Autonomous Discovery of Motor Constraints in an Intrinsically Motivated Vocal Learner

Juan Manuel Acevedo-Valle, Cecilio Angulo, and Clement Moulin-Frier

***Abstract*—This paper introduces new results on the modeling of early vocal development using artificial intelligent cognitive architectures and a simulated vocal tract. The problem is addressed using intrinsically motivated learning algorithms for autonomous sensorimotor exploration, a kind of algorithm belonging to the active learning architectures family. The artificial agent is able to autonomously select goals to explore its own sensorimotor system in regions, where its competence to execute intended goals is improved. We propose to include a somatosensory system to provide a proprioceptive feedback signal to reinforce learning through the autonomous discovery of motor constraints. Constraints are represented by a somatosensory model which is unknown beforehand to the learner. Both the sensorimotor and somatosensory system are modeled using Gaussian mixture models. We argue that using an architecture which includes a somatosensory model would reduce redundancy in the sensorimotor model and drive the learning process more efficiently than algorithms taking into account only auditory feedback. The role of this proposed system is to predict whether an undesired collision within the vocal tract under a certain motor configuration is likely to occur. Thus, compromised motor configurations are rejected, guaranteeing that the agent is less prone to violate its own constraints.**

***Index Terms*—Active learning, early vocal development, Gaussian mixture models (GMMs), intrinsic motivations, sensorimotor exploration.**

I. Introduction

In recent years, there has been an increasing interest in using robots to perform daily life activities in the presence of humans. As robot–human interactions become common then human-like communication systems become more relevant to robotics. Speech is one of the most studied communication systems because it allows human-spoken language. However, as mentioned in [1], the idea that speech is a deeply encrypted "code" prevails among the speech specialists and cracking this code is still an unsolved problem. Some of the mysteries about speech might be solved if we are able to understand all the mechanisms underlying early speech acquisition in children. Thus, this paper, provides new results to contribute to the study of early speech development using machines.

Developmental robotics is a relatively novel approach, it aims at understanding and modeling the role of developmental processes in the emergence of complex behaviors, including social ones. Its goal is twofold, on the one hand it is used to build more efficient cognitive machines applying developmental theories, and on the other hand it also provides insights into human developmental mechanisms, especially during infancy. A deeper understanding of these mechanisms would explain how human beings develop from infancy to functional adults capable of solving highly complex cognitive tasks [2].

Autonomous robot design could notably benefit from the available knowledge of biological science and self-organization theories [3]. Deep understanding of the embodiment paradigm is paramount to integrate that knowledge into robotics. This paradigm is also well represented by the quote "understanding by building" [4]. It states that the behavior of an agent is not only the result of a system control structure, but also a result of complex interactions with its ecological niche, its morphology, and its material properties [3], [4].

In this paper, language emergence is studied according to behavioral and neurophysiological evidence, moreover the role of motor constraints is especially considered. The main assumption is that early vocal development can be studied as a result of embodiment, self-organization, and emergence mechanisms produced by human evolution. In general, studies have shown that infants show preparedness to acquire natural language. Motor, perceptual, social, and learning ability constraints, and their maturation during infant development play a key role in the emergence of language [1].

Equally important, machine learning techniques have rapidly evolved, providing developmental robotics with interesting approaches as active learning. In contrast to the more usual passive learning algorithms, active learning data are collected in order to minimize a given property of the learning process, e.g., the uncertainty [5] or the prediction error [6] of a model. This family of algorithms is of particular interest for developmental robotics. During sensorimotor exploration they allow the agent to focus on parts of the sensorimotor space in which exploration is expected to improve the quality of the learned model [7].

Manuscript received July 19, 2016; revised November 18, 2016 and February 20, 2017; accepted April 14, 2017. This work was supported by the PATRICIA Research Project through the Spanish Ministry of Economy and Competitiveness under Grant TIN2012-38416-C03-01. The work of J. M. Acevedo-Valle was supported by CONACYT under Grant 216872. *(Corresponding author: Juan Manuel Acevedo-Valle.)*

J. M. Acevedo-Valle and C. Angulo are with the GREC Research Group, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain (e-mail: juan.manuel.acevedo.valle@upc.edu).

C. Moulin-Frier was with Flowers team, Inria/ENSTA-Paristech, 33405 Bordeaux, France. He is now with the SPECS Laboratory, Universitat Pompeu Fabra, 08018 Barcelona, Spain.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2017.2699578

The contribution of this paper is extending the study of early language development using intrinsically motivated exploration algorithms. Herein, we provide new simulation results showing the suitability of these algorithms in the self-exploration of sensorimotor vocal spaces. The theoretical basis of the probabilistic models used to represent knowledge is also provided. Furthermore, we propose an architecture that could be used to study the role of constraints during sensorimotor exploration in embodied agents. Finally, it is worth mentioning that the learning algorithm presented herein could be applied to any system subjected to constraints in order to improve learning progress.

The remainder of this paper is organized as follows. Section II introduces related works. Section III highlights the role of intrinsic motivations and proprioceptive feedback in vocal development. The experiment setup is described in Section IV and results are presented and discussed in Section V. Finally, the conclusions are presented in Section VI.

II. Related Work

This paper revisits and expands the investigation introduced in [8] and [9]. In [9], an intrinsically motivated exploration architecture was proposed for the study of the developmental stages emergent during the early vocal development of infants. For the experimentation the simulated ear-vocal tract model DIVA [10] was used. In spite of the relevance of its results, the motor constraints and the somatosensory system were neglected in [9]. However, morphological constraints play a key role in speech acquisition. Therefore, a new exploration algorithm proposed in [8] to incorporate motor constraints awareness using a somatosensory model. In the past, some studies have tried to explain the emergence of developmental stages during the vocal development, assuming their existence, but those stages were bridged using hard-coding for experimentation [10]–[13].

In [14], an approach for inverse kinematics learning in redundant systems was presented. It was demonstrated that goal babbling can be advantageous in learning in the early stages of development, as observed in developmental theories. In parallel, [15] presented an intrinsically motivated goal exploration approach for the active learning of inverse models. This approach was applied to the vocal sensorimotor space exploration in [9] and [16]. The algorithm considered in this paper extends intrinsically motivated exploration in the goal space to include motor constraints. Considering both motor and perceptual constraints during learning and exploration is crucial to design cognitive architectures for motor control.

Among the efforts to model the acquisition of speech there is the DIVA model [10]. It aims to imitate the underlying neurophysiological mechanisms for speech acquisition and production. The cognitive architecture of the system is an artificial neural network. The model includes the premotor, motor, auditory and somatosensory cortical areas, and simulated ear-vocal tract system. In [10], the somatosensory model was effectively integrated into the acquisition and production of speech processes. It was not used as an element to integrate motor constraints but as an extra source of sensory-feedback. The ear-vocal tract component of the DIVA model is used in this paper, as it was in [8] and [9].

Finally, another interesting contribution was the active learning architecture presented in [17] which considered time constraints. This paper proposed a music performance imitation scenario and implemented a learning architecture able to learn a musical instrument model and a body capabilities model; the architecture is also able to imitate a sequence of sound, while simultaneously kinematic errors, due to the control architecture, are corrected. Similar to [9], models employed in [17] were based on Gaussian mixture models (GMMs).

III. Early Vocal Development in Machines

Human speech production is one of the most complex motor acts performed by any living being [18]. Producing a linguistic message that can be understood by another human requires coordinating many degrees of freedom in the respiratory, laryngeal, and supraglottal articulatory system.

How infants acquire the complex ability to control speech production and in general how they learn language remains a matter of research [1]. It has been pointed out that strong regularities can be observed in the structure of the vocal development process independently of interindividual differences [1], [19]. In general, the infant first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds and finally automatically discovers and focuses on babbling with articulated proto-syllables. In [18], some experiments suggested that goals of speech movements are auditory in nature and maintenance of motor command maps to auditory results is performed with auditory feedback.

It is important to inquire into the developmental assumptions considered in the experiments in [8] and [9], as this paper is based on those experiments. Regarding the infant development stages mentioned in [1], our experiments consider the developmental stage known as canonical babbling (CB) [20] and the beginning of language-specific speech production [1]. Results suggest that during CB infants learn to control their ear-vocal tract system based on auditory feedback. Nevertheless, when infants begin to babble they do it regardless of the audibility of their vocalizations. CB could be the result of a natural tendency of infants to move their body parts rhythmically motivated by sensory feedback [20].

Consistent with the theory, we assumed a simplified explanation that the artificial agent is exploring its ear-vocal tract system choosing auditory goals and evaluating the result. Therefore, our cognitive architecture allows the agent to explore regions, where the competence to produce intended sounds is improved. However, we also endow the agent with autonomous mechanisms to discover constraints in order to drive the exploration. To accomplish that objective, previously proposed active learning architectures and the proprioceptive feedback concept are combined.

A. Intrinsically Motivated Exploration Architectures

Among the vast number of active learning architectures, this paper considers the exploration architecture proposed by [15]. This architecture reproduces the formalism of intrinsic motivation inspired by psychological literature as proposed previously in [21] and [22]. Using goal babbling, intrinsically motivated exploration aims to minimize the error of an agent to reach self-generated goals measured according to a competence function. This architecture allows artificial agents to efficiently and actively explore and generate maps from motor capacities to perceived results. Therefore, exploration occurs over regions in which agents perceive they are becoming more competent to reach self-generated goals. Intrinsically motivated exploration architectures were originally designed to actively learn inverse models of high-dimensional input–output spaces. This architecture was later extended by [9] to study self-organization in early vocal development stages in infants and robots.

Intrinsically motivated learning algorithms have shown favorable results in previous experiments to learn sensorimotor coordination skills in redundant nonlinear high-dimensional mappings which share many mathematical properties with vocal spaces. Moulin-Frier *et al.* [9] used a simulated ear-vocal tract system to study the emergence of developmental stages implementing intrinsically motivated exploration. They argued that the development of the agent self-organizes into vocal developmental sequences. The results presented therein opened the door to a new approach in vocal development to be explored. This paper introduces a methodology which enhances intrinsically motivated architectures with constraint awareness.

B. Proprioceptive Feedback

Some of the most adopted theories of speech state that speech production is organized in terms of motor control signals and their associated vocal tract configurations which has been corroborated by several experimental results [23], [24]. Nevertheless, we adopt the simplified hypothesis that speech goals are defined acoustically and maintained by auditory feedback [18]. CB is a rhythmic behavior that, with some differences, emerges in both, normally developing infants and infants with hearing loss. When infants start to babble, they do it regardless of the audibility result (i.e., they produce audible and voiceless vocalizations). However, evidence suggests that, around the onset of CB, infants learn to vocalize based on auditory feedback [1], [20].

How the somatosensory system[1] affects the ear-vocal tract exploration is an open question that was not previously approached in [9]. However, the relevance of the somatosensory system for speech has been shown in different experiments, for instance the results in deaf individuals suggest that somatosensory inputs related to movement play a more important role in speech production than what was thought before [25], [26]. Furthermore, the fact that CB also emerges in deaf infants suggests that somatosensory feedback must play a more relevant role during the prelinguistic vocal development in infants [27].

[1] Strictly, the somatosensory system is also a sensorimotor system. In future works, we will distinguish two sensorimotor systems: 1) the auditory-motor system and 2) the somatosensory-motor system.

Fig. 1. Examples of articulatory configurations that produce collisions in the DIVA vocal tract model.

In [28], a robotic device able to generate patterns of facial skin deformation related to certain speech productions was used. The results showed that when the facial skin is stretched whilst subjects are listening to words, the sounds they hear are altered. Thus, theory and results suggest that the somatosensory system is involved in speech perception. Following this hypothesis, improvements can made to the experiments proposed in [9] by including a somatosensory system to endow the learner with physical constraint awareness.

In [8], the foundations of a simplified architecture were established allowing us to include physical constraints to the learning process through a proprioceptive signal, similar to the ability to feel pain in humans. The open source DIVA model[2] [10] provides a synthesizer that represents the human vocal tract and ear systems. The DIVA model also includes a somatosensory system, but in spite of it, there is a lack of physical constraints in the DIVA vocal tract. The absence of constraints allows the execution of motor commands that lead to collisions or articulatory superpositions. Both circumstances lead to no phonation and moreover, the latter is a contradictory result since it lacks physical sense, as shown in Fig. 1.

To overcome the drawbacks caused by the lack of constraints, we introduce a somatosensory system. This new element, not considered in [9], is based on an area function which is a vector descriptor of the vocal tract shape. It consists of a mechanism that evaluates if an exploratory motor command produces a collision or superposition of articulatory tissues, the system generates a proprioceptive signal. Using the data generated with this mechanism, the agent builds a map from motor commands to proprioceptive results. This map is used to predict which motor commands may lead to undesired collisions, so they may be rejected, forcing the agent to choose a new auditory goal. In the next section, this mechanism is explained in detail.

IV. Proposed Architecture

The experimental architecture proposed in this paper to study the early vocal development in machines is shown in Fig. 2, where five elements interact. These elements are

[2] http://www.bu.edu/speechlab/software/diva-source-code/

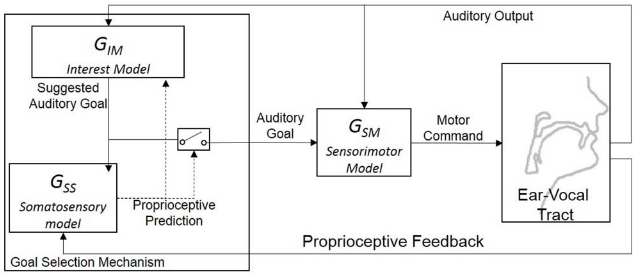


Fig. 2. Experimental architecture. It is composed by five interacting modules, two of them contained within the ear-vocal tract module (the sensorimotor system and the somatosensory system).

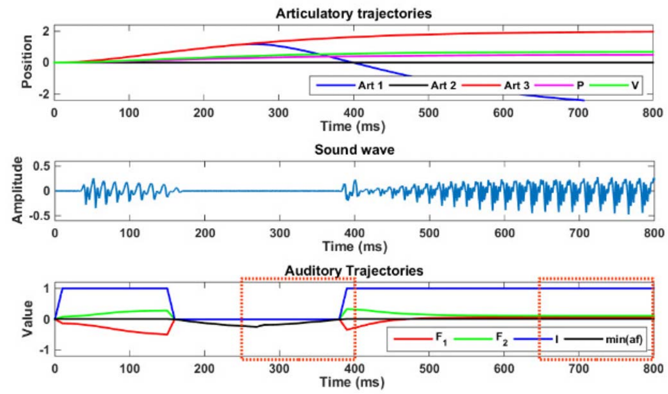


Fig. 3. Vocalization experiment structure. The upper plot shows the articulatory trajectories, from 0 to 250 ms, the commands for Art1, Art2, and Art3 are set to 2, 0, and 2, respectively, whereas the glottal pressure and voicing are both set to 0.5. From 250 to 800 ms, the commands for Art1, Art2, and Art3 are set to -3, 0, and 2, respectively, whereas the glottal pressure and voicing are both set to 0.7. The remaining motor commands are set to zero. The middle plot represents the speech sound wave signal. The bottom plot shows the auditory trajectories. The dotted outlined boxes represent the perception time windows from 250 to 400 ms and the second from 650 to 800 ms. The auditory output s are determined from the average of each trajectories along each one of the time windows. Whereas the proprioceptive feedback p is determined by the average value of $\min(a_f)$.

introduced below and explained in detail in the coming sections.

1) *Sensorimotor system* is a simulated ear-vocal tract. It corresponds to the physical properties of the embodied agent. For the present work the ear-vocal tract system of the DIVA model [10] is used.
2) *Somatosensory system* is a perceptual mechanism that evaluates the shape of the vocal tract. It generates a proprioceptive feedback signal indicating if an undesired contact or collision is produced into the vocal tract.
3) *Sensorimotor model* is a mathematical representation of the vocal tract-ear model. It endows the artificial agent to map motor commands to auditory effects using the data collected from the agent's own vocalizations.
4) *Somatosensory model* is a mathematical representation that maps motor configurations to their likely proprioceptive feedback to acquire self-awareness of its own physical constraints in order to avoid executing motor configurations that produce undesired behaviors.
5) *Interest model* for auditory goals allows the agent to actively choose auditory goals in order to improve the quality of its sensorimotor model based on a certain measure of competence. This model represents the core of the intrinsically motivated sensorimotor self-exploration.

