Social Reinforcement in Artificial Prelinguistic Development: A Study Using Intrinsically Motivated Exploration Architectures

Juan M. Acevedo-Valle*, Verena V. Hafner‡ and Cecilio Angulo*

***Abstract*—This work introduces an intrinsically motivated sensorimotor exploration architecture which considers social reinforcement and motor constraint awareness. The main objective is to study the influence of social interactions during artificial early prelinguistic development. We argue that this architecture contributes to explain development from *voiceless* to *sequence of vowels* vocalizations. A cognitive developmental perspective is considered emphasizing embodied cognition and sensorimotor exploratory behaviors.**

For a new-born agent, motor constraints are unknown. However, the agent is endowed with a somatosensory system that indicates if a motor configuration was reached or not. This information is used to model and predict constraint violations. Furthermore, the architecture considers imitative behaviors that constrain the search space during exploration. Interaction occurs when the learner sensory production is similar to a sensory unit relevant to communication. In that case, the instructor perceives this similitude and reformulates with the relevant sensory unit. When the learner perceives an utterance by the instructor, it attempts to imitate it.

Two systems are considered for experimentation: A toy example and a simulated vocal tract. In general, our results suggest that constraint awareness and social reinforcement contribute to achieve less redundant exploration, lower exploration and evaluation errors, and a clearer picture of developmental transitions.

***Index Terms*—Prelinguistic social development, Vocal development, Sensorimotor exploration, Intrinsic motivations, Somesthetic Senses.**

I. Introduction

Integration of sophisticated social robotic systems to human societies in different environments entails many challenges, e.g. robustness in order to guarantee safety for human beings and robots themselves, communication to achieve efficient interactions and collaborations, and ethical dilemmas. Regarding the communication challenge, robots should be endowed with human-like communication mechanisms. As a consequence, artificial speech and natural language technologies have been largely developed. Advanced Automatic Speech Recognition (ASR) systems have emerged as a popular solution to the communication challenge. Thus, they have been implemented in robots, computers, smartphones and other devices. However, they lack many relevant features of human language and they only consider relatively constrained scenarios.

On the other hand, computer science and robotics have become important tools as a mean for studying the human mind. Machine learning techniques, fast robot prototyping and complex simulators have fostered the appeal of artificial agents for studying the mechanisms of cognitive development. These technologies have also encouraged researchers to follow the idea of "*understanding by building*" [1]. The idea arises from the embodiment paradigm, which states that the behavior of an agent is not only the result of a system control structure, but it is also result of complex interactions of the agent with its ecological niche, its morphology, its material properties and other individuals [1, 2].

There exist some attempts to study artificial early vocal development as a mechanism to understand language emergence from an embodied developmental perspective [3–8]. However, most of the works aiming at studying artificial speech-based communication systems are rather focused on the natural language understanding problem. The lack of focus on early vocal development and, in general, on prelinguistic communication is not surprising. As mentioned in [9], just a couple of decades ago it was still assumed that vocal development was the result of maturational programs, which were independent from environmental influence.

In [3], the simulated model *Elija* was proposed, which goes through three development stages considering pre-programmed transitions: babbling, social learning, and naming objects. In [4], a simulated model was presented which was able to show developmental transitions just as a mere result of intrinsic motivations. The intrinsically motivated architecture was modified in [5]; hence it was endowed with constraint awareness using a somatosensorimotor model. Finally, in [6], the idea of an instructor was included, [4] had done some investigation regarding social emulation as a means of driving development, despite finding that social imitation was a crucial opportunity for development, [4] did not consider interactions neither realistic speech units. In [8], the authors made some progresses in the same line of research of [4], obtaining similar proportions of *single vowels* and *sequence of vowels* vocalizations as in [5]. [8] also studied the possibility of using Mel-frequency cepstral coefficients to model auditory signals instead of using the formant frequencies used here and in [4–7]. Finally, [7] considered a model that was also physically situated in a scenario where tool usage was being learned in parallel to vocal babbling, even though vocal learning was somehow more restricted compared to [4].

Building realistic speech-based communication systems re-

*Knowledge Engineering Research Group, Department of Automatic Control, Universitat Politècnica de Catalunya - BarcelonaTech, Spain. Contact E-mail: {juan.manuel.acevedo.valle, cecilio.angulo}@upc.edu

‡Adaptive Systems Group, Department of Computer Science, Humboldt-Universität zu Berlin, Germany. E-mail: {hafner@informatik.hu-berlin.de}

quires an accurate understanding of the mechanisms used by infants to learn the speech 'code' [10]. Infants show preparedness to master speech and acquire language: from the onset of babbling at 3 months of age, infants achieve to produce full sentences by the age of 3 years. Lack of knowledge underlying this developmental process has been an important obstacle to achieve advanced artificial equivalents to natural language [10].

Babbling, as exploratory sensorimotor behavior, is a milestone in early language development. Exploratory behaviors are sensory-guided motor behaviors that help to form and maintain internal body representations. These representations are used to master sensorimotor control, which is considered by developmental psychologists as a fundamental prerequisite to more complex cognitive and social capabilities [11]. There exist many intertwined processes underlying linguistic development. For instance, one might consider the study of sensory learning, motor control learning, sensorimotor control learning, and social development. In some stages each of those processes seem to go their way, but progress in each process affects the development of the others, integrating them into a complex developmental system. Hence, early sensory and motor development, including their interdependence, are fundamental for a canonical babbling stage where relevant sensorimotor learning occurs [10]. Recent studies [9, 12–17], have found vast evidence suggesting that prelinguistic speech development after the onset of babbling is related to early social development in a mutual relation. Social interactions shape early vocal development in different ways. Studies have found that responses from adults (e.g. touching, smiling, and approaching) change the frequency of vocalization in infants [13].

Since humans are truly social beings, sensorimotor exploration is just one aspect of developmental learning. Often, skills acquired by exploration are reinforced and extended by social mechanisms, e.g. learning by demonstration or imitation learning. An example of such behaviour also discussed in the field of developmental robotics [18, 19] is the onset of pointing behavior. One hypothesis suggests that pointing behavior in young infants initially emerges from the attempt of grasping an object that is out of reach. When the caregiver hands over the requested object to the infant, the pointing gesture is rewarded through social reinforcement [20].

In infants with regular development, there exists an ordered number of typical stages emerging along the progress from newborns to fully functioning adults [21]. A number of works exist attempting to offer an explanation to the emergence of developmental stages during vocal development using artificial intelligence techniques [3, 22–24]. However, those works do not provide any explanation for the onset of developmental stages. A speech acquisition model called *Elija* has been developed in [3]. Using manual mechanisms to onset each stage, it is able to go from babbling to the capability of naming objects using infant-like utterances. Recently, a model of language development stages from the embodied perspective was introduced in [21]. However their efforts are rather directed towards language level development, leaving early vocal development as an open issue.

Intrinsically motivated exploration architectures have been exhaustively studied [4, 25, 26] based on developmental theories. Our work expands this research in [4–6, 27]. In [4] the emergence of stages during vocal development was studied. They found that an intrinsically motivated exploration architecture might be a good candidate to explain the developmental trajectory from *voiceless* to *sequence of vowels* vocalizations. Recently, [5, 6, 27] studied the role of motor constraints awareness in sensorimotor exploration inspired by the role of somesthetic senses in early motor control learning and based on the hypothesis that motor, perceptual, social, and learning ability constraints play a key role in emergence of language during infant development [10]. Besides studying the role of motor constraints, preliminary results are provided in [6] pointing to some evidences that social feedback mechanisms, even considering a simple imitation scenario, drives development more efficiently during intrinsically motivated explorations. It was also demonstrated that constraint awareness and social reinforcement benefit the efficacy of intrinsically motivated exploration architectures, improving both the prediction of self action consequences and the volume of the explored sensorimotor regions.

