Autonomous development of turn-taking behaviors in agent populations: a computational study

Clément Moulin-Frier*, Marti Sanchez-Fibla* and Paul F.M.J. Verschure*†
*SPECS, IUA, Universitat Pompeu Fabra, Carrer de Roc Boronat 138, E-08018 Barcelona, Spain
Email: clement.moulinfrier@gmail.com; marti.sanchez@upf.edu
†ICREA Institució o Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, E-08010 Barcelona
Email: paul.verschure@upf.edu

***Abstract*—We provide an original computational model showing how turn-taking behaviors can self-organize out of sensorimotor interactions between vocalizing agents. These agents are equipped with a cognitive architecture based on two coupled control loops: a reactive one implementing a basic regulatory behavior to maintain vocal listening and an adaptive one learning an action policy to maximize an overall group presence estimation. We show that the reactive process allows to bootstrap the adaptive learning to converge toward a collective turn-taking strategy. This model provides a computational support to the hypothesis that turn-taking can emerge from functional constraints related to group cohesion and vocal signal interferences and suggests future directions of research to understand how social behaviors can result from sensorimotor interactions.**

I. Introduction

A fundamental characteristic of human communication is the ability to take turns during vocal exchanges, resulting in a minimal overlap in time of their vocalizations. A particular advantage of this so-called turn-taking behavior is that it prevents signal interferences to allow the communication content to be properly decoded by the message receiver.

It has been recently shown that this rather complex social behavior is not human-specific but is shared with another primate species, the common marmoset monkeys (*Callithrix jacchus*), which also exhibits cooperative vocal communication by taking turns [1]. This suggests that rather basic cognitive abilities could be sufficient to manage this collective behavior. A simple neural model has been proposed which is indeed able to reproduce some statistical properties of the observed behavior in marmosets [2] and which suggests a crucial role of self-monitoring in this process [3] (both papers [2] and [3] have been published at ICDL/Epirob).

From our analysis, these studies reveal the existence of both a reactive and an adaptive action-perception loops in turn-taking behaviors. First, the analysis of time interval distributions in marmoset vocal exchanges shows that the newborn marmoset vocalizations are not timed independently for each individual but instead according to the perception of both self and other's calls [3]. We interpret this result as an evidence that a reflexive behavior, influenced both by self and other vocalizations, plays a role in vocal interaction. Second, the distribution of time intervals between infant marmoset vocalizations in response to their own call vs in response to other's call are similar, showing that infant marmoset does not distinguish between self vs other calls, i.e. they do not self-monitor. Third, it has been shown that these time interval distributions (self vs other responses) are different in adult marmosets, indicating that, contrarily to the infants, adults possess self-monitoring abilities allowing the distinction of self vs other's vocalizations. The fact that the time interval distributions are different for infants and adults shows that self-monitoring is acquired during infant development, therefore involving a learned adaptive component to control the timing of vocalizations with respect to other individuals (this will be one of the central aspects of the model presented in this paper).

Previous models of turn-taking behavior in vocal agent populations mainly focus on the minimal neural architecture allowing the reproduction of some dynamical properties observed in animal behavior [2], [3]. They show that turn-taking can result from coupled neural architectures which are rather minimal (three neurons) and act as coupled oscillators. However, how learning occurs in an agent population to manage such behavior has not been modeled computationally to our knowledge, although interesting hypotheses about this issue have been proposed in these previous models.

This paper attempts at providing a computational model addressing the following questions. (1) What is the role of reactive vs adaptive control loops in turn-taking behavior? (2) How does learning occur from agent interaction and under what functional constraints? This model is not in divergence of previous ones but is rather taking a different perspective, both to reframe the problem at the functional cognitive level and to address computationally the learning issue. Regarding question (1), we adopt the Distributed Adaptive Control framework (DAC, [4], [5]). DAC proposes that cognition is organized in a number of hierarchical layers of increasing complexities: from reactive (reflex behavior pre-wired from evolution), to adaptive (involving prediction through sensorimotor association learning), to contextual (involving planning and memory). The reactive layer implements a set of homeostatic and allostatic controllers aiming at providing reflex motor responses to incoming stimuli in order to maintain the organism within a comfort zone (related e.g. to feeding, breathing, safety etc. . .). The adaptive layer develops on top of the reactive one to acquire a state space of the agent-environment interaction and to shape action through learning mechanisms, allowing e.g. the anticipation of action to improve the reactive control.

