Using learning to control artificial avatars in human motor coordination tasks

Maria Lombardi, Member, IEEE, Davide Liuzza, and Mario di Bernardo, Fellow, IEEE

Abstract—Designing artificial cyber-agents able to interact with human safely, smartly and in a natural way is a current open problem in control. Solving such an issue will allow the design of cyber-agents capable of co-operatively interacting with people in order to fulfill common joint tasks in a multitude of different applications. This is particularly relevant in the context of healthcare applications. Indeed, the use has been proposed of artificial agents interacting and coordinating their movements with those of a patient suffering from social or motor disorders. Specifically, it has been shown that an artificial agent exhibiting certain kinematic properties could provide innovative and efficient rehabilitation strategies for these patients. Moreover, it has also been shown that the level of motor coordination is enhanced if these kinematic properties are similar to those of the individual it is interacting with. In this paper we discuss, first, a new method based on Markov Chains to confer “human motor characteristics” on a virtual agent, so as that it can coordinate its motion with that of a target individual while exhibiting specific kinematic properties. Then, we embed such synthetic model in a control architecture based on reinforcement learning to synthesize a cyber-agent able to mimic the behaviour of a specific human performing a joint motor task with one or more individuals.

Index Terms—Artificial avatar, human-robot interaction, mirror game, Markov models, movement coordination, non-linear control, reinforcement learning, virtual player.

I. INTRODUCTION

The number of new tasks that involve people coordinating their movement with machines, avatars and robots is expected to experience a fast growth [1]–[3]. Investigating how to enable such artificial agents to co-operatively interact with humans is still an open problem. To achieve this ambitious goal, it is important to study how humans coordinate their motion with each other and then develop control strategies able to drive artificial agents in motor coordination tasks with them. It has been observed that pairs or groups of humans performing a joint task often tend to intentionally or unintentionally synchronize their movement. Examples include studies on people rocking chairs [4], hands clapping [5], team rowing during a race [6], and synchronization of respiratory rhythms [7]. Furthermore in [8], it has been suggested that the motion of each individual exhibits unique kinematic features that can be summarized through a time-invariant individual motor signature (IMS) captured by the distribution of the velocity profile exhibited during the motion (see Sec. II-A for further details), and, more importantly, that when two individuals share similar IMS then their movement coordination is enhanced. Also, it was noted that the IMS of patients suffering from schizophrenia and other social disorders differs significantly from that of healthy individuals providing a new diagnostic tool (biomarker) for this type of disorders [9].

A further conclusion reached in [10], [11], is that rehabilitation of these patients can be performed by making them interact with an artificial avatar endowed with an IMS initially similar to theirs that is then gradually morphed into an healthy human one. In this way, the patient gets used over successive trials to coordinate his/her motion better and better with that of the cyber-agent before starting trials with other humans [12]. This strongly motivated the need for the design of appropriate control frameworks to drive artificial avatars able to coordinate their motion with that of a human while exhibiting a desired IMS (kinematic feature).

As a paradigmatic scenario, the so-called mirror game was selected as a suitable task to study interpersonal human coordination. Introduced in 2011 in [13], it involves two people imitating each other’s hand movements creating spontaneous motion (see Sec. II-A for details). A cognitive architecture to drive a virtual player (VP) in the mirror game was proposed in [14], [15] where optimal control was used to solve a multi-objective control problem aimed at (1) tracking the motion of the end-effector of the human player (HP) (or generating motion to lead the human player); (2) exhibiting a desired IMS in the generated avatar motion. As shown in detail in Sec. II-B the solution proposed therein relies on a deterministic controller solving an optimal control problem on a receding horizon. The cost function is selected so as to minimize the mismatch between the positions of the end-effectors of the virtual player and the human player, while at the same time minimizing the distance between the motion of the VP and a pre-recorded human movement trajectory (providing a reference IMS).

The key disadvantages of this approach are the deterministic nature of the controller often leading to unnatural tracking behaviour and the need of pre-recorded human trajectories. Also, the dual nature of the control objective requires fine tuning of the control parameters in the cost function in order
to achieve an acceptable balance between the two tracking errors.

The aim of this paper is to overcome both of these problems towards a more flexible approach by using learning methods. Specifically,

i. a new data-driven stochastic model based on Markov chains is proposed to generate IMS removing the need of any pre-recorded signal;

ii. a reinforcement learning control algorithm is designed to synthesize a cyber-player (CP) able to play the mirror game while exhibiting the IMS of a reference human player overcoming the limitations of previous control solutions.

In this way, entirely autonomous cyber-agents are obtained who can play the game either between themselves or with other human players. To validate the proposed approach we use experiments together with numerical simulations and a real-time experimental set-up where humans were asked to play with the cyber-agent showing the effectiveness of the suggested control solution.

We wish to emphasize that the problem addressed in this paper is an instance of the larger class of human-in-the-loop problems which are currently the subject of much ongoing research, as for example in human-machine interaction and human-robot interaction research, as for example in social robotics with their bodies. Usually played by musicians, dancers and actors, the mirror game has become a powerful parody to study the complex phenomenon of human motor coordination [13]. In its simplest form, it involves two people mirroring each other’s movements by oscillating two handles horizontally in a mono-dimensional space (see Fig. 1).