The main objective of this work is to formalize a socially reinforced and intrinsically motivated architecture for sensorimotor exploration. This architecture is aimed at studying the impact of social reinforcement according to developmental studies on prelinguistic social development, in particular the influence of *imitation/expansion* maternal responsiveness described in [16]. Different from previous works that considered intrinsic motivations and imitation scenarios to drive learning, e.g., [4], here we consider an instructor, which modifies its behavior according to the learner's behavior, in other words we consider interaction as the driver of development, instead of the sensory ideas that are chosen from internal collections. Moreover, compared to [4], we used realistic vowels in our experiment. Two different scenarios will be considered. Social impact will be measured in terms of explored volume of the auditory space, ratio of *single vowels* and *sequence of vowels* vocalizations, average evaluation error, and percentage of undesired motor configurations violating constraints. In general, our results suggest that constraint awareness and social reinforcement contribute to achieving better results during intrinsically motivated exploration. Achieving less redundant exploration, decreasing exploration and evaluation errors, as well as showing a clearer picture of developmental transitions.

The remainder of this paper is organized as follows. Section II and Section III introduce former results regarding the role of somatosensation and social mechanisms in prelinguistic development. Section IV introduces the intrinsically motivated sensorimotor exploration architecture with social reinforcement. The experimental setup and results are presented in Section IV and Section VI, respectively. Finally, the discussion is completed in Section VII in order to conclude in Section VIII.

II. The Role of Somesthetic Senses in Vocal Development

Besides intrinsic motivations, some works have highlighted the relevance of other mechanisms to sensorimotor exploration. For instance, proprioception is mentioned in [11, 28] and nociception in [29]. Proprioception endows agents with a sense of their own movements. Nociception depends on nociceptors, which are nerve fibers responsible of responding to levels of chemicals, temperature or pressure that might be harmful for the body. The somatosensory system is taken into account in [3, 22] using tactile information in their architectures for speech acquisition. The *Diva* model introduced in [22] includes the premotor, motor, auditory and somatosensory cortical areas, and a simulated auditory-vocal tract system. Therein, a somatosensory model was effectively integrated into the acquisition and production of speech. This somatosensory model integrates tactile and proprioceptive data. However, it was not used as an element to integrate motor constraints but as a part of the sensorimotor system itself.

Proprioception and haptic senses, as mentioned in [30], are essential sensory modalities for the agent to learn how to drive its own movements to reach body states. Evidence provided in [30] states that during the emergence of reaching, as a product of a deeply embodied process, infants first learn how to direct their movement in space using proprioceptive and haptic feedback. In vocal development, these mechanisms must play a key role, as somesthetic senses (i.e., sense of touch, proprioception, nociception, and haptic perception) are a rich source of information available to infants, e.g., driving autonomous exploration through a feedback loop, even long before they are able to control phonation.

During learning of proprioception, nociception, and haptic modalities the agent must discover its own motor limitations. Herein, an architecture accounting for embodied systems with motor constraints is studied. The architecture, introduced in [5, 6], relies on the concept of unreached motor goals due to motor constraints. The embodied agent is endowed with a system that generates a somesthetic signal indicating if a motor configuration was reached or not.

The exploration architecture introduced in this work extends the previous one by considering somatosensory modalities under some simplifications and assumptions. For instance, a somatosensory system based on tactile information is used to generate a somesthetic signal. In other words, tactile information along a simulated vocal tract is encoded into a pain signal emulating the role of nociceptors. The somesthetic signal provides the agent with information of its own body configuration: when tactile information is incoherent, a somesthetic signal is triggered indicating that the desired motor configuration might be 'harmful' or physically unreachable.

III. Prelinguistic Social Interactions Affect Vocal Development

Many studies have been recently completed by developmental psychologist about understanding the effects of social interaction over early vocal development. For instance, it was found in [31] that at six months of age, infants are aware of their vocalizations' social value affecting parental engagement. The present work is based on evidence suggesting that prelinguistic vocalizations are salient signals to parents, who immediately respond. Those parental responses might play an important role in vocal development and language acquisition [14]. The often invoked analogy between human speech and bird song development was studied in [13]. In songbirds, imitation is usually considered the mechanism for vocal development. They found that social contingency provides opportunities for vocal learning in birds, thus vocal development is socially shaped. Testing the ability of infants to use social feedback to facilitate developmental transitions, they observed that contingent interactions foster changes in vocal behavior. Their major conclusion is that, simultaneously, babbling regulates and is regulated by social interaction. They also found that changes in babbling due to social reinforcement might be fostered by different social contingencies as touching, smiling, and approaching. Later, this work was extended in [9]. During naturally occurring interactions, it was found that mothers' vocalizations provide better predictions for infant's vocal utterances compared to other social modalities. In general, adults are sensitive to differences in prelinguistic vocalizations, responding differently to distinctive sounds (e.g. track cries, quasi-voiced vocalizations, voiced 'syllabic', 'vocalic'). Adults can classify vocalizations of children 7-11 months of age, even of unfamiliar infants. Adults see infants as 'real talking' when they produce prelinguistic syllabic sounds and respond to this kind of vocalizations with higher frequency. The fact that adults perceive different infant vocal types suggests that maternal responsiveness plays a role in vocal development.

Evidence suggesting that prespeech vocalizations have a range of pragmatic functions was provided in [15]. However, pragmatic functions were not related by any means to vocalization development. Later, based on experimental results, it was suggested in [16] that maternal responses to infants' directed vocalizations contribute to the emergence of vocal usage and the shaping of vocal development. In general, evidence has shown that mothers respond differently according to infants' vocalization directionality (mother-directed, object-directed, and undirected) and acoustic characteristics. Mother's sensitive responding to mother-directed vocalizations was correlated with increase in developmentally advanced consonant-vowel vocalizations and some language measures [16].

Regarding maternal responsiveness, seven categories of maternal verbal response are distinguished in [16]: acknowledgments, attributions, directives, naming, play vocalizations, questions and imitation/expansions. During the imitation scenario mothers model the word that the sound produced by the infant approximated and expand on it. It was found that imitation in early months of life is a good predictor for an increment in infant mother-directed vocalizations in future months. Infants who received proportionally more responses to their mother-directed vocalizations showed a larger increase in developmentally advanced vocalizations.

Early mother-infant mutual coordinated engagement and its association with more syllabic vocalizations was also evidenced in [12]. Experiments in [16] provided strong support to conclude that maternal response contributes to achieve phonologically advanced consonant-vowel sounds and mother-directed vocalizations. It can be also pointed out that prelinguistic communicative behaviors differentially influence caregivers at the moment and over time, showing that the behaviors of infants and caregivers are deeply intertwined.

Finally, as social feedback to vocalizations is an underlying mechanism for developmental change, it is important to identify its potential for social interactions with social partners in different social environments [16]. In the following, the *imitation/expansions* maternal response as social mechanism is considered and integrated into a sensorimotor exploration architecture. An instructor –expert in vocalizing– is considered: every time a learner produces a vocalization similar to a social relevant vocalization known by the instructor, the latter reformulates and vocalizes it; immediately the learner attempts to imitate it. Even though [16] does not mention the latter imitation mechanism, we proposed it as a means to explain the impact of mothers' responsiveness over vocal development. The mechanism is similar to that used by Howard and colleagues in [3].

IV. Sensorimotor Exploration Architecture

The exploration architecture introduced in this work is based on those presented in [26, 32–34]. It is an active learning architecture that mimics the exploration behaviors observed during sensorimotor exploration in biological agents. The architecture is based on goal babbling, which has been found suitable for learning non linear redundant systems as explained in [35]. During exploration, sensory goals are actively chosen according to a model of interest.