The contextual aspects that relate to the turn-taking and thus the contextual layer of DAC are outside the scope of this paper. Regarding question (2), we propose a multi-agent simulation paradigm inspired by the so called *language games* [6], [7], [8], where a global communication strategy self-organizes out of local interaction between sensorimotor agents through incremental and coupled learning mechanisms. To our knowledge, this is the first time that this paradigm is applied to the modeling of turn-taking behaviors.

In the following Section II, we state our working hypotheses about the functional constraints driving turn-taking behavior emergence from an evolutionary perspective. Then, in Section III, we describe a novel computational model implementing a population of sensorimotor agents equipped both with a reactive and an adaptive control loops and interacting together through vocal production and mutual perception. Section IV analyses simulation results to show how turn-taking behaviors self-organize out of agent vocal interactions under the functional constraints proposed in Section II. Finally, we conclude this study by addressing the two questions asked above with respect to the model results, proposing further extensions of the model to overcome its current limitations as well as applying it to more complex setups.

II. Functional constraints driving turn-taking behavior emergence

Taking inspiration from [1], we make the hypothesis that turn-taking behaviors emerge in agent populations needing to maintain group cohesion for a survival purpose (e.g. because they are not adapted to survive in isolation) and living in a dense environment preventing visual contact, e.g. a dense forest. In this context, a way to maintain group cohesion is through the use of vocalizations to convey information about the presence of each member. We make the assumption that each individual uses vocal identification to ensure that each other member of the group is around. This rather strong hypothesis, which requires that each agent is able to identify a conspecific from the acoustic properties of its calls, is plausible for several social species (see e.g. [9] for the case of adult marmosets). However, such identification abilities can hardly be conceived as pre-wired from infancy and have instead to be tuned during development from inter-agent visual and vocal interactions.

Visual and auditory channels display strong differences in their ability to convey information. In particular, vision is less subject to interferences: whereas one can easily identify individuals from the image of a group, it is generally harder to perform the same task from a mixed acoustic signal merging each individual voice. This is due to stronger interferences between individual signals in the latter case. Therefore, if several group members are vocalizing at the same time, the vocalization sounds will interfere making agent identification harder (i.e. to answer the question: who vocalized?).

We predict that these constraints, vocal group cohesion through vocal identification on the one hand and minimal interferences between calls on the other hand, are sufficient to allow turn-taking behaviors to emerge in an agent population. The model described in the following section attempts at providing a computational support to this prediction.

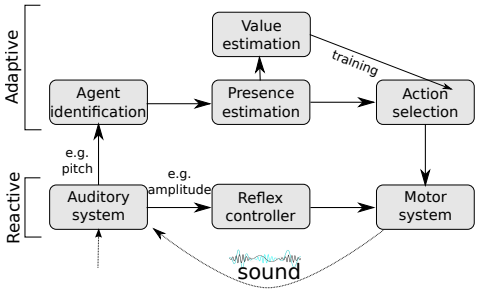

Fig. 1. The agent cognitive architecture is composed of a reactive and an adaptive layers. The auditory system decomposes the sound signals produced by the agents in two features (which can be related to the pitch and the amplitude of the sound signal). The first one indicates whether at least one agent vocalized at time t. It can be conceived as the amplitude of the signal. The second one conveys information about the agent's identity (it could be the pitch of the signal for example, but our model is agnostic in this respect). The control flow is then distributed in both layers. The reactive layer implements a simple reflex controller activating the motor system whenever no vocalization has been heard since a few time steps (described in Section III-B). The adaptive layer relies on agent identification, an ability supposed to be acquired later in the agent development, to maintain an estimation that each member of the population is present. The action selection system learns how to modulate the motor system to maximize the overall group presence estimation through value estimation (described in Section III-C). The motor system sums up the output from both layers, triggering vocalization or not according to the obtained value.

III. Model

A. General architecture

We consider a population of N vocalizing agents. Each agent a_j of the population ($j \in \{1, \ldots, N\}$) implements the cognitive architecture described on Figure 1. The source code related to this model is available open-source at https://github.com/clement-moulin-frier/turntaking_model.

The global simulation loop is the following. At each time step t (time is discrete in the current version of the model), the output value of each agent's motor system indicates if it is vocalizing or not. These vocalizations, possibly overlapping, are in turn received as input by their auditory systems. Each agent receives the exact same input, i.e. we do not consider signal transformations induced by the environmental signal transmission. The auditory system decomposes the signal into two features which are selectively processed in both layers to drive each agent next action at time $t + 1$.

B. Reactive layer

In this section we consider the reactive layer of Figure 1 in isolation. This layer is responsible for agent low-level sensorimotor processing and control and is supposed to be pre-wired (typically from evolutionary processes in a biological perspective). The auditory input at each time step drives a reflex controller which aims at maintaining the agent in a comfort zone with respect to its assumed need to listen to vocalizations of the group members.

