Reinforcement Learning Control of Functional Electrical Stimulation of the upper limb: a feasibility study

D. Di Febbo[1], E. Ambrosini[1], M. Pirotta[2], E. Rojas[1], M. Restelli[3], A. Pedrocchi[1], S. Ferrante[1]

[1]NEuroengineering And Medical Robotics Laboratory, Politecnico di Milano, Italy
[2]SequeL Team, Inria Lille – Nord Europe, France
[3]Artificial Intelligence and Robotics Lab, Politecnico di Milano, Italy

davide.difebbo@mail.polimi.it

Abstract: *Controllers for Functional Electrical Stimulation are still not able to produce natural movements of the paretic arm. In this work, Reinforcement Learning was used to design a non-linear controller for a hybrid upper limb robotic system thought for stroke rehabilitation. The performance of the controller was tested on one healthy subject during elbow extensions in the horizontal plane. Experimental results showed an absolute position error <0.7° for a maximum range of motion of 40° and stability against perturbation induced by simulated muscle spasms. Promising results must be confirmed on a broader population.*

Keywords: *Functional Electrical Stimulation, Reinforcement Learning, Rehabilitation, Hybrid Robotic Systems.*

Introduction

Functional electrical stimulation (FES) is an effective technology used in rehabilitation to restore impaired motor functions in people affected by neurological disorders [1]. Following stroke, spinal cord injuries or multiple sclerosis, patients often have difficulties in performing functional movements with the upper limb as well as in grasping and manipulating objects. Rehabilitative treatments, based on FES, aim at restoring motor functions of the impaired upper extremity [2]. Recently, different robotics solutions have been proposed to be coupled with FES, the so-called "hybrid robotic systems", in order to facilitate the execution of motor exercises and increase the rehabilitative outcomes [3]. However, reliable controllers driving accurate and natural movements through FES are still under investigation. The electrically stimulated human muscle is a nonlinear system whose physiological properties are difficult to be modelled [4]. Moreover, it is a strongly time-variant system: spasticity and fatigue can significantly influence the performance in the short period, while muscle strengthening and motor relearning can improve performance in the medium/long period.
Classical control solutions [5] rely on the accuracy of the model by which the system is described. Due to the complexity of the electrically stimulated muscles response, linear assumptions are frequently made. More advanced techniques, such as non-linear [6] and adaptive control systems [7], have been tested in real environments. However, the increased complexity of the controllers implicates more onerous calibration procedures [8], not suitable for clinical settings and non-technically trained operators. Recently, reinforcement learning (RL) [9] has been investigated to solve upper limb FES control problems in simulation [10]. RL is a sub-field of machine learning which studies how agents can learn from experience collected by interacting with the environment. To achieve desired performances, RL algorithms can directly learn an optimal non-linear control law without prior knowledge about the system.
In this work, we used the Proximal Policy Optimization (PPO) RL algorithm [11] to control a FES-driven elbow extension movements supported by a passive exoskeleton. We identified a common initial position and we trained the controller to reach different target angles. The training was made off-line simulating the subject's arm dynamics with an artificial neural network (ANN) model. The performance of the controller was tested in a real environment, with one healthy subject who was asked to be completely passive. Moreover, we evaluated the stability of the control system in presence of simulated muscles spasms.

Methods

Apparatus: The robotic system for the upper limb [12] consisted of a lightweight passive exoskeleton for the right arm characterized by 3 degrees of freedom, each equipped with an angle sensor (Vert-X 13 E, ConTelec AG) to measure the angle position and an electromagnetic brake to lock a desired target position. The exoskeleton, exploiting the use of the brakes, was configured to allow only elbow flexion/extension movements in the horizontal plane, in the range of 60° (maximum flexion) to 170° (maximum extension). Trains of biphasic pulses were sent through surface patch electrodes (Pals® electrodes, Axelgaard Manufacturing Co., Ltd.) by means of a current-controlled stimulator/EMG recorder device (RehaMovePro, Hasomed GmbH). Only two channels were used, one connected to the biceps brachii and one to the triceps brachii. We controlled the stimulation of the two channels by modulating the pulse width, while the current amplitude and the stimulation frequency were fixed. The raw EMG signals recorded from the two stimulated muscles were filtered using an adaptive linear prediction filter in order to estimate the volitional EMG component [13]. The exoskeleton and the stimulator were controlled by the embedded processor BeagleBoneBlack ®™.