The rest of the paper is organized as follows. After giving the background in Sec. II we derive a data-driven model in Sec. III able to endow the VP with human kinematic features and use it as part of a control architecture that allows VP to play dyadic sessions of the mirror game with a human player as shown in Sec. IV-A. In Sec. IV-B we synthesize a cyber-player through techniques of artificial intelligence. All these methods are validated in Sec. V showing their effectiveness before conclusions are drawn in Sec. VI.

II. BACKGROUND

Next we briefly summarize some fundamental aspects related to the IMS, the mirror game, previous approaches to control a virtual agent in the mirror game, and the reinforcement learning.

A. Mirror game and IMS

The mirror game is a serious game in which two people imitate each other’s movements creating fascinating choreographies with their bodies. Usually played by musicians, dancers and actors, the mirror game has become a powerful paradigm to study the complex phenomenon of human motor coordination [13]. In its simplest form, it involves two people mirroring each other’s movements by oscillating two handles horizontally in a mono-dimensional space (see Fig. 1).

It can be played in three different conditions:

1) Leader - Follower (LF): in this condition one player designated as follower tries to imitate as better as s/he can the trajectory of the other player designated as leader;

2) Joint - Improvising (JI): in which two players play together and coordinating their movements without explicit designation of roles;

3) Solo Condition (SC): in which a player is asked to generate interesting motion in isolation without interacting with each other. This third condition turns to be useful for extracting their individual motor signatures.

![Fig. 1. Illustration representing two subjects playing a dyadic session of mirror game by moving their own spherical handles along a string (adapted from [11]).](image)

In [8] the concept of IMS was introduced as a unique kinematic feature distinguishing the motion of different players in the mirror game. The IMS was defined as the time-invariant probability density function (PDF) of velocity trajectories exhibited by an individual while playing the game and it was shown in [9] to be a valid biomarker for social disorders such as schizophrenia.

B. Cognitive architecture

A cognitive architecture was proposed in [14], [15] to drive a virtual agent, named as Virtual Player (VP), in playing the mirror game with a human player while exhibiting some reference IMS. The architecture is shown in Fig. 2 and is mainly composed of two parts:

- an inner dynamics model: representing how the VP moves in the absence of any interaction with the HP. This was modelled using a nonlinear Haken-Kelso-Bunz oscillator (HKB) [14]–[16];

- a control strategy: that generates the movement of the VP in response to that of the human player while exhibiting the reference IMS (velocity distribution) provided by the “signature generator” block. In [14], [15], [17] an optimal controller was proposed in order to minimize the difference in position between the HP and the VP end-effector. At the same time the cost function adopted therein took into account the error in velocity between the VP motion and the reference signature signal which was pre-recorded from previous human trials.

More specifically, in [14], [15] the motion of the virtual agent is modelled as a controlled nonlinear HKB oscillator of the form:

\[
\ddot{x} + (\alpha x^2 + \beta \dot{x}^2 - \gamma) \dot{x} + \omega^2 x = u,
\]

where \(x, \dot{x}\) and \(\ddot{x}\) are the position, velocity and acceleration of its end-effector, \(u\) is the control input, \(\alpha, \beta, \gamma\) are parameters characterizing the nonlinear damping term while \(\omega\) is the...
natural oscillation frequency of the generated motion when \( u \) is set to zero.

According to [14], [15], the control input \( u \), is chosen as a function of the movement of the HP in order to minimize the following cost function

\[
\min_u J (t_k) = \frac{1}{2} \theta_p (x(t_{k+1}) - r_p(t_{k+1}))^2 + \frac{1}{2} \int_{t_k}^{t_{k+1}} (1 - \theta_p) (\dot{x} (\tau) - \dot{r}_p (\tau))^2 + \eta \dot{u} (\tau)^2 d\tau, \quad (2)
\]

where \( r_p \) is the position time series measured from the HP end-effector, \( \dot{r}_p \) is the reference signal corresponding to the desired motor signature of the VP, \( \eta \) is a positive weight assigned to the minimization of the control energy, \( [t_k, t_{k+1}] \) is the minimization interval. The constant parameter \( \theta_p \in [0, 1] \) is used to determine how much the VP weighs the motion of the human player so that if \( \theta_p = 0 \) the VP acts as a leader (completely ignoring the motion of the other player) while if \( \theta_p = 1 \) as a perfect follower. Any value of \( \theta_p \in [0, 1] \) will make the VP motion a compromise between tracking the HP motion and the reference velocity of the target IMS the VP has to exhibit, allowing to implement different types of leader/follower behaviour.

The proposed control architecture requires two different reference signals, one is the position signal recorded by the HP, the other is a velocity profile that represents the desired IMS the VP has to exhibit. The main drawback is that the IMS is collected off-line, recording several human players performing sessions of the mirror game in solo condition. The use of a pre-registered signature makes the motion of the VP less natural and less variable since the reference is always the same for each session of the game.

In what follows we will take a different approach and use Markov chains to obtain data driven models of HPs performing solo sessions of the mirror game. These models will be integrated in the control scheme of Fig.2 as stochastic signature generators so as to have more flexible VPs. This approach will also be useful to generate synthetic data with different aims, such as the one of complementing human data (if missing) to create an avatar of a desired player, via a reinforcement learning scheme.