An auditory input can either be self-generated, produced by a single agent or the result of overlapping vocalizations produced by several agents at the same time. The reflex controller is driven by the time T since the agent has heard the last vocalization. Its comfort zone is defined with respect to a threshold θ_t below which the agent remains silent ($T < \theta_t$, with $\theta = 4$ in all the simulations we will present). Whenever the threshold is reached, i.e. $T \geq \theta_t$, the agent is out of its comfort zone and activates its motor system, aiming at returning to a comfortable state. The motor output is considered to be probabilistic. It receives an activation value $a \in \mathbb{R}$ from the reflex controller and converts it into a probability of vocalizing $p \in [0, 1]$ through a sigmoid activation function

$$p(a) = \frac{1 + tanh(a)}{2}. \quad (1)$$

The reactive layer is limited to provide two possibles motor activation values:

- Whenever the agent goes out of the comfort zone ($T \geq \theta_t$), the regulatory mechanism of the reflex controller outputs a motor activation $a = 0$, resulting in a probability of vocalizing $p(0) = 0.5$.
- Whenever the agent is in the comfort zone ($T < \theta_t$), the reflex controller outputs a strong motor inhibition $a = -30$, resulting in a quasi-null probability of vocalizing $p(-30) = \epsilon$.

Figure 2 displays a simulation with two interacting agents embedding the reactive layer (the adaptive one being deactivated). Each time the agents go out of the comfort zone, i.e. each time none of them vocalized during the θ_t last time steps, they produce a vocalization with probability 0.5. If at least one of them vocalizes, both return to the comfort zone and stay silent for at least the θ_t subsequent time steps. These basic reactive interactions provide the agents with a primitive level of entrainment where each one acts on the behavior of the other through the mutual auditory perceptions driving their respective reflex controllers. However, we observe that this reactive behavior does not prevent overlapping between the agent vocalizations. The next section describes the processes occurring in the adaptive layer which will allow the agent to converge to a collective turn-taking strategy.

C. Adaptive layer

In this section, we present the control flow of the adaptive layer, which operates on top of the reactive one. The adaptive layer takes as input the extracted features from the auditory signal which allow agent identification and outputs a motor response that is summed up with the output of the reactive layer (see Figure 1). Therefore, the activation provided to the motor system is $a = a_R + a_A$, where a_R and a_A are the motor activations provided by the reactive and the adaptive layers, respectively. This activation is then converted into a probability of vocalizing using the activation function defined in Equation 1.

The adaptive layer is then composed of a number of functional subsystems that we describe below, together with

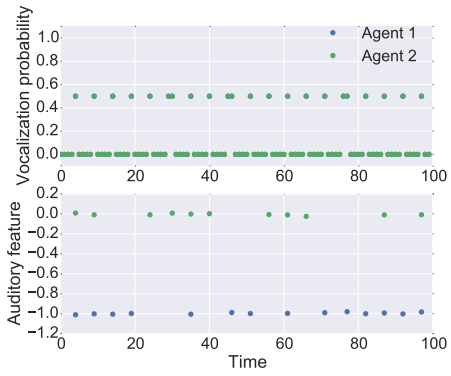

Fig. 2. The reactive layer allows to maintain basic vocal contact but often results in overlapping vocalizations. Here a population of two agents is considered. Top: probability of vocalizing induced by the output motor activation from the reflex controller through Equation 1. This probability is either null if at least one agent vocalized in the $\theta_t = 4$ last time steps and 0.5 otherwise. Because the reflex controller of each agent has the same dynamics and both agents receive exactly the same auditory input, their motor activations are exactly the same (hence only one is visible in the plot). Bottom: vocalizations produced by each agent. The y-axis corresponds to the auditory feature allowing to distinguish each agent, which is used here only for the sake of visualization in order to display which agent is vocalizing at each time step (low value: Agent 1; high value: Agent 2).

the mechanisms allowing to learn an adaptive action policy to maximize vocal contact.

1) Agent identification: The agent identification system makes use of a certain feature extracted from the auditory signal allowing the agents to identify each other. Our model is agnostic about the nature of this feature: it could be for example the pitch of the vocal signals or any other feature allowing to properly distinguish agent vocalizations. Here we simply consider that each population member vocalizes on a certain range allowing the others to identify it from its specific call. This identification ability is supposed to be mature later in the agent development and learned from visuo-auditory inter-agent interactions during infancy. However we do not model this particular learning dynamics here. We rather consider that before being able to identify others, agents behave as described in the previous section, i.e. are only driven by the reactive layer (Figure 2). Then, once agent identification is acquired, it allows the adaptive layer to modulate the activity of the motor system (see Figure 3).