The RL problem statement: Considering a discrete time setting (where i is the discrete time instant), we formally described the episodic RL problem as a Markov decision

process (MDP) with the tuple $\langle S, A, P, R, \gamma, \mu \rangle$, where: S is a continuous set of states $\boldsymbol{s}_i = [\phi_i,\ \dot{\phi}_i,\ \ddot{\phi}_i,\ (\phi^t - \phi_i)]' \in \mathbb{R}^4$ composed of the elbow angular position, ϕ_i, the instantaneous angular velocity, $\dot{\phi}_i$, the instantaneous angular acceleration, $\ddot{\phi}_i$, and the difference between the target angle (ϕ^t) and the actual position; A is the continuous set of actions $\boldsymbol{a}_i = \left[pw_i^{ch1},\ pw_i^{ch2}\right] \in \mathbb{R}^2$ composed of the instantaneous values of pulse width (PW) in the two stimulation channels within the interval $[0, 400]\ \mu s$; P is the transition probability kernel such that $\mathrm{P}[\boldsymbol{s}_{i+1} \mid \boldsymbol{s}_i, a_i]$ represents the probability of reaching the state $\boldsymbol{s}_{i+1}$ from state $\boldsymbol{s}_i$, by performing action a_i, R is the reward function $r(\boldsymbol{s}, a) = E[r \mid \boldsymbol{s},\ a]$; $\gamma = 0.99$ is the discount rate and μ is the initial state distribution. Our scenario can be formalized as an episodic RL problem. An episode defines a finite time interval $I = 60, 0 \leqslant i \leqslant I$: at every time instant i, the agent is in the state $\boldsymbol{s}_i$, and makes an action a_i to reach the next state $\boldsymbol{s}_{i+1}$. At the same time, the environment produces a scalar reward r_i. The agent goal is to find a policy $\pi(a \mid \boldsymbol{s})$ that maximizes the sum of the collected discounted rewards, as described in Eq. 1:

$$J(\pi) = E_\pi[\, G_0 \mid \boldsymbol{s}_0 \sim \mu\,] \qquad (1)$$

where G_0 is the return at the time instant $i = 0$, as described in Eq. 2:

$$G_0 = \sum_{k=0}^{T} \gamma^k r_k \qquad (2)$$

The reward function defines the learning goal, i.e. it defines what the agent has to learn. Hence, we defined the reward function as in Eq. 3:

$$r_{i+1} = -(\phi^t - \phi_i)^2 \qquad (3)$$

The effect of such reward function was that to penalize the distance from the target in each time instant.

To solve the RL problem, we used the PPO policy gradient method, implemented in rllab [15] (available at https://github.com/rll/rllab). We used a Multi-Layer Perceptron (MLP), characterized by two hidden layers with 10 hyperbolic tangent (tanh) neurons and a tanh output layer, to estimate a Gaussian policy. In order to design a controller able to generalize the control strategy to reach different target angles starting from the same initial position, we defined a set of 3 target angles $\Phi^t = [130°, 145°, 160°]$, equally spaced in the elbow extension range of motion, and then we separated the total number of learning iterations (750) in batches of 50. At the beginning of each batch, the state was reset to $\boldsymbol{s}_0 = [\phi_0, 0, 0, (\phi^t - \phi_0)]'$, where ϕ_0 is the initial position, and the target angle was randomly selected from the set Φ^t. In every iteration, the algorithm simulated an episode in which the agent, starting from the initial state $\boldsymbol{s}_0$, has to reach the actual target angle, in I time instants. The policy is then updated using the data collected during a batch.