### C. Reinforcement Learning

Reinforcement learning (RL) is a machine learning technique in which an agent learns how to behave in an external environment through a trial-and-error approach and looking at its successes and failures [18], [19]. RL techniques are well suited when agents are formally modelled as a Markov Decision Process (MDP), composed by the quadruple \((X, U, f, \rho)\) where \( X \) is the set of all possible states in which the environment can be, \( U \) is the set of all possible actions that the agent can take (also termed as action-space), \( f : X \times U \times X \rightarrow [0, 1] \) is the state transition probability function and \( \rho : X \times U \times X \rightarrow \mathbb{R} \) is the reward function (see [20] for more details on MDP). The learning process is made up of the following sequential steps:

1. the agent looks at the environment (process and measurements of interest) and takes an action \( u \) in the set of all possible actions \( U \), which causes the environment to transit into a new state;
2. the agent observes the new state of the environment \( x \in X \) following its action;
3. the agent receives a scalar reward \( r \) that measures how good taking that action in the previous state has been;
4. according to the reward, the agent changes its policy \( \pi : X \rightarrow U \) that maps each environment state \( x \) to an action \( u \);
5. the agent aims at maximizing the sum of all rewards obtained along all the interactions.

Solving a problem of RL mathematically is equivalent to solve a system of \( N_x \times N_u \) non linear equations, where \( N_u \) is the number of all possible actions and \( N_x \) is the number of all possible states. Due to the complexity of such a problem, in particular for huge state space dimensions, several iterative approaches can be used, for which it has been proved that the policy followed by the agent converges to the optimal policy (see [19] for further details). In this work, we use a Temporal Difference Learning approach, in particular the Q-learning algorithm since it does not require the model of the environment and it operates completely online estimating the reward and the state value.

### III. Modelling

Next Markov chains are used to model human movement in solo condition and design an IMS generator. A Markov Chain (MC) is a well-known stochastic model useful to describe randomly changing systems with a finite number of states [21]. It consists of a finite set of possible states in which the system can be, a set of transitions between any two states with their corresponding probability, and a set of possible observations as outputs.

Denoting by \( s_k = i \) with \( i \in [1, \ldots, N] \) that the system is in state \( i \) at time \( k \), a Markov chain is fully characterized by:

- an initial state \( s_0 = \pi \);
• a transition matrix $A := [a_{ij}]$ where $a_{ij} := P(s_{k+1} = j | s_k = i)$ is the probability of being in state $j$ at time $k + 1$ having been in state $i$ at time $k$.

The procedure to construct a MC-based model consists of the following three steps:

1) **Data collection and preprocessing**: input movement data recorded from a human player performing the mirror game in solo condition are preprocessed through short-time Fourier Transform (STFT) and Vector Quantization (VQ) [22].

2) **Markov model training**: the preprocessed data are used to define the coefficients of the transition matrix $A$ of the Markov chain.

3) **Synthetic data movement generation**: the resulting Markov chain is used to generate new synthetic movement data sharing the same IMS as that of the input data.

Next we describe in greater detail how each of these steps was performed.

**A. Step 1: Data collection and preprocessing**

The sampled input signal is first partitioned in a discrete set of frames using a Hamming window [23] of a certain width (see Fig. 3). To prevent loss of information and in order to minimize the distortion of the signal, two consecutive windows are overlapped of $\frac{1}{2}$ of the window’s width. In so doing the sum of the sequence of windows is a resulting flat-top window [23].

Each windowed data segment undergoes a Short Time Fourier Transform (STFT) so that a vector of Fourier transform’s coefficients, or “feature vector”, can be associated to it.

At the end of the process a finite set of feature vectors is obtained that is processed through a vector quantizer in order to get a finite set of symbols. The vector quantizer maps each feature vector to one of the $N$ prototype vectors contained in its code-book (for further details see [18]). In this work, the code-book is generated by means of the Lloyds algorithm [18]. The latter is an iterative algorithm that continuously partitions the Euclidean space into $N$ convex cells (called also Voronoi cells) in order to have each input vector as close as possible to the centroid of one of the cells. The indexes of the array built with the $N$ prototype vectors (integers from 0 to $N - 1$), are used as symbols in the output of the discrete MC.

**B. Step 2: Markov model training**

Once the input data has been transformed into a finite set of symbols, the transition probability between one symbol and any other can be evaluated by looking at the symbol frequency in the data string. This allows to identify the coefficients of the transition matrix $A$ and so to build a MC model in which each state corresponds to a code-book symbol (the model has as many states as symbols). In this paper, the transition coefficients are evaluated over an extensive set of experimental data through the Baum-Welch algorithm [21], which is essentially based on a frequential approach, and finds the maximum likelihood estimate of the coefficients given a set of observed feature vectors.

**C. Step 3: Synthetic data movement generation**

Thanks to its stochastic nature, the MC built in the previous step is able to generate random sequences of symbols according to the probabilities included in the transition matrix $A$. In order to have a position signal over time, the sequence of symbols needs to be reconverted to a continuous signal through a reverse post-processing. First of all, the generated sequence of symbols is de-quantized with the same code-book used in the forward pre-processing in order that each symbol is mapped back onto a prototype feature vector. Then, the inverse STFT is performed of each feature vector and concatenated using the overlap-add (OLA) method [24]. The main advantage of this method is the possibility of reconstructing a smooth position time series, removing the discontinuities obtained by simply concatenating two random symbols.

