalized motion planning, we have integrated RL with ACPGs and introduced iCPGs (see Fig. 1). The iCPG dynamics for encapsulating variations of the overall locomotion frequency $\omega(t)$, oscillation amplitude $\rho(t)$, and phase variation of each joint $\phi_i(t)$ is

$$\begin{aligned}
\ddot{\omega}(t) &= \gamma_\omega \left(\frac{\gamma_\omega}{4}(\Omega + \psi_\omega E_{\text{eff}}(t) - \omega(t)) - \dot{\omega}(t) \right) \\
&\quad + k_\omega u(\omega(t) - \omega_{th_+}) \log\left(\frac{\omega_{\max} - \omega(t)}{\omega_{\max} - \omega_{th_+}} \right) \\
\ddot{\rho}(t) &= \gamma_\rho \left(\frac{\gamma_\rho}{4}(A_\rho + \psi_\rho E_{\text{eff}}(t) - \rho(t)) - \dot{\rho}(t) \right) \\
&\quad + k_\rho u(\rho(t) - \rho_{th_+}) \log\left(\frac{\rho_{\max} - \rho(t)}{\rho_{\max} - \rho_{th_+}} \right) \\
\dot{\phi}_i(t) &= \omega(t) + \sum_{j=1}^{m_i} \eta_{ij} \sin\left(\phi_i(t) - \phi_j(t) - \phi_{ij}\right)
\end{aligned} \quad (2)$$

where m_i is the number of adjacent joints to the joint i, and η_{ij} is the coupling constant between the ith and jth adjacent joints. Ω and A_ρ are the steady-state frequency and amplitude for $\omega(t)$ and $\rho(t)$, and γ_ω and γ_ρ are constant parameters. The parameters ψ_ω and ψ_ρ are constant values for adjusting the effect of physical interaction in iCPG dynamics. The thresholds ω_{th_+} and ρ_{th_+} are the positive threshold of $\omega(t)$ and $\rho(t)$, respectively, that trigger the deceleration term with gains k_ω and k_ρ to avoid reaching the maximum allowable frequency $\omega_{\max}$ and amplitude $\rho_{\max}$. Furthermore, $u(\cdot)$ is a step function that activates the $\log$ functions when the aforementioned thresholds are crossed. In real experiments with the able-bodied person wearing the Indego exoskeleton, these values will be determined based on the users' comfort.

Most notably, and the focus of this paper, is the effective HRI energy, $E_{\text{eff}}(t)$, which is a function of the HRI energy, and is determined via the TD3 algorithm, which will be presented in Sec. III-A [23]. The HRI energy of joint i, $E_i(t)$, is

$$E_i(t) = \int_0^t \tau_{\text{HRI},i}(t)\dot{q}_i(t)\ dt \quad (3)$$

where $\dot{q}_i(t)$ is the velocity of the ith joint and $\tau_{\text{HRI},i}(t)$ is the estimated human torque on the ith joint, which is estimated using a trained neural network based on the method described in Sharifi et al. [12]. The total HRI energy ($E(t)$) is the summation of the interaction energies of all joints.

Using a Fourier series expansion, the described iCPG outputs are transformed into a reference locomotion trajectory, $q_i(t)$, for the ith joint of the exoskeleton:

$$q_i(t) = \xi_i(t) + \rho_i(t) \sum_{k=1}^{N_i} (a_{i_k} \cos k\phi_i(t) + b_{i_k} \sin k\phi_i(t)) \tag{4}$$

where N_i is the number of terms in Fourier's series and a_{i_k} and b_{i_k} are the coefficients of that. The frequency $\omega(t)$, and amplitude $\rho_i(t)$ of walking, and also phase $\phi_i(t)$ of each joint's oscillatory motion (see Eq. (2)) are modified in real-time via the iCPG-based update rules in (2).

III. Implementation of an RL agent to adjust energy contributions for trajectory shaping

Deep reinforcement learning was used to modify pHRI energy ($E(t)$) and determine effective energy values (E_{eff}) in (2) based on the physical interaction behaviour of lower-limb exoskeleton users. The RL algorithm employed in this project and the reward function used for determining the E_{eff} are introduced in the following subsections.