The ANN model of the subject arm: Our RL environment consisted of an ANN model of the electrically stimulated subject's arm. We chose a feedforward architecture and we trained it with data collected during a 20-minute acquisition session. The model estimated the state transition dynamics of the environment given the agent's action, as defined in Eq. 4:

$$\mathbf{s}_{n+1} = \mathrm{f}_{\mathbf{w}}\left(\underline{\mathbf{s}}_n, \mathrm{a}_n\right) \qquad (4)$$

where $\mathrm{f}_{\mathbf{w}}(\cdot)$ is the estimated dynamics of the environment, $\mathbf{w}$ is the vector of the neural network weights, $\underline{\mathbf{s}}_n = [\phi_n,\ \dot{\phi}_n,\ \ddot{\phi}_n,\ \mathrm{h}_n^{ch1},\ \mathrm{h}_n^{ch2}]'$ is the enlarged state, and n is a discrete time index. To increase the amount of input information, and also considering the whole stimulation sequence and muscle fatigue, we extended the state s_n with two additional signals, defined as in Eq. 5:

$$\mathrm{h}_n^{ch} = \begin{cases} \sum_{c=1}^{n} \mathrm{a}_c, & \text{if } \mathrm{a}_n \neq 0 \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

where $ch \subset \{1, 2\}$ indicates the stimulation channel.

We designed an ad-hoc experimental protocol to collect data for ANN training. During the acquisition, the subject wore the exoskeleton and elbow flex-extension movements in the horizontal plane were only allowed. The subject was asked to remain passive, while stimulation sequences were sent to both muscles. We used the volitional EMG signals, estimated from the raw EMG recordings, to monitor the capability of the subject to remain passive during data collection [13]. The current amplitude of the two stimulation channels were identified at comfortable values for the subject able to induce visible muscle contractions (8 and 10 mA for the biceps and the triceps, respectively) and the stimulation frequency was set at 25 Hz. The pulse width was modulated with a predefined sequence of ramps, ranging between 0 and 400 μs, with an inter-ramp interval of $5s$. The stimulation ramps were alternated between the two channels, but we also stimulated the two channels simultaneously to collect information related to the dynamics of co-contraction. The acquisition procedure lasted 20 minutes and a total number (N) of 35000 samples were collected. We used the data sampled at each time instant n to compute the extended state $\underline{\mathbf{s}}_n$. Then, we defined the inputs matrix X [N $\times$ 8], whose n^{th} row is $\mathrm{x}_n = \left[\underline{\mathbf{s}}_n',\ \mathrm{a}_n\right]$, and the targets matrix Y [N $\times$ 3], whose n^{th} row is $\mathrm{y}_n = \mathbf{s}_{n+1}'$, for training the ANN model.

We implemented and trained a single-layer feedforward ANN with 9 hidden tanh neurons with Keras (https:/keras.io). The mean squared error (MSE) was set as performance function and the Adam optimization algorithm was chosen [15].

Experimental protocol and performance measures: We identified the initial position ϕ_0, equal to 120°, as the position at which the subject was completely relaxed. The interval [120°, 170°] was considered as the range of motion for the elbow extension movement and the angles 130°, 140°, 150° and 160° were chosen as target positions for testing the controller. The trained control system was used in the real environment (one healthy male subject, 26 years old). The subject was asked to remain completely passive while executing 10 repetitions of four target elbow extensions movements, always starting from ϕ_0.