In previous works [5, 6] modifications have been proposed to include the concept of somatic senses and social reinforcement in intrinsically motivated exploration architectures. Firstly, based on biological evidence regarding the role of motor constraints in early development of cognitive skills, the original architecture from [34] was modified in [5] to consider a somatosensorimotor model in order to acknowledge constraints during the active selection of interesting goals. Next, a similar architecture to that represented in Figure 1 was introduced in [6]. This architecture includes a social instructor, expert in sensory units relevant to communication, that interacts with the developmental agent. Interaction occurs when the learner production is 'enough' similar to one relevant to communication. In that case, the instructor perceives this similitude and reformulates with the relevant sensory unit. When the learner perceives an utterance by the instructor, it attempts to imitate it. This reformulation mechanism is similar to the one used in the *Elija* model [3], which was motivated by the episodes of vocal imitation observed in mother-child interaction.

The exploration architecture in this work consists of the following elements:

- **Physical Embodiment** consists of a sensorimotor system producing sensory outcomes perceivable by other agents. It includes a somatosensory system, which produces a somesthetic signal indicating whether a motor command has been successfully executed or not.
- **Sensorimotor Model** is an internal representation that maps motor commands to sensor results. It is used to solve the inverse problem of inferring motor commands from provided sensory goals.

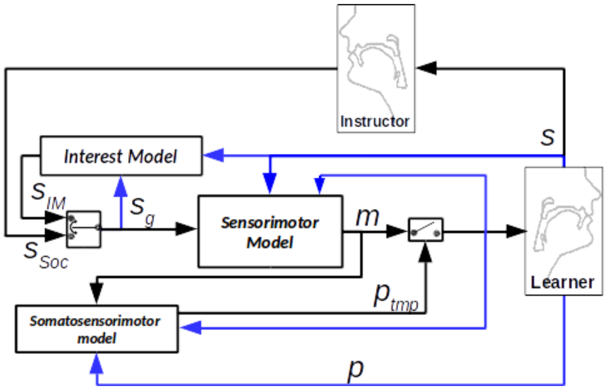

Fig. 1: Diagram of the exploration architecture. Black lines represent the flow of data during each vocalization. Blue lines represent signals used to update the models. Notice that the selector switch indicates that s_g could be generated either directly from the social reinforcement (prioritized) or from the intrinsic motivations mechanism. The simple switch indicates that the somatosensory model might accept or reject a proposed interesting goal.

- **Interest Model** is the core of the intrinsic motivation mechanism. This model keeps information, based on a competence function c, about how well the agent is performing in reaching intended goals through time. Thus, the agent can choose goals that are likely to improve its sensorimotor control skills.
- **Somatosensorimotor Model** is an internal representation that maps motor commands to somesthetic results, thus it can predict somesthetic results from motor commands before they are executed.
- **Embodied Instructor** is an agent that shares a similar physical embodiment with the learner. However, it is assumed to have mastered sensory units relevant to communication purposes.

The pseudo-code associated to the exploration architecture is shown in Algorithm 1. The first part of the pseudo-code corresponds to the learner. On the other hand, the $interaction$ function represents the instructor behavior, which is able to produce sensory units from a set of sensory units **S** relevant to communication purposes.

As shown in Algorithm 1, the learner starts without any experience producing intended goals nor self constraints knowledge. It is assumed that the agent is randomly initialized accordingly (lines 1 and 2). Then, looking at Figure 1, the interest model M_{IM} in the diagram is able to propose new goals that are likely to foster the progress of the competence function $c_i = e^{-|s_{g,i} - s_i|}$, which measures the ability of the agent to reach sensory goals.

Suppose that M_{IM} proposes a goal s_{IM} and the instructor is not currently interacting with the learner. In this case the selector switches to the signal $s_g = s_{IM}$. The proposed goal is sent to the sensorimotor model M_{SM}. Then, the sensorimotor model infers which is the motor action m_i that would produce s_g according to the current agent's knowledge. Finally, m_i is received by the somatosensorimotor model M_{SS} and this model predicts the somesthetic outcome p_{tmp} of executing

Algorithm 1 Self-exploration with goal babbling, motor constraint awareness and social reinforcement.

```
Set $\{n_e,\ randomseed\}$
Initialize $M_{SM}$ and $M_{SS}$
Initialize $M_{IM}$ and $i \leftarrow 1$
while $i \leq n_e$ do
    $p_{tmp} \leftarrow 1$
    while $p_{tmp}$ do
        $s_{g,i} \leftarrow sample\left(M_{IM}\right)$
        $m_i \leftarrow M_{SM}\left(s_{g,i}\right)$
        $p_{tmp} \leftarrow M_{SS}\left(m_i\right)$
    $s_i \leftarrow f\left(m_i\right) + \sigma$ and $p_i \leftarrow g\left(m_i\right)$
    $c_i \leftarrow e^{-|s_{g,i}-s_i|}$
    $i \leftarrow i + 1$
    $train\_models()$
    $s_{g,i} \leftarrow interaction(s_i)$
    if $s_{g,i} \neq null$ then
        $m_i \leftarrow M_{SM}\left(s_{g,i}\right)$
        $p_{tmp} \leftarrow M_{SS}\left(m_i\right)$
        if $!p_{tmp}$ then
            $s_i \leftarrow f\left(m_i\right) + \sigma$ and $p_i \leftarrow g\left(m_i\right)$
            $c_i \leftarrow e^{-|s_{g,i}-s_i|}$
            $i \leftarrow i + 1$
            $train\_models()$

function $interaction(s)$
Define $\mathbf{S}$, $\mathbf{th_S}$, $\alpha_t h$
$dist = |s - \mathbf{S}[i]|$
if $min(dist) < \mathbf{th_S}[argmin(dist)]$ then
    $\mathbf{th_S}[argmin(dist)] = \alpha_t h * \mathbf{th_S}[argmin(dist)]$
    return $\mathbf{S}[argmin(dist)]|$
else return $null$
```

action m_i.

If the nociceptive prediction indicates that the somesthetic signal p will be triggered when executing, then the simple switch is open and the motor command is not executed by the agent, thus the interest model proposes a new goal and the prediction process is repeated. On the other hand, if the somesthetic prediction suggests that there is no risk when executing the motor action m_i, then the simple switch is closed and the agent executes m_i.

When the motor action is executed, sensory outcomes are produced. The salient outcomes s are observed by the instructor whereas the somesthetic outcome p is an internal sense of the agent. The signals are used to train the models. The instructor perceives the sensory outcome s and compares it to the set of sensory units relevant to communication in $\mathbf{S}$. The instructor selects the more similar unit $\mathbf{S}[i] \in \mathbf{S}$. If the Euclidean distance between $\mathbf{S}[i]$ and s is lower than a predefined threshold $\mathbf{th_S}[i]$, then the instructor produces $s_{IM} = \mathbf{S}[i]$ reformulating s and directed to the learner. At that point the double switch selects s_{IM} as the new sensory goal $s_g = s_{IM}$. The somesthetic prediction mechanisms is then activated as explained before. If the somatosensorimotor model determines that it is possible to imitate the instructor reformulation without risk of reaching undesired configurations, then the imitation action is executed, finishing the imitation episode. Otherwise, if there exists a risk when imitating s_{IM} according to p_{tmp}, then the interest model start proposing intrinsically motivated goals again to continue with the exploration. It is important to notice that every time the instructor produces a reformulation, it decreases the social threshold for that sensory unit, i.e., $\mathbf{th_S}[i]$, multiplying by a scaling factor $\alpha_t \in [0, 1]$. After each vocalization, models are updated unless they are configured to be trained with data batches after certain number of vocalizations in order to save computational resources.