A. Deep reinforcement learning

RL is a learning strategy that attempts to model an agent interacting with its environment while learning reward-maximizing behaviour. At each time step t in a given state $s \in \mathcal{S}$, an RL agent selects an action $a \in \mathcal{A}$ with respect to a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, and receives a reward r and transitions to a new state $s' \in \mathcal{S}$ in its environment. The return, r_t, is defined as the discounted sum of rewards $r_t = \sum_{k=t+1}^{T} \gamma^{k-t} r(s_k, r_k)$, where γ is a discount factor determining the relative importance of future rewards and T is the end of an episode. The objective in reinforcement learning is to find the optimal policy π_ϕ, with parameters ϕ, which maximizes the expected return $J(\phi) = \mathbb{E}_{\tau\sim\pi_\phi}[r(\tau)]$, where τ is the probability of a trajectory $(s_0, a_0, \ldots, s_{T+1})$ following π_ϕ. To optimize $J(\phi)$, the agent learns a value function Q, which maps the agent's state and action to a reward, $Q : \mathcal{S} \times \mathcal{A} \rightarrow R$.

A TD3 strategy was used to formulate the RL problem in this paper. The characteristics of the TD3 algorithm make it a good fit for the personalized trajectory generation problem in this paper. In particular, TD3 uses double critic networks to approximate the reward from a given state and action using the Bellman equation in terms of the discounted sum of expected TD errors:

$$\begin{aligned} Q_\theta(s,a) &= r_t + \gamma \mathbb{E}[Q_\theta(s_{t+1}, a_{t+1} - \delta_t] \\ &= r_t + \gamma \mathbb{E}[r_{t+1} + \gamma \mathbb{E}[Q_\theta(s_{t+2}, a_{t+2})] - \delta_t \\ &= \mathbb{E}_{\tau\sim\pi_\phi} \left[\sum_{i=t}^{T} \gamma^{i-t} (r_i - \delta_i) \right] \end{aligned} \tag{5}$$

where $\sum_{i=t}^{T} \gamma^{i-t}(r_i - \delta_i)$ is the discounted sum of returns, $Q_\theta(s,a)$ is the differentiable function approximator with the parameter θ, and $\mathbb{E}[\cdot]$ is the expectation from a sequence of states and actions τ following the policy π_ϕ. During training, an actor network and two critic networks are initialized with random parameters $(\phi, \theta_1, \theta_2)$. Because training a policy using both actor and critic networks can result in divergence of the agent behaviour and cause instability, target networks with parameters $(\phi', \theta'_1, \theta'_2)$ are initialized. A replay buffer $\mathcal{B}$ is also initialized to record a subset of tuples of the agent's experiences (s_t, a_t, r_t, s_{t+1}), which is later randomly sampled for learning. At each timestep, an action is taken with an added exploration noise to prevent overfitting:

$$a \sim \pi_\phi(s) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma) \tag{6}$$

where ϵ is the exploration noise sampled from a normal distribution with standard deviation σ. From the action taken, the resulting transition tuple (s_t, a_t, r_r, s_{t+1}) is stored in the replay buffer $\mathcal{B}$. Next, an action is selected with target policy smoothing applied. The action is clipped to the action space, and the noise is clipped between constants $\pm c$ to keep the target close to the original action:

$$\tilde{a} \leftarrow \text{clip}(\pi_{\phi'}(s') + \text{clip}(\epsilon, -c, c), a_{\text{low}}, a_{\text{high}}), \quad \epsilon \sim \mathcal{N}(0, \tilde{\sigma}) \tag{7}$$

Using this estimate of $\tilde{a}$, the target Q values from the double critic networks are computed using the smaller value of the two networks to prevent the maximization bias. Next, the loss function is computed for the two critic networks by computing the mean squared error between each critic and the target Q value. The networks are then optimized using backpropagation.