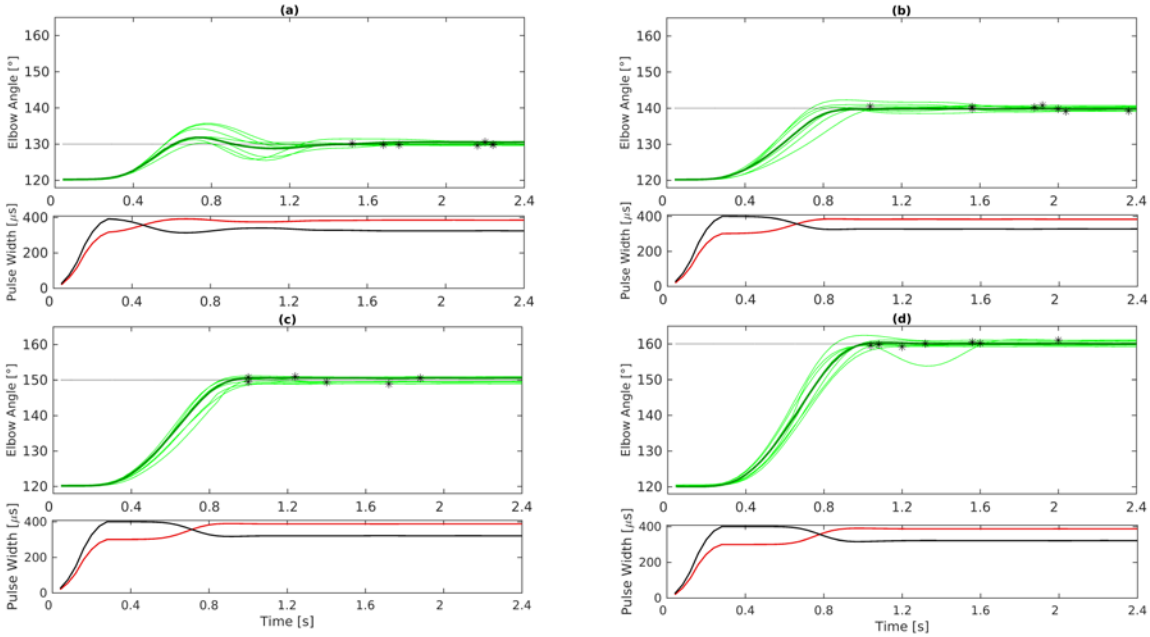

Figure 1: Controller performances in reaching 4 different targets of the elbow extension.

As for the data acquisition protocol (described in the previous section), the volitional EMG signal was monitored to check the capability of the subject to remain passive.

Each repetition was evaluated in terms of time needed to reach the target and stop the arm in that position, i^{set}, the absolute position error, e^{abs}, and the smoothness, sm as in Eq. 6:

$$sm = \frac{\dot{\phi}_{mean}}{\dot{\phi}_{max}} \qquad (6)$$

Where $\dot{\phi}_{mean}$ and $\dot{\phi}_{max}$ are the mean and the maximum instantaneous velocity within the repetition, from the beginning to the time i^{set}.

We also simulated muscle spasms during the executions of the elbow extension (from 120° to 140°), with the aim to verify the stability of the controller against intra-subject dynamical disturbances. Once reached the target position, we asked the subject to make a quick voluntary muscle contraction which moved the arm away from the target position. Then the subject had to relax again and let the control system bring back the arm to the target. We checked the elbow angle after $2s$ from the disturbance occurrence.

Results

We used the standardized data, collected from the subject, to train the ANN model of the arm dynamics and we achieved MSEs values of 0.001, 0.016 and 0.021 for the three outputs of the model, respectively. The model was then used as the simulated environment for the RL experiment. The optimal policy was obtained by running the PPO algorithm for 750 iterations, as described in the Methods section. Such number of epochs was enough to let the average returns (over a batch) converge to acceptable values of -944.28, -2471.21 and -14305.84, considering 130°, 145° and 160° respectively as the target angle to reach in the episodes simulated in that batch.