A. Sensorimotor System

The DIVA vocal tract configuration is determined by the position of ten articulators and three phonation parameters. Along this paper, only seven articulators and two phonation parameters (voicing and glottal pressure) are considered [9]. Articulators and voicing parameter motor dynamics are modeled as overdamped second order systems

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

with $\zeta = 1.01$ and $\omega = (2\pi/0.8)$ representing the damping factor and the natural frequency, respectively. The duration of each vocal experiment in seconds is 0.8, whereas m and x represents the desired articulator position (motor command) and the current articulator position, respectively. During each vocal experiment two different motor commands are introduced for each of the seven articulators and the two voicing parameters: one for the 0–250 ms window and another for the remaining time. Thus, each motor command is represented by an 18-D vector. The auditory output of a human vocalization can be described by its formant frequencies. We consider the first two formant frequencies, F_1 and F_2, along with an intonation signal I. The intonation signal is 1 when phonation occurs and 0 otherwise, two conditions are required for phonation to occur: 1) the area function a_f of the vocal tract must be positive elsewhere and 2) the voicing and pressure parameters must be positive. The area function is a vector function that describes the transversal shape of the vocal tract.

During the vocalization the auditory output of the system is observed along two time windows, the first from 250 to 400 ms and the second from 650 to 800 ms. The value of each auditory output is averaged for each time window, the result is a 6-D output signal (two formants and the intonation, hence three values, per each of the two time windows). In Fig. 3, we reproduce the vocalization representation shown in [9]. To be consistent with the co-articulated nature of speech, only two perceptual windows are used [1]. However, since only two portions of the vocalization are considered, a lot of information is lost. For instance, it is shown in [23] the continuum of co-articulated gestures. Therefore, future works should consider studying the continuum of speech gestures and self-structuring of vocalizations.

B. Somatosensory System

In Fig. 3, it is shown that the area function a_f is observed during both perception time windows. The minimal value of the area function $\min(a_f)$ would be zero when the vocal tract is closed at any point and negative values mean that some tissues are overlapped, which does not have physical meaning. However, in some cases it might be interpreted as the tongue being bitten. In other cases it might represent high pressure between the tongue and the palate, which might be interesting to the learner in a realistic scenario, where motor constraints are not violated. In general, we made a strong assumption that any motor constraint violation over a threshold is uncomfortable or painful. Hence, the average value of $\min(a_f)$ in each perception time window is used to generate a proprioceptive feedback signal p: if the average of $\min(a_f)$ is lower than a threshold for any perception window, then the configuration is evaluated as a undesired collision with $p = 1$, and $p = 0$ otherwise.

C. Sensorimotor Model

GMMs are linear combinations of multivariate Gaussian distributions that represent clusters of data. They have been previously used to represent nonlinear redundant maps [17], [21], [29] in order to solve the inverse problem of inferring input motor commands from desired sensory outputs. GMMs can be learned using an online variant of the expectation-maximization (EM) algorithm in order to learn incrementally from incoming data [30]. Here, the algorithms used to train GMMs are based on the open source tools[3] associated with [30], and modified according to our problem requirements. The three models in the experimental setup are probabilistic representations in the form of GMMs, obtained using data collected from experiments with the DIVA ear-vocal tract. A detailed explanation of the GMMs training is provided below.

We assume that an n-dimensional input command space $X \in \mathbb{R}^n$ is mapped to an m-dimensional output space $Y \in \mathbb{R}^m$, through a transform function $y = f(x) + \varepsilon$, where $y \in Y$, $x \in X$ and ε is random noise. When a dataset of couples (x, y) is available, the EM-algorithm is used to obtain a GMM which is defined by the parameters $\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}$, where π_j, μ_j, and Σ_j are, respectively, the prior probability, the distribution centroid and the covariance matrix of the jth Gaussian, for $j = 1, 2, \ldots, K$, being K the number of Gaussian components. From [30], Gaussian mixture regression (GMR) is applied to compute the conditional probability distribution $P(X\,|\,y)$ in the input space X given a desired output y. Once it is computed, the value $x^* \in X$ is selected such that it maximizes $P(X\,|\,y)$.

To obtain the input x that maximizes the probability to produce the output y, the GMR process first defines the partitioned vector $z \in X \times Y$, where

$$z = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{2}$$

For each Gaussian j in the GMM the partitions

$$\mu_j = \begin{pmatrix} \mu_j^x \\ \mu_j^y \end{pmatrix} \quad \text{and} \quad \Sigma_j = \begin{pmatrix} \Sigma_j^x & \Sigma_j^{xy} \\ \Sigma^{yx} & \Sigma_j^y \end{pmatrix} \tag{3}$$

are considered to compute the conditional probability distribution $P_j(X\,|\,y) \sim N_j(\hat{\mu}_j, \hat{\Sigma}_j)$ in the input space X given a desired output y, where

$$\hat{\mu}_j = \mu_j^y + \Sigma_j^{yx}\left(\Sigma_j^x\right)^{-1}\left(x - \mu_j^x\right), \hat{\Sigma}_j = \Sigma_j^y + \Sigma_j^{yx}\left(\Sigma_j^x\right)^{-1}\Sigma_j^{xy}. \tag{4}$$

Considering that $P(X\,|\,y)$ is at its maximum when $x = \hat{x}_j = \hat{\mu}_j$, then a natural selection for x in order to produce y is $\hat{x}_j$. But we have K candidates for x, hence it is necessary to compute the probability of the vector $\hat{z}_j = [\hat{x}_j,\ y]^T$ belonging to its generator Gaussian as

$$P\left(\hat{z}_j\right) = \pi_j \frac{1}{\sqrt{(2\pi)^K |\Sigma_j|}} e^{-\frac{1}{2}\left(\left(\hat{z}_j - \mu_j\right)^T \Sigma_j^{-1}\left(\hat{z}_j - \mu_j\right)\right)} \tag{5}$$

and finally the point $z^* = \hat{z}_j$ that maximizes $P(\hat{z}_j)$ is selected as the point that better fits the model. In other words, according to our prior knowledge of $f(x)$, $z^* \in f(x)$, we infer that the output y is generated by $\hat{x}_j$.

Taking into account the above for the sensorimotor model, an 18-D motor command space M, with $m \in M$, is defined for the vocal tract articulatory configuration. A 6-D auditory output space S, with $s \in S$, is also defined, the agent being able to observe s according to $s = f(m) + \sigma$, where $\sigma \sim N(0, 0.01)$ is Gaussian noise. The aim is to find a GMM that solves the inverse problem $m = f^{-1}(s_g)$, where s_g is an auditory goal.

We define a GMM, G_{SM}, to model the sensorimotor system, with $X = M$ and $Y = S$. Such a model allows computation of the inverse model $P(M\,|\,s_g)$ using GMR. At the beginning of the experiment, m is selected either, randomly or according to the interest for initializing the inverse sensorimotor model $m \sim f^{-1}(s_g) \sim P(M\mid s_g)$ around a specific region of the sensorimotor space. After the initialization stage, the agent starts to select new auditory goals, according to the interest model explained below. In order to reduce memory storage requirements, we consider a generative method for the training stage, which means that the model is trained using the last N_{SM} samples obtained from experimentation along with

$$N_{\mathrm{old}} = \left\lceil \frac{(1-\alpha)N_{\mathrm{SM}}}{\alpha} \right\rceil \ \text{samples} \tag{6}$$

generated using G_{SM}, where $\alpha \in [0,\ 1]$ is the forgetting rate.

D. Interest Model for Auditory Goals

The interest model for auditory goals endows the learner the ability to select goals that maximize the expected competence progress in order to improve the quality of its sensorimotor model, resulting in better control over it. It is derived from the model proposed in [9]. The competence value for a goal is defined by

$$c = e^{-|s_g - s|} \tag{7}$$

where s_g is the auditory goal and s is the actual auditory production after executing a motor command $m \sim P(M\,|\,s_g)$. To construct the interest model, the auditory goal space is augmented with two extra dimensions: 1) the competence $c \in C$ and 2) time tag $t \in T$. The number of vocalizations N_{IM} considered to build the interest model is fixed. A GMM, G_{IM}, with K_{IM} components will be computed from the 8-D data set with N_{IM} samples of the augmented goal space. To initialize this model, some auditory results from the initialization of G_{SM} are selected as the first auditory goals s_g.

Those Gaussian components in G_{IM} that, according to the covariance matrices Σ_j, contain goals that will likely increase

[3] http://www.calinon.ch/sourcecodes.php

the competence progressively are considered to build a probabilistic distribution $P(S)$ over the auditory space. In order to build $P(S)$, the components in G_{IM} are weighted according to their time-competence covariance magnitudes. Thus, $P(S)$ will prioritize goals in regions, where competence is expected to increase. Finally, a sample s_g is drawn from $P(S)$ for the next vocalization experiment. Model training is performed every time the agent has performed n_{IM} experiments, using the last N_{IM} vocalizations.

E. Somatosensory Model

For the somatosensory model we consider the 18-D motor command space M, with $m \in M$, and a new binary proprioceptive output space $P = \{0, 1\}$, with $p \in P$. If a vocal production leads to undesired contacts, then $p = 1$, otherwise $p = 0$. A map g is assumed to exist such that $p = g(m)$ and the agent can observe p for each vocal experiment. Thus, it is possible to find a GMM G_{SS}, with $X = M$ and $Y = P$, that allows computation of the probability distribution $P(P \mid m)$ applying GMR, and determine when a motor command m is likely to lead to an undesired collision in the vocal tract.

The inverse sensorimotor model G_{SM} and the somatosensory model G_{SS} are initialized together. When an auditory goal s_g has been selected, m is computed using $P(M \mid s_g)$. Next, to predict the value of p, $P(P \mid m)$ is used. If the prediction suggests that m will produce $p = 1$ then s_g is rejected, otherwise s_g and m are accepted. If s_g is rejected, then $G_{\mathrm{IM}}(S)$ is recomputed without considering the Gaussian component in G_{IM} that generated s_g, this mechanism decreases the prior of the conflicting Gaussian in G_{IM}. The new $G_{\mathrm{IM}}(S)$ is used to select a new goal s_g, and the process is repeated until s_g is accepted. During the agent's life, the model G_{SS} is trained when G_{SM} is trained using the previously described generative mechanism.

F. Self-Exploration Algorithm

The self-exploration architecture with motor constraints self-awareness, first proposed in [8] for ear-vocal tract exploration, is an extended version of [9]. The algorithm associated with the cognitive architecture is shown in Algorithm 1. Our extended self-exploration algorithm with goal babbling and motor constraints self-awareness starts with the learner having no experience in vocalizing. Models G_{SM} and G_{SS} are initialized using random vocalizations with small values around the neutral position of the articulators. The neutral position of the pressure and voicing parameters are set to -0.25 to produce no phonation, whereas for the articulators it is considered 0, i.e., the rest position. Model G_{IM} is also initialized.

Then, in line 6 of Algorithm 1 the vocal learner agent selects a goal s_g for the next experiment according to the probabilistic distribution $P(S)$ and the motor command m is obtained using the inverse model G_{SM} in line 7. The main feature of this algorithm, different from similar architectures, is that in line 8 $P(P \mid m)$ provides a prediction for p that indicates if the selected motor command is likely to produce an undesired collision. From line 9 to 11, if $p \approx 1$, the goal is rejected and the probabilistic distribution $P(S)$ is updated, ignoring the

Algorithm 1 Self-Exploration With Goal Babbling and Self-Constraints Awareness

```
1:  Initialize G_SM and G_SS
2:  Initialize G_IM and i ← 1
3:  while i in [1, 1e5] do
4:      p_tmp ← 1
5:      while p_tmp do
6:          s_{g,i} ← G_IM(S)
7:          m_i ← G_SM(M|s_{g,i})
8:          p_tmp ← G_SS(P|m_i)
9:          if p_tmp then
10:             update(G_IM(S))
11:         end if
12:     end while
13:     s_i ← f(m_i) + σ and p_i ← g(m_i)
14:     c_i ← e^{-|s_{g,i} - s_i|}
15:     i ← i + 1
16:     if i mod N_SM = 0 then
17:         train(G_SM, m_{(i-N_SM+1:i)}, s_{(i-N_SM+1:i)})
18:         train(G_SS, p_{(i-N_SM+1:i)}, s_{(i-N_SM+1:i)})
19:     end if
20:     if i mod n_IM = 0 then
21:         train(G_IM, s_{g,(i-N_IM+1:i)}, c_{(i-N_IM+1:i)})
22:     end if
23: end while
```

Gaussian component in G_{IM} that generated s_g and the algorithm goes back to line 6. Otherwise, $p \approx 0$, both, s_g and m are accepted. Next, the motor command is executed with the vocal tract and the agent observes s and p in line 13. In line 14, the learner evaluates the competence value c. It also checks if we are at the end of a learning episode, so models G_{SM}, G_{SS}, and G_{IM} are updated in lines 17, 18, and 21, respectively. To provide objective evaluation elements, some experiments without considering the somatosensory model for choosing goals are also presented. In this later case, s_g is always accepted, thus line 4 is substituted with $p_{\mathrm{tmp}} = 0$ in Algorithm 1.

V. Experimental Results

Eighteen independent simulations using Algorithm 1 were run. All simulations consisted of half a million of vocalizations, including an initial vocalization set of 1000 random samples. Nine different random seeds were considered to generate the same number of motor command sets from a uniform distribution. The limits for those motor commands related to the vocal tract articulators were $[-1, 1]$, whereas for motor commands related to the phonation parameters were $[0, 0.7]$. Each set was used twice to initialize simulations of Algorithm 1, first without using the somatosensory model and second with it.

Considering as a reference the parameters used for simulations in [8] and [9], a few variations in their values were tested. First, when values for K_{SM} or K_{SS} are increased the inference error decreases slightly but the computation time grows considerably. On the other hand if these values are decreased, the inference error increases considerably. Second, if the training

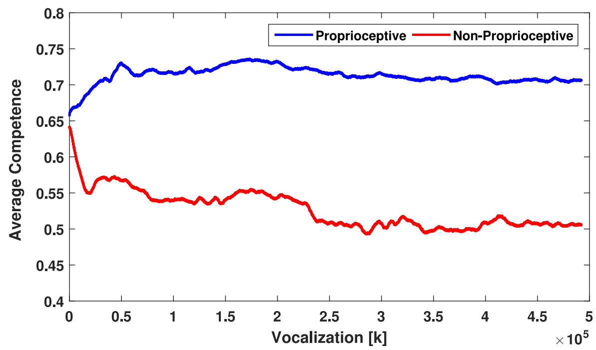


Fig. 4. Mean competence evolution during simulations using Algorithm 1 for nine different initialization data sets. Moving average of 5000 samples are considered to filter the results of each simulation. Results are shown in the case of proprioceptive and nonproprioceptive agents.

steps are increased for the somatosensory model and the sensorimotor model, then the training computational time increases as N_{old} in (6) increases proportionally, but no improvement is obtained in the inference error. However, if these values are decreased beyond the values used in [9], the mean inference error increases. Third, when α_{SM} and α_{SS} are larger than 0.1 the competence progress is slower. Finally, the parameters linked to the interest model allow a wider range of values to be chosen obtaining similar results. For K_{IM} we observed that values greater than or equal to 12 worked similarly, but smaller values negatively impacted the competence progress. Thus, the main parameters for all the simulations were kept as in [8] and [9] as they performed better than other simulations in terms of exploration results and simulation time. Summarizing, values were set to $K_{\mathrm{SM}} = 28$, $N_{\mathrm{SM}} = 400$, $K_{\mathrm{SS}} = 28$, $K_{\mathrm{IM}} = 12$, $N_{\mathrm{IM}} = 4800$, $n_{\mathrm{IM}} = 12$, and the continuous sampling time used for the DIVA ear-vocal tract was $t_s = 10$ ms. The forgetting rate parameter α_{SM} for G_{SM} starts from 0.1 and decreases logarithmically to 0.05 after half a million of vocalizations. On the other hand, α_{SS} for G_{SS} was chosen to be 0.05 through the whole simulation.[4]

During the simulation, G_{SM} and G_{SS} are initialized as indicated in line 1 of Algorithm 1 with the initial motor command sets. Then, all the initial phonatory productions are used as auditory goals to initialize the interest model G_{IM} as indicated in line 2 of Algorithm 1. In this stage, G_{SM} is used to infer the motor commands that will likely produce the initial auditory goals. These commands are executed without considering the proprioceptive prediction p.