V. Experimental Setup

Two sensorimotor setups are considered for experimentation. First, a simple non-linear model represented by a parabolic shaped constrained region is used as an illustrative example to evaluate the implementation of the proposed architecture and discuss the results. Next, the exploration architecture is used to study prelinguistic communication and early vocal-development using the speech synthesizer (vocal tract) from the Diva Model[1] [22]. Regarding the models used to implement the architecture, the somatosensorimotor and interest models are built using the open-source library `explauto` [36]. The sensorimotor model is built using incremental learning of Gaussian Mixture Models (ilGMM), as explained in [37]. In the following, details are provided about the experimental setup for each of the systems.

A. Parabolic Shaped Constrained Region

Figure 2 shows the toy example model, a parabolic shaped region described by the equations:

$$s_1 = m_1,\ \ s_2 = (m_2 - 3)^2,\ \text{ and } p = \begin{cases} 1 & \text{if } s \in \text{constraints} \\ 0 & \text{elsewhere} \end{cases}$$

where s_i are the components of the sensor space, m_i the components of the motor space and p is the somesthetic signal indicating if constraints are violated or not. Both motor components are constrained to the interval $[0, 6]$, whilst sensor dimensions are constrained to the white region and its blue borders in Figure 2[2]. If after executing a motor action the sensor result lies in the constrained region, then the sensor result is relocated to the closest point in the allowed region. An obvious consequence of relocation is the increment of sensorimotor redundancy. Regarding the blue marks in Figure 2, they represent sensory units laying close to the system constraints. An instructor able to produce those sensory units is assumed and units are assumed to be relevant to communication.

B. Vocal Tract

The auditory-vocal tract component of the Diva model is used in this work as simulated physical embodiment to study early speech development [22]. In this vocal tract, based on Maeda's synthesizer, the shape of the vocal tract is determined by the position of ten articulators, whereas voicing is controlled by three phonation parameters. Changing some of the parameters, the vocalization structure is kept as in [5].

[1] An implementation of the synthesizer running purely in Python has been developed. It is available on https://github.com/yumilceh/divapy.

[2] Python codes with examples for this system and ilGMM are available on https://github.com/yumilceh/igmm/

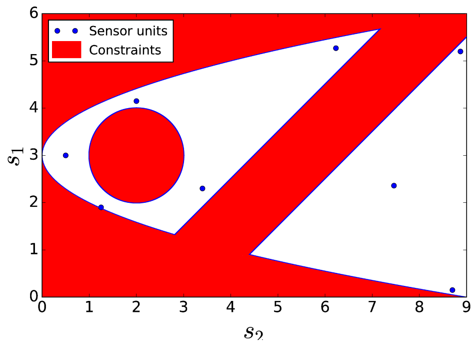

Fig. 2: Parabolic shaped constrained region.

The motor dynamics of articulators and voicing parameters are modeled as second order systems

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2\,(x - m) = 0, \qquad (1)$$

with $\zeta = 1.01$ and $\omega = \frac{2\pi}{0.01}$ representing the damping factor and the natural frequency, respectively. The duration of each vocal experiment is 800 ms, whereas m and x represent the motor command for the articulator and the current articulator position, respectively. The structure of a vocalization experiment is shown in Figure 3. As two motor commands are executed sequentially during 400 ms for each of the thirteen articulators, the result is a motor command vector of 26 dimensions. On the other hand, four sensor channels are observed: the first two formant frequencies F_1 and F_2, the intonation signal I indicating whether there is sound ($I = 1$) or not ($I = 0$), and a somatosensory signal $min(a_f)$ consisting of the minimum value of the transverse section of the vocal tract. Regarding the three first channels, representing the auditory result of the vocalization, each of them is averaged in two perceptual windows in accordance with the execution of the motor commands, thus the sensor result is a vector of 6 dimensions composing the sensory outcome.

The shape of the vocal tract is described by the area function a_f. The minimal value of the area function $\min(a_f)$ is zero when the vocal tract is closed at any point and negative when tissues are overlapping, which lacks of physical sense. Thus, $\min(a_f)$ is an indicator for the occurrence of configurations which lack physical sense. When the average of $\min(a_f)$ is negative during either one of the two perception windows then the somesthetic signal is $p = 1$, and $p = 0$ otherwise. When $p = 1$, it is assumed that the motor command is unreachable or 'harmful'. This signal is used to build the somatosensorimotor model mapping motor commands m to somesthetic outcomes p. Later, the model can be used to predict somesthetic outcomes before a motor action is executed.

For the social interaction mechanisms, in the case of prelinguistic development, this work considers an instructor with an identical embodiment as the one explained above. The instructor is capable of producing vocalizations using vowels very similar to German vowels. The seventeen German vowels were synthesized using optimization methods and considering as a reference the formant frequencies from [38], shown in Table I. The seventeen vowels are recombined to generate 289 Vowel-To-Vowel (VTV) articulatory movements (*sequence of vowels* vocalizations), following the same pattern of vocalizations shown in Figure 3. Each vowel was used as well to generate unarticulated Silence-vowel (SV) and Vowel-silence (VS) (*single vowels*). Thus summing up the articulated VTV gestures with the unarticulated SV and VS gestures, the total number of sensory units relevant to communication is 323.

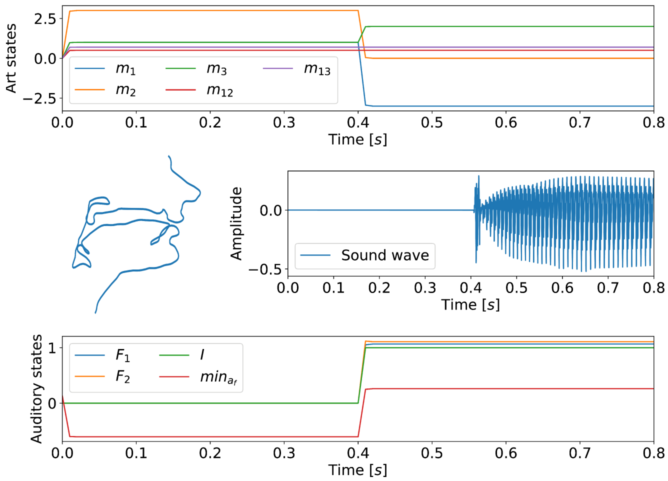

Fig. 3: Vocalization experiment example. The upper plot shows the articulatory trajectories. From 0 to 400 ms, the commands m_1, m_2 and m_3 are set to 1, 3 and 1, respectively, whereas the glottal pressure (m_{12}) and voicing (m_{13}) are set to 0.5 and 0.7, respectively. From 400 to 800 ms, the commands m_1, m_2 and m_3 are set to -3, 0 and 2, respectively, whereas m_{12} and m_{13} keep their value. The remaining motor commands are set to zero. The middle plot represents the speech sound wave signal. The bottom plot shows the auditory trajectories. There are two perception time windows, one from 0 to 400 ms and the second from 400 to 800 ms. The auditory output s are determined from the average of each trajectories along each one of the perception windows. Auditory output, includes the two first formant frequencies, F_1 and F_2, and an intonation parameter I. Finally, the somesthetic feedback p is determined from the average value of the somatosensory signal $\min(a_f)$.

TABLE I:
Formant frequencies of German vowels (Hz).