Computationally, a vocalization can be perceived as belonging to a particular agent a_j if and only if it has been produced in isolation, i.e. if it does not overlap with the vocalization of another agent at the same time step. This reflects the fact that due to interferences in the sound signals, overlapping vocalizations impair agent identification (see Section II). Thus, if only one agent was vocalizing at time t, the output of the agent identification system is a discrete value encoding the agent identity a_j. If no agent vocalized or if several of them did, it does not output any value. Note that an agent can identify its own vocalization as being produced by itself.

2) Presence estimation: The output of the agent identification system is used to estimate the presence of each member of

the population. Each agent maintains a set of presence values $\{p_j(t)\}_{j\in\{1,\ldots,N\}}$ which can be considered as the estimated probabilities that each member is present. Note that this set does not reflect a probability distribution, the N values not necessarily summing up to 1: in the two extreme cases all agents are estimated as surely present (all p_j equal 1) or surely absent (all p_j equal 0).

The presence value of each agent, including the one of itself, decrease exponentially following the dynamics:

$$p_j(t+1) = 0.9p_j(t).$$

Whenever a particular agent a_j is identified at time t (see agent identification above), the presence is increased by:

$$p_j(t+1) = p_j(t) + 0.3,$$

and then stays at the same value $p_j(t+1)$ during the 8 subsequent time steps. Moreover these presence estimation values are programmatically bounded between 0 and 1.

Thus, the presence probability decreases at each time step in an exponential way when no identification is performed and is increased by a constant value whenever an agent is properly identified. In this latter case, the corresponding presence estimation starts to decrease again only after 8 time steps.

3) Action selection: Action selection uses the current presence estimation values $\{p_j(t)\}$ to infer a motor activation. This involves a parametrized motor policy which is learned according to each agent experience.

The action policy is modeled as a linear combination of the presence estimation vector $\{p_j(t)\}_{j\in\{1,\ldots,N\}}$. By noting $w_A(t) = [w_0(t), w_1(t), \ldots, w_N(t)]$ the weights of the linear combination at time t, the activation sent to the motor system is given by:

$$\begin{aligned} a_A(t) &= w_A(t)[1, p_1(t), \ldots, p_N(t)]^T \\ &= \sum_{i=0}^{N} w_i(t)p_i(t), \end{aligned} \quad (2)$$

where $p_0(t) = 1$ and $w_0(t)$ is the value of x when all the p_j are null.

The output value $a_A(t)$ will be summed with the output from the reactive layer in the motor system to modulate the motor activation from the estimated agent presences ($a = a_R + a_A$ in Equation 1, where a_R is the motor activation from the reactive layer). How the weights are learned is explained below.

4) Value estimation and learning: The weights $w_A(t)$ parameterizing the action selection system are adaptively learned by each agent. At each time step, the agent estimates how well its last action contributed to the overall group presence, i.e. to the probability that each member of the group is present. The action policy parameters are iteratively adapted through a learning rule to maximize the cumulative overall presence.

Computationally, we use a classical actor-critic method widely use in reinforcement learning (see e.g. [10]). In this context, the action selection system corresponds to the actor (parametrized by $w_A(t)$) and the value estimation system corresponds to the critic (evaluating the actor performance and updating the policy parameters in the direction of performance improvement). For this aim, the critic approximates a value function mapping the input of the actor (i.e. the presence activation vector $[p_0(t), \ldots, p_N(t)]$, analog to the state in the classical reinforcement learning terminology) to the expected discounted sum of rewards of applying the policy from that state.

The reward at each time step corresponds to the overall group presence and is defined by:

$$r(t) = \prod_{i=1}^{N} p_i(t). \quad (3)$$

It can be viewed as the estimated probability that all agents of the population are present.

Function approximation is performed analogously to the action selection system, i.e. using a linear combination:

$$v(t) = w_V(t)[1, p_1(t), \ldots, p_N(t)]^T,$$

where $w_V(t)$ are the weights of the linear combination and $v(t)$ is the expected value of applying the policy from the current state $[p_1(t), \ldots, p_N(t)]$.

This value function and the reward at time t are then used to compute a temporal-difference error $e(t)$:

$$e(t) = r(t) + \gamma v(t) - v(t-1),$$

where γ is a discount factor that is set to 0.9.