$$\begin{aligned} y &\leftarrow r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a}) \\ \theta_i &\leftarrow \text{argmin}_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s,a))^2 \end{aligned} \tag{8}$$

The actor policy is optimized periodically when $t \mod d = 1$, where d is the number of steps before an update. The mean of the Q values from the critic networks are used in the backpropagation of the actor networks:

$$\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s,a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s) \tag{9}$$

where $\nabla_\phi J(\phi)$ is the gradient of the expected return $J(\phi)$ following the target policy π_ϕ. Finally, the target networks are updated using a soft update as follows:

$$\begin{aligned} \theta'_i &\leftarrow \tau\theta_i + (1-\tau)\theta'_i \\ \phi' &\leftarrow \tau\phi + (1-\tau)\phi' \end{aligned} \tag{10}$$

where τ is the soft update coefficient selected to provide stable updates in the policy network.

B. Interaction energy modification via RL

The objective of the TD3 algorithm in our study was to control the effective HRI energy, $E_{\text{eff}}(t)$, in (2) to facilitate reaching the user's desired locomotion trajectory via iCPGs. In particular, we designed a reward function, $\mathcal{R}$, which the RL agent attempted to maximize. In this section, we will

first outline the reward function for the TD3 algorithm and then discuss how the proposed reward function can address challenges in different scenarios that the agent may face.

The state space for the RL agent is the HRI energy ($E(t)$) in Eq. (3) and its average ($E_{\text{avg}}(t)$); and frequency (ω) in Eq. (2) and its average (ω_{avg}) which represents RL estimation about user's desired frequency. The action space is the effective HRI energy ($E_{\text{eff}}(t)$) used in (2).

The following reward function was used to identify the optimal value of effective interaction energy for the user's desired walking pattern via an iCPGs:

$$\mathcal{R} = -\Big(\mathcal{K}_E[E(t) - E_{\text{avg}}(t)]^2 + \mathcal{K}_\omega[\omega(t) - \omega_{\text{avg}}(t)]^2 + \mathcal{K}_{\ddot{\omega}}[\ddot{\omega}(t)]^2 + \mathcal{R}_p + \mathcal{R}_E\Big) \tag{11}$$

where $\mathcal{K}_E$, $\mathcal{K}_\omega$, and $\mathcal{K}_{\ddot{\omega}}$ are constant values and $\mathcal{R}_P$ and $\mathcal{R}_E$ are defined as follows

$$\mathcal{R}_p = \begin{cases} \mathcal{P}_p, & \omega \notin [\omega_{\min}, \omega_{\max}] \text{ or } \rho \notin [\rho_{\min}, \rho_{\max}] \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$\mathcal{R}_E = \begin{cases} \mathcal{P}_E, & (E(t) - E(t - \tau_E) > \epsilon_E \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $\mathcal{P}_P > 0$ and $\mathcal{P}_E > 0$ are constant values and $\omega_{\max}$, $\omega_{\min}$, ρ_{max}, and ρ_{min} are pre-defined safety thresholds to provide safe locomotion patterns. $E(t)$ is the current HRI energy state, $E(t - \tau_E)$ is a delayed version of HRI energy and τ_E represents the amount of delay. A step is detected if the difference $E(t) - E(t - \tau_E)$ is greater than a threshold ϵ_E. Note that this threshold value will be determined by trial and error in real experiments with the exoskeleton and able-bodied user.

$\mathcal{R}_P$ is the safety term that encourages RL agent to avoid transitions to unsafe states. The terms of difference between actual energy and frequency with their average values ($E(t) - E_{\text{avg}}(t)$ and $\omega(t) - \omega_{\text{avg}}(t)$), and the acceleration of the frequency ($\ddot{\omega}(t)$), play an important role when the frequency is close to the user's desired value and system is almost in steady state. However, they can make the system less responsive by introducing lower E_{eff}, which has been resolved by adding the term $\mathcal{R}_E$, which penalizes based on the number of interactions that a user has applied.