A. *On Competence and Contacts*

First of all, Fig. 4 represents the evolution of the competence parameter c in (7) for self-generated auditory goals. To obtain this plot, first the result of each simulation is filtered using a 5000 samples moving average window. Next, simulations are divided into two general groups: 1) proprioceptive agents and 2) nonproprioceptive agents. Finally, the mean of all the filtered results for each group is computed. The same mechanism is considered to obtain the percentage of contacts observed in Fig. 5(a).

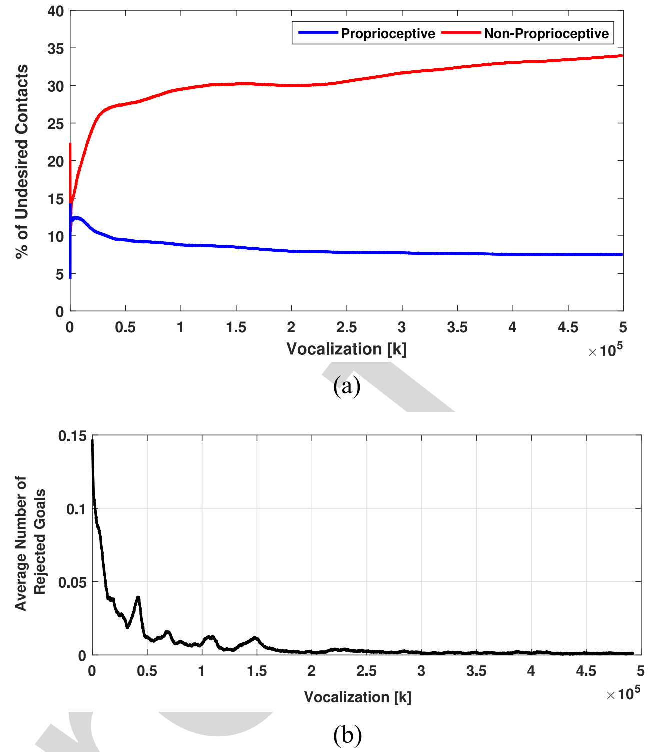


Fig. 5. Algorithm 1 simulation results. (a) Mean percentage of vocalizations producing undesired collisions considering all the simulated agents. Agents are grouped by proprioceptive and nonproprioceptive. The results of each agent are prefiltered considering a 5000 samples moving average. (b) Mean number of rejected goals using the proprioceptive prediction and considering all the simulated proprioceptive agents. The results of each agent are prefiltered considering a 5000 samples moving average.

TABLE I
RESULTS CONSIDERING ALL DATA FROM EXPLORATION

Experiment	Non-Proprioceptive			Proprioceptive		
	Vol.	mean(c)	% Contacts	Vol	mean(c)	% Contacts
1	0.58	0.50	41.67%	0.49	0.80	3.42%
2	0.47	0.50	37.50%	0.49	0.76	3.88%
3	0.49	0.50	43.53%	0.43	0.61	5.28%
4	0.47	0.54	30.83%	0.52	0.73	9.03%
5	0.47	0.54	29.17%	0.44	0.68	15.55%
6	0.56	0.59	23.01%	0.52	0.70	8.18%
7	0.49	0.55	30.86%	0.41	0.71	6.76%
8	0.57	0.54	35.49%	0.43	0.68	8.57%
9	0.49	0.50	32.42%	0.63	0.74	6.74%
Average	0.51	0.53	33.83%	0.48	0.71	7.49%
Min	0.47	0.50	23.01%	0.41	0.61	3.42%
Max	0.58	0.59	43.53%	0.63	0.80	15.55%

Note: Experiments with different vocalization initial sets for Proprioceptive and Non-Proprioceptive Agents. The volume of a the convex-hull described by the explored data, the mean value for the competence c, and the final percentage of contacts along the simulations are shown.

Tables I and II show the volume of a convex hull covering the explored auditory region. They also display the mean competence and the percentage of undesired contacts at the end of the simulation. First, in Table I, descriptors are computed considering all the vocalization during each simulation. Second, in Table II, figures were computed considering

[4]Supplementary downloadable material provided by the authors is available at https://dx.doi.org/10.6084/m9.figshare.c.3676645.v1. After each experiment, 20 random samples from the last 1000 vocalizations were drawn to generate videos with audio. Videos of the experiments 1, 5, and 9 are provided.

TABLE II
RESULTS WITHOUT CONSIDERING VOCALIZATIONS WITH UNDESIRED CONTACTS

Experiment	Non-Proprioceptive		Proprioceptive	
	Vol.	mean(c)	Vol.	mean(c)
1	0.48	0.69	0.39	0.81
2	0.38	0.67	0.39	0.78
3	0.37	0.71	0.33	0.63
4	0.40	0.65	0.42	0.77
5	0.36	0.66	0.37	0.76
6	0.44	0.67	0.42	0.74
7	0.36	0.69	0.32	0.74
8	0.45	0.67	0.35	0.71
9	0.38	0.64	0.49	0.78
Average	0.40	0.67	0.39	0.75
Min	0.36	0.64	0.32	0.63
Max	0.48	0.71	0.49	0.81

Note: Experiments with different vocalization initial sets for Proprioceptive and Non-Proprioceptive Agents. The volume of a the convex-hull described by the explored data and the mean value for the competence c along the simulations are shown.

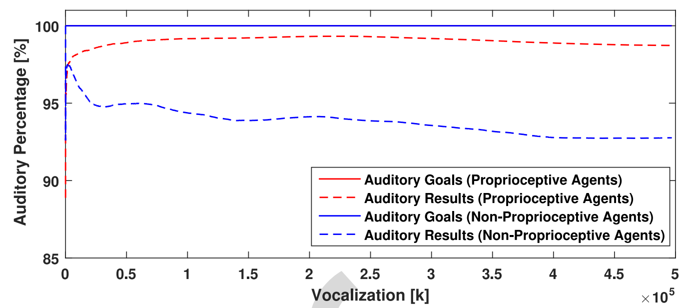


Fig. 6. Mean percentage of vocalizations producing phonatory result along all the simulations per each group, proprioceptive and nonproprioceptvie agents. The percentage of phonatory auditory goal is also shown per each group. *Note: red and blue solid lines are overlapped.*

vocalizations without undesired collisions. Convex hulls provide an insight regarding the size of the explored regions in the auditory-space. They are computed considering formant frequency dimensions F_{11}, F_{21}, F_{12}, and F_{22}.

In Fig. 4, results suggest that proprioceptive agents perform better than those which are not endowed with proprioception. We observe that at the beginning of the exploration the mean average competence is very similar for both groups. However, after the initialization the nonproprioceptive agents suffer an important decrement of the competence, which also coincides with a significant increment of the percentage of contacts in Fig. 5(a). Table I also confirms the expected results according to our hypothesis: using proprioceptive feedback drives artificial agents to produce significantly fewer undesired contacts and also increases the competence to reach self-generated auditory goals.

Some observers might ask the reason of high competence values at the beginning of the simulations. We argue that it is an expected result as the competence computation begins when G_{IM} is initialized with self-generated goals drawn from the initial auditory productions of the agent. In other words, G_{SM} and G_{SS} models are initialized around a set of initial vocalizations and auditory productions. The initial auditory productions are selected as auditory goals, then motor commands are computed with a sensorimotor model that represents very well those initialization samples. Later, as the agent explores the auditory space and it moves toward farther regions from those of initialization, the competence values might slightly decrease. This is due to the incremental learning of probabilistic models and depends on the values assigned to the forgetting rates α_{SM} and α_{SS}. If these forgetting rates are close to zero, then the agent is less prone to update its knowledge when new data are far from the current knowledge. On the other hand, if the forgetting rates are high, then the agent will adapt its model to the new data very fast but also it will forget faster its previous knowledge since it is not reinforced.

We also argue that one of the reasons the proprioceptive agents perform better is the null competence produced by nonphonatory vocalizations. The nonproprioceptive agent produces much more undesired contacts (four times more undesired contacts on average) and, therefore, has more nonphonatory vocalizations as corroborated in Fig. 6. Fig. 6 was obtained using the same procedure than Fig. 4. It shows the mean percentage of phonatory goals and actual phonatory vocalizations considering all the simulations per each group of artificial agents. In general, all the auditory goals through the simulations are phonatory. The reason is that nonphonatory goals become uninteresting very early in the artificial agent's life as they are very easy to be produced. Thus, all those nonphonatory vocalizations which produce null competence impact negatively the average competence. We also might think about all those nonphonatory vocalizations as a waste of energy during the exploration. Knowing which regions of the motor space are leading to collisions might be a relevant knowledge for the agent. However, as the nonproprioceptive agents keep exploring conflicting regions, the proprioceptive agents avoid exploitation of these regions due to their ability to predict somatosensory results from a given motor command.

Additionally, discussion about the tradeoff between exploration and exploitation can be detailed. We argue that proprioceptive agents show a better performance with respect to exploitation, as agents avoid exploring uninteresting regions with high number of contacts. In other words, proprioception, and in general constraint awareness, contributes to the agent finding regularities faster and then fosters specialization in regions of the auditory space, where the agent competence to reach self-generated goals is higher. It is worth mentioning that we are aware that it is also important to include the social factor in the learning development of the artificial agent, in order to better understand the role of proprioception in social learning. In social learning, exploration is not just driven by the progress in competence and discovery of constraints, but also by the relevance of auditory goals for socialization purposes. These studies leading to more exploring behaviors is left for future work.

Furthermore, Fig. 5(b) shows the mean number of goals rejected by the proprioceptive mechanism, represented in lines 5–12 of Algorithm 1. In this plot, we prefilter the results for proprioceptive agents considering a 5000 samples moving average for visualization purposes. It can be observed in Fig. 5(b) that in general the proprioceptive mechanism is more active at the beginning of the simulations presumably due to the quantity of contacts along the initial set of vocalizations.

Thus, we might deduce that proprioception prevents the agent from further exploration in regions that are producing undesired contacts especially in the early stages. In the next, we introduce some figures in order to show the implications over the shape of the explored auditory region when proprioception is considered.

B. On Explored Regions

Regarding the volume of the explored region, Table I indicates that the ratio of average volume of convex hulls described by the explored regions in the frequency space is 0.51/0.48 between the nonproprioceptive and proprioceptive agents, whereas the ratio of the mean competence is 0.53/0.71. In other words, whereas proprioceptive agents explore a 5.88% tighter region than the nonproprioceptive, their performance is 25.35% better than the later ones. On the other hand, Table II considers only the vocalizations without undesired contacts. A shrinkage of the convex hulls is observed, the ratio of average volumes is, in this case, 0.40/0.39 while the mean competence ratio is 0.67/0.75. From these numbers we observe that, in general, the competence to vocalizations without undesired contacts is higher for both kinds of agents. However, regarding competence the proprioceptive agents still perform 11.94% better than the nonproprioceptive agents.

Based on Table I, we selected three different initial sets given their simulation results in order to produce Figs. 7–9. First, we select initial set 1, as it performs better in terms of competence when proprioception is considered. Second, to contrast with initial set 1 we select initial set 5, as its proprioceptive agent performs the worst with regards to the percentage of undesired contacts. Finally, we select initial set 9, as its proprioceptive agent produced the largest convex hull volume.

Figs. 7–9 show some projections of vocalizations distribution maps of the auditory productions generated along the simulation. Points in the plots are colored according to the percentage of undesired contacts produced in its neighborhood. Specifically, three projections are shown for each of the selected sets of initial vocalizations. Projection $F_{1,1}F_{2,1}$ represents the auditory fingerprint of the vocalizations in the first perceptual window. Projections $F_{1,2}F_{2,2}$ is similar to the first projection but for the second perceptual window. Finally, projection I_1I_2 represents the value of the intonation parameter in the first perceptual window against the same parameter in the second perceptual window.

Distributions in Figs. 7–9 indicate that the intonation parameter projection I_1I_2 is the most influenced sensory-output due to the proprioceptive feedback. Recall that I_1 and I_2 depend on the average audibility of the vocalization which is null in two cases: 1) when the voicing parameters are lower than zero or 2) when the area function of the vocal tract is nonpositive elsewhere. Therefore, keeping in mind the latter case, those vocalizations producing an intonation parameter (either I_1 or I_2) lower than one indicate that a contact has likely occurred. If a contact occurs, there are two possible results, the average of the minimum value of the area function might be negative or not. If it is negative, then the contact is classified as an undesired contact and the proprioceptive signal takes the value one. Thus, having both values lower than one at the same time is even more likely to produce undesired contacts. That is the reason proprioceptive agents explore less intensively the middle of the region in the intonation space. However, we argue that in spite of the low density of vocalizations in that region, proprioceptive agents succeed in finding more vocalizations that produces nonconflicting articulatory configurations in that region. For instance, looking at the projections in Fig. 9, the proprioceptive agent almost covers all the intonation space with low density of contacts.

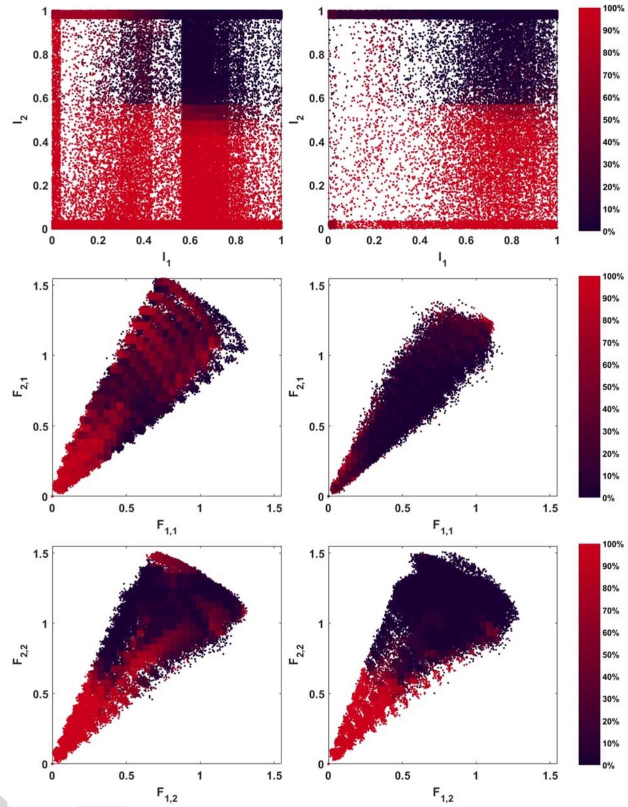


Fig. 7. Projections of vocalizations distribution along simulations using initial set 1 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

Moreover, comparing proprioceptive and nonproprioceptive agents in Figs. 7–9, we observe that the area of the explored regions varies slightly due to the proprioceptive mechanism. This fact is supported by Tables I and II. In general, in most of the cases using the proprioceptive feedback results in a slightly smaller explored region but this is not a conservative fact. For instance, Table I indicates that the convex hull volumes described in the auditory space by the experiments 2, 4, and 9 were larger when proprioception was considered. In general, besides a certain degree of randomness due to our probabilistic approach, we argue that there are three main elements that determine the shape of the explored region: 1) the initial set of vocalizations; 2) evolution of competence; and 3) proprioception.

Regarding the initial set, our criteria to choose random vocalizations close to the neutral positions produces rich sets of phonatory vocalizations, either with contacts or without.