The temporal-difference error $e(t)$ indicates if the policy has performed better ($e(t) > 0$) or worse ($e(t) < 0$) than expected given the last observed reward $r(t)$. It is used as a learning signal to update both the value function and the policy (respectively parametrized by $w_V(t)$ and $w_A(t)$) using the following formulas:

$$w_V(t+1) = w_V(t) + lr(t)e(t)\tilde{p}(t-1) \quad (4)$$

$$w_A(t+1) = w_A(t) + lr(t)e(t)m(t-1)\tilde{p}(t-1), \quad (5)$$

where $lr(t)$ is a decreasing learning rate given by $lr(t) = 0.1 * t^{-0.1}$, $m(t-1)$ indicates whether the agent vocalized at that time step ($m(t-1) = 1$) or not ($m(t-1) = -1$), and $\tilde{p}(t-1)$ is a short notation for $[1, p_1(t-1), \ldots, p_N(t-1)]$.

IV. Results

In Section III-B we have shown the basic behavior resulting from agents equipped with only the reactive layer (Figure 2). In the current section we thus limit our analysis to the behavior resulting from agents equipped with the entire cognitive architecture of Figure 1.

A. Adaptive behavior

Figure 3 displays the result of a 2-agent population equipped with the full cognitive architecture of Figure 1 (both the reactive and the adaptive layers are activated). At the beginning of their interaction (time steps 0 to 100, left column), their behavior is similar to when using the reactive layer alone

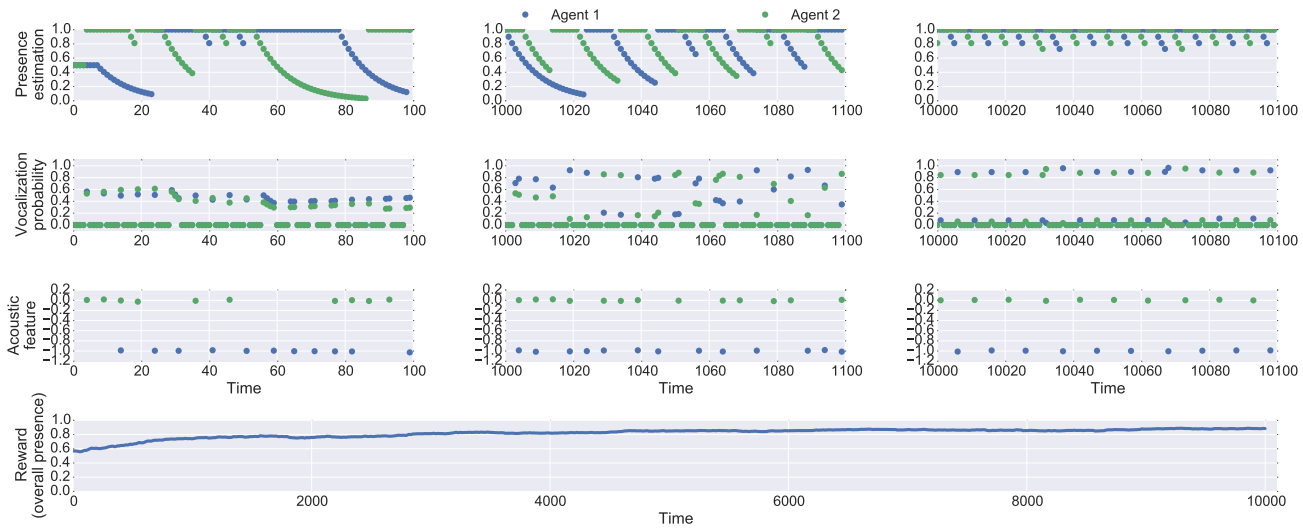

Fig. 3. The adaptive layer allows the learning of an action policy maximizing the overall presence estimation. Two agents embedding the full cognitive architecture of Figure 1 interact as described in Section III-A during 10000 time steps. Each one is represented by a particular color (top legend). Each of the three columns refers to the agent states at three different time windows (100 time steps each indicated in the x-axis). First row: presence estimation following the dynamics described in Section III-C2. The evolution of these values is the same for both agents because they receive the exact same auditory input and implements the same deterministic agent identification process. Thus, each point color in the first-row plots corresponds to the estimate that each agent is present, these values being computed the same way by each agent in the population. For example, around the 20^{th} time step, both agents estimate a low presence of Agent 1 (blue) and a high presence of Agent 2 (green). Second row: probability of vocalizing induced by the summed motor activations from both the reactive and the adaptive layers to the motor system, through the activation function of Equation 1. Whenever only one point is visible at a given time step, it means that both agents have the same value (this is in particular the case for the low level activations which correspond to cases where the reflex controller strongly inhibits the motor system as described in Section III-B). Third row: vocalizations actually produced by each agent, resulting from the motor system output (i.e. from a probabilistic choice) and expressed as the agent specific auditory feature (y-axis). Bottom row: overall group presence estimation during the entire simulation, corresponding here to the running mean over 1000 time step of the reward computed by Equation 3. As for the presence estimation from which it is computed, this value is the same for both agents.