IV. Results and Discussion

The hyperparameters for the TD3 algorithm were set experimentally, and they included a random seed of 10, starting exploration time steps of 64 on a random policy, standard deviation of 0.1 from a Gaussian distribution for exploration noise, batch size of 512, γ of 0.99, τ of 0.005, policy noise of 0.2 from a Gaussian distribution for critic updates, and policy update frequency of 2. The averages were calculated with a moving average with a window size of 10 s, and τ_E 0.05 s. The iCPG parameters were all set based on Sharifi et al. [10], except ψ_ω and ψ_ρ were set 0.0072 and 0.0096, respectively in simulations and $\psi_\omega = 0.0007$ and $\psi_\rho = 0.0009$ in experiments.

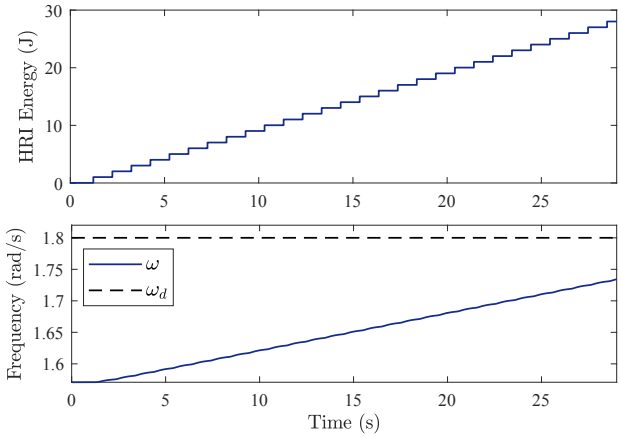


Fig. 2: Variation of HRI energy and frequency of walking for the weak muscle scenario without RL refinements.

A. Simulations

The simulation environment was created in MATLAB Simulink R2022a. The environment consisted of a frequency-dependent HRI energy input in (3), which increases or decreases in rectangular steps at fixed intervals. The desired frequency was manually set and hidden from the RL agent. If the current frequency was below the desired frequency, the HRI energy input increased until the desired frequency was obtained, and vice versa. The HRI energy input remained constant while the frequency is close enough to the desired frequency. The RL agent receives the current state from the environment and takes an action that modulates the HRI energy (E) to determine E_{eff}. Simulations were divided into a training and testing phase, where the RL agent modulated the HRI energy to meet the desired frequency. In both training and testing, the desired frequency was changed every 10 s. The training phase involved five episodes of training, with each episode lasting 30 s. Each episode consisted of 300 time steps corresponding to a sampling rate of 10 Hz. The testing phase consisted of 30 trials with the trained model, again for 30 s.

In real applications were the iCPG was initialized using experimental user data, and then the dynamics were updated when a new user interacted with the exoskeleton. Therefore there were two possible scenarios that need to be considered in the simulations. Firstly, there was the case of a new user with weaker muscles generating smaller interaction torques than the user with whom the initialization was performed. This is the main scenario we are trying to address in this research, as people with mobility impairments often have weaker muscles and are more easily fatigued by interacting with the exoskeleton than neurologically-intact individuals. The second case was when a new user has stronger muscles than the person for whom the exoskeleton was initialized.

1. Weak muscles (small stepwise τ_{HRI}): This scenario represents individuals with weak muscles who apply rectangular pulses of τ_{HRI} insufficient alone to reach their desired frequency. Note that the amplitude of walking is a function

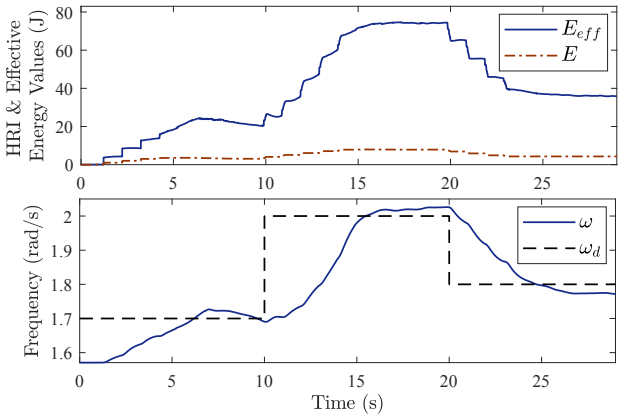


Fig. 3: Variation of HRI energy due to an RL agent selecting an effective energy to modify the frequency of walking for the weak muscle scenario.