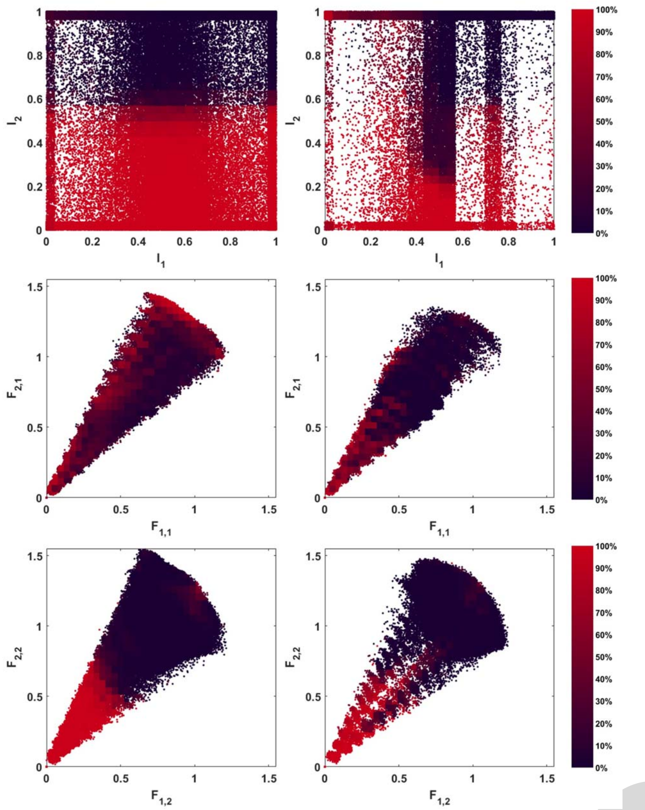


Fig. 8. Projections of vocalizations distribution along simulations using initial set 5 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

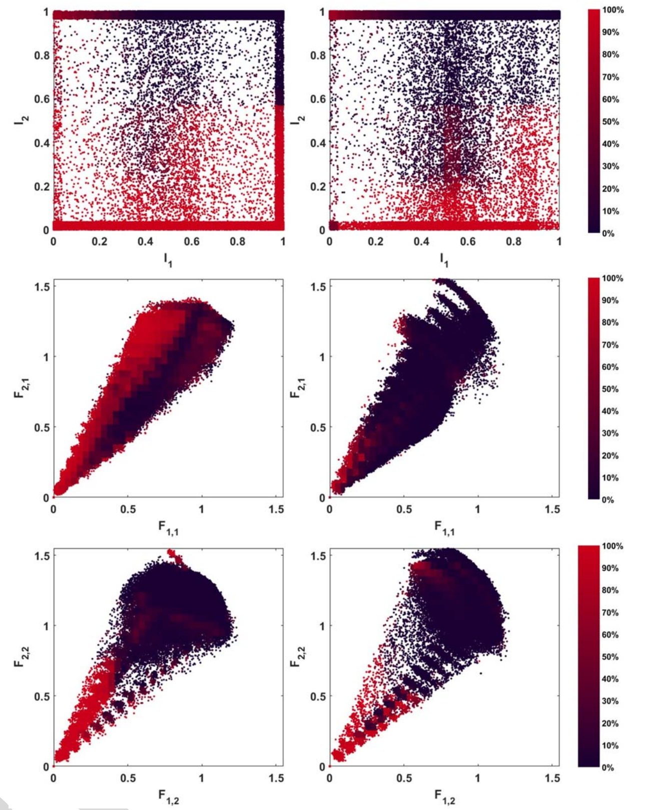


Fig. 9. Projections of vocalizations distribution along simulations using initial set 9 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

As we are working with self-generated goals, the agent is expected to be very good at the beginning at reaching goals when auditory goals are close to the initialization region. In the case of nonproprioceptive agents, the only parameter that drives the exploration is the evolution of competence, which is why we observe plenty of areas with a huge amount of undesired contacts in the plots of nonproprioceptive agents with respect to the proprioceptive agents. Furthermore, we observe that proprioception might lead toward two different situations: 1) an unexplored auditory region or 2) explored region but with nonconflicting articulatory configurations. For instance, in Fig. 7, specifically in the projection $F_{1,1}F_{2,1}$, we observe that in general the nonproprioceptive agents produce a lot of undesired contacts over almost the whole explored region. On the contrary, the proprioceptive agent explores a smaller region over the same projection, however it achieves a considerably lower density of undesired contacts; results also supported by the convex hull volume displayed in Table I. In addition, projection $F_{1,2}F_{2,2}$ shows similar explored regions for both agents. Indeed, the proprioceptive agent explored a wider region in that projection and was capable of finding nonconflicting vocalizations for some of the regions, where the nonproprioceptive agent produces a lot of undesired contacts. We observe, in general, for all the agents in Figs. 7–9, that producing auditory results for the projection $F_{1,2}F_{2,2}$ close to the origin is hard without producing contacts.

In Fig. 8, corresponds to the agent with the worst results using proprioception regarding the number of undesired contacts, the projection $F_{1,1}F_{2,1}$ indicates that the proprioceptive agent has explored a smaller region than the nonproprioceptive agent. However, if we observe the boundaries of the explored region with proprioception, they coincide with regions where the nonproprioceptive agent produces a high amount of undesired contacts. Thus, the proprioceptive mechanism does not allow the proprioceptive agent to exploit those regions. In spite of less exploration, we observe that in the explored region where both agents intersect, the proprioceptive agent produces less undesired contacts. Looking at Table II, the results for the agent corresponding to Fig. 8 (experiment 5), we observe that the explored regions with and without proprioception are described by convex hulls with similar volume. This suggests that the conflicting region explored by the nonproprioceptive agent prevents the agent from exploiting regions without undesired contacts. Thus, the agent achieve lower competence values over the latter regions. On the other hand, the proprioceptive agent avoids conflicting regions, in consequence it produces 15% less contacts and achieve a higher competence average.

In addition, looking at the projection $F_{1,1}F_{2,1}$ Fig. 9, we observe that the proprioceptive agent explores a larger region. Moreover the density of contacts along the explored region is significantly lower. This is also supported by the numerical results in Tables I and II. On the other hand, in the projection

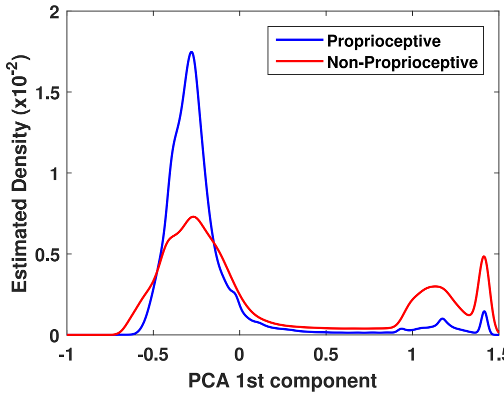


Fig. 10. Density distribution computed using Gaussian-kernels over all the data obtained along the simulations considering the first principal component with a variance contribution ratio of 0.68.

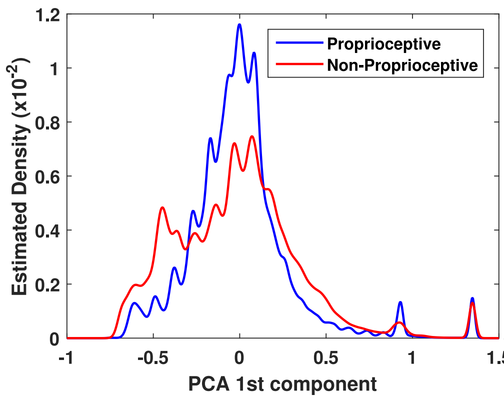


Fig. 11. Density distribution computed using Gaussian-kernels over the data, excluding vocalizations with undesired contacts, obtained along the simulations considering the first principal component with a variance contribution ratio of 0.61.

$F_{1,2}F_{2,2}$ the exploration close to the origin of that projection is less intensive in the proprioceptive agent, which reinforces the previous observation of the difficulties to produce auditory results in that region without contacts, similar results are observed in Fig. 8. Future work also must focus in the study of what is happening in that region and how relevant it is to language, as well as modify the system accordingly.

Finally, in order to observe the differences between the vocalization distributions obtained using the different exploration algorithms, we perform a sample density analysis over the formant frequency dimensions. In order to make the results easier to visualize we perform a principal component analysis (PCA) procedure. We consider analyzing the data twice, first, considering all the data collected along the exploration and second, without considering the vocalizations with undesired contacts. The PCA is done considering the dimensions $F_{1,1}$, $F_{2,1}$, $F_{1,2}$, and $F_{2,2}$, the data of all the 18 simulations are concatenated and used to perform the PCA. The PCA considering all the samples is performed and the first component is kept, which contributes to the variance with a ratio of 0.61. A second PCA is performed considering only the non conflicting vocalizations, again only the first component is kept, since it contributes to the variance with a ratio of 0.68. Once PCA transforms 4-D data into 1-D data, kernel-distribution estimation is performed using Gaussian-kernels according to [31] for the proprioceptive and nonproprioceptive cases.

In Figs. 10 and 11, we can observe the density distributions obtained separately with all the proprioceptive and the nonproprioceptive agents. First in Fig. 10, the distribution considering all the data obtained from all the experiments is shown. In general, it is observed that the agents explored similar regions, but with different intensity. In Fig. 11, we observe the distribution of the first component obtained from the PCA when only nonconflicting vocalizations are considered. In the latter case it is observed that regarding the regions which are of interest, in other words the regions where physical constraints are not violated, both kinds of agents explore with a similar density shape, which means that even though both agents explore similar interesting regions, the proprioceptive agents achieve in general higher competence.

VI. Conclusion

An application of active learning techniques applied to the study of vocal exploration considering motor constraints has been introduced. It has been presented as an intrinsically motivated sensorimotor self-exploration architecture with motor constraints self-awareness. Constraints awareness is achieved by providing a proprioceptive mechanism which endows an artificial agent with the capacity to autonomously generate a somatosensory model. This model is then used to predict the consequences of a motor action and to avoid its execution if it is expected to generate an undesired proprioceptive result.

The proprioceptive mechanism improved the quality of learning according to a competence function. However, we observe a tradeoff between exploration and exploitation, predominantly nonproprioceptive agents achieve greater exploration in the auditory space. In contrast, we observe a more intensive exploitation in interesting regions driving to the higher competence values achieved by proprioceptive agents. In general, vocal-auditory spaces are high dimensional redundant spaces, thus an auditory output may be produced by different articulatory configurations. Some of these articulatory configurations may lead to undesired contacts. Hence, we argue that sensorimotor redundancy is reduced when proprioception is included in the system allowing the agent to focus on exploitation of nonconflicting vocalizations. In consequence, the sensorimotor model generated through the exploration does not include conflicting regions, where constraint violations are likely to happen. For that reason, sensorimotor models achieve better fitting to the regions of interest where constraints are met. In this way, we showed how sensorimotor exploration, and in general sensorimotor knowledge, can be shaped by constraints.

Regarding the advance toward vocal exploration, we have showed the suitability of the presented architecture to learn vocal spaces in interesting and less redundant regions as children might do. However, in order to continue our research on early vocal development, we must study in greater depth the first period of vocalization development. A deeper analysis of the learning processes underlying the nonauditory development related to mastication, deglutition, and crying from the cognitive and developmental perspectives should be completed in order to generate more complex somatosensory architectures. Finally, the next step of this paper should be directed toward the self-structuring of vocalization and social learning.

REFERENCES

[1] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nat. Rev. Neurosci.*, vol. 5, no. 11, pp. 831–843, 2004.

[2] M. Asada *et al.*, "Cognitive developmental robotics: A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

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

[4] R. Pfeifer and C. Scheier, *Understanding Intelligence*. Cambridge, U.K.: MIT Press, 1999.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, 1996.

[6] S. Thrun, "Exploration in active learning," in *Handbook of Brain Science and Neural Networks*, 1995, pp. 381–384.

[7] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: A unified probabilistic framework," in *Proc. Int. Conf. Develop. Learn. (ICDL/Epirob)*, Osaka, Japan, 2013, pp. 1–6.

[8] J. M. Acevedo-Valle, C. Angulo, N. Agell, and C. Moulin-Frier, "Proprioceptive feedback and intrinsic motivations in early-vocal development," in *Proc. 18th Int. Conf. Catalan Assoc. Artif. Intell.*, 2015, pp. 9–18.

[9] 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," *Front. Psychol.*, vol. 4, pp. 1006–1025, Jan. 2014, doi: 10.3389/fpsyg.2013.01006.

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

[11] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Netw.*, vol. 38, pp. 64–75, Feb. 2013.

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

[13] 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.

[14] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Auton. Mental Develop.*, vol. 2, no. 3, pp. 216–229, Sep. 2010.

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

[16] 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 *Proc. 1st Multidiscipl. Conf. Reinforcement Learn. Decis. Making (RLDM)*, Princeton, NJ, USA, Oct. 2014, Art. no. hal-00922537. [Online]. Available: https://hal.inria.fr/hal-00922537/document

[17] A. Ribes, J. Cerquides, Y. Demiris, and R. Lopez de Mántaras, "Active learning of object and body models with time constraints on a humanoid robot," *IEEE Trans. Cogn. Develop. Syst.*, vol. 8, no. 1, pp. 26–41, Mar. 2016, doi: 10.1109/TAMD.2015.2441375.

[18] J. Perkell *et al.*, "The sensorimotor control of speech production," in *Proc. 1st Int. Symp. Meas. Anal. Model. Human Functions*, 2001, pp. 359–365.

[19] D. K. Oller and R. E. Eilers, "The role of audition in infant babbling," *Child Develop.*, vol. 59, no. 2, pp. 441–449, 1988.

[20] K. Ejiri, "Relationship between rhythmic behavior and canonical babbling in infant vocal development," *Phonetica*, vol. 55, no. 4, pp. 226–237, 1998.

[21] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evol. Comput.*, vol. 11, no. 2, pp. 265–286, Apr. 2007.

[22] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: Computational and neural mechanisms," *Trends Cogn. Sci.*, vol. 17, no. 11, pp. 585–593, 2013.

[23] B. Galantucci, C. A. Fowler, and M. T. Turvey, "The motor theory of speech perception reviewed," *Psychonomic Bull. Rev.*, vol. 13, no. 3, pp. 361–377, 2006.

[24] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception," *J. Neurolinguist.*, vol. 25, no. 5, pp. 336–354, 2012.

[25] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, no. 6942, pp. 866–869, 2003.

[26] S. M. Nasir and D. J. Ostry, "Speech motor learning in profoundly deaf adults," *Nature Neurosci.*, vol. 11, no. 10, pp. 1217–1222, 2008.

[27] S. N. Iyer and D. K. Oller, "Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss," *Volta Rev.*, vol. 108, no. 2, pp. 115–138, 2008.

[28] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 4, pp. 1245–1248, 2009.

[29] C. Moulin-Frier and P.-Y. Oudeyer, "The role of intrinsic motivations in learning sensorimotor vocal mappings: A developmental robotics study," in *Proc. Interspeech*, Lyon, France, 2013, pp. 1268–1272.

[30] S. Calinon, *Robot Programming by Demonstration*. Lausanne, Switzerland: EPFL Press, 2009.

[31] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ, USA: Wiley, 2015.

Juan Manuel Acevedo-Valle, photograph and biography not available at the time of publication.

Cecilio Angulo, photograph and biography not available at the time of publication.

Clement Moulin-Frier, photograph and biography not available at the time of publication.