**IV. Control synthesis**

**A. MC in the loop**

As a first step, we embed the MC model developed above to replace the “signature generator” block in the control schematic depicted in Fig. 2 and described in Sec. [1, 3] By using the MC model trained on data acquired from a target HP, the virtual player driven by the MC endowed control architecture will play the mirror game with a human player while exhibiting the IMS of the target HP the MC was modelled upon. The quality of the tracking and of the IMS exhibited by the virtual agent will still depend upon the many parameters of the optimal control approach described in Sec. [1] and [15] that require careful off-line tuning and numerous
Fig. 4. Architecture used to train a CP to mimic a given specific player. Two HPs involved in a dyadic session of mirror game are directly provided to a reinforcement learning scheme or simulated by two VPs. The latter are designed with the proposed control architecture embedding the MC which captures the individual behaviour of the HP to be replaced. While the two HPs or VPs are playing together, a CP driven by a reinforcement learning algorithm learns how to interact.

trials-and-errors. To overcome this problem we take next a radically different approach based on the use of reinforcement learning techniques which will lead to a fully data-driven control algorithm.

B. A reinforcement learning approach

The goal of the reinforcement learning approach is to develop a virtual agent (from here on referred to as a Cyber-Player (CP)) to distinguish it from the VP already described in previous literature) able to play sessions of the mirror game as a follower in a LF condition while exhibiting the IMS of a target human player (a similar approach could be used to make the CP act as a leader but it is not reported here for the sake of brevity). Note that the aim is for the CP to emulate the way in which a specific human player would interact with the leader in a mirror game session with all of his/her “human imperfections” rather than achieving perfect tracking of the leader position.

To achieve this goal, the reinforcement learning algorithm needs to be provided with data from leader-follower sessions of the mirror game so that the CP could learn to mimic the behaviour of the follower player in these sessions. Specifically, particularizing the learning process described in Sec. II-C to our specific case, we consider as system’s state \( x := [x, \dot{x}, x_l, \dot{x}_l] \), where \( x, \dot{x} \) are position and velocity of the CP, while \( x_l, \dot{x}_l \) are position and velocity of the leader player. The reward function as \( \rho := -(x - x_f)^2 - 0.1(\dot{x} - \dot{x}_f)^2 - \eta u^2 \) where \( x_f, \dot{x}_f \) are position and velocity of the follower player (while playing with the leader), \( \eta \) is a positive weight for the minimization of the control energy \( u \). The choice of such reward function is motivated to the fact that we want the CP to mimic the follower behaviour, thus behaving as him/her synthetic avatar. To maximize this function, the action-space consists of a set of acceleration values that can be imparted to the end-effector of the cyberplayer. From simulations, we found that 9 different actions (negative and positive values of the acceleration) represent a good compromise between the quality of the resulting motion and the learning time.

The data used to train the CP should be the position and velocity time series extracted from two real human players playing the mirror game. As this might require a large enough dataset, we propose here to use synthetic data obtained by making two VPs play against each other, each driven by the control architecture shown in Fig. 4 with the signature generator block being a different Markov chain trained as described in Sec. IV-A

To implement reinforcement learning we use the Q-learning algorithm. As anticipated in Sec. II-C this is an iterative approach, in which the CP adapts its behaviour according to the measures it receives from the two players (leader and follower) and trying to find the best action that it can perform for each state to emulate the follower. As a first step we define a matrix \( Q := [q_{ij}] \) where the states are listed on the rows and the actions on the columns. Each element \( q_{ij} \) is the “value” given to the corresponding state-action pair, also termed q-value. At the beginning of the learning process, the matrix \( Q \) is initialized with random values. Then, each iteration is structured as follows (see Fig. 5):

- the CP observes the state \( x_k \) (the subscript \( k \) denotes the sampled value of the state at time instant \( k \)) and the follower state needed to evaluate the reward function;
- the CP chooses an action \( u_k \) at a time instant \( k \) according to a policy rule \( \pi \). In this way we use an \( \epsilon \)-greedy policy
Specifically, the CP takes the best known action, i.e., the action with the highest q-value (exploitation) with $(1 - \epsilon)$ probability, whereas with $\epsilon$ probability it takes a random action (exploration). The value $\epsilon$ follows a monotonic decreasing function, since as time increases the exploration phase is replaced by the exploitation phase;

- the CP evolves in a new state at time $k + 1$ and observes the state $x_{k+1}$ and the reward $r_{k+1} = \rho(x_k, u_k, x_{k+1})$ following the action $u_k$ taken at the previous time $k$. For the sake of brevity, we will simply denote the obtained reward as $r_{k+1}$, omitting the dependence from the state and the actions;

- according to the reward received, the CP updates the value of the entry of the matrix $Q$ corresponding to the pair $(x_k, u_k)$ following the rule:

$$q_{k+1}(x_k, u_k) = q_k(x_k, u_k) + \alpha \left[ r_{k+1} + \gamma \max_{u_{k+1} \in U} q_k(x_{k+1}, u_{k+1}) + q_k(x_k, u_k) \right],$$

where $\alpha$ is the learning rate and $\gamma$ is a discount factor;

- a new iteration is performed until convergence is achieved.