of the gait frequency. A constant stepwise torque input is simulated using the rect function with a positive unity gain for frequencies below the desired frequency and a positive unity gain for those greater than the desired values. To aid the user, a penalty term, $\mathcal{R}_E$, is applied for jumps in E_{HRI} to incentivize the RL agent to choose actions which minimize the number of jumps. This penalty plays the most critical role in cases where the user's muscles are weak, so the RL agent amplifies the interaction energy to reach the user's desired walking speed faster. Also, the other elements of the reward function improve the agent's behaviour when it is close to the desired values.

The results for weak muscle scenario in the absence of RL modifications show that the user could not reach the desired frequency ($1.8\,\text{rad/s}$) after $30\,\text{s}$ (Fig. 2). However, the user reached the desired frequencies in less than five seconds by integrating an RL agent introduced in the effective energy term in the iCPG structure. As seen in Fig. 3, the RL agent amplified the user's interaction energy, $E(t)$ (brown dashed-dot line), and suggested higher values for the effective energy, $E_{\text{eff}}(t)$ (solid blue line). This amplification rate is lower when the user is close to the steady-state behaviour (7-10 s, 18-20 s, and 27-30 s). This is because of fewer jumps in this period (i.e., fewer jump penalties, $\mathcal{P}_E$), which forces the agent to pay more attention to the other elements of the reward function. Note that the desired frequencies for the training phase for the weak muscle case was $2\,\text{rad/s}$, $1.7\,\text{rad/s}$, and $2\,\text{rad/s}$ in this order. The testing phase had desired frequencies of $1.7\,\text{rad/s}$, $2\,\text{rad/s}$, and $1.8\,\text{rad/s}$, and acceptable frequency range of $\pm 0.05\,\text{rad/s}$ around the desired frequency. The control case had the same acceptable frequency range.

2. Strong muscles (large stepwise τ_{HRI}): This scenario represents users with strong muscles who apply rectangular pulses of τ_{HRI} to reach their desired frequency. The estimation of the desired values in RL were chosen as the average value over a constant time window. Our approach for the reward function was to minimize the sum of the

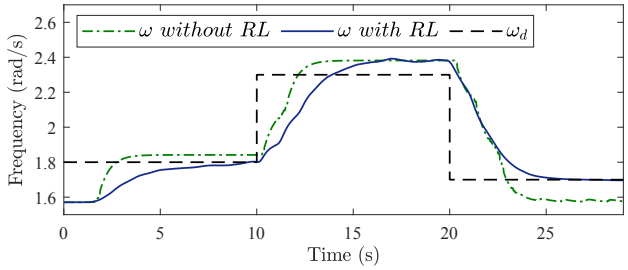


Fig. 4: Variations of gait frequency in the absence and presence of RL modification for the strong muscle scenario.

mean squared error between the actual and desired E_{HRI} and iCPG frequency, i.e., maximizing the reward function. For the strong muscle scenario, the desired frequencies for the training phase were $2.2\,\text{rad/s}$, $1.7\,\text{rad/s}$, and $2.2\,\text{rad/s}$ in this order. The testing phase had desired frequencies of $1.8\,\text{rad/s}$, $2.3\,\text{rad/s}$, and $1.7\,\text{rad/s}$. An acceptable frequency range of $\pm 0.15\,\text{rad/s}$ around the desired frequency was implemented to prevent oscillations about the desired frequency. The control case had the same acceptable frequency range. The results showed that the trained agent could facilitate reaching desired frequency values by adjusting the effective energy over time (see Fig. 4). As it can be seen in Fig. 4, the integral of error between the user's desired frequency and iCPGs output in the steady-state period (7-10 s, 18-20 s, and 27-30 s) was decreased by 65% for the case of using effective energy values which was determined via RL. Note that the desired frequency is hidden from the RL agent.