Autonomous Discovery of Motor Constraints in an Intrinsically Motivated Vocal Learner

Juan Manuel Acevedo-Valle, Cecilio Angulo, and Clement Moulin-Frier

***Abstract*—This paper introduces new results on the modeling of early vocal development using artificial intelligent cognitive architectures and a simulated vocal tract. The problem is addressed using intrinsically motivated learning algorithms for autonomous sensorimotor exploration, a kind of algorithm belonging to the active learning architectures family. The artificial agent is able to autonomously select goals to explore its own sensorimotor system in regions, where its competence to execute intended goals is improved. We propose to include a somatosensory system to provide a proprioceptive feedback signal to reinforce learning through the autonomous discovery of motor constraints. Constraints are represented by a somatosensory model which is unknown beforehand to the learner. Both the sensorimotor and somatosensory system are modeled using Gaussian mixture models. We argue that using an architecture which includes a somatosensory model would reduce redundancy in the sensorimotor model and drive the learning process more efficiently than algorithms taking into account only auditory feedback. The role of this proposed system is to predict whether an undesired collision within the vocal tract under a certain motor configuration is likely to occur. Thus, compromised motor configurations are rejected, guaranteeing that the agent is less prone to violate its own constraints.**

***Index Terms*—Active learning, early vocal development, Gaussian mixture models (GMMs), intrinsic motivations, sensorimotor exploration.**

I. Introduction

IN RECENT years, there has been an increasing interest in using robots to perform daily life activities in the presence of humans. As robot–human interactions become common then human-like communication systems become more relevant to robotics. Speech is one of the most studied communication systems because it allows human-spoken language. However, as mentioned in [1], the idea that speech is a deeply encrypted "code" prevails among the speech specialists and cracking this code is still an unsolved problem. Some of the mysteries about speech might be solved if we are able to understand all the mechanisms underlying early speech acquisition in children. Thus, this paper, provides new results to contribute to the study of early speech development using machines.

Developmental robotics is a relatively novel approach, it aims at understanding and modeling the role of developmental processes in the emergence of complex behaviors, including social ones. Its goal is twofold, on the one hand it is used to build more efficient cognitive machines applying developmental theories, and on the other hand it also provides insights into human developmental mechanisms, especially during infancy. A deeper understanding of these mechanisms would explain how human beings develop from infancy to functional adults capable of solving highly complex cognitive tasks [2].

Autonomous robot design could notably benefit from the available knowledge of biological science and self-organization theories [3]. Deep understanding of the embodiment paradigm is paramount to integrate that knowledge into robotics. This paradigm is also well represented by the quote "understanding by building" [4]. It states that the behavior of an agent is not only the result of a system control structure, but also a result of complex interactions with its ecological niche, its morphology, and its material properties [3], [4].

In this paper, language emergence is studied according to behavioral and neurophysiological evidence, moreover the role of motor constraints is especially considered. The main assumption is that early vocal development can be studied as a result of embodiment, self-organization, and emergence mechanisms produced by human evolution. In general, studies have shown that infants show preparedness to acquire natural language. Motor, perceptual, social, and learning ability constraints, and their maturation during infant development play a key role in the emergence of language [1].

Equally important, machine learning techniques have rapidly evolved, providing developmental robotics with interesting approaches as active learning. In contrast to the more usual passive learning algorithms, active learning data are collected in order to minimize a given property of the learning process, e.g., the uncertainty [5] or the prediction error [6] of a model. This family of algorithms is of particular interest for developmental robotics. During sensorimotor exploration they allow the agent to focus on parts of the sensorimotor space in which exploration is expected to improve the quality of the learned model [7].

Manuscript received July 19, 2016; revised November 18, 2016 and February 20, 2017; accepted April 14, 2017. This work was supported by the PATRICIA Research Project through the Spanish Ministry of Economy and Competitiveness under Grant TIN2012-38416-C03-01. The work of J. M. Acevedo-Valle was supported by CONACYT under Grant 216872. *(Corresponding author: Juan Manuel Acevedo-Valle.)*

J. M. Acevedo-Valle and C. Angulo are with the GREC Research Group, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain (e-mail: juan.manuel.acevedo.valle@upc.edu).

C. Moulin-Frier was with Flowers team, Inria/ENSTA-Paristech, 33405 Bordeaux, France. He is now with the SPECS Laboratory, Universitat Pompeu Fabra, 08018 Barcelona, Spain.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2017.2699578

The contribution of this paper is extending the study of early language development using intrinsically motivated exploration algorithms. Herein, we provide new simulation results showing the suitability of these algorithms in the self-exploration of sensorimotor vocal spaces. The theoretical basis of the probabilistic models used to represent knowledge is also provided. Furthermore, we propose an architecture that could be used to study the role of constraints during sensorimotor exploration in embodied agents. Finally, it is worth mentioning that the learning algorithm presented herein could be applied to any system subjected to constraints in order to improve learning progress.

The remainder of this paper is organized as follows. Section II introduces related works. Section III highlights the role of intrinsic motivations and proprioceptive feedback in vocal development. The experiment setup is described in Section IV and results are presented and discussed in Section V. Finally, the conclusions are presented in Section VI.

II. Related Work

This paper revisits and expands the investigation introduced in [8] and [9]. In [9], an intrinsically motivated exploration architecture was proposed for the study of the developmental stages emergent during the early vocal development of infants. For the experimentation the simulated ear-vocal tract model DIVA [10] was used. In spite of the relevance of its results, the motor constraints and the somatosensory system were neglected in [9]. However, morphological constraints play a key role in speech acquisition. Therefore, a new exploration algorithm proposed in [8] to incorporate motor constraints awareness using a somatosensory model. In the past, some studies have tried to explain the emergence of developmental stages during the vocal development, assuming their existence, but those stages were bridged using hard-coding for experimentation [10]–[13].

In [14], an approach for inverse kinematics learning in redundant systems was presented. It was demonstrated that goal babbling can be advantageous in learning in the early stages of development, as observed in developmental theories. In parallel, [15] presented an intrinsically motivated goal exploration approach for the active learning of inverse models. This approach was applied to the vocal sensorimotor space exploration in [9] and [16]. The algorithm considered in this paper extends intrinsically motivated exploration in the goal space to include motor constraints. Considering both motor and perceptual constraints during learning and exploration is crucial to design cognitive architectures for motor control.

Among the efforts to model the acquisition of speech there is the DIVA model [10]. It aims to imitate the underlying neurophysiological mechanisms for speech acquisition and production. The cognitive architecture of the system is an artificial neural network. The model includes the premotor, motor, auditory and somatosensory cortical areas, and simulated ear-vocal tract system. In [10], the somatosensory model was effectively integrated into the acquisition and production of speech processes. It was not used as an element to integrate motor constraints but as an extra source of sensory-feedback. The ear-vocal tract component of the DIVA model is used in this paper, as it was in [8] and [9].

Finally, another interesting contribution was the active learning architecture presented in [17] which considered time constraints. This paper proposed a music performance imitation scenario and implemented a learning architecture able to learn a musical instrument model and a body capabilities model; the architecture is also able to imitate a sequence of sound, while simultaneously kinematic errors, due to the control architecture, are corrected. Similar to [9], models employed in [17] were based on Gaussian mixture models (GMMs).

III. Early Vocal Development in Machines

Human speech production is one of the most complex motor acts performed by any living being [18]. Producing a linguistic message that can be understood by another human requires coordinating many degrees of freedom in the respiratory, laryngeal, and supraglottal articulatory system.

How infants acquire the complex ability to control speech production and in general how they learn language remains a matter of research [1]. It has been pointed out that strong regularities can be observed in the structure of the vocal development process independently of interindividual differences [1], [19]. In general, the infant first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds and finally automatically discovers and focuses on babbling with articulated proto-syllables. In [18], some experiments suggested that goals of speech movements are auditory in nature and maintenance of motor command maps to auditory results is performed with auditory feedback.

It is important to inquire into the developmental assumptions considered in the experiments in [8] and [9], as this paper is based on those experiments. Regarding the infant development stages mentioned in [1], our experiments consider the developmental stage known as canonical babbling (CB) [20] and the beginning of language-specific speech production [1]. Results suggest that during CB infants learn to control their ear-vocal tract system based on auditory feedback. Nevertheless, when infants begin to babble they do it regardless of the audibility of their vocalizations. CB could be the result of a natural tendency of infants to move their body parts rhythmically motivated by sensory feedback [20].

Consistent with the theory, we assumed a simplified explanation that the artificial agent is exploring its ear-vocal tract system choosing auditory goals and evaluating the result. Therefore, our cognitive architecture allows the agent to explore regions, where the competence to produce intended sounds is improved. However, we also endow the agent with autonomous mechanisms to discover constraints in order to drive the exploration. To accomplish that objective, previously proposed active learning architectures and the proprioceptive feedback concept are combined.

A. Intrinsically Motivated Exploration Architectures

Among the vast number of active learning architectures, this paper considers the exploration architecture proposed by [15]. This architecture reproduces the formalism of intrinsic motivation inspired by psychological literature as proposed previously in [21] and [22]. Using goal babbling, intrinsically motivated exploration aims to minimize the error of an agent to reach self-generated goals measured according to a competence function. This architecture allows artificial agents to efficiently and actively explore and generate maps from motor capacities to perceived results. Therefore, exploration occurs over regions in which agents perceive they are becoming more competent to reach self-generated goals. Intrinsically motivated exploration architectures were originally designed to actively learn inverse models of high-dimensional input–output spaces. This architecture was later extended by [9] to study self-organization in early vocal development stages in infants and robots.

Intrinsically motivated learning algorithms have shown favorable results in previous experiments to learn sensorimotor coordination skills in redundant nonlinear high-dimensional mappings which share many mathematical properties with vocal spaces. Moulin-Frier *et al.* [9] used a simulated ear-vocal tract system to study the emergence of developmental stages implementing intrinsically motivated exploration. They argued that the development of the agent self-organizes into vocal developmental sequences. The results presented therein opened the door to a new approach in vocal development to be explored. This paper introduces a methodology which enhances intrinsically motivated architectures with constraint awareness.

B. *Proprioceptive Feedback*

Some of the most adopted theories of speech state that speech production is organized in terms of motor control signals and their associated vocal tract configurations which has been corroborated by several experimental results [23], [24]. Nevertheless, we adopt the simplified hypothesis that speech goals are defined acoustically and maintained by auditory feedback [18]. CB is a rhythmic behavior that, with some differences, emerges in both, normally developing infants and infants with hearing loss. When infants start to babble, they do it regardless of the audibility result (i.e., they produce audible and voiceless vocalizations). However, evidence suggests that, around the onset of CB, infants learn to vocalize based on auditory feedback [1], [20].

How the somatosensory system[1] affects the ear-vocal tract exploration is an open question that was not previously approached in [9]. However, the relevance of the somatosensory system for speech has been shown in different experiments, for instance the results in deaf individuals suggest that somatosensory inputs related to movement play a more important role in speech production than what was thought before [25], [26]. Furthermore, the fact that CB also emerges in deaf infants suggests that somatosensory feedback must play a more relevant role during the prelinguistic vocal development in infants [27].

[1] Strictly, the somatosensory system is also a sensorimotor system. In future works, we will distinguish two sensorimotor systems: 1) the auditory-motor system and 2) the somatosensory-motor system.

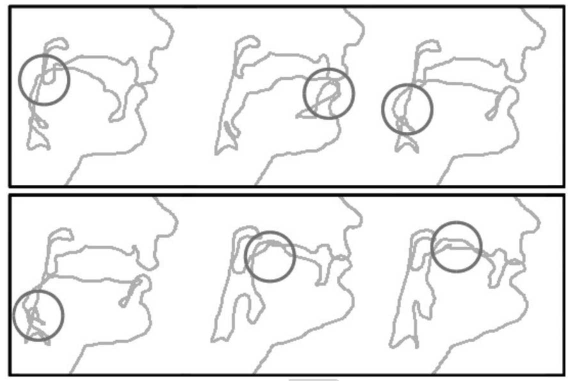


Fig. 1. Examples of articulatory configurations that produce collisions in the DIVA vocal tract model.

In [28], a robotic device able to generate patterns of facial skin deformation related to certain speech productions was used. The results showed that when the facial skin is stretched whilst subjects are listening to words, the sounds they hear are altered. Thus, theory and results suggest that the somatosensory system is involved in speech perception. Following this hypothesis, improvements can made to the experiments proposed in [9] by including a somatosensory system to endow the learner with physical constraint awareness.

In [8], the foundations of a simplified architecture were established allowing us to include physical constraints to the learning process through a proprioceptive signal, similar to the ability to feel pain in humans. The open source DIVA model[2] [10] provides a synthesizer that represents the human vocal tract and ear systems. The DIVA model also includes a somatosensory system, but in spite of it, there is a lack of physical constraints in the DIVA vocal tract. The absence of constraints allows the execution of motor commands that lead to collisions or articulatory superpositions. Both circumstances lead to no phonation and moreover, the latter is a contradictory result since it lacks physical sense, as shown in Fig. 1.

To overcome the drawbacks caused by the lack of constraints, we introduce a somatosensory system. This new element, not considered in [9], is based on an area function which is a vector descriptor of the vocal tract shape. It consists of a mechanism that evaluates if an exploratory motor command produces a collision or superposition of articulatory tissues, the system generates a proprioceptive signal. Using the data generated with this mechanism, the agent builds a map from motor commands to proprioceptive results. This map is used to predict which motor commands may lead to undesired collisions, so they may be rejected, forcing the agent to choose a new auditory goal. In the next section, this mechanism is explained in detail.

IV. Proposed Architecture

The experimental architecture proposed in this paper to study the early vocal development in machines is shown in Fig. 2, where five elements interact. These elements are

[2] http://www.bu.edu/speechlab/software/diva-source-code/

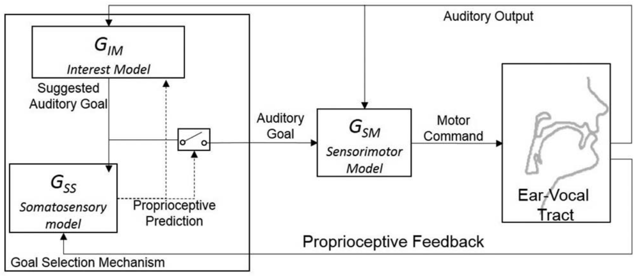


Fig. 2. Experimental architecture. It is composed by five interacting modules, two of them contained within the ear-vocal tract module (the sensorimotor system and the somatosensory system).

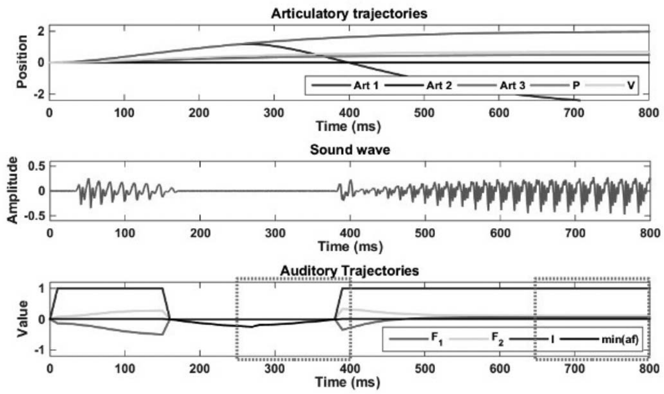


Fig. 3. Vocalization experiment structure. The upper plot shows the articulatory trajectories, from 0 to 250 ms, the commands for Art1, Art2, and Art3 are set to 2, 0, and 2, respectively, whereas the glottal pressure and voicing are both set to 0.5. From 250 to 800 ms, the commands for Art1, Art2, and Art3 are set to −3, 0, and 2, respectively, whereas the glottal pressure and voicing are both set to 0.7. The remaining motor commands are set to zero. The middle plot represents the speech sound wave signal. The bottom plot shows the auditory trajectories. The dotted outlined boxes represent the perception time windows from 250 to 400 ms and the second from 650 to 800 ms. The auditory output s are determined from the average of each trajectories along each one of the time windows. Whereas the proprioceptive feedback p is determined by the average value of $\min(a_f)$.

introduced below and explained in detail in the coming sections.

1) *Sensorimotor system* is a simulated ear-vocal tract. It corresponds to the physical properties of the embodied agent. For the present work the ear-vocal tract system of the DIVA model [10] is used.
2) *Somatosensory system* is a perceptual mechanism that evaluates the shape of the vocal tract. It generates a proprioceptive feedback signal indicating if an undesired contact or collision is produced into the vocal tract.
3) *Sensorimotor model* is a mathematical representation of the vocal tract-ear model. It endows the artificial agent to map motor commands to auditory effects using the data collected from the agent's own vocalizations.
4) *Somatosensory model* is a mathematical representation that maps motor configurations to their likely proprioceptive feedback to acquire self-awareness of its own physical constraints in order to avoid executing motor configurations that produce undesired behaviors.
5) *Interest model* for auditory goals allows the agent to actively choose auditory goals in order to improve the quality of its sensorimotor model based on a certain measure of competence. This model represents the core of the intrinsically motivated sensorimotor self-exploration.