V. Validation

To validate our methodology, we carried out experiments with the following features:

- Participants: a total of 6 people took part in the experiments (4 females and 2 males). All participants were right-handed and none of them had physical or mental disabilities. All of them participated voluntarily, signing an informed consent form in accordance with the Declaration of Helsinki. The players have been numbered from 1 to 6.

- Experimental set-up: the set-up is the one based on the use of a leap-motion controller (position sensor) presented in [25] and depicted in Fig. 6.

- Experimental task: each participant was asked to carry out 30 different trials of the mirror game in solo condition. Each trial was 30 seconds long and performed through Chronos, a software tool recently developed by some of the authors to study movement coordination and described in [25]. The given instruction was to oscillate the index of the preferred hand in a spontaneous way from left to right.
A. Validation of the MC modelling approach

To validate the MC model synthesis process presented in Sec. III, we computed the following quantities:

- **Probability Density Function (PDF)** of the VP motion velocity which is indicative of the exhibited IMS.
- **Earth Mover’s Distance (EMD)**: to evaluate how close the PDF of the velocity signal generated by the MC matches that of the human player on which it was trained [26]. In the case of univariate probability distribution, the EMD is given by the area of the difference between their Cumulative Distribution Functions (CDF). Formally we have

\[
EMD(PDF_{HP}(z), PDF_{VP}(z)) = \int_{\mathbb{R}} |CDF_{HP}(z) - CDF_{VP}(z)| dz. \quad (4)
\]

- **Similarity space**: through multidimensional scaling (MDS) [11], [26], it is possible to represent the player’s velocity profiles as points in an abstract geometric space, called “similarity space”. Points corresponding to different trials of the same players can be encircled by an ellipse defining a “characteristic region” for each player. The MDS is a technique that allows to reduce the dimensionality of the data, preserving as much information as possible (see [26] for more details). The similarity space was highlighted as a valuable tool to analyse IMS in movement data in [11]. Specifically, the Euclidean distance between two points in the similarity space is a good approximation of the EMD between the corresponding PDFs; the closer the points are, the more similar are the corresponding IMS they associated with.

- **Kurtosis** and **skewness**: to show that the trajectories generated by the MC have the same features of a human movement, as described in [13]. They are the \(3^{rd}\) and the \(4^{th}\) moment of a velocity curve respectively. In particular, given a generic smooth curve \(f(t)\) with mean \(\mu\) and standard deviation \(\sigma\) defined on the interval \(T = [t_{1}, t_{2}]\), its skewness and kurtosis are defined respectively as

\[
s = \frac{1}{\sigma^3} \int_{t_{1}}^{t_{2}} (t - \mu)^3 f(t) \, dt, \quad (5)
\]

\[
\kappa = \frac{1}{\sigma^4} \int_{t_{1}}^{t_{2}} (t - \mu)^4 f(t) \, dt. \quad (6)
\]

Roughly speaking, skewness indicates the asymmetry in acceleration and deceleration, whereas kurtosis provides information about the uniformity of the maximal velocity. Low kurtosis means that an object is quickly accelerating and decelerating and keeps constant the velocity in between, conversely high kurtosis means that the object is accelerating and decelerating slowly keeping the maximal velocity for a short time [8].

In our experiments, each position time series, also termed as a “trial”, was captured at a sampling rate of 10Hz. It was then interpolated to 100 Hz and windowed with a 60 samples long Hamming window with an overlapping of 45 samples. The vector quantizer was chosen to have 256 levels (each level corresponds to a symbol). At the end six MC models of 256 states each were derived, one model per player. To assess the quality of the signatures generated by the MC, a total of 30 new motion signals were generated and compared with those of the corresponding human players. For a better graphic visualization, the PDFs, the skewness and the kurtosis are shown for only one player out of the six involved while data for all six players was used to obtain the similarity space pictures described below.

Fig. 7 shows that the velocity PDF (IMS) of the motion signals generated by the MC well approximates that of the human player it was trained upon.

Fig. 8 shows that the velocity PDF (IMS) of the motion signals generated by the MC well approximates that of the human player it was trained upon.

![Velocity PDF](image-url)

**Fig. 7.** Velocity PDF of the signatures recorded by the human player (in dark blue) and velocity PDF of the signatures generated by the MC trained on the same human player (dash-dotted line in light blue).

![Skewness - Kurtosis](image-url)

**Fig. 8.** Skewness-kurtosis plane having the skewness on x-axis and the kurtosis on y-axis. Two different plots are reported for negative (a) and positive (b) segments. The velocity segments belonging to the human players are represented as points in dark blue, whereas the generated ones as diamonds in light blue. A dotted line shows the bound for the values of \(s\) and \(\kappa\) given by the theoretical relationship \(\kappa \leq s^2 + 1\) [27].

To evaluate the skewness and the kurtosis of the generated signals, the velocity time series were divided into segments; each corresponding to one direction of motion (left-to-right or...
right-to-left). Each segment was rescaled to a common support ([0, 1]) and normalised so that its underneath area is unitary. Velocity segments with less than 20 samples were ignored. The skewness and kurtosis of each segment is represented as a point in Fig. 8 both for the MC generated data and that measured from the HP. We observe that both are located in the same area of the skewness-kurtosis plane.