B. Experimental evaluations

The experimental set-up in Fig. 6(a) was used to evaluate the effectiveness of our proposed iCPG for lower-limb exoskeletons. A 29-year-old able-bodied user wore the Indego lower-limb exoskeleton (Parker Hannifin Corporation, Macedonia, OH). The user was asked to apply physical interactions to the exoskeleton joints to change the walking frequency to the user's desired values. The desired frequency was hidden from the RL agent, and the agent used the aver-

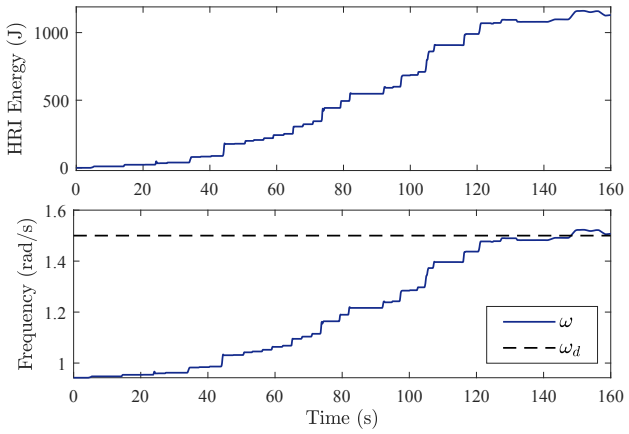
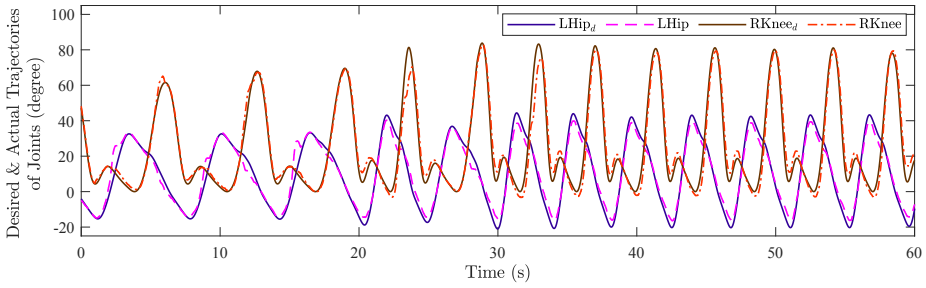


Fig. 5: HRI energy and walking frequency variations for a user interacting with a lower-limb exoskeleton in the absence of RL modifications.

(a)



(b)

Fig. 6: a) Experimental set-up: A 29-year-old neurologically-intact user wearing the Indego lower-limb exoskeleton, b) Desired and actual trajectories generated via iCPGs for the left hip and right knee joints.

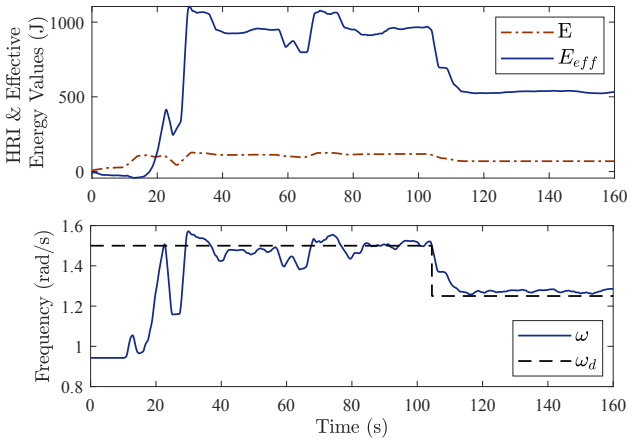


Fig. 7: HRI energy, effective energy, and walking frequency variations for a user interacting with a lower-limb exoskeleton in the presence of RL modifications.

age value of iCPGs frequency to estimate the user's desired frequencies. Three different experiments were performed. The first experiment was the control case, which used only ACPGs without an RL agent. The second experiment trained the RL agent. In this experiment, the user interacted with the exoskeleton for $150\,\mathrm{s}$ to reach different desired frequency values. In addition, the actor-critic networks in the TD3 algorithm (see Sec. III) were trained in this period and used in the final experiments for tuning the effective energy values. Note whole process for training RL can also be performed in our developed simulation environment for safety reasons.