Tab. 1 reports the performance measures achieved by the RL control for the four targets. The low values of the absolute position error showed the capability of RL control to drive accurate movements. However, the amount of dispersion around the mean suggested a consistent variability in the performance of the controller. Overall, the four tasks achieved low values of smoothness, which were previously found in healthy subjects who performed similar movements with the help of a passive exoskeleton for weight relief [16].

Table 1: Performance measures of the RL control.

Target Angle	e_{abs} [°]	i_{set} [s]	sm
130°	0.29 ± 0.17	1.93 ± 0.22	0.12 ± 0.03
140°	0.47 ± 0.29	1.69 ± 0.31	0.25 ± 0.09
150°	0.67 ± 0.24	1.61 ± 0.37	0.30 ± 0.10
160°	0.53 ± 0.38	1.52 ± 0.34	0.31 ± 0.08

In Fig. 1, we reported the angular trajectories during the execution of the movements. Each panel represents the performance of the controller for a different target angle: 130° in (a), 140° in (b), 150° in (c) and 160° in (d). The upper panels show the superimposed repetitions to reach the same target, shown with a dashed line. The control actions, corresponding to the highlighted (dark green) single repetition, are displayed in the lower panels (red line for the biceps and black line for the triceps). The black asterisks indicate the setting time of each movement.

Fig. 2 shows an experiment in which a muscles spasm was simulated by the subject's volitional activation. The angular trajectory is plotted in panel (a), where the target angle (140°) is indicated with a dashed line. Panel (b) displays the volitional EMG signals of the subject (red for the biceps and black for the triceps) during the experiment. At the onset of the stimulation, the subject's arm is in the initial position (120°) and the volitional EMG signals of both muscles are lower than 20 μV, which can be considered as the subject's relax threshold. After 1.7s, the subject made a

quick muscle contraction which moved the arm away from the target. The volitional activation is clearly visible in the rapid and consistent increase of the EMG signals. Following the induced disturbance, the subject relaxed his muscles and the arm was driven again by the controller in the target position, with a final absolute position error equal to 0.2°.

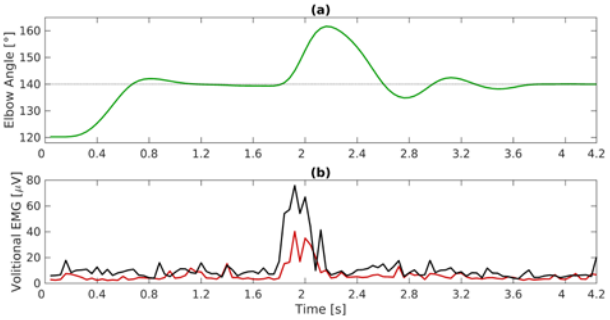

Figure 2: Robustness of the controller to "simulated" muscle spasms (at about 2s).

Discussion and conclusions

Our results showed the ability of reinforcement learning to drive very accurate flexion extension angles in a real environment combining an exoskeleton for weight relief with upper limb FES. The control system revealed good stability and disturbance rejection properties in reaching different target angles within an overall range of motion of 40° on a single healthy subject.

The control actions (pulse width of the biceps and triceps muscles) were modulated to rapidly reach the angle target. Then, to hold the position, high values of pulse width, resulting a massive co-contraction, were used. This control strategy cannot be considered efficient form a physiological point of view, since the overstimulation can accelerate the muscle fatiguing process. The obtained overstimulation might depend on the reward function choice. Indeed, we penalized the distance of the arm from the target in each time step. Therefore, the goal of the agent was to find the optimal policy which drove the arm to the target position as fast as possible, also increasing the stiffness to avoid large overshoots. To modify this behaviour, additional penalty terms for the speed or the stimulation integral should be investigated. Exoskeleton brakes can also be exploited to avoid co-contractions to keep the target position once achieved. Based on the good accuracy achieved by the RL controller in reaching different target angles, methods based on machine learning seem to be promising in overcoming the typical problems of more traditional control solutions. Further investigations on both healthy and impaired subjects need to be done to ascertain its performance and its future suitability in clinics.