Fig. 9 shows the mapping of the IMS of the MC models and the corresponding human players in the similarity space. The overlapping between the ellipses corresponding to human player trials with those generated by the Markov chain confirms the effectiveness of the modelling approach and the ability of the MC to generate human-like synthetic movement signals.

B. Validation of the MC in the loop approach

To evaluate the performance of the virtual player presented in Sec. IV-A when it is engaged in a dyadic session of the mirror game, the following metrics are used:

1) **Root mean square error** (RMSE) between the position time series of the two players given by

\[\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_{pk} - x_k)^2}, \tag{7}\]

where \(N\) is the number of samples in the simulation, \(r_{pk}\) and \(x_k\) are the position values at the \(k\)th time instant of the HP and the VP respectively.

2) **Relative phase** (RP) defined as the difference between the phases of the players: \(\Delta \Phi = \Phi_H - \Phi_V\). The phase was estimated from the data as done in [28].

3) **Circular variance** (CV): to quantify the coordination level between the two players and computed as follows [29]

\[CV = \left\| \frac{1}{N} \sum_{k=1}^{N} e^{i \Delta \Phi_k} \right\| \in [0, 1], \tag{8}\]

where \(N\) is the number of collected samples, \(\Delta \Phi_k\) is the relative phase between the two players at a time instant \(k\) and \(\| \cdot \|\) denote the 2-norm. The higher is the CV, the more the players are coordinated.

The VP was validated both in a follower and in a leader condition playing a session of mirror game with a human player. For the sake of brevity, we report here only the case when the MC model, used for the control, is that belonging to player 2. Similarly, any other MC model can be used. The parameters of the proposed control architecture are tuned heuristically as follows: \(\alpha = 1, \beta = 2, \gamma = -1, \eta = 10^{-4}\) and with \(T = 0.03s\). For the VP to act as a leader, \(\theta_p = 0.8\), whereas for it to act as a follower, \(\theta_p = 0.9\) and \(\omega = 0.1\).

When the VP acts as a follower [panels (a) and (c) in Fig. 10], the CV between the two players is 0.933 which indicates a high level of coordination between them. Coherently, the RMS of the position error is only 0.112. According to the definition of phase leadership in [30], [31], the positive phase in panel (a) confirms that the VP is acting as a follower while the HP as a leader (\(\Delta \Phi = 0.394 \pm 0.408\)). When the VP acts as a leader [panels (b) and (d) in Fig. 10], the CV reaches 0.868 and the RMS of the position error is 0.122. In this case the phase is negative, indicating that the VP is effectively acting as leader (\(\Delta \Phi = -0.664 \pm 0.574\)).
As depicted in Fig. 10(c) and Fig. 10(d), the velocity PDF of the VP changes according to its role. When the VP is leader it tends to be more similar to the PDF generated by the MC model ($EMD(Ref, VP) = 0.006; EMD(HP, VP) = 0.018$), whereas when it plays as a follower it is closer to that of the HP ($EMD(Ref, VP) = 0.0263; EMD(HP, VP) = 0.011$).

C. Validation of the learning approach

Next, we validate the approach presented in Sec. IV-B. As mentioned therein, synthetic data is used for training which is obtained by running the mirror game between two virtual players controlled using the MC embedded strategy discussed earlier.

Specifically, a VP endowed with the MC model of player 1 was set as the virtual leader (VL) and a VP endowed with MC model of player 5 as the virtual follower (VF). For both players the control parameters were selected as before.

Training was performed in real-time while the two VPs played against each other for approximately 7 hours (approximately 2000 trials). Once training was completed, the performance of the cyberplayer was validated by comparing 20 game sessions where the virtual leader played against the virtual follower used for the training, with 20 game sessions where the same virtual leader played against the cyberplayer. The results are summarized in Fig. 12. The characteristic regions of the VF (in red) and of CP (in blue) are shown while playing with the virtual leader (whose characteristic region is depicted in black). It is possible to see that the region of the CP mostly overlaps with that of the VF, meaning that it is able to emulate its signature. Quantitatively, we registered an $EMD(Leader, VP) = 0.0024$, an $EMD(Leader, CP) = 0.0033$, and a $RMSE(Leader, VP) = 0.104 \pm 0.006$ and $RMSE(Leader, CP) = 0.1088 \pm 0.015$ which confirms the tracking ability of the CP.

So far we have been able to show that the CP is able to play the game with the VL which was also used for its training. To further validate the effectiveness of the RL approach we trained the CP using an alternative strategy. Namely we performed the training using different VLs based on MC models of players 1, 2, 3 and 4 with a VF based on the MC model of player 5. The training took 27 hours (approximately 8000 trials). All control parameters were selected as before. We tested the CP performance by comparing its behaviour against that of the VF with both following the same VL. The results are summarized in Fig. 11 where the outcomes are shown of the CP and VF playing against a VL used in the training (player 2) [panel (a)] and a VL which was not used...
in the training set (player 6) [panel (b)]. In both cases the CP successfully tracks the leader showing that it is able to follow also leaders whose data was not included in the training set.