As shown in Fig. 5, the user reached their desired frequency of $1.5\,\mathrm{rad/s}$ after about $150\,\mathrm{s}$ and increased the HRI energy level to more than $1000\,\mathrm{J}$ by applying continuous interaction energies. Continuous energy inputs were necessary because the ACPG initialization was performed with a different user with much stronger muscles. However, the RL agent and iCPGs resolved this issue by adjusting the HRI energy value and introducing effective energy in Fig. 7. As observed in Fig. 7, the maximum HRI energy applied by the user (brown dashed-dot line) was about $125\,\mathrm{J}$. However, the RL agent amplified that value (solid blue line) to about $1000\,\mathrm{J}$, which facilitated reaching the user's desired frequency. The results for the frequency showed that users could reach their desired frequencies on average in $10\,\mathrm{s}$ with iCPGs, only by modifying their effective energy. Furthermore, comparing the rate of amplification of HRI energy shows that the RL agent introduced a lower energy amplification rate for the period that the user tended to walk at a constant frequency, which provided a smoother walking experience for the user.

The amplitude, frequency, and phase values determined by iCPGs were translated to the desired trajectories of joints via Fourier series in (4). Fig. 6(b) shows the results for the desired and actual trajectories of joints for the first $60\,\mathrm{s}$ of walking. The RL & iCPGs-based generated desired trajectories have been commanded to a PD position control to be tracked. As depicted in Fig. 6(b), the maximum error between desired and actual trajectories was about 6° for the knee joint and 4° for the hip joint, which shows an appropriate tracking performance.

V. Conclusion

This study introduced iCPGs, which combined reinforcement learning with ACPGs to generate user-specific gait trajectories. The previously introduced ACPG algorithm could change gait trajectories in response to a user's physical interaction. However, the effectiveness of ACPGs was limited to precise parameter identification and a lack of considerable change in the interaction behaviour of users. The proposed iCPGs employed RL to learn a user's interaction behaviour in real-time and adjusted the HRI energy to facilitate reaching a user's desired gait pattern. The simulation results showed that the proposed RL agent could modify HRI energy and introduce an effective energy term to the iCPGs, removing the need for precise parameter identification and fixed interaction behaviour. Furthermore, the results provided evidence for the effectiveness of the proposed iCPGs in scenarios of having weaker or stronger muscles than the user that has been used for identifying the parameters. Finally, the experimental results showed that the method could be used for personalized motion planning of lower-limb exoskeletons.

REFERENCES

[1] S. A. Murray, R. J. Farris, M. Golfarb, C. Hartigan, C. Kandilakis, and D. Truex, "Fes coupled with a powered exoskeleton for cooperative muscle contribution in persons with paraplegia," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2788–2792.

[2] K. A. Inkol and J. McPhee, "Assessing control of fixed-support balance recovery in wearable lower-limb exoskeletons using multibody dynamic modelling," in *8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, 2020, pp. 54–60.

[3] G. Zeilig, H. Weingarden, M. Zwecker, I. Dudkiewicz, A. Bloch, and A. Esquenazi, "Safety and tolerance of the rewalk^{TM} exoskeleton suit for ambulation by people with complete spinal cord injury: a pilot study," *The Journal of Spinal Cord Medicine*, vol. 35, pp. 96–101, 2012.

[4] O. Jansen, D. Grasmuecke, R. C. Meindl, M. Tegenthoff, P. Schwenkreis, M. Sczesny-Kaiser, M. Wessling, T. A. Schildhauer, C. Fisahn, and M. Aach, "Hybrid assistive limb exoskeleton hal in the rehabilitation of chronic spinal cord injury: proof of concept; the results in 21 patients," *World neurosurgery*, vol. 110, pp. 73–78, 2018.