A. Sensorimotor System

The DIVA vocal tract configuration is determined by the position of ten articulators and three phonation parameters. Along this paper, only seven articulators and two phonation parameters (voicing and glottal pressure) are considered [9]. Articulators and voicing parameter motor dynamics are modeled as overdamped second order systems

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

with $\zeta = 1.01$ and $\omega = (2\pi/0.8)$ representing the damping factor and the natural frequency, respectively. The duration of each vocal experiment in seconds is 0.8, whereas m and x represents the desired articulator position (motor command) and the current articulator position, respectively. During each vocal experiment two different motor commands are introduced for each of the seven articulators and the two voicing parameters: one for the 0–250 ms window and another for the remaining time. Thus, each motor command is represented by an 18-D vector. The auditory output of a human vocalization can be described by its formant frequencies. We consider the first two formant frequencies, F_1 and F_2, along with an intonation signal I. The intonation signal is 1 when phonation occurs and 0 otherwise, two conditions are required for phonation to occur: 1) the area function a_f of the vocal tract must be positive elsewhere and 2) the voicing and pressure parameters must be positive. The area function is a vector function that describes the transversal shape of the vocal tract.

During the vocalization the auditory output of the system is observed along two time windows, the first from 250 to 400 ms and the second from 650 to 800 ms. The value of each auditory output is averaged for each time window, the result is a 6-D output signal (two formants and the intonation, hence three values, per each of the two time windows). In Fig. 3, we reproduce the vocalization representation shown in [9]. To be consistent with the co-articulated nature of speech, only two perceptual windows are used [1]. However, since only two portions of the vocalization are considered, a lot of information is lost. For instance, it is shown in [23] the continuum of co-articulated gestures. Therefore, future works should consider studying the continuum of speech gestures and self-structuring of vocalizations.

B. Somatosensory System

In Fig. 3, it is shown that the area function a_f is observed during both perception time windows. The minimal value of the area function $\min(a_f)$ would be zero when the vocal tract is closed at any point and negative values mean that some tissues are overlapped, which does not have physical meaning. However, in some cases it might be interpreted as the tongue being bitten. In other cases it might represent high pressure between the tongue and the palate, which might be interesting to the learner in a realistic scenario, where motor constraints are not violated. In general, we made a strong assumption that any motor constraint violation over a threshold is uncomfortable or painful. Hence, the average value of $\min(a_f)$ in each perception time window is used to generate a proprioceptive feedback signal p: if the average of $\min(a_f)$ is lower than a threshold for any perception window, then the configuration is evaluated as a undesired collision with $p = 1$, and $p = 0$ otherwise.

C. Sensorimotor Model

GMMs are linear combinations of multivariate Gaussian distributions that represent clusters of data. They have been previously used to represent nonlinear redundant maps [17], [21], [29] in order to solve the inverse problem of inferring input motor commands from desired sensory outputs. GMMs can be learned using an online variant of the expectation-maximization (EM) algorithm in order to learn incrementally from incoming data [30]. Here, the algorithms used to train GMMs are based on the open source tools[3] associated with [30], and modified according to our problem requirements. The three models in the experimental setup are probabilistic representations in the form of GMMs, obtained using data collected from experiments with the DIVA ear-vocal tract. A detailed explanation of the GMMs training is provided below.

We assume that an n-dimensional input command space $X \in \mathbb{R}^n$ is mapped to an m-dimensional output space $Y \in \mathbb{R}^m$, through a transform function $y = f(x) + \varepsilon$, where $y \in Y$, $x \in X$ and ε is random noise. When a dataset of couples (x, y) is available, the EM-algorithm is used to obtain a GMM which is defined by the parameters $\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}$, where π_j, μ_j, and Σ_j are, respectively, the prior probability, the distribution centroid and the covariance matrix of the jth Gaussian, for $j = 1, 2, \ldots, K$, being K the number of Gaussian components. From [30], Gaussian mixture regression (GMR) is applied to compute the conditional probability distribution $P(X\,|\,y)$ in the input space X given a desired output y. Once it is computed, the value $x^* \in X$ is selected such that it maximizes $P(X\,|\,y)$.

To obtain the input x that maximizes the probability to produce the output y, the GMR process first defines the partitioned vector $z \in X \times Y$, where

$$z = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{2}$$

For each Gaussian j in the GMM the partitions

$$\mu_j = \begin{pmatrix} \mu_j^x \\ \mu_j^y \end{pmatrix} \quad \text{and} \quad \Sigma_j = \begin{pmatrix} \Sigma_j^x & \Sigma_j^{xy} \\ \Sigma^{yx} & \Sigma_j^y \end{pmatrix} \tag{3}$$

are considered to compute the conditional probability distribution $P_j(X\,|\,y) \sim N_j(\hat{\mu}_j, \hat{\Sigma}_j)$ in the input space X given a desired output y, where

$$\hat{\mu}_j = \mu_j^y + \Sigma_j^{yx}\left(\Sigma_j^x\right)^{-1}\left(x - \mu_j^x\right), \hat{\Sigma}_j = \Sigma_j^y + \Sigma_j^{yx}\left(\Sigma_j^x\right)^{-1}\Sigma_j^{xy}. \tag{4}$$

[3]http://www.calinon.ch/sourcecodes.php

Considering that $P(X\,|\,y)$ is at its maximum when $x = \hat{x}_j = \hat{\mu}_j$, then a natural selection for x in order to produce y is $\hat{x}_j$. But we have K candidates for x, hence it is necessary to compute the probability of the vector $\hat{z}_j = [\hat{x}_j,\ y]^T$ belonging to its generator Gaussian as

$$P(\hat{z}_j) = \pi_j \frac{1}{\sqrt{(2\pi)^K |\Sigma_j|}} e^{-\frac{1}{2}\left((\hat{z}_j-\mu_j)^T \Sigma_j^{-1} (\hat{z}_j-\mu_j)\right)} \tag{5}$$

and finally the point $z^* = \hat{z}_j$ that maximizes $P(\hat{z}_j)$ is selected as the point that better fits the model. In other words, according to our prior knowledge of $f(x)$, $z^* \in f(x)$, we infer that the output y is generated by $\hat{x}_j$.

Taking into account the above for the sensorimotor model, an 18-D motor command space M, with $m \in M$, is defined for the vocal tract articulatory configuration. A 6-D auditory output space S, with $s \in S$, is also defined, the agent being able to observe s according to $s = f(m) + \sigma$, where $\sigma \sim N(0, 0.01)$ is Gaussian noise. The aim is to find a GMM that solves the inverse problem $m = f^{-1}(s_g)$, where s_g is an auditory goal.

We define a GMM, $G_{\rm SM}$, to model the sensorimotor system, with $X = M$ and $Y = S$. Such a model allows computation of the inverse model $P(M\,|\,s_g)$ using GMR. At the beginning of the experiment, m is selected either, randomly or according to the interest for initializing the inverse sensorimotor model $m \sim f^{-1}(s_g) \sim P(M \mid s_g)$ around a specific region of the sensorimotor space. After the initialization stage, the agent starts to select new auditory goals, according to the interest model explained below. In order to reduce memory storage requirements, we consider a generative method for the training stage, which means that the model is trained using the last $N_{\rm SM}$ samples obtained from experimentation along with

$$N_{\rm old} = \left\lceil \frac{(1-\alpha)N_{\rm SM}}{\alpha} \right\rceil \ \text{samples} \tag{6}$$

generated using $G_{\rm SM}$, where $\alpha \in [0,\ 1]$ is the forgetting rate.

D. Interest Model for Auditory Goals

The interest model for auditory goals endows the learner the ability to select goals that maximize the expected competence progress in order to improve the quality of its sensorimotor model, resulting in better control over it. It is derived from the model proposed in [9]. The competence value for a goal is defined by

$$c = e^{-|s_g - s|} \tag{7}$$

where s_g is the auditory goal and s is the actual auditory production after executing a motor command $m \sim P(M\,|\,s_g)$. To construct the interest model, the auditory goal space is augmented with two extra dimensions: 1) the competence $c \in C$ and 2) time tag $t \in T$. The number of vocalizations $N_{\rm IM}$ considered to build the interest model is fixed. A GMM, $G_{\rm IM}$, with $K_{\rm IM}$ components will be computed from the 8-D data set with $N_{\rm IM}$ samples of the augmented goal space. To initialize this model, some auditory results from the initialization of $G_{\rm SM}$ are selected as the first auditory goals s_g.

Those Gaussian components in $G_{\rm IM}$ that, according to the covariance matrices Σ_j, contain goals that will likely increase the competence progressively are considered to build a probabilistic distribution $P(S)$ over the auditory space. In order to build $P(S)$, the components in G_{IM} are weighted according to their time-competence covariance magnitudes. Thus, $P(S)$ will prioritize goals in regions, where competence is expected to increase. Finally, a sample s_g is drawn from $P(S)$ for the next vocalization experiment. Model training is performed every time the agent has performed n_{IM} experiments, using the last N_{IM} vocalizations.

E. Somatosensory Model

For the somatosensory model we consider the 18-D motor command space M, with $m \in M$, and a new binary proprioceptive output space $P = \{0, 1\}$, with $p \in P$. If a vocal production leads to undesired contacts, then $p = 1$, otherwise $p = 0$. A map g is assumed to exist such that $p = g(m)$ and the agent can observe p for each vocal experiment. Thus, it is possible to find a GMM G_{SS}, with $X = M$ and $Y = P$, that allows computation of the probability distribution $P(P \mid m)$ applying GMR, and determine when a motor command m is likely to lead to an undesired collision in the vocal tract.

The inverse sensorimotor model G_{SM} and the somatosensory model G_{SS} are initialized together. When an auditory goal s_g has been selected, m is computed using $P(M \mid s_g)$. Next, to predict the value of p, $P(P \mid m)$ is used. If the prediction suggests that m will produce $p = 1$ then s_g is rejected, otherwise s_g and m are accepted. If s_g is rejected, then $G_{\mathrm{IM}}(S)$ is recomputed without considering the Gaussian component in G_{IM} that generated s_g, this mechanism decreases the prior of the conflicting Gaussian in G_{IM}. The new $G_{\mathrm{IM}}(S)$ is used to select a new goal s_g, and the process is repeated until s_g is accepted. During the agent's life, the model G_{SS} is trained when G_{SM} is trained using the previously described generative mechanism.

F. Self-Exploration Algorithm

The self-exploration architecture with motor constraints self-awareness, first proposed in [8] for ear-vocal tract exploration, is an extended version of [9]. The algorithm associated with the cognitive architecture is shown in Algorithm 1. Our extended self-exploration algorithm with goal babbling and motor constraints self-awareness starts with the learner having no experience in vocalizing. Models G_{SM} and G_{SS} are initialized using random vocalizations with small values around the neutral position of the articulators. The neutral position of the pressure and voicing parameters are set to -0.25 to produce no phonation, whereas for the articulators it is considered 0, i.e., the rest position. Model G_{IM} is also initialized.

Then, in line 6 of Algorithm 1 the vocal learner agent selects a goal s_g for the next experiment according to the probabilistic distribution $P(S)$ and the motor command m is obtained using the inverse model G_{SM} in line 7. The main feature of this algorithm, different from similar architectures, is that in line 8 $P(P \mid m)$ provides a prediction for p that indicates if the selected motor command is likely to produce an undesired collision. From line 9 to 11, if $p \approx 1$, the goal is rejected and the probabilistic distribution $P(S)$ is updated, ignoring the

Algorithm 1 Self-Exploration With Goal Babbling and Self-Constraints Awareness

1: Initialize G_{SM} and G_{SS}
2: Initialize G_{IM} and $i \leftarrow 1$
3: **while** i in [1, 1e5] **do**
4: $\quad p_{tmp} \leftarrow 1$
5: $\quad$ **while** p_{tmp} **do**
6: $\quad\quad s_{g,i} \leftarrow G_{IM}(S)$
7: $\quad\quad m_i \leftarrow G_{SM}\big(M|s_{g,i}\big)$
8: $\quad\quad p_{tmp} \leftarrow G_{SS}(P|m_i)$
9: $\quad\quad$ **if** p_{tmp} **then**
10: $\quad\quad\quad update(G_{IM}(S))$
11: $\quad\quad$ **end if**
12: $\quad$ **end while**
13: $\quad s_i \leftarrow f(m_i) + \sigma$ and $p_i \leftarrow g(m_i)$
14: $\quad c_i \leftarrow e^{-|s_{g,i}-s_i|}$
15: $\quad i \leftarrow i + 1$
16: $\quad$ **if** i mod $N_{SM} = 0$ **then**
17: $\quad\quad train\big(G_{SM}, m_{(i-N_{SM}+1:i)}, s_{(i-N_{SM}+1:i)}\big)$
18: $\quad\quad train\big(G_{SS}, p_{(i-N_{SM}+1:i)}, s_{(i-N_{SM}+1:i)}\big)$
19: $\quad$ **end if**
20: $\quad$ **if** i mod $n_{IM} = 0$ **then**
21: $\quad\quad train\big(G_{IM}, s_{g,(i-N_{IM}+1:i)}, c_{(i-N_{IM}+1:i)}\big)$
22: $\quad$ **end if**
23: **end while**

Gaussian component in G_{IM} that generated s_g and the algorithm goes back to line 6. Otherwise, $p \approx 0$, both, s_g and m are accepted. Next, the motor command is executed with the vocal tract and the agent observes s and p in line 13. In line 14, the learner evaluates the competence value c. It also checks if we are at the end of a learning episode, so models G_{SM}, G_{SS}, and G_{IM} are updated in lines 17, 18, and 21, respectively. To provide objective evaluation elements, some experiments without considering the somatosensory model for choosing goals are also presented. In this later case, s_g is always accepted, thus line 4 is substituted with $p_{\mathrm{tmp}} = 0$ in Algorithm 1.

V. Experimental Results

Eighteen independent simulations using Algorithm 1 were run. All simulations consisted of half a million of vocalizations, including an initial vocalization set of 1000 random samples. Nine different random seeds were considered to generate the same number of motor command sets from a uniform distribution. The limits for those motor commands related to the vocal tract articulators were $[-1, 1]$, whereas for motor commands related to the phonation parameters were $[0, 0.7]$. Each set was used twice to initialize simulations of Algorithm 1, first without using the somatosensory model and second with it.

Considering as a reference the parameters used for simulations in [8] and [9], a few variations in their values were tested. First, when values for K_{SM} or K_{SS} are increased the inference error decreases slightly but the computation time grows considerably. On the other hand if these values are decreased, the inference error increases considerably. Second, if the training

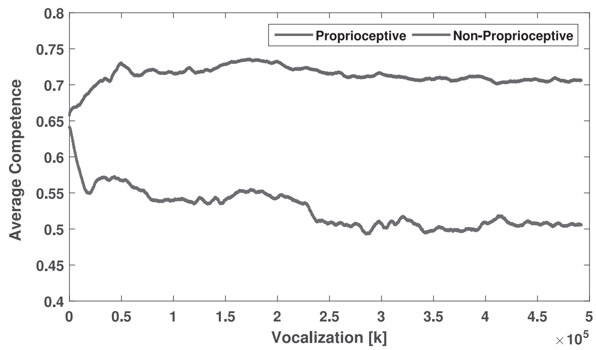


Fig. 4. Mean competence evolution during simulations using Algorithm 1 for nine different initialization data sets. Moving average of 5000 samples are considered to filter the results of each simulation. Results are shown in the case of proprioceptive and nonproprioceptive agents.