Further evidence is provided in Fig. 13 where the circular variance and the RMS of the position error between the CP and VLs modelled on players 1, 2, 3, 4 and 6 are given and compared to those obtained when the VF modelled on player 5 is used (similar performance was observed when a different set of players was chosen as VF).

**VI. Conclusion**

We addressed the problem of designing control architectures to construct artificial agents able to engage with humans in motor coordination tasks. The mirror game was chosen as a paradigmatic example where two individuals have to perform a joint oscillatory task. To overcome the limitations of previous approaches, we proposed the use of Markov chains to model human behaviour in the game and remove the need of pre-recorded human trajectories in previous control approaches in the literature. We showed that Markov chain models can be obtained that reproduce the unique kinematic features (IMS) that distinguish the motion of different people. We then embedded such Markov chain models in a control architecture that achieves the goal of making a virtual player able to coordinate its motion with that of a human player in a completely autonomous manner. Still the problem remained of having to fine tune the parameters of the control algorithm which was based on solving an optimal control problem on a receding horizon. To further improve the control performance,
we introduced a reinforcement learning approach to the problem using data generated by virtual agents playing against each other to perform the training stage. The resulting cyberplayer was shown both numerically and experimentally to be able to coordinate its motion with that of a human player while exhibiting the IMS of a target individual.

We wish to emphasize that the proposed cyberplayer can be an invaluable tool to be used in the human dynamic clamp setting proposed in [32] as a method to study social interaction and movement coordination among humans. It can also be effectively used for the implementation of innovative rehabilitation strategies for patients suffering from social disorders as highlighted in [10], [11]. For example, an avatar directly trained to mimic the kinematic features of the patient motion may result beneficial in the initial stages of the therapy, allowing an easier interaction with the patient himself/herself and gaining his/her trust, while simultaneously performing online diagnosis [9], [12], [33], [34].

ACKNOWLEDGEMENT

The authors wish to thank Dr. Piotr Słowiński at University of Exeter (UK) for his support and for providing the data used to test the classification of patients. We also wish to thank all people who took part in the experiments and Mr. Alberto Ingenito who carried out some of the simulations for used to test the classification of patients. We also wish to thank The Royal Society Interface for his support and for providing the data online diagnosis [9], [12], [33], [34].

REFERENCES

[1] T. Iqbal, S. Rack, and L. D. Rieck, “Movement Coordination in Human-Robot Teams: A Dynamical Systems Approach,” *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 909–919, 2016.

[2] C. G. Atkeson, J. G. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaal, T. Shibata, G. Tsvatia, A. Ude, S. Vijayakumar, E. Kawato, and M. Kawato, “Using humanoid robots to study human behavior,” *IEEE Intelligent Systems and their Applications*, vol. 15, no. 4, pp. 46–56, 2000.

[3] A. Shukla and A. Billard, “Coupled dynamical system based arm-hand grasping model for learning fast adaptation strategies,” *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 424–440, 2012.

[4] M. J. Richardson, K. L. Marsh, R. W. Isenhower, J. R. Goodman, and R. C. Schmidt, “Rocking together: Dynamics of intentional and unintentional interpersonal coordination,” *Human Movement Science*, vol. 26, no. 6, pp. 867–891, 2007.

[5] Z. Nédé, E. Ravaux, Y. Brecqet, T. Vicese, and A. L. Barabási, “The sound of many hands clapping,” *Nature*, vol. 403, no. 6772, pp. 849–850, 2000.

[6] A. M. Wing and C. Woodburn, “The coordination and consistency of rowers in a racing eight,” *Journal of Sports Sciences*, vol. 13, no. 3, pp. 187–197, 1995.

[7] E. Codrons, N. F. Bernardi, M. Vandoni, and L. Bernardi, *PLoS ONE*, no. 9, p. e107538.

[8] P. Słowiński, C. Zhai, F. Alderisio, R. Salesse, M. Gueugnon, L. Marin, B. G. Bardy, M. di Bernardo, and K. Tsaneva-Atanasova, “Unravelling socio-motor biomarkers in schizophrenia,” *npj Schizophrenia*, vol. 3, no. 1, p. 8, 2017.

[9] (2013) Alterego: Enhancing socio-motor interactions using information technology. [Online]. Available: http://www.euromov.eu/alterego/homepage

[10] L. Noy, E. Dekel, and U. Alon, “The mirror game as a paradigm for studying the dynamics of two people improvising motion together,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20947–20952, 2011.

[11] C. Zhai, F. Alderisio, P. Słowiński, K. Tsaneva-Atanasova, and M. di Bernardo, “Design of a Virtual Player for Joint Improvisation with Humans in the Mirror Game,” *PLoS ONE*, vol. 11, no. 4, p. e0154361, 2016.

[12] ——, “Design and Validation of a Virtual Player for Studying Interpersonal Coordination in the Mirror Game,” *IEEE Transactions on Cybernetics*, vol. 48, pp. 1018–1029, 2017.

[13] H. Haken, J. A. S. Kelso, and H. Bunz, “A Theoretical Model of Phase Transitions in Human Hand Movements,” *Biol. Cybern.*, vol. 51, pp. 347–356, 1985.

[14] C. Zhai, F. Alderisio, K. Tsaneva-Atanasova, and M. di Bernardo, “A model predictive approach to control the motion of a virtual player in the mirror game,” in *54th IEEE Annual Conference on Decision and Control (CDC)*, 2015, pp. 3175–3180.