(Figure 2). This is because the initialization of the action selection system parameters $w_A(0)$ are initially sampled around 0. However, we observe that the learning process at work in the adaptive layer has an effect quite early as indicated by the diverging motor activations of both agents (remember that the reactive layer alone always result in the same motor activation for both agents as shown in Figure 2). Due to the motor term $m(t-1)$ of the action policy learning rule in Equation 5, which can differ between both agents according to the probabilistic nature of the motor system, each agent converges towards different action policies. We observe that this allows the overall presence probability to continuously increase during agent interactions (bottom panel spanning the three columns). From time step 1000 to time step 1100 (middle column), we observe that their motor activations are diverging: when one vocalizes the other often remains silent and vice-versa (second row). From time step 10000 to time step 10100 (right column), we observe that they have converged to a near-optimal policy, resulting in a turn-taking behavior with no overlapping. Note that the overall group presence estimation in the bottom panel cannot reach the value 1 due to the motor inhibition performed by the reactive layer, preventing the agents to vocalize whenever one of them did it in the recent past and preventing the presence estimation values to stay at the maximal level (upper-right panel).

B. Action policy learning

To analyze this result in more detail, Figure 4 shows the evolution of the acquired action policy by the two agents during the same simulation as in Figure 3. These action policies correspond to the linear combination weights of the action selection system which are learned by the agent (Equation 2). Remember that at each time step t, the action selection system takes as input the estimated presence probabilities $\{p_j(t)\}_{j\in\{1,\dots,N\}}$ (noted $P(a1)$ and $P(a2)$) on the figure axes) and returns a motor activation $a_A(t) \in \mathbb{R}$ through Equation 2. This motor action is then summed with the one from the reflex controller of the reactive layer, $a_R(t)$, and a probability of vocalizing is obtained through the sigmoid activation function (with $a = a_R + a_A$ in Equation 1). This is the probability which is displayed in the plots of Figure 4, which corresponds to the cases where the reactive layer does not inhibit the motor system (i.e. whenever $a_R = 0$ resulting in a probability of vocalizing of $p(a_A)$)). Note that whenever the reactive layer inhibits the motor system (i.e. whenever $T < \theta_t$, resulting in $a_R = -30$) the probability of vocalizing is always quasi-null, the motor activation from the adaptive layer being unable to compensate this inhibitory effect (that case is not shown in Figure 4).

We observe that the incremental learning process occurring in the adaptive layer through the value estimation allows the

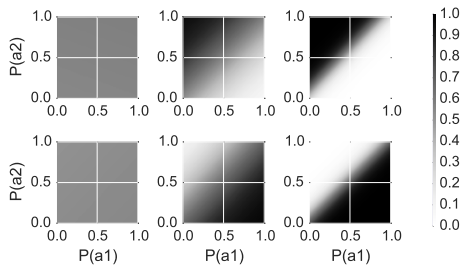

Fig. 4. Adaptive agents converge to opposite action policies. Data from the same simulation as in Figure 3. Each of the six subplots represents the policy learned by the two agents (top row: Agent 1; bottom row: Agent 2) at three different time steps (left column: $t = 0$; middle column: $t = 1000$; right column: $t = 10000$). These action policies map the estimated presence probabilities $\{p_j(t)\}_{j\in\{1,\dots,N\}}$ (noted $P(a1)$ and $P(a2)$ on the axes) to the resulting probability of vocalizing (color map from white to black) provided by Equation (2) and (1). We assume here that $a = a_A$, i.e. that the reflex controller does not inhibits the motor system ($a_R = 0$, see Section III-B). For example, at time $t = 10000$ (third column), Agent 1 (first row) has converged to a policy resulting in a high probability of vocalization (black color) whenever its own estimated presence $P(a1)$ is lower than the one of its partner $P(a2)$.