	F_1	F_2		F_1	F_2
/a:/	716	1184	/a/	694	1294
/e:/	346	2222	/E:/	526	1918
/i:/	265	2179	/y:/	274	1704
/o:/	337	605	/O/	534	929
/u:/	288	628	/U/	405	951
/2:/	316	1311	/@/	435	1614
/I/	406	1854	/E/	532	1859
/Y/	396	1302	/9/	501	1334
/6/	639	1388			

VI. Experimental Results

This section introduces the results for both experimental scenarios explained in the previous section.

A. Parabolic Shaped Constrained Region

In order to minimize randomness in the results, a large number of simulations were run considering 200 random seeds. Thus, for each set of chosen parameters 200 exploration scenarios changing the random seed were run. The considered design parameters were the exploration mode (autonomous or socially reinforced) and the scaling factor for the social threshold $\alpha_t = [0.8, 0.9, 0.95, 0.98, 0.99, 0.999999, 1]$. Each simulation consists of 100 experiments to initialize M_{SM} and M_{SS}, 100 experiments to initialize M_{IM} and 10K exploratory experiments. The 200 simulations are subdivided in two groups of 100 simulations each. For the first subgroup, the relevant social sensory units shown in Figure 2 are used to evaluate the sensorimotor model every 500 samples during exploration. For the second group, a set of 441 points evenly distributed along the allowed region of the parabolic shaped region are considered to perform the evaluation every 500 samples.

Regarding the remaining parameters of the model, the initial threshold for all the sensor social units was set to 0.3, following the results in [6]. The minimum and maximum number of Gaussians in the sensorimotor model M_{SM}, which is an ilGMM, are 3 and 20, respectively. The model is trained every 120 experiments, and the maximum number of Gaussians that can be added to the model at each training step is 5. The forgetting rate for the sensorimotor model is set to 0.2 at the beginning but decreases logarithmically up to 0.05 after 10K experiments. The somatosensorimotor model M_{SS} is a weighted k-Nearest Neighbor (wNN), with $k = 3$. Finally, the interest model M_{IM} is the "discretized progress" model from the `explauto` library.

At the end of each simulation, the final sensorimotor model is evaluated against both datasets, the social dataset and the dataset covering the whole reachable parabolic region (whole dataset). For each simulation the mean evaluation error and the ratio of undesired configurations (when the somesthetic outcome is $p = 1$) are computed. Next, the average mean error e_{av} and the average ratio of undesired configurations $r_{uc,av}$ are computed by grouping the 200 simulations performed for each combination of parameters using different random seeds. In an equation form, the mean error average is written as:

$$e_{av} = \frac{1}{n_{rs}} \sum_{i=0}^{n_{rs}} \left[\frac{1}{n_{es}} \sum_{j=0}^{n_{es}} |s_{g,i,j} - s_{i,j}| \right], \qquad (2)$$

where n_{rs} is the number of random seeds considered, n_{es} is the number of evaluation samples in the dataset and the evaluation error of the j-th evaluation sample for simulation with the i-th random seed is $|s_{g,i,j} - s_{i,j}|$. Whereas, the average ratio of undesired configurations is written as:

$$r_{uc,av} = \frac{1}{n_{rs}} \sum_{i=0}^{n_{rs}} \left[\frac{1}{n_{es}} \sum_{j=0}^{n_{es}} p_{i,j} \right], \qquad (3)$$

where $p_{i,j}$ is the somesthetic outcome of the j-th evaluation sample for simulation with the i-th random seed.

The results of the simulations according to Equation (2) and Equation (3) are displayed in Table II. Moreover, Figure 4 and Figure 5 show the average results for the simulations in which the social and whole datasets were used for evaluation, respectively. The pictures in the figures show the evolution for the evaluation error, the exploration error, and the collisions ratio for the autonomous case and the best values of $\alpha_t = [0.8, 0.99]$, according to the numerical results in Table II. The size of the markers in the average evaluation error in the upper-left graphics is proportional to the standard deviation of the averaged values. Figure 5 only shows the evolution of the average evaluation error because the remaining plots were very similar to those of Figure 4, thus they do not provide new information for further discussion.

TABLE II:
Average results for the parabolic shaped area.

	Average error		Ratio of collisions	
	Social	Whole	Social	Whole
Autonomous	0.1359	0.0981	0.1525	0.1734
Social/α_t	Social	Whole	Social	Whole
1.0	0.1191	0.0878	0.1263	0.1699
0.999999	0.1204	0.0899	0.1350	0.1716
0.99	0.1153	**0.0788**	0.1412	0.168
0.98	0.1216	0.0804	0.1288	**0.1589**
0.95	0.1245	0.0828	0.1300	0.1688
0.9	0.1188	0.0841	0.1450	0.1701
0.8	**0.1144**	0.0806	**0.1225**	0.1663

Note: Results averaged over simulations considering two groups of simulations. The first group is evaluated against a **Social** sensory units, and the second group is evaluated against a dataset distributed over the whole reachable sensory space (**Whole**). α_t is the social threshold scaling factor, 'Average error' is the evaluation error, 'Ratio of collisions' is the ratio of undesired motor configurations. The best results are written in bold fonts.

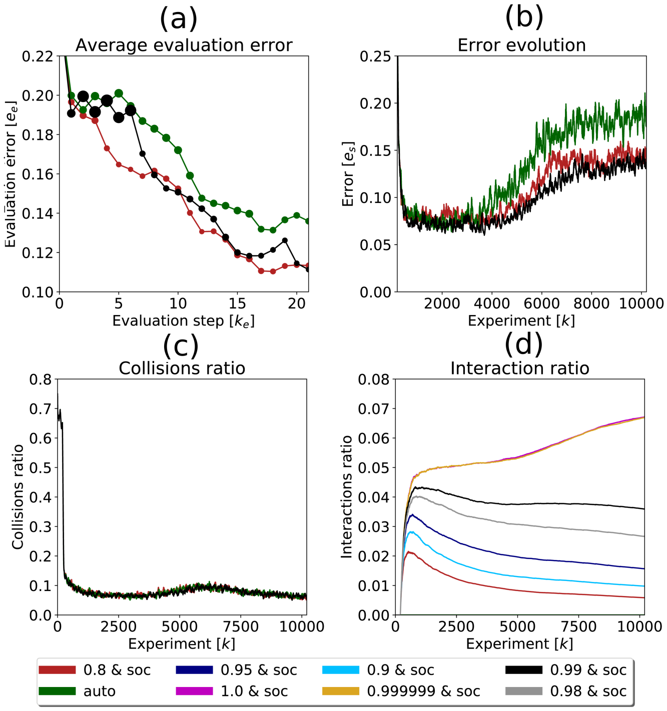

Fig. 4: Simulation results for the parabolic shaped region. (a) Average evaluation error over simulations with the same parameters considering the social relevant sensory units. (b) Average exploratory error over simulations with the same parameters. (c) Average ratio of collisions (constraint violations) over simulations with the same parameters. (d) Average ratio of interactions over simulations with the same parameters for social learners. It is observed that the autonomous agents produce higher average evaluation and exploration errors.

TABLE III:
Average results of explorations with the vocal tract.

n_v	α_t	e_{av}	$r_{uc,av}$	Vol.	V–/–V	VTV	% in.
50	-	1.51	0.71	1.029	**0.58**	0.10	-
50	0.999999	**1.31**	**0.62**	**1.039**	0.36	**0.35**	16.0
123	-	1.61	0.77	**1.029**	**0.58**	0.10	-
123	0.93	**1.49**	**0.72**	0.976	0.49	**0.13**	0.5
223	-	1.62	0.78	**1.029**	**0.58**	0.10	-
223	0.99	**1.47**	**0.69**	0.988	0.42	**0.32**	4.1
323	-	1.61	0.78	**1.027**	**0.57**	0.10	-
323	0.96	**1.39**	**0.73**	0.965	0.50	**0.16**	1.5

Note: Results averaged over simulations. n_v is the number of vowel combinations known by the instructor, α_t is the social threshold scaling factor, e_{av} is the evaluation error, $r_{uc,av}$ is the ratio of undesired motor configurations, 'Vol.' is the volume of the convex-hull described by the explored region over the formant frequency dimensions, 'V–/–V' is the final proportion of *single vowels* vocalizations, 'VTV' is the final proportion of *sequence of vowels* vocalizations and '% in.' is the final percentage of interactions. The best results are written in bold fonts.