To improve the system, the generalization properties of the policy could be enhanced by adding more initial conditions and on-line learning can be exploited to let the controller follow the physiologic time-varying dynamics.

Acknowledgement

The project was funded by the EU project RETRAINER (GA no. 644721). We thank Axelgaard Manufacturing Ltd. for donating us the stimulation electrodes.

References

[1] J. H. Burridge, M. Ladoucer, Clinical and therapeutic applications of neuromuscular stimulation: a review of current use and speculation into future developments. Neuromodulation, 2001. 4:147-154.

[2] French, L. H. Thomas, J. Coupe, N. E. Mcmahon, L. Connell, J. Harrison, C. J. Sutton, S. Tishkovskaya, and C. L. Watkins, "Repetitive task training for improving functional ability after stroke," Cochrane Database Syst. Rev., vol. 2016, no. 11, pp. 102–104.

[3] Klauer, C., Schauer, T., Reichenfelser et Al. Feedback control of arm movements using Neuro-Muscular Electrical Stimulation (NMES) combined with a lockable, passive exoskeleton for gravity compensation (2014) Frontiers in Neuroscience, 8 (SEP), art. no. 262,

[4] Bolsterlee B, Veeger DH, Chadwick EK. Clinical applications of musculoskeletal modelling for the shoulder and upper limb. Med Biol Eng Comput. 2013;51(9):953:63.

[5] Vette AH, Masani K, Kim JY, Popovic MR. Closed-loop control of functional electrical stimulation-assisted arm-free standing in individuals with spinal cord injury: a feasibility study. Neuromodulation. 2009;12(1):22-32.

[6] Kirsch N., Alibeji N., Sharma N., Nonlinear model predictive control of functional electrical stimulation. Control Engineering Practice. Vol 58, Jan 2017, pp. 319-331.

[7] Abbas JJ, Triolo RJ. Experimental evaluation of an adaptive feedforward controller for use in functional neuromuscular stimulation systems. IEEE Trans Rehabil Eng. 1997;5(1):12:22.

[8] Lynch C. L., Popovic M. R. Functional Electrical Stimulation. ClosedLoop Control of Induced Muscle Contractions. IEEE Control Systems Magazine, April 2008.

[9] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA. 1998.

[10] Thomas P, Branicky M,1 van den Bogert A, Jagodnik K. Application of the Actor-Critic Architecture to Functional Electrical Stimulation Control of a Human Arm. Proc Innov Appl Artif Intell Conf. 2009;165:172.

[11] Schulman J., Wolski F., Dhariwal P., Radford, Klimov O., Proximal Policy Optimization Algorithms. 2017. Submitted 20 Jul 2017 (v1), last revised 28 Aug 2017. OpenAI.com

[12] Ambrosini E, Ferrante S, Zajc J. The combined action of the passive exoskeleton and EMG-controlled neuroprosthesis for upper limb stroke rehabilitation: First results of the RETRAINER project. IEEE Int Conf Rehabil Robot. 2017 Jul;2017:56-61.

[13] Ambrosini E., Ferrante S., Schauer T. et al. A myocontrolled neuroprosthesis integrated with a passive exoskeleton to support upper limb activities. Journal of Electromyography and Kinesiology 24 (2014) 307–317.

[14] Duan Y., Chen X., Houthooft R., Schulman J., Abbeel P., Benchmarking Deep Reinforcement Learning for Continuous Control. Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

[15] Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization. Submitted 22 Dec 2014 (v1), last revised 30 Jan 2017.

[16] Ambrosini E, Ferrante S, Rossini M, Molteni F, Gföhler M, Reichenfelser W, Duschau-Wicke A, Ferrigno G, Pedrocchi A. Functional and usability assessment of a robotic exoskeleton arm to support activities of daily life. *Robotica*. 2014; 32(08):1213-1224.