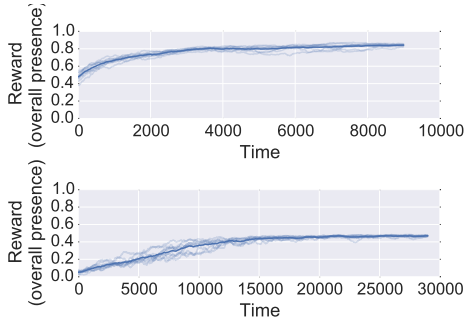

Fig. 5. The performances of the model are robust both across simulations and with more than two agents. Overall presence probability over 10 independent simulations with different random seeds (light blue: individual simulations; dark blue: mean over all the 10 simulations). Top: two-agent simulations. Bottom: three-agent simulations. Note the different time scales on the x-axes due to the slower convergence of the 3-agent simulations. The y-axis corresponds to the running mean over 1000 time step of the reward computed by Equation 3.

agents to progressively converge towards opposite action policies: the agent with the lowest estimated presence yields a high probability of vocalizing, whereas the one with the greatest estimated presence yields a low probability of vocalizing. This emerging strategy is actually optimal to maximize the overall group presence (i.e. the reward $r(t)$) because it allows to increase the lowest estimated presence each time the agents are able to vocalize (i.e. each time the reflex controller does not completely inhibit the motor system).

C. Model robustness

Figure 5 shows the model performances across a number of independent simulations in 2-agent and 3-agent populations. We observe that the populations improve the overall group presence in each situation, thus showing the robustness of the model performances. Note that the maximal overall group presence decreases with the number of agents in the population (around 0.8 for two agents and around 0.5 for three agents, according to the figure) due to the reflex controller which prevents agents to vocalize whenever they have listen to a vocalization in the last θ_t time steps. As θ_t is fixed, this implies lower average presence values for populations with more agents.

V. Conclusion

In this paper, we proposed a computational model of vocalizing agents where the control flow is distributed into two layers: a reactive one based on a simple reflex controller driven by a need to listen to vocalizations (Section III-B) and an adaptive one allowing action policy learning driven by the maximization of the overall group presence (Section III-C). The reflex controller we proposed, which is regulated through the possibility of increasing/decreasing motor activity depending on whether there was auditory input in the recent past, allows the agents to display a basic level of entrainment where each one is influenced by the vocalizations of the others. However this behavior is clearly too simple to allow by itself the convergence to a collective turn-taking strategy (as shown by the overlapping vocalizations in the bottom panel of Figure 2). A reason for this is that the agents do not self-monitor, i.e. they are unable to distinguish their own vocalizations from those of the other group members, what is coherent with the results of a previous model [3] published at ICDL/Epirob. Nevertheless, we showed that the sensorimotor data collected through this reactive control loop allows to appropriately bootstrap the learning process occurring at the adaptive level, in the sense that the agents converge robustly to a turn-taking collective strategy maximizing the overall group presence. This is performed through the iterative learning of an action policy by each agent, which triggers vocalizations according to presence estimation values. The learned action policy we observed (Figure 4) is also coherent with the mentioned previous model in the sense that the vocalization rate is increased by the other vocalizations and decreased by those of oneself. Our own contribution is to show how such a policy can be learned by each individuals from agent interactions.

The proposed model has obvious limitations: time is considered as discrete, the reflex controller is based on a fixed parameter θ_t, all agents receive the exact same auditory input, the acquisition of the agent identification system is not modeled and assumed that each agent is aware of the number N of populations members. Nevertheless, this model still provides a first computational basis to address the questions we asked in the introduction: (1) What is the role of reactive vs adaptive control loops in turn-taking behaviors? (2) How does learning occur from agent interaction and under what functional constraints?

Our tentative answer to (1) is that the reactive control loop, which can be easily conceived as pre-wired from evolutionary processes due to its simplicity, provides the necessary sensorimotor constraints to bootstrap the autonomous development of

a more efficient strategy through learning in the adaptive layer. In biological terms, the reflex controller could be considered as analogous to an homeostatic loop regulating the need for vocal listening of each agent. This particular need could be related to physiological states (e.g. arousal, security or social cohesion). With the addition of the adaptive layer and with the possibility of performing agent identification, a presence value of each group member is maintained that has similar dynamics than the reactive loop. It could be interesting to relate this multi-loop regulation to the physiological concept of allostasis as we see it in [11] (a meta-regulation of homeostatic loops) and thus we foresee a possible unifying principle for regulating the loops of both reactive and adaptive layers.