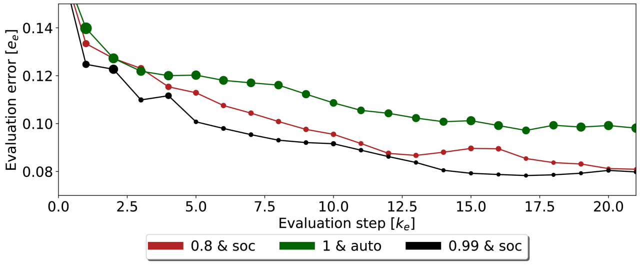

Fig. 5: Average evaluation error over simulations with the same parameters for evaluation considering the whole reachable sensor space of the parabolic shaped region.

B. Vocal Tract

Due to computational costs, in the case of the vocal tract only 6 random seeds were considered for each parameter set simulated. However, in this case three design parameters were considered: the exploration mode, the scaling factor for the social threshold $\alpha_t = [0.93, 0.96, 0.99, 0.999999, 1]$, and the number of vowel combinations $n_v = [50, 123, 223, 323]$ in the social relevant sensor set $\mathbf{S}$. Each simulation consists of 1K experiments to initialize M_{SM} and M_{SS}, 1K experiments to initialize M_{IM} and 100K exploratory experiments. Evaluation against $\mathbf{S}$ is performed every 2.5K samples during each simulation.

Regarding the remaining parameters of the model, the initial threshold for all the sensor social units was set to 0.5, following the results in [6]. The minimum and maximum number of Gaussian components in the sensorimotor model M_{SM}, which is an ilGMM, are 3 and 30, respectively. The model is trained every 400 experiments, and the maximum number of Gaussians that can be added to the model at each training step is 10. The forgetting rate for the sensorimotor model is set to 0.2 at the beginning but decreases logarithmically up to 0.01 after 100K experiments. The somatosensorimotor model M_{SS} is a weighted k-Nearest Neighbor (wNN), with $k = 3$. Finally, the interest model M_{IM} is the "tree" model from the `explauto` library.

At the end of each simulation, the final sensorimotor model is evaluated against $\mathbf{S}$, the ratio of undesired configurations r_{uc} is computed. The next step is to group the six simulations performed for each combination of the three design parameters and compute the average mean error e_{av} and the average ratio of undesired configurations $r_{uc,av}$ as indicated in Equation (2) and Equation (3). Looking at the results for different n_v, simulations with the lowest error average evaluation error e_{av} are collected in Table III.

Table III also contains a column indicating the ratio of undesired collisions for the evaluation set $\mathbf{S}$; the volume of a convex hull computed over the explored sensor space over the formant frequencies dimensions; the ratio of *single vowels* vocalizations (vocalization in which one of the perceptual windows was voiceless) and the ratio of *sequence of vowels* (vocalization in which sound occurs in both perceptual windows). Finally, the average percentage of interactions along the explorations for each set of design parameter is also shown.

In order to analyze the results in terms of exploration and vocal development as in [5, 6, 34], Figure 6 and Figure 7 provide relevant information. First of all, as in [5], Principal Component Analysis (PCA) is performed over all the formant frequencies dimensions of the sensory space considering all the data obtained during all the performed explorations. The two first principal components contributing with 97.3% of the information, 50.6% and 46.7%, respectively. The PCA transformation over the first two principal component dimensions is performed for the scenarios described in Table III. For the autonomous agents, as all the four scenarios produced similar results, PCA transformation is performed with the results when evaluating with $n_v = 123$. Once the transformation are performed, Gaussian Kernel Density Estimation (GKDE) is performed in order to observe the distribution of the explored samples over the principal components which are shown in Figure 6.

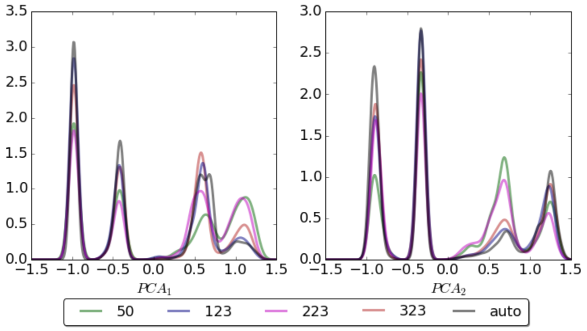

Fig. 6: Density distribution computed using Gaussian KDE. First, all the data obtained during simulations is concatenated and Principal Component Analysis performed. Secondly, KDE is performed over the first two PCA components. (Left) First Component. (Right) Second component.

In order to observe possible developmental transitions, Figures 7-9 were generated. In Figure 7, the average proportion of the three different kinds of vocalizations over explorations with the same simulation parameters are shown for the autonomous learners and the best results for the social agents with $n_v = [50, 223]$ because those are the cases with larger final ratios of *sequence of vowels* vocalizations. In 'Silence' or 'Voiceless' vocalizations, phonation does not occur. In *single vowels* vocalizations, phonation occurs only in one of the perception windows, and in *sequence of vowels* vocalizations, phonation occurs in both perception windows. Figure 8 shows as well the proportion of each type of vocalization but for individual simulations considering $n_v = 50$ and $\alpha_{th} = 0.999999$ as it is the case with larger ratio of interactions. Figure 9 is generated as well for individual simulations considering the same parameters, the figure displays the ratio of collisions and interactions in order to analyze their relation with the proportion of vocalization types.

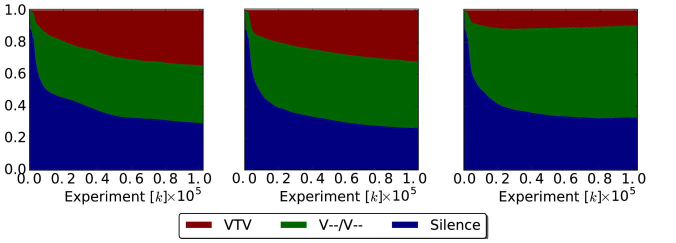

Fig. 7: Proportions of vocalization classes. (Left) $n_v = 50$, $\alpha_t = 0.999999$. (Center) $n_v = 223$, $\alpha_t = 0.99$. (Right) autonomous.

VII. Discussion

A. Parabolic Shaped Constrained Region

The first fact which can be observed in Table II is that any socially reinforced scenario returns better results than autonomous scenarios, considering both evaluation scenarios in terms of average mean errors and collision ratios. A correlation test was performed to determine whether the scaling factor α_t was relevant for the final results. Using the Pearson's correlation test between α_t and the average error when evaluating with the social and the whole datasets, the results suggested statistical insignificance. In general, $\alpha_t = 0.8$ produced the best results in terms of the social evaluation scenario. Even thought the results for all the scenarios were very similar, the large number of simulation runs guarantees a certain degree of conservativeness. Despite finding a good value for α_t, we did not find a mechanism to tune this value other than trial and error.