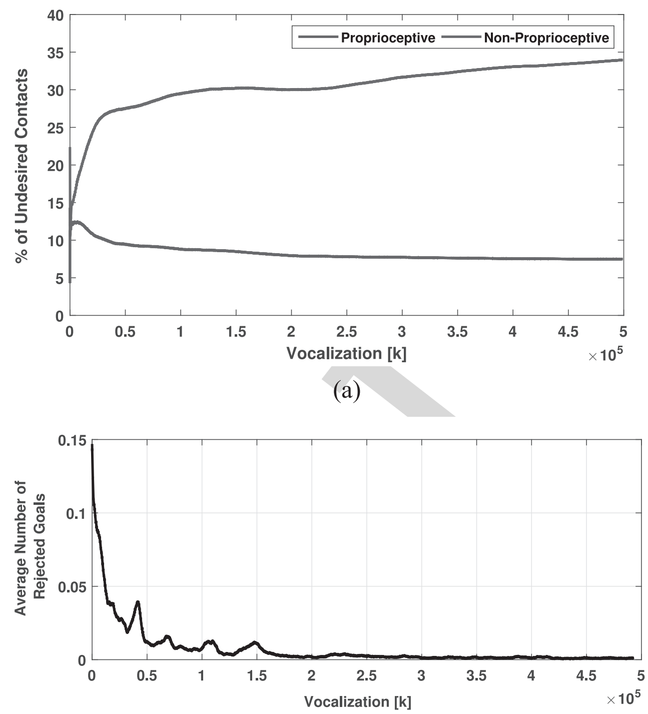
steps are increased for the somatosensory model and the sensorimotor model, then the training computational time increases as N_{old} in (6) increases proportionally, but no improvement is obtained in the inference error. However, if these values are decreased beyond the values used in [9], the mean inference error increases. Third, when α_{SM} and α_{SS} are larger than 0.1 the competence progress is slower. Finally, the parameters linked to the interest model allow a wider range of values to be chosen obtaining similar results. For K_{IM} we observed that values greater than or equal to 12 worked similarly, but smaller values negatively impacted the competence progress. Thus, the main parameters for all the simulations were kept as in [8] and [9] as they performed better than other simulations in terms of exploration results and simulation time. Summarizing, values were set to $K_{\text{SM}} = 28$, $N_{\text{SM}} = 400$, $K_{\text{SS}} = 28$, $K_{\text{IM}} = 12$, $N_{\text{IM}} = 4800$, $n_{\text{IM}} = 12$, and the continuous sampling time used for the DIVA ear-vocal tract was $t_s = 10$ ms. The forgetting rate parameter α_{SM} for G_{SM} starts from 0.1 and decreases logarithmically to 0.05 after half a million of vocalizations. On the other hand, α_{SS} for G_{SS} was chosen to be 0.05 through the whole simulation.[4]

During the simulation, G_{SM} and G_{SS} are initialized as indicated in line 1 of Algorithm 1 with the initial motor command sets. Then, all the initial phonatory productions are used as auditory goals to initialize the interest model G_{IM} as indicated in line 2 of Algorithm 1. In this stage, G_{SM} is used to infer the motor commands that will likely produce the initial auditory goals. These commands are executed without considering the proprioceptive prediction p.

A. On Competence and Contacts

First of all, Fig. 4 represents the evolution of the competence parameter c in (7) for self-generated auditory goals. To obtain this plot, first the result of each simulation is filtered using a 5000 samples moving average window. Next, simulations are divided into two general groups: 1) proprioceptive agents and 2) nonproprioceptive agents. Finally, the mean of all the filtered results for each group is computed. The same mechanism is considered to obtain the percentage of contacts observed in Fig. 5(a).

Tables I and II show the volume of a convex hull covering the explored auditory region. They also display the mean competence and the percentage of undesired contacts at the end of the simulation. First, in Table I, descriptors are computed considering all the vocalization during each simulation. Second, in Table II, figures were computed considering



(b)

Fig. 5. Algorithm 1 simulation results. (a) Mean percentage of vocalizations producing undesired collisions considering all the simulated agents. Agents are grouped by proprioceptive and nonproprioceptive. The results of each agent are prefiltered considering a 5000 samples moving average. (b) Mean number of rejected goals using the proprioceptive prediction and considering all the simulated proprioceptive agents. The results of each agent are prefiltered considering a 5000 samples moving average.

TABLE I
RESULTS CONSIDERING ALL DATA FROM EXPLORATION

Experiment	Non-Proprioceptive			Proprioceptive		
	Vol.	mean(c)	% Contacts	Vol	mean(c)	% Contacts
1	0.58	0.50	41.67%	0.49	0.80	3.42%
2	0.47	0.50	37.50%	0.49	0.76	3.88%
3	0.49	0.50	43.53%	0.43	0.61	5.28%
4	0.47	0.54	30.83%	0.52	0.73	9.03%
5	0.47	0.54	29.17%	0.44	0.68	15.55%
6	0.56	0.59	23.01%	0.52	0.70	8.18%
7	0.49	0.55	30.86%	0.41	0.71	6.76%
8	0.57	0.54	35.49%	0.43	0.68	8.57%
9	0.49	0.50	32.42%	0.63	0.74	6.74%
Average	0.51	0.53	33.83%	0.48	0.71	7.49%
Min	0.47	0.50	23.01%	0.41	0.61	3.42%
Max	0.58	0.59	43.53%	0.63	0.80	15.55%

Note: Experiments with different vocalization initial sets for Proprioceptive and Non-Proprioceptive Agents. The volume of a the convex-hull described by the explored data, the mean value for the competence c, and the final percentage of contacts along the simulations are shown.

[4]Supplementary downloadable material provided by the authors is available at https://dx.doi.org/10.6084/m9.figshare.c.3676645.v1. After each experiment, 20 random samples from the last 1000 vocalizations were drawn to generate videos with audio. Videos of the experiments 1, 5, and 9 are provided.

TABLE II
RESULTS WITHOUT CONSIDERING VOCALIZATIONS WITH UNDESIRED CONTACTS

Experiment	Non-Proprioceptive		Proprioceptive	
	Vol.	mean(c)	Vol.	mean(c)
1	0.48	0.69	0.39	0.81
2	0.38	0.67	0.39	0.78
3	0.37	0.71	0.33	0.63
4	0.40	0.65	0.42	0.77
5	0.36	0.66	0.37	0.76
6	0.44	0.67	0.42	0.74
7	0.36	0.69	0.32	0.74
8	0.45	0.67	0.35	0.71
9	0.38	0.64	0.49	0.78
Average	0.40	0.67	0.39	0.75
Min	0.36	0.64	0.32	0.63
Max	0.48	0.71	0.49	0.81

Note: Experiments with different vocalization initial sets for Proprioceptive and Non-Proprioceptive Agents. The volume of a the convex-hull described by the explored data and the mean value for the competence c along the simulations are shown.

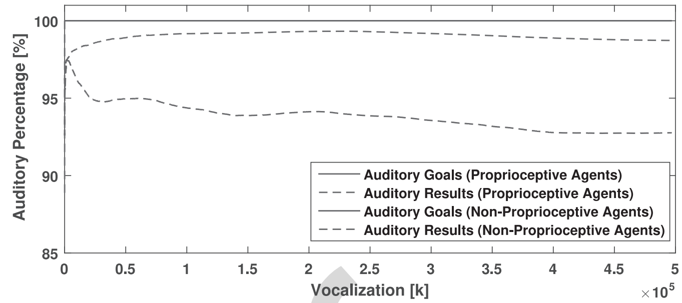


Fig. 6. Mean percentage of vocalizations producing phonatory result along all the simulations per each group, proprioceptive and nonproprioceptvie agents. The percentage of phonatory auditory goal is also shown per each group. *Note: red and blue solid lines are overlapped.*

vocalizations without undesired collisions. Convex hulls provide an insight regarding the size of the explored regions in the auditory-space. They are computed considering formant frequency dimensions F_{11}, F_{21}, F_{12}, and F_{22}.

In Fig. 4, results suggest that proprioceptive agents perform better than those which are not endowed with proprioception. We observe that at the beginning of the exploration the mean average competence is very similar for both groups. However, after the initialization the nonproprioceptive agents suffer an important decrement of the competence, which also coincides with a significant increment of the percentage of contacts in Fig. 5(a). Table I also confirms the expected results according to our hypothesis: using proprioceptive feedback drives artificial agents to produce significantly fewer undesired contacts and also increases the competence to reach self-generated auditory goals.

Some observers might ask the reason of high competence values at the beginning of the simulations. We argue that it is an expected result as the competence computation begins when G_{IM} is initialized with self-generated goals drawn from the initial auditory productions of the agent. In other words, G_{SM} and G_{SS} models are initialized around a set of initial vocalizations and auditory productions. The initial auditory productions are selected as auditory goals, then motor commands are computed with a sensorimotor model that represents very well those initialization samples. Later, as the agent explores the auditory space and it moves toward farther regions from those of initialization, the competence values might slightly decrease. This is due to the incremental learning of probabilistic models and depends on the values assigned to the forgetting rates α_{SM} and α_{SS}. If these forgetting rates are close to zero, then the agent is less prone to update its knowledge when new data are far from the current knowledge. On the other hand, if the forgetting rates are high, then the agent will adapt its model to the new data very fast but also it will forget faster its previous knowledge since it is not reinforced.

We also argue that one of the reasons the proprioceptive agents perform better is the null competence produced by nonphonatory vocalizations. The nonproprioceptive agent produces much more undesired contacts (four times more undesired contacts on average) and, therefore, has more nonphonatory vocalizations as corroborated in Fig. 6. Fig. 6 was obtained using the same procedure than Fig. 4. It shows the mean percentage of phonatory goals and actual phonatory vocalizations considering all the simulations per each group of artificial agents. In general, all the auditory goals through the simulations are phonatory. The reason is that nonphonatory goals become uninteresting very early in the artificial agent's life as they are very easy to be produced. Thus, all those nonphonatory vocalizations which produce null competence impact negatively the average competence. We also might think about all those nonphonatory vocalizations as a waste of energy during the exploration. Knowing which regions of the motor space are leading to collisions might be a relevant knowledge for the agent. However, as the nonproprioceptive agents keep exploring conflicting regions, the proprioceptive agents avoid exploitation of these regions due to their ability to predict somatosensory results from a given motor command.

Additionally, discussion about the tradeoff between exploration and exploitation can be detailed. We argue that proprioceptive agents show a better performance with respect to exploitation, as agents avoid exploring uninteresting regions with high number of contacts. In other words, proprioception, and in general constraint awareness, contributes to the agent finding regularities faster and then fosters specialization in regions of the auditory space, where the agent competence to reach self-generated goals is higher. It is worth mentioning that we are aware that it is also important to include the social factor in the learning development of the artificial agent, in order to better understand the role of proprioception in social learning. In social learning, exploration is not just driven by the progress in competence and discovery of constraints, but also by the relevance of auditory goals for socialization purposes. These studies leading to more exploring behaviors is left for future work.

Furthermore, Fig. 5(b) shows the mean number of goals rejected by the proprioceptive mechanism, represented in lines 5–12 of Algorithm 1. In this plot, we prefilter the results for proprioceptive agents considering a 5000 samples moving average for visualization purposes. It can be observed in Fig. 5(b) that in general the proprioceptive mechanism is more active at the beginning of the simulations presumably due to the quantity of contacts along the initial set of vocalizations.

Thus, we might deduce that proprioception prevents the agent from further exploration in regions that are producing undesired contacts especially in the early stages. In the next, we introduce some figures in order to show the implications over the shape of the explored auditory region when proprioception is considered.

B. On Explored Regions

Regarding the volume of the explored region, Table I indicates that the ratio of average volume of convex hulls described by the explored regions in the frequency space is 0.51/0.48 between the nonproprioceptive and proprioceptive agents, whereas the ratio of the mean competence is 0.53/0.71. In other words, whereas proprioceptive agents explore a 5.88% tighter region than the nonproprioceptive, their performance is 25.35% better than the later ones. On the other hand, Table II considers only the vocalizations without undesired contacts. A shrinkage of the convex hulls is observed, the ratio of average volumes is, in this case, 0.40/0.39 while the mean competence ratio is 0.67/0.75. From these numbers we observe that, in general, the competence to vocalizations without undesired contacts is higher for both kinds of agents. However, regarding competence the proprioceptive agents still perform 11.94% better than the nonproprioceptive agents.

Based on Table I, we selected three different initial sets given their simulation results in order to produce Figs. 7–9. First, we select initial set 1, as it performs better in terms of competence when proprioception is considered. Second, to contrast with initial set 1 we select initial set 5, as its proprioceptive agent performs the worst with regards to the percentage of undesired contacts. Finally, we select initial set 9, as its proprioceptive agent produced the largest convex hull volume.

Figs. 7–9 show some projections of vocalizations distribution maps of the auditory productions generated along the simulation. Points in the plots are colored according to the percentage of undesired contacts produced in its neighborhood. Specifically, three projections are shown for each of the selected sets of initial vocalizations. Projection $F_{1,1}F_{2,1}$ represents the auditory fingerprint of the vocalizations in the first perceptual window. Projections $F_{1,2}F_{2,2}$ is similar to the first projection but for the second perceptual window. Finally, projection I_1I_2 represents the value of the intonation parameter in the first perceptual window against the same parameter in the second perceptual window.

Distributions in Figs. 7–9 indicate that the intonation parameter projection I_1I_2 is the most influenced sensory-output due to the proprioceptive feedback. Recall that I_1 and I_2 depend on the average audibility of the vocalization which is null in two cases: 1) when the voicing parameters are lower than zero or 2) when the area function of the vocal tract is nonpositive elsewhere. Therefore, keeping in mind the latter case, those vocalizations producing an intonation parameter (either I_1 or I_2) lower than one indicate that a contact has likely occurred. If a contact occurs, there are two possible results, the average of the minimum value of the area function might be negative or not. If it is negative, then the contact is classified as an undesired contact and the proprioceptive signal takes the value

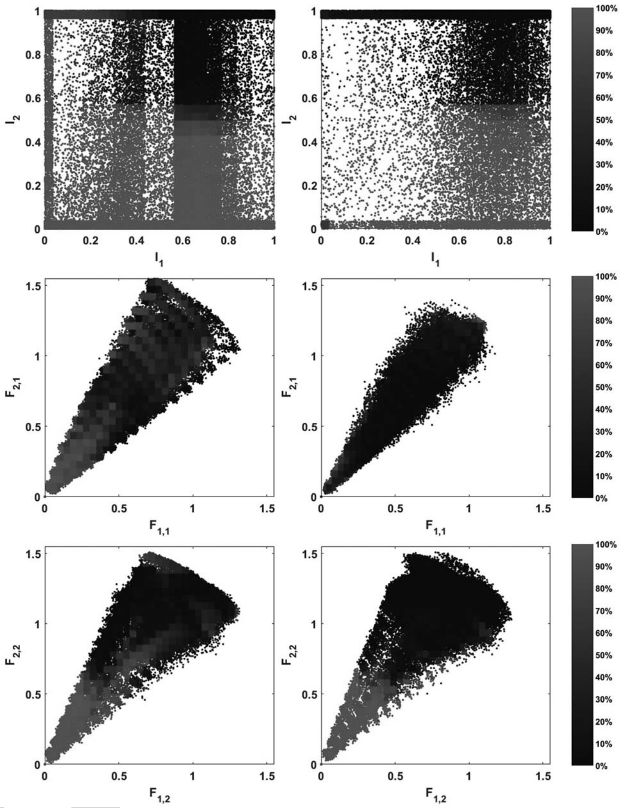


Fig. 7. Projections of vocalizations distribution along simulations using initial set 1 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

one. Thus, having both values lower than one at the same time is even more likely to produce undesired contacts. That is the reason proprioceptive agents explore less intensively the middle of the region in the intonation space. However, we argue that in spite of the low density of vocalizations in that region, proprioceptive agents succeed in finding more vocalizations that produces nonconflicting articulatory configurations in that region. For instance, looking at the projections in Fig. 9, the proprioceptive agent almost covers all the intonation space with low density of contacts.

Moreover, comparing proprioceptive and nonproprioceptive agents in Figs. 7–9, we observe that the area of the explored regions varies slightly due to the proprioceptive mechanism. This fact is supported by Tables I and II. In general, in most of the cases using the proprioceptive feedback results in a slightly smaller explored region but this is not a conservative fact. For instance, Table I indicates that the convex hull volumes described in the auditory space by the experiments 2, 4, and 9 were larger when proprioception was considered. In general, besides a certain degree of randomness due to our probabilistic approach, we argue that there are three main elements that determine the shape of the explored region: 1) the initial set of vocalizations; 2) evolution of competence; and 3) proprioception.