Regarding (2), we showed how this process can occur at the population level in a decentralized way, where the learning processes of the agents are coupled through the mutual perception of their vocal productions. We saw how a collective turn-taking behavior can self-organize out of basic sensorimotor interactions, in a way similar to the language-game paradigm previously used to model the emergence of lexical [12] or phonological [8], [7] systems in agent populations. In the current model each agent is driven by a need to maintain vocal contact and we showed that the interaction of these individual needs through mutual auditory perception allows the convergence toward a globally efficient strategy.

Starting from these preliminary results, our more general goal is to understand how entrainment between individuals can emerge at various levels of sensorimotor and cognitive interaction. For this aim, we adopt the Distributed Adaptive Control (DAC, [4], [5]) framework, which proposes that cognition is organized in a number of layers. In this paper, we focused on agent interactions at the reactive and adaptive levels. The bottom layer of DAC (in the sense of Figure 1) is called *Soma* and is the interface between the *Reactive* layer and the environment through exosensing (sensing of the world through vision, audition, etc…), endosensing (sensing of the self from physiological needs) and action (motor execution). The top layer of DAC is *Contextual* and is involved in memory and planning to regulate the activity of the *Adaptive* layer, e.g. to store or recall successful sequences of actions to reach more abstract goals. We want to study how entrainment between individuals can emerge at each level of this hierarchy from their sensorimotor interactions under certain environmental constraints.

In order to achieve this goal, we are now extending this model toward a neuromorphic implementation, by identifying the possible neural correlates of the proposed subsystems in the brain, both from a computational (noting that the proposed subsystems could be easily implemented in spiking neural networks) and a structural perspectives (based on DAC which is strongly grounded in brain theory [5]). This will allow to propose experimental predictions to be confronted against animal behavior data (e.g. [1]) in order to validate or invalidate parts of the model.

Finally, this kind of modeling could also be applied to the conception of original interactive systems for the cooperative production of a musical performance between humans and machines. An interesting line of research, for which we have encouraging first results, is to use this model for rhythmical sequence learning, where each element of the sequence is represented by an agent which adaptively times its motor actions with respect to those of the others (resulting in a decentralized implementation of former human-machine musical synchronization systems, e.g. [13]).

ACKNOWLEDGMENT

This work is supported by the Socialising Sensori-Motor Contingencies project socSMC-641321H2020-FETPROACT-2014.

REFERENCES

[1] D. Y. Takahashi, D. Z. Narayanan, and A. A. Ghazanfar, "Coupled oscillator dynamics of vocal turn-taking in monkeys," *Current Biology*, vol. 23, no. 21, pp. 2162–2168, 2013.

[2] D. Y. Takahashi, D. Narayanan, and A. A. Ghazanfar, "A computational model for vocal exchange dynamics and their development in marmoset monkeys," in *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–2.

[3] ——, "Development of self-monitoring essential for vocal interactions in marmoset monkeys," in *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*. IEEE, 2013, pp. 1–2.

[4] P. F. Verschure, T. Voegtlin, and R. J. Douglas, "Environmentally mediated synergy between perception and behaviour in mobile robots," *Nature*, vol. 425, no. 6958, pp. 620–624, 2003.

[5] P. F. Verschure, C. M. Pennartz, and G. Pezzulo, "The why, what, where, when and how of goal-directed choice: neuronal and computational principles," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1655, p. 20130483, 2014.

[6] L. Steels, "The synthetic modeling of language origins," *Evolution of Communication*, vol. 1, no. 1, pp. 1–34, 1997.

[7] P. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, no. 3, pp. 435–449, Apr. 2005.

[8] C. Moulin-Frier, J. Schwartz, J. Diard, and P. Bessire, *Primate communication and human language: Vocalisations, gestures, imitation and deixis in humans and non-humans*. Advances in Interaction Studies' series by John Benjamins Pub. Co., 2011, ch. Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework.

[9] C. T. Miller and A. W. Thomas, "Individual recognition during bouts of antiphonal calling in common marmosets," *Journal of Comparative Physiology A*, vol. 198, no. 5, pp. 337–346, 2012.

[10] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 1291–1307, 2012.

[11] M. Sanchez-Fibla, U. Bernardet, E. Wasserman, T. Pelc, M. Mintz, J. C. Jackson, C. Lansink, C. Pennartz, and P. F. Verschure, "Allostatic control for robot behavior regulation: a comparative rodent-robot study," *Advances in Complex Systems*, vol. 13, no. 03, pp. 377–403, 2010.

[12] L. Steels, "Emergent adaptive lexicons," in *SAB96*, P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, Eds. Cambridge, MA: MIT Press, 1996.

[13] A. Cont, "Antescofo: Anticipatory synchronization and control of interactive parameters in computer music." in *International Computer Music Conference (ICMC)*, 2008.