Regarding Figure 4, a first look at the error evolution (upper, right) allows to observe that after the agent takes advantage of regions were learning might seem easy, it is pushed to move to regions where the error increases in order to keep the progress in competence. This part of the exploration also coincides with a slight but perceptible increase of collisions (see bottom, left), indicating that, as could be expected, regions close to constraints are harder to learn and competence progress is slow there, thus the agent first explores other regions. Another evidence that supports this last conclusion is the fact that the agent with $\alpha_t = 1$, keeps increasing the level of interactions (bottom, right). As social relevant units are close to constraints, increments in imitations suggests that the agent is exploring close to constrained regions. Looking also at the evolution of interaction ratios, it is possible to observe the drastic effect of α_t on the ratio of interactions, also expected. However, more interestingly to remark, is the fact that experiments with $\alpha_t = 0.8$ obtained the best results when evaluated with the social dataset, but they are the ones with less interactions. This result suggests that the quantity of interactions, at least in this simple example, is not the most relevant factor for social development.

Finally, looking at Figure 4 and Figure 5, it is observed that the average evaluation error decreases until reaching a sort of minimum level. Compared to the social dataset, the agents reach a minimum value faster for the whole dataset and stay around that minimum value through the rest of the agent's life. In this case, marks indicating the standard deviation magnitude offer information, showing that those architectures with social reinforcement achieved more conservative results. The fact that the best results were found with $\alpha_t = 0.8$ and $\alpha_t = 0.99$ does not lead to any conclusions regarding the role of α_t, even considering the correlation test. However, one conclusion from the obtained results is that it is possible to find a value for α_t that generates better results than when this parameter is not considered, as for the architecture in [6].

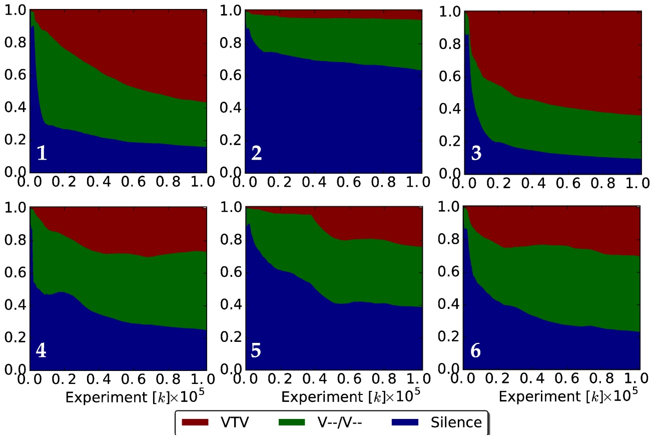

Fig. 8: Proportions of vocalization classes for individual simulations considering $n_v = 50$ and $\alpha_t = 0.999999$.

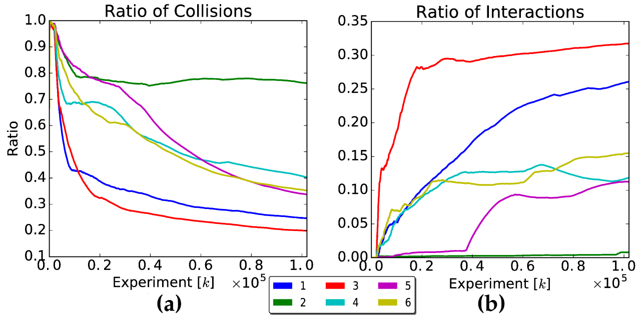

Fig. 9: Trajectories for ratio of collisions and ratio of interactions along individual simulations for $n_v = 50$ and $\alpha_t = 0.999999$.

B. Vocal Tract

First of all, looking at Table III, it can be observed that, similarly to the parabolic shaped region system, socially reinforced agents outperform autonomous systems, obtaining lower errors, less undesired configurations, larger explored volumes, and also larger proportion of *sequence of vowels* vocalizations. However, it is hard to establish a direct relation between α_t and performance. Nevertheless, it is possible to find a value for this parameter that outperforms the case for $\alpha_t = 1$, considered in [6]. Regarding the percentage of interactions, a large number of interactions is not necessarily leading to better results in the social task. For instance, for $n_v = [123, 323]$ the best results, in terms of average evaluation error, were obtained for values of α_t that led to small ratios of interactions. However, small ratios of interaction also were found to lead to explored regions with smaller volumes. In fact, for $n_v = [50, 223]$ and $n_v = [123, 323]$ with other values of α_t, when larger ratios of interaction were produced, convex hulls describing explored regions were larger.In general, socially reinforced agents explored smaller regions than autonomous agents, or at most regions with similar volumes.

Looking at Figure 6, it can be observed that the autonomous agent is the system leading to the highest peaks. Whereas the exploration occurs along the same regions, it is more uniform when $n_v = [50, 223]$ is considered. This uniformity might be seen as the larger number of *sequence of vowels* vocalizations. For instance, results in Table III with $n_v = [123, 323]$ show a proportion of *sequence of vowels* vocalizations similar to that of the autonomous systems, and Figure 6 shows a similar pattern of exploration between the three scenarios.

In [4], the authors proposed an emulation mechanism to study environmental language influence over intrinsically motivated exploration. They consider that the learner had awareness of two adult auditory productions. If the intrinsic motivations suggested that social learning was better to improve competence, then the learner chose one of the two auditory units randomly and attempted to generate it. It was observed that at the beginning, the learner preferred autonomous exploration, but after a while a social guided stage emerged, which vanishes as the agent learns to produce those adult vocalizations. Different to [4], in our architecture we consider the social interaction, as the way to drive social learning, but intrinsic motivations indirectly define which sensory regions to explore, regions where social learning is likely to occur or regions without social interest. In [6], a behavior was found in socially reinforced agents that suggested the onset of a socially guided developmental stage. There, it was observed that at the beginning, learners endowed with the somesthetic mechanism do not imitate too frequently the instructor response, due to the lack of knowledge on how to imitate without reaching undesired configurations. As far as the learner continues exploring and discovers regions where attempting imitation is not likely to produce undesired configurations, then the amount of interactions increases dramatically. In this work, the same effect is observed, however analysis focuses on the developmental processes from voiceless vocalizations to *sequence of vowels* vocalizations. Regarding the evolution of proportions of vocalization classes, an interesting fact is observed.

Figure 7 shows the average tendency of proportion of vocalizations through simulations. In general, it is observed that autonomous learners decrease the proportion of silent vocalizations but predominately produce *single vowels* vocalizations. On the other hand, socially reinforced learners achieve larger improvements in the production of *sequence of vowels* vocalizations. In order to have a clear image of developmental transitions we look at individual simulations in order to observe possible abrupt transitions in proportion of vocalizations, ratio of collisions and ratio of interactions. The scenario with $n_v = 50$ and $\alpha_{th} = 0.999999$ is described by Figures 8-9, which were generated considering the six simulations with different random seeds.

Figures 8-9 help us to corroborate the findings of [6] regarding the ratio of interactions and its relation with the ratio of collisions and the proportion of each type of vocalization. In individual simulations, there is a clear evidence of the onset of social imitation when constraint awareness is considered in the social reinforced architecture. At the beginning learners are not imitating too frequently the instructor response. Under our current setup, this unwillingness to imitate may be attributed to the somesthetic mechanism. In other words, if the somesthetic prediction in line 16 of Algorithm 1 indicates that an undesired configuration is likely to occur, then imitation does not occur. However, as the agent continues exploring and discovers regions where attempting imitation is not likely to produce undesired configurations, then the amount of interactions increases dramatically, the most significant examples are the learners 5 and 6 in Figures 8-9. For the instructor, who is unaware of the learner internal cognitive processes, this increment of interactions might be seen just as a spontaneous desire of the learner for social interaction, in other words as the onset of a socially guided developmental stage. This developmental transitions also suggest that, as it is seen in mother-children imitation scenarios, interaction is a good predictor in the emergence of *sequence of vowels* vocalizations.