Regarding the initial set, our criteria to choose random vocalizations close to the neutral positions produces rich sets of phonatory vocalizations, either with contacts or without.

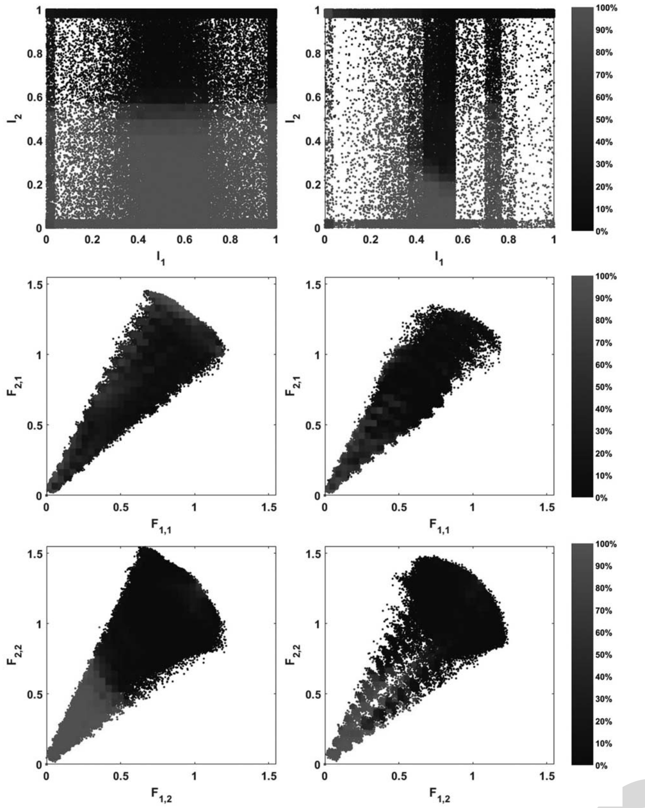


Fig. 8. Projections of vocalizations distribution along simulations using initial set 5 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

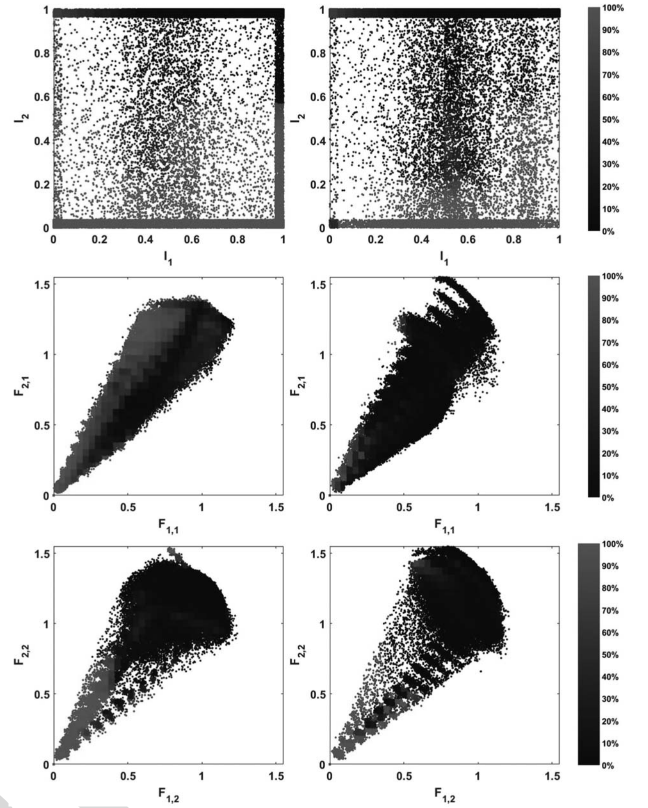


Fig. 9. Projections of vocalizations distribution along simulations using initial set 9 with Algorithm 1. Results for nonproprioceptive agent (left) and proprioceptive agent (right). Points are colored according to the total percentage of undesired contacts in their neighborhood.

As we are working with self-generated goals, the agent is expected to be very good at the beginning at reaching goals when auditory goals are close to the initialization region. In the case of nonproprioceptive agents, the only parameter that drives the exploration is the evolution of competence, which is why we observe plenty of areas with a huge amount of undesired contacts in the plots of nonproprioceptive agents with respect to the proprioceptive agents. Furthermore, we observe that proprioception might lead toward two different situations: 1) an unexplored auditory region or 2) explored region but with nonconflicting articulatory configurations. For instance, in Fig. 7, specifically in the projection $F_{1,1}F_{2,1}$, we observe that in general the nonproprioceptive agents produce a lot of undesired contacts over almost the whole explored region. On the contrary, the proprioceptive agent explores a smaller region over the same projection, however it achieves a considerably lower density of undesired contacts; results also supported by the convex hull volume displayed in Table I. In addition, projection $F_{1,2}F_{2,2}$ shows similar explored regions for both agents. Indeed, the proprioceptive agent explored a wider region in that projection and was capable of finding nonconflicting vocalizations for some of the regions, where the nonproprioceptive agent produces a lot of undesired contacts. We observe, in general, for all the agents in Figs. 7–9, that producing auditory results for the projection $F_{1,2}F_{2,2}$ close to the origin is hard without producing contacts.

In Fig. 8, corresponds to the agent with the worst results using proprioception regarding the number of undesired contacts, the projection $F_{1,1}F_{2,1}$ indicates that the proprioceptive agent has explored a smaller region than the nonproprioceptive agent. However, if we observe the boundaries of the explored region with proprioception, they coincide with regions where the nonproprioceptive agent produces a high amount of undesired contacts. Thus, the proprioceptive mechanism does not allow the proprioceptive agent to exploit those regions. In spite of less exploration, we observe that in the explored region where both agents intersect, the proprioceptive agent produces less undesired contacts. Looking at Table II, the results for the agent corresponding to Fig. 8 (experiment 5), we observe that the explored regions with and without proprioception are described by convex hulls with similar volume. This suggests that the conflicting region explored by the nonproprioceptive agent prevents the agent from exploiting regions without undesired contacts. Thus, the agent achieve lower competence values over the latter regions. On the other hand, the proprioceptive agent avoids conflicting regions, in consequence it produces 15% less contacts and achieve a higher competence average.

In addition, looking at the projection $F_{1,1}F_{2,1}$ Fig. 9, we observe that the proprioceptive agent explores a larger region. Moreover the density of contacts along the explored region is significantly lower. This is also supported by the numerical results in Tables I and II. On the other hand, in the projection

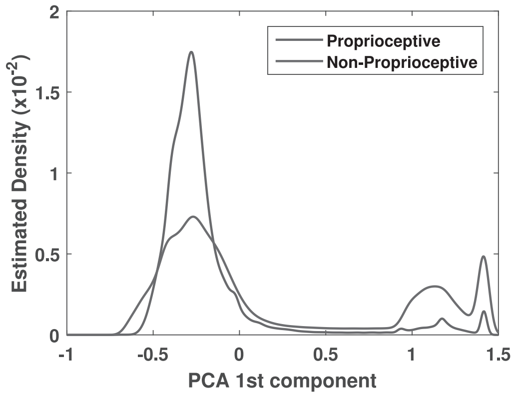


Fig. 10. Density distribution computed using Gaussian-kernels over all the data obtained along the simulations considering the first principal component with a variance contribution ratio of 0.68.

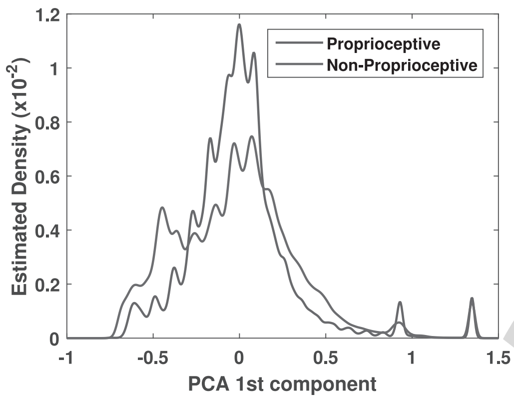


Fig. 11. Density distribution computed using Gaussian-kernels over the data, excluding vocalizations with undesired contacts, obtained along the simulations considering the first principal component with a variance contribution ratio of 0.61.

$F_{1,2}F_{2,2}$ the exploration close to the origin of that projection is less intensive in the proprioceptive agent, which reinforces the previous observation of the difficulties to produce auditory results in that region without contacts, similar results are observed in Fig. 8. Future work also must focus in the study of what is happening in that region and how relevant it is to language, as well as modify the system accordingly.

Finally, in order to observe the differences between the vocalization distributions obtained using the different exploration algorithms, we perform a sample density analysis over the formant frequency dimensions. In order to make the results easier to visualize we perform a principal component analysis (PCA) procedure. We consider analyzing the data twice, first, considering all the data collected along the exploration and second, without considering the vocalizations with undesired contacts. The PCA is done considering the dimensions $F_{1,1}$, $F_{2,1}$, $F_{1,2}$, and $F_{2,2}$, the data of all the 18 simulations are concatenated and used to perform the PCA. The PCA considering all the samples is performed and the first component is kept, which contributes to the variance with a ratio of 0.61. A second PCA is performed considering only the non conflicting vocalizations, again only the first component is kept, since it contributes to the variance with a ratio of 0.68. Once PCA transforms 4-D data into 1-D data, kernel-distribution estimation is performed using Gaussian-kernels according to [31] for the proprioceptive and nonproprioceptive cases.

In Figs. 10 and 11, we can observe the density distributions obtained separately with all the proprioceptive and the nonproprioceptive agents. First in Fig. 10, the distribution considering all the data obtained from all the experiments is shown. In general, it is observed that the agents explored similar regions, but with different intensity. In Fig. 11, we observe the distribution of the first component obtained from the PCA when only nonconflicting vocalizations are considered. In the latter case it is observed that regarding the regions which are of interest, in other words the regions where physical constraints are not violated, both kinds of agents explore with a similar density shape, which means that even though both agents explore similar interesting regions, the proprioceptive agents achieve in general higher competence.

VI. Conclusion

An application of active learning techniques applied to the study of vocal exploration considering motor constraints has been introduced. It has been presented as an intrinsically motivated sensorimotor self-exploration architecture with motor constraints self-awareness. Constraints awareness is achieved by providing a proprioceptive mechanism which endows an artificial agent with the capacity to autonomously generate a somatosensory model. This model is then used to predict the consequences of a motor action and to avoid its execution if it is expected to generate an undesired proprioceptive result.

The proprioceptive mechanism improved the quality of learning according to a competence function. However, we observe a tradeoff between exploration and exploitation, predominantly nonproprioceptive agents achieve greater exploration in the auditory space. In contrast, we observe a more intensive exploitation in interesting regions driving to the higher competence values achieved by proprioceptive agents. In general, vocal-auditory spaces are high dimensional redundant spaces, thus an auditory output may be produced by different articulatory configurations. Some of these articulatory configurations may lead to undesired contacts. Hence, we argue that sensorimotor redundancy is reduced when proprioception is included in the system allowing the agent to focus on exploitation of nonconflicting vocalizations. In consequence, the sensorimotor model generated through the exploration does not include conflicting regions, where constraint violations are likely to happen. For that reason, sensorimotor models achieve better fitting to the regions of interest where constraints are met. In this way, we showed how sensorimotor exploration, and in general sensorimotor knowledge, can be shaped by constraints.

Regarding the advance toward vocal exploration, we have showed the suitability of the presented architecture to learn vocal spaces in interesting and less redundant regions as children might do. However, in order to continue our research on early vocal development, we must study in greater depth the first period of vocalization development. A deeper analysis of the learning processes underlying the nonauditory development related to mastication, deglutition, and crying from the cognitive and developmental perspectives should be completed in order to generate more complex somatosensory architectures. Finally, the next step of this paper should be directed toward the self-structuring of vocalization and social learning.

References

[1] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nat. Rev. Neurosci.*, vol. 5, no. 11, pp. 831–843, 2004.

[2] M. Asada *et al.*, "Cognitive developmental robotics: A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

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

[4] R. Pfeifer and C. Scheier, *Understanding Intelligence*. Cambridge, U.K.: MIT Press, 1999.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, 1996.

[6] S. Thrun, "Exploration in active learning," in *Handbook of Brain Science and Neural Networks*, 1995, pp. 381–384.

[7] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: A unified probabilistic framework," in *Proc. Int. Conf. Develop. Learn. (ICDL/Epirob)*, Osaka, Japan, 2013, pp. 1–6.

[8] J. M. Acevedo-Valle, C. Angulo, N. Agell, and C. Moulin-Frier, "Proprioceptive feedback and intrinsic motivations in early-vocal development," in *Proc. 18th Int. Conf. Catalan Assoc. Artif. Intell.*, 2015, pp. 9–18.

[9] 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," *Front. Psychol.*, vol. 4, pp. 1006–1025, Jan. 2014, doi: 10.3389/fpsyg.2013.01006.

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

[11] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Netw.*, vol. 38, pp. 64–75, Feb. 2013.

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

[13] 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.

[14] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Auton. Mental Develop.*, vol. 2, no. 3, pp. 216–229, Sep. 2010.

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

[16] 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 *Proc. 1st Multidiscipl. Conf. Reinforcement Learn. Decis. Making (RLDM)*, Princeton, NJ, USA, Oct. 2014, Art. no. hal-00922537. [Online]. Available: https://hal.inria.fr/hal-00922537/document

[17] A. Ribes, J. Cerquides, Y. Demiris, and R. Lopez de Mántaras, "Active learning of object and body models with time constraints on a humanoid robot," *IEEE Trans. Cogn. Develop. Syst.*, vol. 8, no. 1, pp. 26–41, Mar. 2016, doi: 10.1109/TAMD.2015.2441375.

[18] J. Perkell *et al.*, "The sensorimotor control of speech production," in *Proc. 1st Int. Symp. Meas. Anal. Model. Human Functions*, 2001, pp. 359–365.

[19] D. K. Oller and R. E. Eilers, "The role of audition in infant babbling," *Child Develop.*, vol. 59, no. 2, pp. 441–449, 1988.

[20] K. Ejiri, "Relationship between rhythmic behavior and canonical babbling in infant vocal development," *Phonetica*, vol. 55, no. 4, pp. 226–237, 1998.

[21] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evol. Comput.*, vol. 11, no. 2, pp. 265–286, Apr. 2007.

[22] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: Computational and neural mechanisms," *Trends Cogn. Sci.*, vol. 17, no. 11, pp. 585–593, 2013.

[23] B. Galantucci, C. A. Fowler, and M. T. Turvey, "The motor theory of speech perception reviewed," *Psychonomic Bull. Rev.*, vol. 13, no. 3, pp. 361–377, 2006.

[24] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception," *J. Neurolinguist.*, vol. 25, no. 5, pp. 336–354, 2012.

[25] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, no. 6942, pp. 866–869, 2003.

[26] S. M. Nasir and D. J. Ostry, "Speech motor learning in profoundly deaf adults," *Nature Neurosci.*, vol. 11, no. 10, pp. 1217–1222, 2008.

[27] S. N. Iyer and D. K. Oller, "Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss," *Volta Rev.*, vol. 108, no. 2, pp. 115–138, 2008.

[28] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 4, pp. 1245–1248, 2009.

[29] C. Moulin-Frier and P.-Y. Oudeyer, "The role of intrinsic motivations in learning sensorimotor vocal mappings: A developmental robotics study," in *Proc. Interspeech*, Lyon, France, 2013, pp. 1268–1272.

[30] S. Calinon, *Robot Programming by Demonstration*. Lausanne, Switzerland: EPFL Press, 2009.

[31] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ, USA: Wiley, 2015.

Juan Manuel Acevedo-Valle, photograph and biography not available at the time of publication.

Cecilio Angulo, photograph and biography not available at the time of publication.

Clement Moulin-Frier, photograph and biography not available at the time of publication.