[15] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall, 2003.

[16] R. S. Sutton, A. G. Barto et al., *Reinforcement learning: An introduction*. MIT press, 1998.

[17] C. C. White III and D. J. White, “Markov decision processes,” *European Journal of Operational Research*, vol. 39, no. 1, pp. 1–16, 1989.

[18] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[19] N. Ahmed and K. R. Rao, *Orthogonal transforms for digital signal processing*. Springer Science & Business Media, 2012.

[20] G. Heinzel, A. Rüdiger, R. Schilling, and T. Hannover, “Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top,” Tech. Rep., 2002. [Online]. Available: http://hdl.handle.net/11858/00-001M-0000-0013-557A-5

[21] L. Rabiner and B. Gold, *Theory and application of digital signal processing*. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975.

[22] F. Alderisio, M. Lombardi, G. Fiore, and M. di Bernardo, “A novel computer-based set-up to study movement coordination in human ensembles,” *Frontiers in Psychology*, vol. 8, p. 967, 2017.

[23] I. Borg, P. J. Groenen, and P. Mair, *Applied multidimensional scaling*. Springer Science & Business Media, 2012.

[24] J. Shohat, “Inequalities for moments of frequency functions and for various statistical constants,” *Biometrika*, vol. 21, no. 1/4, pp. 361–375, 1929.

[25] B. Kralemann, L. Cimponeriu, M. Rosenblum, A. Pikovsky, and R. Mrowka, “Phase dynamics of coupled oscillators reconstructed from data,” *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 77, no. 6, pp. 1–16, 2008.

[26] T. Krezu, F. Mormann, R. G. Andrzejak, A. Kraskov, K. Lehnertz, and P. Grassberger, “Measuring synchronization in coupled model systems: A comparison of different approaches,” *Physica D: Nonlinear Phenomena*, vol. 225, no. 1, pp. 29–42, 2007.

[27] F. Dorfler, M. Chertkov, and F. Bullo, “Synchronization in complex networks and smart grids,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 6, pp. 2005–2010, 2013.

[28] C. Zhai, F. Alderisio, P. Slowinski, K. Tsaneva-Atanasova, and M. di Bernardo, “Dynamic clamp as a paradigm for social interaction,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 35, pp. E3726–E3734, 2014.

[29] M. Pirro, C. Zhai, F. Alderisio, and M. di Bernardo, “Exergaming and rehabilitation: A methodology for the design of effective and safe therapeutic exergames,” *Entertainment Computing*, vol. 14, pp. 55–65, 2016.

[30] M. Lombardi, D. Liuza, and M. di Bernardo, in *2018 European Control Conference (ECC)*.
Maria Lombardi (M'18) received her M.Sc. degree in Computer Engineering in 2016 from the University of Naples “Federico” (Italy) and she is currently a Ph.D student in Engineering Mathematics at University of Bristol (UK) under the supervision of Prof. Mario di Bernardo.

Her current research interests include human-robot interaction, design of artificial agents interacting with humans through machine learning as well as modeling coordination and leadership both in human groups and in groups mixed with virtual agents.

She is a member of Alumni and Friends association at University of Bristol and of IEEE Circuits and System Society.

Davide Liuzza received the PhD. degree in Automation Engineering from the University of Naples “Federico II” (Italy).

He was a visiting Ph.D student in the Department of Applied Mathematics at University of Bristol (UK) in 2012, and at the ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm (Sweden) from 2012 to 2013. From 2013 to 2015 he was a post-doctoral researcher at Automatic Control Laboratory at KTH, currently he is a post doc at the department of Engineering at University of Sannio in Italy.

His research interests are in networked control systems, coordination of multi-agent systems, incremental stability of nonlinear systems, as well as nonlinear control and hybrid systems.

Mario di Bernardo (SM’06–F’12) is Professor of Automatic Control at the University of Naples “Federico II”, Italy and Professor of Nonlinear Systems and Control at the University of Bristol, U.K.

On 28th February 2007 he was bestowed the title of Cavaliere of the Order of Merit of the Italian Republic for scientific merits from the President of Italy. He was elevated to the grade of Fellow of the IEEE in January 2012 for his contributions to the analysis, control and applications of nonlinear systems and complex networks. He was elected to the BoG of the IEEE Circuits and System Society in 2006 and then again in 2009. He is currently VP Financial Activities of IEEE CASS and President of the Italian Society for Chaos and Complexity.

His research interests include the analysis, synchronization and control of complex network systems; piecewise-smooth dynamical systems; nonlinear dynamics and nonlinear control with applications to engineering and computational biology.

He authored or co-authored more than 220 international scientific publications including more than 110 papers in scientific journals, a research monograph and two edited books. He is Associate Editor of the IEEE Transactions on Control of Network Systems; Nonlinear Analysis: Hybrid Systems; CEBs of the IEEE Control System Society and the European Control Association (EUCA). From 1st January 2014 he is Deputy Editor-in-Chief of the IEEE Transactions on Circuits and Systems I. He is regularly invited as Plenary Speakers in Italy and abroad.

He has been organizer and co-organizer of several scientific initiatives and events and received funding from several funding agencies and industry including the European Union, the UK research councils the Italian Ministry of Research and University.