[5] R. W. Evans et al., "Robotic locomotor training leads to cardiovascular changes in individuals with incomplete spinal cord injury over a 24-week rehabilitation period: a randomized controlled pilot study," *Archives of Physical Medicine and Rehabilitation*, 2021.

[6] S. Qiu, W. Guo, D. Caldwell, and F. Chen, "Exoskeleton online learning and estimation of human walking intention based on dynamical movement primitives," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 1, pp. 67–79, 2020.

[7] S. Guo, Y. Ding, and J. Guo, "Control of a lower limb exoskeleton robot by upper limb semg signal," in *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2021, pp. 1113–1118.

[8] Y. He, X. Wu, Y. Ma, W. Cao, N. Li, J. Li, and W. Feng, "Gc-igtg: A rehabilitation gait trajectory generation algorithm for lower extremity exoskeleton," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 2031–2036.

[9] X. Wu, D.-X. Liu, M. Liu, C. Chen, and H. Guo, "Individualized gait pattern generation for sharing lower limb exoskeleton robot," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1459–1470, 2018.

[10] M. Sharifi, J. K. Mehr, V. K. Mushahwar, and M. Tavakoli, "Autonomous locomotion trajectory shaping and nonlinear control for lower limb exoskeletons," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 2, pp. 645–655, 2022.

[11] A. Sproewitz, R. Moeckel, J. Maye, and A. J. Ijspeert, "Learning to move in modular robots using central pattern generators and online optimization," *The International Journal of Robotics Research*, vol. 27, no. 3-4, pp. 423–443, 2008.

[12] M. Sharifi, J. K. Mehr, V. K. Mushahwar, and M. Tavakoli, "Adaptive cpg-based gait planning with learning-based torque estimation and control for exoskeletons," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8261–8268, 2021.

[13] A. J. Ijspeert, A. Crespi, D. Ryczko, and J.-M. Cabelguen, "From swimming to walking with a salamander robot driven by a spinal cord model," *science*, vol. 315, no. 5817, pp. 1416–1420, 2007.

[14] J. Fang, Y. Ren, and D. Zhang, "A robotic exoskeleton for lower limb rehabilitation controlled by central pattern generator," in *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*. IEEE, 2014, pp. 814–818.

[15] S. O. Schrade, Y. Nager, A. R. Wu, R. Gassert, and A. Ijspeert, "Bio-inspired control of joint torque and knee stiffness in a robotic lower limb exoskeleton using a central pattern generator," in *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2017, pp. 1387–1394.

[16] R. Luo, S. Sun, X. Zhao, Y. Zhang, and Y. Tang, "Adaptive cpg-based impedance control for assistive lower limb exoskeleton," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2018, pp. 685–690.

[17] K. Gui, H. Liu, and D. Zhang, "A generalized framework to achieve coordinated admittance control for multi-joint lower limb robotic exoskeleton," in *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2017, pp. 228–233.

[18] J. K. Mehr, M. Sharifi, V. K. Mushahwar, and M. Tavakoli, "Intelligent locomotion planning with enhanced postural stability for lower-limb exoskeletons," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7588–7595, 2021.

[19] Y. Shen, Y. Wang, Z. Zhao, C. Li, and M. Q.-H. Meng, "A modular lower limb exoskeleton system with rl based walking assistance control," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2021, pp. 1258–1263.

[20] R. Huang, H. Cheng, H. Guo, X. Lin, Q. Chen, and F. Sun, "Learning cooperative primitives with physical human-robot interaction for a human-powered lower exoskeleton," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 5355–5360.

[21] P. Zhang and J. Zhang, "Motion generation for walking exoskeleton robot using multiple dynamic movement primitives sequences combined with reinforcement learning," *Robotica*, pp. 1–16, 2022.

[22] Z. Li, H. Ma, Y. Ding, C. Wang, and Y. Jin, "Motion planning of six-dof arm robot based on improved ddpg algorithm," in *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 3954–3959.

[23] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, 2018, pp. 1582–1591.

[24] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.