Finally, interesting results can be observed at the best imitation scenarios. During the final evaluation, the best 10 imitation scenarios are considered and videos are generated. For each n_v value, videos are separated in social and autonomous exploration groups[3]. Videos show that many scenarios of imitation are good when considering acoustic features, nevertheless when observing the motor configuration, it can be observed that different vocal tract shapes lead to similar auditory results[4]. Thus, the redundancy of the system is demonstrated as well as the ability of the sensorimotor exploration architecture to deal with it.

VIII. Conclusions and Future Work

Maternal responses variations to different types of prelinguistic vocalizations as a function of context, responding mainly to speech-like sounds, suggest that mothers responding as if children were approximating a word may support language development (*imitation/expansion* responses) [17]. Herein, inspired in these responses during social prelinguistic development, a socially reinforced intrinsically motivated exploration architecture was introduced. Results from experimentation suggest that social reinforcement is crucial to the emergence of *sequence of vowels* vocalizations. The novel architecture is compared with those presented in former works, where somatosensation and intrinsic motivation roles were studied. A need for studying mechanisms of social development in parallel to vocal development mechanisms has been established.

The study of language must evolve in two directions, more realistic speech architecture and social scenarios. Even though

[3] Videos are available on https://doi.org/10.6084/m9.figshare.c.3921718.v2

[4] E.g., look at the 25th, 31st, and 33rd video of social agents considering $n_v = 123$.

the new architecture has advantages over previous ones, future works should consider unstructured vocalizations that would allow to study canonical babbling that requires the production of supragottal consonants and more realistic speech perception. Shorter perception windows, using for example Mel Frequency Cepstral Coefficient as [8], must be considered. Finally, investigations must consider more realistic social scenarios attempting to cover other categories of maternal response and infants' vocalization directionality as defined in [16].

Acknowledgment

J.M. Acevedo-Valle acknowledges the financial support provided by CONACYT grant 216872. Authors acknowledge the contributions made to this project by Karla Trejo, Jennifer Nguyen and Clément Moulin-Frier.

References

[1] R. Pfeifer and C. Scheier, *Understanding intelligence*. MIT press, 1999.

[2] R. Pfeifer, M. Lungarella, and F. Iida, "Self-organization, embodiment, and biologically inspired robotics," *Science*, vol. 318, no. 5853, pp. 1088–1093, 2007.

[3] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011.

[4] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in psychology*, vol. 4, 2013.

[5] J. M. Acevedo-Valle, C. Angulo, and C. Moulin-Frier, "Autonomous discovery of motor constraints in an intrinsically motivated vocal learner," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 314–325, June 2018.

[6] J. M. Acevedo-Valle, V. V. Hafner, and C. Angulo, "Social reinforcement in intrinsically motivated sensorimotor exploration for embodied agents with constraint awareness," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Sept 2017, pp. 255–262.

[7] S. Forestier and P.-Y. Oudeyer, "A unified model of speech and tool use early development," in *39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, 2017.

[8] S. Najnin and B. Banerjee, "A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production," *Speech Communication*, vol. 92, pp. 24–41, September 2017.

[9] J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King, "Mothers provide differential feedback to infants' prelinguistic sounds," *International Journal of Behavioral Development*, vol. 30, no. 6, pp. 509–516, 2006.

[10] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

[11] G. Schillaci, V. V. Hafner, and B. Lara, "Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents," *Frontiers in Robotics and AI*, vol. 3, p. 39, 2016.

[12] H.-C. Hsu and A. Fogel, "Infant vocal development in a dynamic mother-infant communication system," *Infancy*, vol. 2, no. 1, pp. 87–109, 2001.

[13] M. H. Goldstein, A. P. King, and M. J. West, "Social interaction shapes babbling: Testing parallels between birdsong and speech," *Proceedings of the National Academy of Sciences*, vol. 100, no. 13, pp. 8030–8035, 2003.

[14] M. H. Goldstein, J. A. Schwade, and M. H. Bornstein, "The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers," *Child development*, vol. 80, no. 3, pp. 636–644, 2009.

[15] D. K. Oller, E. H. Buder, H. L. Ramsdell, A. S. Warlaumont, L. Chorna, and R. Bakeman, "Functional flexibility of infant vocalization and the emergence of language," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6318–6323, 2013.

[16] J. Gros-Louis, M. J. West, and A. P. King, "Maternal responsiveness and the development of directed vocalizing in social interactions," *Infancy*, vol. 19, no. 4, pp. 385–408, 2014.

[17] J. Gros-Louis, M. J. West, and A. King, "The influence of interactive context on prelinguistic vocalizations and maternal responses," *Language Learning and Development*, vol. 12, no. 3, pp. 280–294, 2016.

[18] V. V. Hafner and G. Schillaci, "From field of view to field of reach - could pointing emerge from the development of grasping?" *Frontiers in Computational Neuroscience*, no. 17, 2011.

[19] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, no. 2, pp. 135–169, 2006.

[20] V. C. Ramenzoni and U. Liszkowski, "The social reach," *Psychological Science*, vol. 27, no. 9, pp. 1278–1285, 2016, PMID: 27481910.

[21] A. F. Morse and A. Cangelosi, "Why are there developmental stages in language learning? a developmental robotics model of language development," *Cognitive Science*, vol. 41, pp. 32–51, 2017.

[22] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and language*, vol. 96, no. 3, pp. 280–301, 2006.

[23] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Networks*, vol. 38, no. 0, pp. 64 – 75, 2013.

[24] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793–809, 2009.

[25] C. Moulin-Frier and P.-Y. Oudeyer, "Learning how to reach various goals by autonomous interaction with the environment: unification and comparison of exploration strategies," in *1st Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM2013), Princeton University, New Jersey*, Princeton, United States, October 2014.

[26] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[27] J. M. Acevedo-Valle, C. Angulo, N. Agell, and C. Moulin-Frier, "Proprioceptive feedback and intrinsic motivations in early-vocal development," in *18th International Conference of the Catalan Association for Artificial Intelligence*. IOS Press, 2015.

[28] D. Luo, F. Hu, Y. Deng, W. Liu, and X. Wu, "An infant-inspired model for robot developing its reaching ability," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, September 2016, pp. 310–317.

[29] N. Navarro-Guerrero, R. J. Lowe, and S. Wermter, "Improving robot motor learning with negatively valenced reinforcement signals," *Frontiers in neurorobotics*, vol. 11, 2017.

[30] D. Corbetta, S. L. Thurman, R. F. Wiener, Y. Guan, and J. L. Williams, "Mapping the feel of the arm with the sight of the object: on the embodied origins of infant reaching," *Frontiers in psychology*, vol. 5, p. 576, 2014.

[31] B. Franklin, A. S. Warlaumont, D. Messinger, E. Bene, S. Nathani Iyer, C.-C. Lee, B. Lambert, and D. K. Oller, "Effects of parental interaction on infant vocalization rate, variability and vocal type," *Language Learning and Development*, vol. 10, no. 3, pp. 279–296, 2014.

[32] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic Motivation Systems for Autonomous Mental Development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, April 2007.

[33] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in Cognitive Sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[34] C. Moulin-Frier and P.-Y. Oudeyer, "The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study," in *Interspeech*, 2013.

[35] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[36] C. Moulin-Frier, P. Rouanet, and P.-Y. Oudeyer, "Explauto: an open-source python library to study autonomous exploration in developmental robotics," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2014, pp. 171–172.

[37] J. M. Acevedo-Valle, K. Trejo, and C. Angulo, "Multivariate regression with incremental learning of gaussian mixture models," in *20th International Conference of the Catalan Association for Artificial Intelligence*, 2017.

[38] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.